PRIVACY PROTECTION OF USER PROFILES IN PERSONALIZED INFORMATION SYSTEMS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF TELEMATICS ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF UNIVERSITAT POLITÈCNICA DE CATALUNYA
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Javier Parra Arnau
October 2013

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Jordi Forné Muñoz
(Principal Co-Adviser)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

David Rebollo Monedero
(Principal Co-Adviser)

Approved for the University Committee on Graduate Studies.

A Elisa,
mis padres,
y mi hermano.

Abstract

In recent times we are witnessing the emergence of a wide variety of information systems that tailor the information-exchange functionality to meet the specific interests of their users. Most of these personalized information systems capitalize on, or lend themselves to, the construction of profiles, either directly declared by a user, or inferred from past activity. The ability of these systems to profile users is therefore what enables such intelligent functionality, but at the same time, it is the source of serious privacy concerns.

Although there exists a broad range of privacy-enhancing technologies aimed to mitigate many of those concerns, the fact is that their use is far from being widespread. The main reason is that there is a certain ambiguity about these technologies and their effectiveness in terms of privacy protection. Besides, since these technologies normally come at the expense of system functionality and utility, it is challenging to assess whether the gain in privacy compensates for the costs in utility. Assessing the privacy provided by a privacy-enhancing technology is thus crucial to determine its overall benefit, to compare its effectiveness with other technologies, and ultimately to optimize it in terms of the privacy-utility trade-off posed.

Considerable effort has consequently been devoted to investigating both privacy and utility metrics. However, most of these metrics are specific to concrete systems and adversary models, and hence are difficult to generalize or translate to other contexts. Moreover, in applications involving user profiles, there are a few proposals for the evaluation of privacy, and those existing are not appropriately justified or fail to justify the choice.

The first part of this thesis approaches the fundamental problem of quantifying user privacy. Firstly, we present a theoretical framework for privacy-preserving systems, endowed with a unifying view of privacy in terms of the estimation error incurred by an attacker who aims to disclose the private information that the system is designed to conceal. Our theoretical analysis shows that numerous privacy metrics emerging from a broad spectrum of applications are bijectively related to this estimation error, which permits interpreting and comparing these metrics under a common perspective.

Secondly, we tackle the issue of measuring privacy in the enthralling application of personalized information systems. Specifically, we propose two information-theoretic quantities as measures of the privacy of user profiles, and justify these metrics by building on Jaynes' rationale behind entropy-maximization methods and fundamental results from the method of types and hypothesis testing.

Equipped with quantifiable measures of privacy and utility, the second part of this thesis investigates privacy-enhancing, data-perturbative mechanisms and architectures for two important classes of personalized information systems. In particular, we study the elimination of tags in semantic-Web applications, and the combination of the forgery and the suppression of ratings in personalized recommendation systems. We design such mechanisms to achieve the optimal privacy-utility trade-off, in the sense of maximizing privacy for a desired utility, or vice versa. We proceed in a systematic fashion by drawing upon the methodology of multiobjective optimization. Our theoretical analysis finds a closed-form solution to the problem of optimal tag suppression, and to the problem of optimal forgery and suppression of ratings. In addition, we provide an extensive theoretical characterization of the trade-off between the contrasting aspects of privacy and utility. Experimental results in real-world applications show the effectiveness of our mechanisms in terms of privacy protection, system functionality and data utility.

Acknowledgments

Son muchas las personas a las que querría agradecer su apoyo a lo largo de estos casi cuatro años de doctorado. En primer lugar, querría expresar mi más sincera gratitud a mis dos directores de tesis, Jordi y David, por haberme guiado, aconsejado y asistido en todo momento. Ha sido un verdadero honor poder trabajar con vosotros y aprender de vuestra sabiduría y experiencia. Sois un referente para mí.

En segundo lugar, estaré siempre agradecido a Miquel Soriano por haberme brindado la oportunidad de formar parte del Information Security Group, y por haberme ayudado en los momentos difíciles. También querría mostrar mi agradecimiento a todos los miembros del grupo, en especial a José L. Muñoz, Óscar Esparza y Esteve Pallarès, y a la profesora del Departamento de Ingeniería Telemática Mónica Aguilar. Asimismo, ha sido todo un privilegio poder colaborar con Elena Ferrari y Andrea Perego de la Universidad de Insubria, y con Claudia Díaz de la Universidad Católica de Leuven. Del mismo modo, también querría agradecer a Félix Gómez-Mármol por haberme ofrecido la posibilidad de realizar una estancia en NEC Laboratories Europe.

Por otro lado, querría darles las gracias a mis compañeros de despacho August, Carlos, Carolina, Elisabeth, Ernesto, Juan, Juan Felipe, Sergi, Victoria y Xavi, con los he compartido el día a día durante esta estimulante y enriquecedora experiencia que es el doctorado.

Finalmente, querría dedicar esta tesis a Elisa, a mis padres y a mi hermano. A Elisa, mi compañera de viaje, por estar a mi lado, en los buenos y en los malos momentos, dispuesta siempre a escucharme, animarme y apoyarme. A mis padres y hermano, por todo, pero sobre todo por creer en mí. Sin todos vosotros, nada de esto hubiera sido posible.

Contents

v

Abstract **vi**

Acknowledgments **viii**

1 Introduction **1**

1.1 Objectives 4

1.2 Summary of Contributions 5

1.3 Related Publications 7

1.4 Outline of this Thesis 10

2 Background and Related Work **12**

2.1 Privacy Issues in Personalized Information Systems 12

2.1.1 Impact of Personalization 14

2.1.2 Privacy Risks 16

2.1.3 Privacy and Utility Metrics 20

2.1.4 Data-Perturbative Mechanisms and Privacy-Utility Trade-Off 21

2.2 Statistical and Information-Theoretic Preliminaries 23

2.3 Privacy Protection in Personalized Information Systems 26

2.3.1 Trust Models 27

2.3.2 Privacy-Enhancing Technologies 28

2.4 Privacy Metrics 41

2.4.1 Statistical Disclosure Control 42

2.4.2 Anonymous-Communication Systems 46
2.4.3 Personalized Information Systems 48

I Privacy Metrics 50

3 Measuring Privacy as an Attacker's Estimation Error 52
3.1 Introduction 52
3.2 Background on Bayes Decision Theory 55
3.3 Measuring Privacy as an Attacker's Estimation Error 57
3.3.1 Mathematical Assumptions and Notation 57
3.3.2 Adversary Model 59
3.3.3 Privacy-Metric Definition 61
3.3.4 Example 62
3.4 Theoretical Analysis 64
3.4.1 Hamming Distortion 66
3.4.2 Non-Hamming Distortion 73
3.5 Numerical Example 78
3.5.1 Data Perturbation in Location-Based Services 79
3.5.2 Crowds-like Protocol for Anonymous Communications 81
3.6 Guide for Designers of SDC and ACSs 85
3.7 Conclusion 89

4 Measuring the Privacy of User Profiles 92
4.1 Introduction 92
4.2 User Profiling 93
4.2.1 Construction and Application of Profiles 94
4.2.2 Individual and Group Profiling 95
4.2.3 Definition of Profiling 96
4.3 Adversary Model 96
4.3.1 Scenario 97
4.3.2 User-Profile Model 99

4.3.3 Attacker's Objective . . . 102
4.4 Privacy Metric against Individuation . . . 104
4.4.1 Rationale behind the Maximum-Entropy Method . . . 106
4.4.2 Measuring the Privacy of User Profiles . . . 108
4.5 Privacy Metric against Classification . . . 111
4.6 Connection with Other Privacy Metrics . . . 114
4.7 Conclusion . . . 118

II Data-Perturbative Mechanisms and Privacy-Utility Trade-Off 121

5 Tag Suppression in the Semantic Web 123
5.1 Introduction . . . 123
5.2 Privacy-Enhancing Mechanism . . . 127
5.2.1 Tag Suppression vs. Other Privacy-Protecting Techniques . . . 128
5.3 Adversary Model and Privacy Metric . . . 135
5.4 Architecture . . . 137
5.5 Trade-Off between Privacy and Tag-Suppression Rate . . . 142
5.6 Theoretical Analysis . . . 143
5.6.1 Monotonicity and Quasiconcavity . . . 144
5.6.2 Critical Suppression . . . 145
5.6.3 Closed-Form Solution . . . 148
5.6.4 Low-Suppression Case . . . 154
5.6.5 High-Privacy Case . . . 155
5.6.6 Numerical Example . . . 156
5.7 Experimental Analysis . . . 159
5.7.1 Data set . . . 159
5.7.2 Tag Categorization . . . 160
5.7.3 Results . . . 162
5.8 Conclusions . . . 164

6 Privacy-Preserving Enhanced Collaborative Tagging 169
6.1 Introduction . . . 169
6.2 Overview of the Proposed Approach . . . 172
6.3 Tag Suppression at the Privacy Layer . . . 175
6.4 Reference Scenarios . . . 177
6.5 Experimental Analysis . . . 180
6.5.1 Data Set . . . 180
6.5.2 Tag Categorization and Methodology . . . 181
6.5.3 Results . . . 186
6.6 Conclusions . . . 200

7 Forgery and Suppression of Ratings in Recommendation Systems 202
7.1 Introduction . . . 202
7.2 Privacy-Enhancing Mechanism . . . 205
7.3 Adversary Model and Privacy Metric . . . 207
7.4 Architecture . . . 209
7.5 Trade-Off among Privacy, Forgery and Suppression . . . 216
7.6 Theoretical Analysis . . . 218
7.6.1 Closed-Form Solution . . . 219
7.6.2 Orthogonality, Continuity and Proportionality . . . 236
7.6.3 Critical-Privacy Region . . . 240
7.6.4 Case of Low Forgery and Suppression . . . 241
7.6.5 Pure Strategies . . . 244
7.6.6 Numerical Example . . . 246
7.7 Experimental Analysis . . . 249
7.7.1 Data set . . . 250
7.7.2 Results . . . 251
7.8 Conclusions . . . 258

8 Conclusions and Future Work 262
8.1 Conclusions . . . 263
8.1.1 Privacy Metrics . . . 263

8.1.2 Data-Perturbative Mechanisms and Privacy-Utility Trade-Off 266
8.2 Future Work 269

Acronyms **277**

Bibliography **278**

List of Tables

3.1 Notation of the theoretical framework. 58
3.2 Notation of the theoretical framework in the special case of SDC. . . 59
3.3 Notation of the theoretical framework in the special case of mixes. . . 61
3.4 Guide for designers of SDC and ACSs. 89

4.1 Outline of the adversary model assumed in personalized information systems. 104
4.2 KL divergence, Shannon's entropy and other criteria for measuring the privacy of user profiles. 116

5.1 Tag suppression versus other privacy-protecting approaches. 134
5.2 Notation for the analysis of the tag-suppression mechanism. 146

7.1 Notation for the analysis of the forgery and the suppression of ratings. 219
7.2 Categories of Movielens sorted for one particular user. 252

List of Figures

2.1 Illustration of privacy risks in personalized information systems. . . . 19
2.2 Data perturbation as an alternative to traditional methods based on access control and encryption. 23
2.3 Conceptual depiction of Chaumian mixes. 34
2.4 Database perturbation to achieve the k-anonymity requirement. . . . 43

3.1 Attacker's estimation error in the context of mixes. 60
3.2 Example of privacy-utility trade-off curve. 63
3.3 State-of-the-art privacy metrics as an attacker's estimation error. . . . 64
3.4 Example of microdata set meeting the requirement of 3-anonymity. . 68
3.5 Example of microdata set satisfying the principle of 2-diversity, 4-anonymity. 70
3.6 Example of microdata set vulnerable to skewness attacks. 75
3.7 Perturbation of location information. 79
3.8 Anonymous-communication protocol inspired by Crowds. 82
3.9 Probability distribution of the possible senders of a given message. . . 83

4.1 User profile modeled as a tag cloud in a collaborative tagging system. 100
4.2 User profile modeled as a histogram of absolute frequencies of ratings in personalized recommendation systems. 101
4.3 Query forgery in personalized Web search. 102
4.4 Outline of our interpretations of KL divergence and Shannon's entropy as measures of privacy. 105
4.5 Profile perturbation to hinder an adversary from individuating users. 110

4.6 Profile perturbation to hinder an attacker who aims to classify users. 114
4.7 Example of reidentification through profile linkage. 117

5.1 Adversary model assumed in collaborative tagging systems. 136
5.2 Block diagram of the tag-suppression architecture. 139
5.3 Interaction between a user suppressing tags and a semantic-Web server. 142
5.4 Conceptual plot of the privacy-suppression function. 147
5.5 User's tag distribution and their corresponding optimal apparent profile. 153
5.6 Optimal trade-off curve between privacy and tag-suppression rate. . . 157
5.7 Probability simplices showing the user's tag distribution and their optimal apparent profile. 158
5.8 User's actual tag distribution and optimal tag suppression strategy. . 162
5.9 Privacy-suppression function and suppression thresholds. 163
5.10 User's apparent tag distribution in BibSonomy. 164
5.11 Probability distributions of suppression thresholds. 165
5.12 Percentiles curves of relative privacy gain for a common tag-suppression rate in BibSonomy. 166

6.1 Extended architecture of a collaborative tagging system. 173
6.2 Tag categorization. 184
6.3 User's apparent tag distribution in Delicious. 187
6.4 Trade-off between privacy and tag-suppression rate. 188
6.5 User's actual tag profile and optimal suppression strategy. 189
6.6 Percentiles curves of relative privacy gain for a common tag-suppression rate in Delicious. 190
6.7 Average semantic loss versus tag-suppression rate. 191
6.8 Loss in semantic functionality when all users eliminate tags. 192
6.9 Curves of semantic loss for different fractions of the population suppressing tags. 193
6.10 Distribution of tags in Delicious before and after perturbation. 194
6.11 Number of false positives in a content-filtering application. 196
6.12 Number of false negatives in a content-filtering application. 198

6.13 Precision in a content-filtering application. 199
6.14 Recall in a content-filtering application. 200

7.1 User-profile model as a histogram of absolute frequencies of ratings. . 208
7.2 Adversary model assumed in personalized recommendation systems. . 210
7.3 Block diagram of the architecture implementing the forgery and the suppression ratings. 213
7.4 Optimal forgery and suppression strategies, user's item distribution and the population's. 234
7.5 Proportionality relationship between the optimal user's apparent item distribution and the population's profile. 239
7.6 Conceptual plot of the critical and noncritical privacy regions for $n = 5$ categories. 241
7.7 Contour lines of the privacy-forgery-suppression function. 246
7.8 Probability simplices showing the user's item distribution and their optimal apparent profile. 248
7.9 Example of optimal forgery and suppression strategies in Movielens. . 253
7.10 Optimal trade-off surface among privacy, forgery rate and suppression rate for one particular user of Movielens. 254
7.11 Example showing the proportionality relationship between the optimal user's apparent item distribution and the population's distribution. . 256
7.12 Percentiles surfaces of relative reduction in privacy risk for a common forgery rate and a common suppression rate. 257
7.13 Probability distribution of the relative decrement factors of forgery and suppression. 258
7.14 Probability distribution of the critical-forgery rate and the critical-suppression rate. 259

8.1 Profile of a Twitter user by hours and days of the week. 274

Chapter 1

Introduction

Recent years have witnessed the accelerated growth of a rich variety of information systems of unparalleled sophistication, whose aim is to help users deal with information overload [(a)]. The key enabling technology of these systems is *personalization*, a research field that has received great attention lately and which strives to tailor information-exchange functionality to the specific interests of their users. Examples of *personalized information systems* comprise resource tagging in the semantic Web, multimedia recommendation systems and personalized Web search.

The advent of personalization technologies is not only changing how people access information these days, but it is also leading a profound transformation of the traditional business model. To a large extent, this is because companies are increasingly approaching users in a personalized manner, attending their specific and particular needs more effectively. However, this is not the only reason for such substantial change: collecting information about the user's tastes and preferences has created additional opportunities with respect to monetizing and commercializing these personal data. The upshot is that personalized information systems are contributing to

[(a)] IBM claims that "90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals" [1].

unprecedented performance improvements in large business and small and medium enterprises (SMEs). For example, Amazon [(b)], who invented item-to-item collaborative-filtering algorithms [2], one of the most widely used personalization techniques, had more than 115 million average monthly unique visitors during the fourth quarter of 2012 [3]. Another example that illustrates this transformation is Facebook [(c)], which reported $1.46 billion in revenue for the first quarter of 2013. An 86 % of total income came from selling access to their data so that marketers could deliver targeted advertising messages to Facebook users [4]. Pushed by these personalization techniques, online advertising is expected to reach $139.8 billion in 2018, with an annual growth rate of 7.3% during the period 2013-2018 [5].

The impact of personalized information systems on the economy is evident as it is on *user privacy*. Most of these systems capitalize on, or lend themselves to, the construction of *profiles*, either directly declared by a user, or inferred from past activity, not only of the user in question, but also from the profiles of users with whom social relationships are known to the information system. Personalization allows users to deal with the overwhelming overabundance of information, but inevitably at the expense of privacy, especially when profiling is conducted across several information systems. In a nutshell, the ability of these systems to profile users based on their queries, clicks, tags, ratings and any other digital evidence and trace they leave in the online world is what enables such desired personalized service, but at the very same time, it poses evident privacy and security risks.

A variety of privacy-enhancing technologies (PETs) have been proposed to protect user privacy. Anonymous-communication networks [6–15], anonymous credentials [16–18], anonymous electronic cash [19], multiparty computation [20] and oblivious transfer protocols [21] are some examples of general-purpose PETs whose development roughly originates from the fields of security and cryptography. Unfortunately, PETs have not yet gained wide adoption. This is mainly because it remains unclear whether their overall benefits outweigh the operational costs caused by their use [22];

[(b)] `http://www.amazon.com`
[(c)] `https://www.facebook.com`

PETs typically come with penalties in terms of system functionality and data utility, which pose a trade-off between privacy protection and said penalties.

Evaluating the privacy provided by a PET is therefore crucial to determine its benefit, compare its effectiveness with other technologies, and ultimately to improve it. Further, quantifiable measures of the privacy gained and the cost incurred enable system designers to devise and optimize privacy-enhancing mechanisms in terms of the aforementioned privacy-utility trade-off; say, maximizing privacy for a given, acceptable cost.

Consequently, it is not surprising that a great deal of research has been devoted to the investigation of both privacy and utility metrics. The vast majority of these metrics have emerged from the mature fields of statistical disclosure control (SDC) [23–32] and anonymous-communication systems (ACSs) [9,33–44]. The problem is that most of them build upon different adversary models, capture diverse privacy threats and are intended to be used in specific settings. This restricts the scope of application of these metrics and precludes their generalization to other contexts such as personalized information systems. Besides, system designers often have several privacy metrics to choose from in a concrete application. In those cases, there is no guidelines that help designers decide which is the most appropriate approach for their privacy requirements.

In personalized information systems, the literature of privacy criteria is still in its infancy. There exist several proposals for assessing user privacy in this context, but they fail to justify the choice and are often defined to evaluate just the effectiveness of a concrete PET. This calls for the formalization of adversary models, the specification of the privacy risks considered under such models, and the rigorous justification of the privacy and utility metrics. Only in this way, quantitative measures of privacy and utility will contribute to the widespread adoption of PETs.

In the context of personalization, it is of special importance for PETs to deal with the compelling case when the intended recipient of sensitive information, i.e., the personalized information system, is not fully trusted and may thus be construed as a privacy attacker. Traditional encryption techniques offer the possibility of either fully delivering or completely obfuscating user information, by either providing or not

a cryptographic key permitting its deciphering. In the case of intended yet untrusted recipients, however, we are faced with a dilemma of great practical relevance.

Among a myriad of alternative PETs, those relying on *data perturbation* allow users to expose portions of their data, or somewhat modified versions of it, without the requirement of trusted intermediaries. By slightly perturbing confidential data locally prior to its disclosure, users attain a certain level of privacy in the presence of an untrusted information system, but at the expense of slightly degrading the utility of the data received by such system. Naturally, any perturbation introduced in the data will translate into a degradation of the quality of the personalized services. In a nutshell, we are confronted with the inescapable compromise between the contrasting aspects of privacy and utility.

The existence of this inherent compromise is a strong motivation to systematically develop quantifiable metrics of privacy and utility, and ultimately to design practical privacy-enhancing, data-perturbative mechanisms achieving serviceable points of operation in this privacy-utility trade-off.

1.1 Objectives

The objective of this dissertation is twofold. On the one hand, we tackle the issue of quantifying user privacy, first, in a general context that embraces the fields of SDC, ACSs and location-based services (LBSs); and secondly, in the specific and fascinating application of personalized information systems. On the other hand, we aim to solve the fundamental problem of privacy protection in these information systems. We approach this problem by means of data-perturbative mechanisms engineered to achieve a formally optimal trade-off encompassing the contrasting aspects of privacy and utility.

The scientific and technical objectives of this thesis may be more precisely described as follows:

- **Privacy metrics**. We shall derive quantitative measures of privacy that are meaningful in the context of each of the applications we consider. Our study of metrics extends beyond the measurement of the privacy of user profiles in

personalized information systems, and contemplates other applications such as SDC and anonymous communications. We do not expect a single set of measures to apply to every case; instead, each application is bound to call for different measures, and perhaps even optimization approaches. Moreover, such measures and optimization approaches will have to take into account multiple factors, including the adversaries' abilities and the nature of the private information to be protected. Mathematical measures of privacy will build upon mathematical models of user profiles, in the form of relative histograms of activity across predetermined sets of categories of interest.

- **Data-perturbative mechanisms and privacy-utility trade-off**. We shall design novel privacy-enhancing mechanisms and architectures for the privacy protection of user profiles in personalized information systems. These mechanisms and architectures will be devised in the form of parameterized models, allowing, in particular, the perturbation of the data by means of suppression and forgery. Building upon the previous objective, we shall measure the privacy gained by those data-perturbative strategies, and any possible data utility loss incurred. This will enable us to engineer such mechanism, specifically by modeling them as privacy-utility, multiobjective optimization problems. In order to achieve this goal, we shall develop and adapt theoretical procedures to solve those optimization problems. More generally, we intend to capitalize on rich concepts and powerful techniques from the mature fields of information theory, statistics and convex optimization.

1.2 Summary of Contributions

Next, we give an overview of the major contributions of this dissertation.

- We investigate a theoretical framework that enables system designers, first, to comprehend the relationships among state-of-the-art privacy criteria from SDC, ACSs and LBSs; secondly, to grasp the privacy properties and the underlying adversary models associated with each of these metrics; and ultimately, to assess their suitability for a given application. Our framework permits interpreting

such metrics as particular cases of our more general and unifying view of privacy, namely the attacker's estimation error. The arguments presented in our interpretations capitalize on fundamental results from the fields of information theory, probability theory and Bayes decision theory.

- We propose two information-theoretic quantities as measures of the privacy of user profiles in the context of personalized information systems. The proposed criteria build on Jaynes' rationale behind entropy-maximization methods, and some results from large deviation theory and hypothesis testing. We contemplate two adversary models, each one capturing different objectives for the attacker. These objectives are defined consistently with the technical literature of profiling, thus connecting notations of this field with information theory.

- In the context of the semantic Web, we design *tag suppression*, a privacy-enhancing mechanism that leverages on the principles of data perturbation and data minimization. Such mechanism lends itself to be implemented as a software application running on the user's computer. We devise an architecture that provides high-level functional specifications to implement this software. Like any data-perturbative approach, privacy protecting comes at the cost of data utility. This privacy-utility trade-off is formulated as a multiobjective optimization problem. We find a closed-form solution to said problem and mathematically characterize the trade-off curve of our privacy-utility optimized mechanism. Experimental results show how tag suppression may enhance user privacy in a semantic-Web application.

- The architecture of current collaborative tagging services is extended to include a policy layer and a privacy layer. The former allows users to explicitly denote resources of interest and to specify which resources should be blocked while browsing the Web. The latter implements the tag-suppression mechanism. We assess the impact that tag suppression would have on the services enabled by such policy layer. In particular, our performance evaluation shows the effectiveness of the extended architecture in terms data utility and filtering capabilities for the applications of resource recommendation and parental control.

- Finally, in the context of personalized recommendation applications we engineer a privacy-protecting technology that simultaneously combines the forgery and the suppression of ratings. The details of a practical implementation of this technology are presented in the form of a modular architecture. The trade-off between privacy risk on the one hand, and on the other, loss in the accuracy of the recommendations, is modeled as a multiobjective optimization problem. We find an explicit closed-form solution, which allows us to configure the operating point of our mechanism within the optimal privacy-utility trade-off surface. Further, we provide an extensive theoretical analysis that investigates several compelling properties of this trade-off, including its behavior at low rates of forgery and suppression, and some results showing when suppression is more convenient than forgery. Lastly, we apply the forgery and the suppression of ratings to a real-world recommendation system and evaluate to which degree the proposed mechanism may effectively protect the privacy of its users.

1.3 Related Publications

This thesis has been developed within the framework of several Spanish R&D projects, in particular, TSI2007-65393-C02-02 "ITACA", TEC2010-20572-C02-02 "CONSEQUENCE" and CONSOLIDER 2010 CSD2007-00004 "ARES". Most of the research results presented in this dissertation have been published in journals and conferences. In this section we provide a list of such publications, together with their complete bibliographic information. Further, we include other complementary articles that are not directly related with the research topic of this thesis, but which are especially significant from the state-of-the-art perspective.

Journal publications:

1. J. Parra-Arnau, A. Perego, E. Ferrari, J. Forné and D. Rebollo-Monedero, "Privacy-preserving enhanced collaborative tagging," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*. ISSN: 1041-4347. Impact factor 2012: 1.892. To appear [45].

2. J. Parra-Arnau, D. Rebollo-Monedero and J. Forné, "Measuring the privacy of user profiles in personalized information systems," *(Elsevier) Future Generation Computer Systems (FGCS), Special Issue on Data Knowledge and Engineering*. ISSN: 0167-739X. Impact factor 2012: 1.864. To appear [46].

3. J. Parra-Arnau, D. Rebollo-Monedero, J. Forné, J. L. Muñoz and O. Esparza, "Optimal tag suppression for privacy protection in the semantic Web," *(Elsevier) Data & Knowledge Engineering (DKE)*, vol. 81-82, November-December 2012. ISSN: 0169-023X. Impact factor 2012: 1.519 [47].

4. D. Rebollo-Monedero, J. Parra-Arnau, C. Diaz and J. Forné, "On the measurement of privacy as an attacker's estimation error," *(Springer) International Journal of Information Security (IJIS)*, vol. 12, no. 2, pp. 129-149, April 2012. ISSN: 1615-5262. Impact factor 2012: 0.480 [48].

5. J. Parra-Arnau, D. Rebollo-Monedero and J. Forné, "Optimal forgery and suppression of ratings for privacy enhancement in recommendation systems," submitted. [Online]. Available: `http://arxiv.org/abs/1302.2501` [49].

6. J. Parra-Arnau, D. Rebollo-Monedero and J. Forné, "A privacy-protecting architecture for recommendation systems via the suppression of ratings," *(Science & Engineering Research Support Society) International Journal of Security and Its Applications (IJSIA)*, vol. 6, no. 2, pp. 61-80, April 2012. ISSN: 1738-9976 [46].

Conference publications:

7. D. Rebollo-Monedero, J. Parra-Arnau and J. Forné, "An information-theoretic privacy criterion for query forgery in information retrieval," in Proceedings of the *International Conference on Security Technology (SecTech)*, Communications in Computer and Information Science (Springer), vol. 259, Jeju Island, South Korea, December 2011, pp. 146-154. ISBN: 978-3-642-27188-5. Best paper award [50].

8. J. Parra-Arnau, D. Rebollo-Monedero and J. Forné, "A privacy-preserving architecture for the semantic Web based on tag suppression," in Proceedings of the *7th International Conference on Trust, Privacy & Security in Digital Business (TrustBus)*, Lecture Notes in Computer Science (Springer), vol. 6264, Bilbao, Spain, August 2010, pp. 58-68. ISBN: 978-3-642-15151-4 [51].

9. J. Parra-Arnau, D. Rebollo-Monedero and J. Forné, "A privacy-protecting architecture for collaborative filtering via forgery and suppression of ratings," in Proceedings of the *6th International Workshop on Data Privacy Management (DPM)*, Lecture Notes in Computer Science (Springer), vol. 7122, Leuven, Belgium, September 2011, pp. 42-57. ISBN: 978-3-642-28878-4 [52].

10. D. Rebollo-Monedero, J. Parra-Arnau and J. Forné, "Un criterio de privacidad basado en teoría de la información para la generación de consultas falsas," in Proceedings of the *XI Reunión Española sobre Criptología y Seguridad de la Información (RECSI)*, Tarragona, Spain, September 2010, pp. 129-134. ISBN: 978-84-693-3304-4 [50].

Finally, we list the complementary articles mentioned at the beginning of this section.

11. D. Rebollo-Monedero, J. Forné, E. Pallarès and J. Parra-Arnau, "A modification of the Lloyd algorithm for k-anonymous quantization," *(Elsevier) Information Sciences*, vol. 222, February 2013, pp. 185-202. ISSN: 0020-0255. Impact factor 2011: 3.643 [53].

12. C. Tripp-Barba, L. Urquiza, M. Aguilar, J. Parra-Arnau, D. Rebollo-Monedero, J. Forné, E. Pallarès, "A collaborative protocol for anonymous reporting in vehicular ad hoc networks," *(Elsevier) Computer Standards & Interfaces*, vol. 36, November 2012, pp. 188-197. ISSN: 0920-5489. Impact factor 2012: 0.978 [54].

13. D. Rebollo-Monedero, J. Forné, E. Pallarès, J. Parra-Arnau, C. Tripp-Barba, L. Urquiza, M. Aguilar, "On collaborative anonymous communications in lossy networks," *(Wiley) Security and Communication Networks, Special Issue on Security in a Completely Interconnected World.* ISSN: 1939-0114. Impact factor 2012: 0.311. To appear [55].

1.4 Outline of this Thesis

The structure of this dissertation is in line with the research objectives defined in Sec. 1.1. In particular, this thesis is organized into two parts. The first part focuses on the investigation of privacy criteria as well as other aspects intimately related to them, such as adversary models and models of user profiles. The second part focuses on PETs based on data perturbation and studies how to optimize them in terms of the privacy-utility trade-off they pose. These two parts are called *privacy metrics* and *data-perturbative mechanisms and privacy-utility trade-off*, respectively.

Chapter 2 illustrates the privacy risks inherent in personalized information systems and reviews the state of the art relevant to this dissertation. Several concepts from information theory are also examined in this chapter.

The first part starts right after this. Chapter 3 proposes quantifying privacy in terms of the attacker's estimation error. The theoretical framework developed in this chapter is then shown to provide a unifying view of numerous privacy criteria emerging from a wide range of applications. Chapter 4 focuses on measuring user privacy in the context of personalized information systems. In particular, it describes our assumptions about the potential privacy attackers and presents a mathematically, tractable model of user interests. Building upon this adversary model, Chapter 4 proposes and justifies several privacy metrics stemming from information-theoretic concepts. Such metrics lay the foundation for the investigation of novel privacy-enhancing mechanisms in the next chapters.

The second part of this dissertation begins with Chapter 5. This chapter explores a mechanism aimed at protecting user privacy in the semantic Web, and describes how it could be implemented in practice. Chapter 5 also includes a theoretical characterization of the privacy-utility trade-off posed by our approach. Afterwards, Chapter 6 provides an extension of the current architecture of collaborative tagging systems. Such extension incorporates, on the one hand, the privacy-preserving technology explored in Chapter 5, and on the other, additional services such as resource recommendation and parental-control filtering. This chapter examines how our data-perturbative mechanism would degrade these two additional services. Lastly, Chapter 7 proposes

the combination of two data-perturbative strategies as a means for safeguarding user privacy in the context of personalized recommendation systems. This chapter analyzes the privacy gained and the cost incurred by the proposed mechanism and reports experimental results in a popular recommender system.

Chapter 2

Background and Related Work

The first part of this chapter begins with an introduction to personalized information systems. We explore the privacy risks inherent in such systems and emphasize the importance of privacy and utility metrics. Then, we present data perturbation as a compelling approach to enhance user privacy.

The second part of this chapter, namely Sec. 2.2, recalls some information-theoretic concepts that will be used throughout this dissertation. Afterwards, Secs. 2.3 and 2.4 review the state of the art in privacy-protecting mechanisms and privacy measures. A great portion of this review is adapted from [46–49, 53–55].

2.1 Privacy Issues in Personalized Information Systems

Selecting and directing information are crucial in every aspect of our modern lives, including areas as diverse as health, leisure, marketing and research. In the past, these processes were largely manual, but due to the exponential improvements in computation and memory, sophistication of software and the gradual ubiquity of mobile and fixed Internet access, they are now becoming increasingly automated.

The automation of these processes clearly facilitates effective handling of information. In a world where online information systems, society and economics have become inextricably entangled, the automated, personalized filtering and selection of an otherwise overwhelming overabundance of information is indispensable. To put

this continuous bombardment of information in numbers, every minute 6 600 pictures are uploaded to Flickr [(a)], 600 videos are submitted to YouTube [(b)], 70 new Internet domains are registered, 98 000 tweets are generated on the social networking site Twitter [(c)], 20 000 new posts are published on the micro-blogging platform Tumblr [(d)] and 12 000 new ads are posted on Craigslist [(e)] [56].

Endowing the above systems with intelligent processes for the selection and direction of such tremendous flow of information increases their usability and guarantees their effectiveness. Said processes of information filtering and targeting can be built on the basis of user profiles, either explicitly declared by a user, or derived from past activity. Automated information filtering may, for example, help tailor a Google search to the personal preferences of a user, by leveraging on their search history. When searching in Facebook for a name of a person we would like to become virtual friends with, the site takes into account numbers of common friends to recommend the most likely person with that name. Under a conceptual, abstract perspective, personalized search and social networks are really a special case of recommendation systems, which encompass functionality of a growing variety of information services, predominantly multimedia recommendation systems such as YouTube, Netflix [(f)], Spotify [(g)], the Genius function of iTunes or Pandora Radio [(h)], to name just a few.

As for automated information targeting, the market of personalized online marketing, lavishly illustrated by Google AdSense or Yahoo! Advertising, is yet another critical aspect of modern life, to the point that the success of most competitive economic activities is largely dependent on advertising. In a scenario with hundreds of TV channels, Internet and spam filters, the competitiveness in the process of advertising itself is of paramount importance.

[(a)] `http://www.flickr.com`
[(b)] `http://www.youtube.com`
[(c)] `https://twitter.com`
[(d)] `http://www.tumblr.com`
[(e)] `http://www.craigslist.org`
[(f)] `http://www.netflix.com`
[(g)] `https://www.spotify.com`
[(h)] `http://www.pandora.com`

2.1.1 Impact of Personalization

We now underpin the arguments of the previous section with a few, albeit sufficiently illustrative, quantitative, economic and social data.

During the last two decades, the Internet and the World Wide Web have been gradually integrating into people's daily lives and have enabled new forms of communication such as e-mail and instant messaging. The so-called network of networks [57] not only has become an essential communication channel but also has transformed people's habits: online shopping, electronic voting and streaming media are just other examples of services and applications built upon this network. In recent years, we have also witnessed the emergence of mobile phones with advanced computing and connectivity, allowing users to access the Internet everywhere and enabling a myriad of new applications such as LBSs. Last but not least, we have seen how social networks are changing the way we socialize, create and share information with friends and colleagues. A clear example of this is Facebook, which nowadays is the greatest exponent of social networking with more than 1.11 billion users around the world, including numerous firms which provide information about their products and services [58].

The dimension of this transformation is still not appreciated in its full extent. As the Internet is expanding from the current 2.4 billion users to the 5 billion users predicted in 2020 [59], a recent survey indicates that the Internet has a strong influence on economic growth rates across a range of large and developed countries [60]. The report in question shows that the Internet represents, on average, 3.4% of GDP across the large economies that make up 70% of global GDP. Should the Internet consumption and expenses be deemed a sector, its magnitude in terms of GDP would be greater than education or agriculture sectors. Another significant figure is the steadily growing market penetration of smartphones, with 153.9 million units sold worldwide in the second quarter of 2012 [61].

Nevertheless, breathing new life into traditional activities is possibly the Internet's most relevant impact. The network of networks has led to key business changes embracing the whole value chain in almost all sectors and companies. These changes have had an impact not only on how products are sold but also, and more importantly,

on how companies approach users in a personalized manner, taking into account their unique preferences. With the emergence of mobile devices, this paradigm shift is even more exacerbated, as smartphones and tablets enable marketers to stay close by, literally in one's hand. An example that gives an idea of this transformation is Pandora Radio, the biggest automated music recommendation system, which streamed more than 3.9 billion hours of music in 2012, and generated $83.9 million revenue in mobile platforms during the last quarter of 2013 [62].

Consequently, the Internet and the technologies enabling personalization as a solution to the one-size-fit-all paradigm are contributing to performance improvements in large businesses; but their influences are also essential to SMEs: now it is feasible for a small company to be a global company from the very beginning, spanning geographies, cultures and nearly all conceivable domains, capabilities that once were in the hands of big corporations. In this respect, a study on SMEs showed that 75% of the economic impact of the Internet was found in traditional companies that would not consider themselves as being Internet's players [63]. Another report shows that, among the more than 4 800 SMEs surveyed, those firms using Web technologies grew more than twice as fast as those with a minimal Web presence [60]. In a nutshell, this just reinforces the fact that these information technologies are also contributing to the transformation of the business model.

On the other hand, personalization is having a great impact on those technologies that allow users to navigate and retrieve information from the Web. In the current context of information overload, where the amount of information available to users grows exponentially, search engines can help them separate the wheat from the chaff by exploiting their search histories or location. The relevance of personalized information is also stressed in the way search engines capitalize on the data available on social networks to improve search results. For instance, Facebook's users can use the button "like" to indicate interest in some content they find on the Web. Afterwards, when a user submits a query, those pages classified as "liked" by their friends may be used to rank search results.

The upshot is that today we are witnessing the advent of a number of services in the Internet where personalization plays a prime role. Google Search and News [(i)], Digg [(j)], YouTube, Netflix and FourSquare [(k)] are just a few examples of those services, with billions of users worldwide. Although these services are leading to a profound transformation both in numerous aspects of people's lives and in economy, we would like to emphasize that we are still in an early stage, incapable of discerning the changes these technologies will foster. As the Internet grows and many more enabling technologies arise, the capability of providing many more users with enhanced personalized services will continue to increase exponentially. As a result, our society should be ready to embrace the myriad of opportunities that personalized information systems can create, but without losing sight of the privacy challenges these technologies pose.

2.1.2 Privacy Risks

At the heart of personalized information systems is *profiling*. From a home computer or a smartphone, users submit queries to Google, search for news on Digg, rate movies at IMDb [(l)] and tag their favorite Web pages on Delicious [(m)]. Over time, the collection and processing of all these actions allow such systems to extract an accurate snapshot of their interests or *user profile*, without which personalized services could not be provided. Profiling is therefore what enables those systems to determine what information is relevant to users, but at the same time, it is the source of serious privacy concerns.

These concerns become more serious and difficult to manage when user profiles are cross-referenced among a number of information services. An illustrative example is [64], which demonstrates that it is feasible to unveil private information about a person from their movie rating history by cross-referencing data from other sources. The cited work analyzed the Netflix Prize data set [65], which contained anonymous

[(i)] `http://news.google.com`
[(j)] `http://digg.com`
[(k)] `https://foursquare.com/`
[(l)] `http://www.imdb.com`
[(m)] `https://delicious.com`

movie ratings of around half a million users of Netflix, and was able to uncover the identity, political leaning and even sexual orientation of some of those users, by simply correlating their ratings with reviews they posted on the popular movie Web site IMDb.

Moreover, the enrichment of these information services with data from social networks creates additional opportunities with respect to information sharing, but inevitably aggravates the user privacy risks. User profiles may reveal sensitive information such as health-related issues, political preferences, salary and religion, not only about the user in question, but also about other users with whom social relationships are available to the service provider.

Further, the advent of cloud computing makes information and communication technologies more interconnected: a single online transaction may involve multiple business partners and create multiple pieces of digital evidence at various service providers. A major current trend is the provisioning of applications of increasing complexity and sophistication to a standard Web browser. While this eliminates the need for the user to locally maintain software, it further increases privacy risks because all data are necessarily stored in the cloud.

All these environments favor the collection, exchange and processing of personal information about the users. As a consequence, there is pressing need that the systems and applications which entail such processing of personal data take into account the existing European legal and regulatory framework on privacy and data protection.

As the intrinsic privacy risks of personalized information systems become clearer to society, legal compliance and social acceptance will become an increasingly important success factor. Privacy protection may even become a competitive business advantage in the design of such systems. Simultaneously searching for the terms “privacy” and “Facebook” in the New York Times search tool, for example, retrieves over four million articles [(n)]; the progressive integration of everyday activities into the Internet can only increase both the risk and its social awareness.

[(n)] As of May 28, 2013, resulting from the query `http://query.nytimes.com/search/sitesearch/#/privacy+facebook`

Use Case

In this subsection, we motivate and put in perspective the privacy risks posed by the personalized information systems that proliferate these days in the Internet.

Jane Doe is about to finish a long day of work in the patent department of her law firm in New York City. It has been a pretty hectic week, due to the forthcoming, albeit still unannounced, release of a spanking new model of smartphone by Apple. This patent is by far her favorite legal case, as she enjoys keeping herself up to date on the latest technological gadgets, often browsing for them via Google search and YouTube. She also loves how, these days, online tools retrieve both intelligent search results and videos, almost anticipating her interests, undoubtedly learning from her past activity. Unsurprisingly, after health, she rated technology highest when customizing her preferences in Google News, which she accesses almost religiously every morning. Her boyfriend, a computer scientist, keeps telling her that the future of information systems lies in their personalization, by means of automated compilation of user profiles, implicitly from behavior or explicitly from declared interests. Sounds about right.

Jane is aware that her company may be tracking her work habits by monitoring the use of applications and Internet access, with tools such as Track4Win. Still, before turning off her desktop computer at work, she quickly checks a friend's post in Twitter confirming a meeting this Friday evening to chat about tomorrow's protest, organized by the Occupy Wall Street movement, against the budget cuts planned by the government. She promptly responds, and adds a link to an intriguing article on the subject in The New York Times, an American newspaper with left-wing views.

They are meeting at "Café Lalo", a famous café on Upper West Side. During the half-hour bus ride to that location, Jane uses her iPhone to log into Facebook, to find the lovely pictures of her cousin's newborn baby. She politely types a cheerful comment in the album congratulating the happy family. Over the last few months, she and her boyfriend have been seriously considering having a baby, although she wishes her job at the law firm would offer a better work-life balance. Still a few bus stops to go, giving her ample time to discover a couple of new Web sites on childbearing, one of them showing Facebook's "like" button, which she immediately presses almost

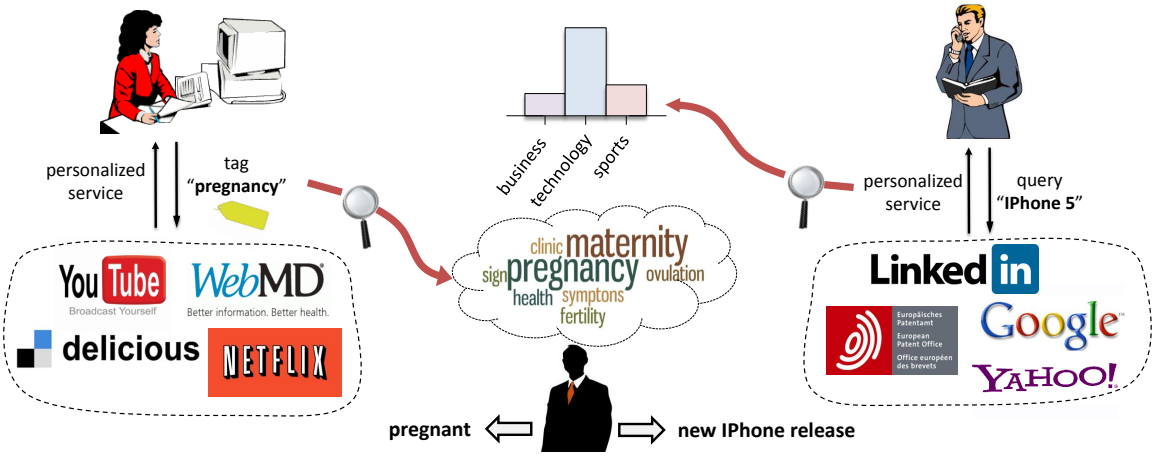


Figure 2.1: During an Internet session through various personalized information systems, users leave innumerable traces of sensitive information which, especially in combination, pose serious risks, not only to their own privacy, but also to the privacy of others.

as a reflex response. Of course, her action will be diligently reflected back in her profile. In a way, social networks are personalized information systems, reactively and proactively providing media tailored to their users' profiles of interests, built on the basis of their social interactions. She also notes a new friend request in Facebook, coming from a coworker in the human resources department. Even though their relationship is strictly professional, she finally accepts the request out of courtesy.

Comfortably seated in the café, while waiting for her friend, Jane continues using her smartphone to turn to Delicious, a social Web service where millions share and tag their favorite bookmarks. Luckily, she comes across a bookmark pointing to a site advertising an interesting job opportunity, also in the area of patents, in a law firm with more flexible hours, which she tags with the description "work-life balance", having her plans to get pregnant in mind. However, she is not sure whether she has to seriously consider this job opportunity since she is unfamiliar with both the law firm and the bookmark's author. Her friend arrives a few minutes late, but they both have a pleasant evening.

Little does Jane know that, during her Internet expedition from Google search to Delicious, passing by Google News, Twitter and Facebook, among other sites, she has left innumerable traces of sensitive information which, especially in combination,

pose a serious risk, not only to her own privacy, but also to the privacy of others. Hypothetically speaking, Google could correlate queries on smartphones with patents and Jane's declared interests on technology news, with Internet protocol (IP) addresses, presumably targeting her computer at work, from which she recently posted a detailed CV in LinkedIn [(o)], and thus learning about her occupation. Gathering additional evidence confirming a surge in query activity on the subject from similar sources, Google could be led to infer that Apple is likely to release the new iPhone 5, and retaliate by moving forward the new Android version.

Also hypothetically, someone in the department of human resources in Jane's law firm, which has started considering her promotion, could have attempted to become friends and inspect her Facebook profile to deduce the existence of a statistical chance of her having pregnancy plans. Further, her Twitter account is indicative of leftist views that might conflict with the political convictions of the company management. The fact that she uses a pseudonym in Delicious may not prevent the computer specialist in the human resources department from correlating users with tags related to law, patents, smartphones, pregnancy and the political Occupy Wall Street movement to guess her actual identity, and find out about her interest in job positions with better work-life balance. Not to mention the monitoring of her work habits and activity profile with Track4Win. Any of this could presumably endanger her promotion or even her current position. Some of these privacy risks are conceptually depicted in Fig. 2.1.

2.1.3 Privacy and Utility Metrics

The use case of Sec. 2.1.2 illustrates the privacy risks of the inherent need for profiles in personalized information systems. A large number of mechanisms have been developed to mitigate such risks. Some examples comprise anonymizers, pseudonymizers, ACSs [6–15], cryptography-based methods [16–18, 66, 67] and protocols relying on user collaboration [68–70]. All these mechanisms are PETs that can be applied to different scenarios and situations. The field of SDC, where an entity wants to publish aggregate information about a population but without compromising the privacy of

[(o)] `http://www.linkedin.com`

the individuals in that population, has also engineered numerous mechanisms that capitalize on data perturbation [53, 71–74].

Despite the great diversity of technologies available and their broad scope of application, the fact is that their use is far from being widespread. The main reason is that PETs are seen as an "expensive innovation with unclear benefits" [22]. This is because there is a certain ambiguity about PETs and their effectiveness in terms of privacy protection. Besides, since these technologies frequently come at the expense of system functionality and utility, it is difficult to evaluate whether the gain in privacy compensates for the costs in utility. It is worth to mention that the operational and deployment costs these technologies impose are often *perceived* as higher than those of traditional security mechanisms [22].

Consequently, measuring the privacy provided by a PET and the associated costs goes a long way in determining its actual overall benefit. It is therefore no surprise that much previous research has been dedicated to this topic. For instance, in the context of SDC, some of the best-known privacy metrics are k-anonymity [23, 24], l-diversity [27, 28], t-closeness [29] and differential privacy [31]. In ACSs, two information-theoretic privacy criteria are Shannon's entropy and Hartley's entropy. In personalized information systems, there are a few proposals to measure the privacy of user profiles. Section 2.4 examines these and more approaches for quantifying user privacy.

2.1.4 Data-Perturbative Mechanisms and Privacy-Utility Trade-Off

Two main conclusions follow from the previous subsection. First, if there is a chance to create a successful technology for privacy protection it is with a holistic approach, treating privacy on the one hand, and utility on the other, as two sides of the same coin. Secondly, a formal approach to evaluate, compare, improve and optimize novel and existing mechanisms entails the definition of quantitative measures of privacy and utility, contrasting aspects inherent in the design of practical, usable PETs for personalized information systems.

As we shall discuss in Sec. 2.3.2, a major portion of research initiatives build upon extensively studied privacy-enhancing mechanisms related to data access control, anonymization and pseudonymization. However, more recent studies, reviewed also in that section, contemplate the perturbation of sensitive data, while modeling incurred losses in data usability.

The perturbation of user data in the context of personalized information systems represents a completely different approach to more conventional privacy and security strategies. In traditional approaches to privacy, users or designers decide whether certain sensitive information, such as the user profile, is to be made available or not. However, in practice, the intended recipient of sensitive information may not be fully trusted. The availability of this data enables certain functionality, for example a personalized recommendation. Its unavailability, traditionally attained by means of access control or encryption, produces the highest level of privacy. In this dissertation we do not only consider these two extremes, but the interesting continuum in between enabled by data-perturbative mechanisms. Namely, we contemplate the possibility of exposing only portions of the data, or somewhat distorted versions of it, to gain privacy at the cost of data utility.

Inherent to data perturbation is therefore the existence of a trade-off between privacy and utility. Throughout this thesis, when we refer to the privacy-utility trade-off, the term *utility* will denote a quantification of the degree of functionality maintained with respect to that intended by the information system, despite the implementation of privacy mechanisms that may hide or perturb part of the data, along with the degree of quality of service maintained, despite processing, storage and communication overheads incurred by such mechanisms.

Data-perturbative techniques thus come at the expense of utility, but have three important features that make them particularly interesting to the application at hand. First, these techniques can be implemented as a software program running on the user's computer. This is without the need for deploying any infrastructure, one of the reasons that currently impede the adoption of PETs [22]. Secondly, as a result of the above, users need not trust the personalized information system, nor the Internet

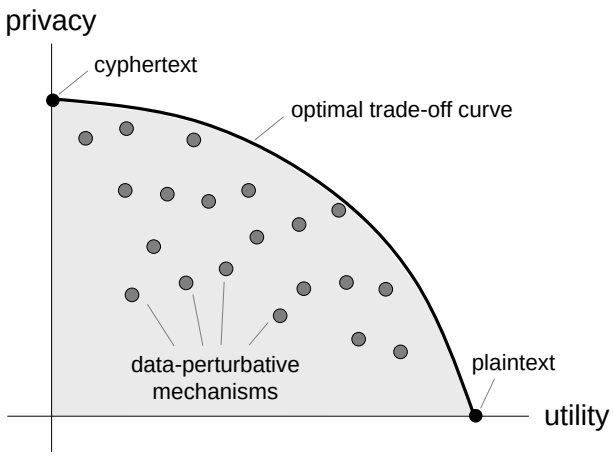


Figure 2.2: In contrast to traditional methods based on access control and encryption, where data is transmitted in the clear (plaintext) or encrypted (cyphertext), we contemplate the entire fascinating gray area in between. In particular, we consider the perturbation of some data, or the transmission of certain parts of it. In doing so, users enhance their privacy to a certain extent, although clearly at the cost of utility. In this dissertation, we investigate mechanisms based on user profile perturbation and optimize them in terms of the privacy-utility trade-off they pose.

service provider (ISP), nor any other external entity. And thirdly, data-perturbative mechanisms can be combined synergically with other PETs.

Equipped with quantitative measures of privacy and utility, we may strive to conceive such mechanisms modeled and engineered to attain the optimal privacy-utility trade-off, in the sense of maximizing privacy for a desired utility, or vice versa, with the aid of convex optimization techniques [75]. Fig. 2.2 conceptually illustrates such trade-off.

2.2 Statistical and Information-Theoretic Preliminaries

This section establishes notational aspects and recalls key information-theoretic concepts assumed to be known in the remainder of this work.

The measurable space in which a *random variable* (r.v.) takes on values will be called an *alphabet*. With a mild loss of generality, we shall always assume that the alphabet is discrete. We shall follow the convention of using uppercase letters for r.v.'s, and lowercase letters for particular values they take on. The probability mass function (PMF) p of an r.v. X is a function that maps the values taken by X to their probabilities. Conceptually, a PMF is *histogram* of relative frequencies across

the possible values determined by the alphabet of the r.v. in question. Throughout this dissertation, PMFs will be subindexed by their corresponding r.v.'s in case of ambiguity risk. Accordingly, both $p(x)$ and $p_X(x)$ denote the value of the function p_X at x. Occasionally, we shall refer to the function p by its value $p(x)$. We use the notations $p_{X|Y}$ and $p(x|y)$ equivalently.

The *expectation* of an r.v. X will be written as $\mathrm{E}\,X$, concisely denoting $\sum_x x\,p(x)$, where the sum is taken across all values of x in its alphabet. We adopt the same notation for information-theoretic quantities used in [76]. Concordantly, entropy, Kullback-Leibler (KL) divergence and mutual information will be denoted by the symbols H, D and I, respectively. We briefly recall these concepts for the reader not intimately familiar with information theory.

The *Rényi entropy* of order α of a discrete r.v. X with PMF p_X and alphabet $\mathcal{X}$ is defined as

$$\mathrm{H}_\alpha(X) = \frac{1}{1-\alpha} \log \sum_x p_X(x)^\alpha.$$

Regarded as a measure of the uncertainty of an r.v., Rényi's entropy may more conceptually be defined as

$$\mathrm{H}_\alpha(X) = -\log \mathrm{M}_{\alpha-1}[p_X(X)],$$

where $\mathrm{M}_{\alpha-1}$ denotes the power mean with exponent $\alpha - 1$ of the values of the distribution p_X, weighted by itself. In the important case when $\alpha = 0$, Rényi's entropy is essentially given by the support set of p_X, that is,

$$\mathrm{H}_0(X) = \log \left|\{x \in \mathcal{X} : p_X(x) > 0\}\right|.$$

In this particular case, Rényi's entropy is referred to as *Hartley's entropy*. Evidently, if p_X is strictly positive, then $\mathrm{H}_0(X) = \log|\mathcal{X}|$. On the other hand, in the limit when α approaches 1, Rényi's entropy reduces to *Shannon's entropy*,

$$\mathrm{H}_1(X) = -\sum_x p_X(x)\, \log p_X(x).$$

We shall also use the notation $\mathrm{H}_1(p)$ whenever we wish to emphasize the dependence of such entropy on the PMF of X [(p)]. Lastly, in the limit as α goes to ∞, the Rényi entropy approaches the *min-entropy*

$$\mathrm{H}_\infty(X) = \min_x -\log p_X(x) = -\log \max_x p_X(x).$$

Above, all logarithms are taken to base 2, and subsequently the entropy units are *bits*. If the base is e, we denote the natural logarithm by ln, and entropy is measured in *nats*. We use the convention that $0 \log 0 = 0$, which can be justified by continuity arguments.

Given two probability distributions $p(x)$ and $q(x)$ over the same alphabet, the *KL divergence* is defined as

$$\mathrm{D}(p \,\|\, q) = \mathrm{E}_p \, \log \frac{p(X)}{q(X)} = \sum_x p(x) \log \frac{p(x)}{q(x)},$$

where the expectation is taken over the distribution p. The KL divergence is often referred to as *relative entropy*, as it may be regarded as a generalization of the Shannon entropy of a distribution, relative to another. Conversely, Shannon's entropy is a special case of KL divergence, as for a uniform distribution u on a finite alphabet of cardinality n,

$$\mathrm{D}(p \,\|u) = \log n - \mathrm{H}_1(p). \tag{2.1}$$

Although the KL divergence is not a distance function, because it is neither symmetric nor satisfies the triangle inequality, it does provide a measure of discrepancy between distributions, in the sense that $\mathrm{D}(p \,\|\, q) \geqslant 0$, with equality if, and only if, $p = q$. On account of this fact, relation (2.1) between entropy and KL divergence implies that $\mathrm{H}_1(p) \leqslant \log n$, with equality if, and only if, $p = u$.

Consider two r.v. X and Y, with joint PMF p_{XY} and marginal distributions p_X and p_Y. The mutual information of these two r.v.'s is defined as the KL divergence

[(p)] From Chapter 4 onwards, our use of Rényi's entropies will be limited only to Shannon's. In those chapters, we shall drop the subindex and write $\mathrm{H}(X)$ to denote the Shannon entropy of the r.v. X.

between the joint distribution and the product distribution $p_X\, p_Y$,

$$\mathrm{I}(X;Y) = \mathrm{D}(p_{XY} \,\|\, p_X\, p_Y) = \sum_x \sum_y p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x)\, p_Y(y)},$$

and may be interpreted as a measure of the mutual dependence of the two r.v.'s. Another information-theoretic quantity is the *cross entropy* between the distributions $p(x)$ and $q(x)$, which is defined as

$$\mathrm{H}(p \,\|\, q) = -\,\mathrm{E}_p\, \log q(X) = -\sum_x p(x) \log q(x),$$

from whence it follows that

$$\mathrm{H}(p \,\|\, q) = \mathrm{H}_1(p) + \mathrm{D}(p \,\|\, q).$$

On the other hand, we shall follow the notation in [76] to specify that two sequences a_k and b_k are approximately equal in the exponent if $\lim_{k\to\infty} \frac{1}{k} \log \frac{a_k}{b_k} = 0$. To illustrate this, consider for example the sequences $a_k = 2^{3k+\sqrt{k}}$ and $b_k = 2^{3k}$, and check that $\lim_{k\to\infty} \frac{1}{k} \log \frac{a_k}{b_k} = \lim_{k\to\infty} \frac{1}{\sqrt{k}} = 0$, which implies that they agree to first order in the exponent. Still in the case of sequences, we shall use the abbreviated notation x^n to denote $x_1, x_2, \ldots, x_n$.

Last but not least, consider the variables x, y to be categorical or numerical data, vectors, tuples or sequences of such data. Accordingly, the *Hamming distance* between these two variables is defined as

$$d_{\mathrm{Hamming}}(x,y) = \begin{cases} 0, & x = y \\ 1, & x \neq y \end{cases}.$$

2.3 Privacy Protection in Personalized Information Systems

In this section, we shall examine the main proposals aimed at protecting user privacy in the scenario this thesis focuses on, namely personalized information systems. Before proceeding, Sec. 2.3.1 will introduce several *trust models*, essentially assumptions about the level of trust that users place in the entities they communicate with. The next subsection, Sec. 2.3.2, will survey the approaches of the state of the art in this scenario, showing in each case the level of trust assumed by users.

2.3.1 Trust Models

A number of actors are involved in the provision of personalized services. Among these actors, we obviously find users and the information systems themselves, but also we have the ISP, routers, switches, firewalls and any other networking infrastructure placed between the service provider and the end user.

Any of these entities may be considered as an attacker. To hinder these attackers in their efforts to compromise user privacy, users have a wide variety of PETs at their disposal, such as the technologies based on proxy systems, protocols exploiting collaboration among users, or mechanisms capitalizing on data perturbation. In some of these cases, users must place all their trust in these technologies. In other cases, however, it is not necessary that users trust the underlying privacy-protecting mechanism. In this section we define three models that specify this degree of trust. Such levels will allow us to identify the assumptions upon which the mechanisms surveyed in Sec. 2.3.2 build.

In the *trusted model*, users entrust an external entity or trusted third party (TTP) to safeguard their privacy. That is, users put their trust in an entity which will hereafter be in charge of protecting their private data. In the literature, numerous attempts to protect user privacy have followed the traditional method of anonymous communications, which is fundamentally based on the suppositions of our trusted model. Additional examples of PETs assuming this model are anonymizers and pseudonymizers. The idea behind these TTP-based approaches is conceptually simple. Their main drawbacks are that they come at the cost of infrastructure and suppose that users are willing to trust other parties. However, even in those cases where we could trust an entity completely, that entity could eventually be legally enforced to reveal the information they have access to [77]. The AOL search data scandal of 2006 [78] is another example that shows that the trust relationship between users and TTPs may be broken. In short, whether privacy is preserved or not depends on the trustworthiness of the data controller and its capacity to effectively manage the entrusted data.

On the other extreme is the *untrusted model*, where users mistrust any of the aforementioned actors. Since users just trust themselves, it is their own responsibility

to protect their privacy. Examples of mechanisms relying on the assumptions of our untrusted model are those based on data perturbation and operating on the user side. In this kind of data-perturbative approaches, users need not trust any entity but, as argued in Sec. 2.1.4, privacy protection comes at the cost of system functionality and data utility.

On a middle ground lies the *semi-trusted model*, where trust is distributed among a set of peers that collaborate to protect their privacy against a set of untrusted entities. An example of this trust model is found in the collaborative or peer-to-peer (P2P) approaches examined later in Sec. 2.3.2. In these approaches, users trust other peers and typically participate in the execution of a protocol aimed at guaranteing their privacy. Users clearly benefit from this collaboration, but nothing can prevent a subset of those peers from colluding and compromising the privacy of other users.

2.3.2 Privacy-Enhancing Technologies

In this section we review the state of the art in PETs in the context of personalized information systems. Partly inspired by [79], we classify these technologies into five categories: basic anti-tracking technologies, cryptography-based methods from private information retrieval (PIR), TTP-based approaches, collaborative mechanisms and data-perturbative techniques. We would like to stress that many of the technologies reviewed, far from being mutually exclusive, may in fact be combined synergically.

Basic Anti-Tracking Technologies

A key element in the provision of personalized services are *tracking technologies*. Thanks to these technologies, personalized information systems can identify users across different visits or sessions as well as multiple Web domains. Tracking mechanisms are therefore a means of driving personalization, as they allow these systems to follow users over time, thus enabling profiling.

The inherent operation of the Internet does permit tracking users. As many other data-communication networks, the Internet requires that every user [(q)] be identified

[(q)] Technically, machines, not users, are identified by addresses.

by a unique address, in order for messages to be routed through the network. ISPs are precisely in charge of allocating addresses to users and keeping the correspondence between user identifiers and addresses. In this manner, users wishing to communicate through the Internet just need to attach the source and destination addresses to the message to be sent. On the one hand, these addresses enable the intermediary entities (switches, routers, firewalls) involved in the communication process to forward these messages until the destination address is reached. But on the other hand, since the addresses are transmitted in the clear, the entities themselves or any adversary capable of intercepting the messages may ascertain who is communicating with whom and therefore may track user activity.

Employing dynamic IP addresses and rejecting hypertext transfer protocol (HTTP) cookies are two basic methods to prevent an attacker, possibly the service provider itself, from tracking users. The identification of users through IP addresses actually fails when a large number of users share a single IP address. This is the case of the users of a private network who resort to network address translation [80] and share a static IP address. The use of the dynamic host configuration protocol [81] also provides a means to hinder privacy attackers in their efforts to monitor user behavior. The main drawback of dynamic IP addresses is that the assignment and renewal of these addresses are controlled by ISPs. On the other hand, rejecting HTTP cookies may be an alternative to avoid tracking. The problem of this approach is that it can disable other Web services.

The result of the application of these basic mechanisms is clear: the attacker cannot build a profile of the user in question, but this is at the expense of a nonpersonalized service; if the service provider is unable to profile users based, for example, on their search or tag history, no personalization is possible. We would like to note that if these methods were completely effective, users would achieve the maximum level of privacy protection, but the worst level in terms of utility. In terms of performance, these mechanisms would be comparable to those more conventional techniques based on access control or encryption. As we shall see in the remainder of this state-of-the-art section, other PETs aimed at preserving user privacy in the context of personalized information systems assume that users are tracked and, in a way, identified.

The aim of some these approaches is then to thwart the attacker from *accurately* profile users.

Private Information Retrieval

In this subsection we briefly touch upon a few early proposals in the field of PIR. Afterwards, we review other mechanisms relying also on cryptography. As we shall see, the PETs reviewed in this subsection and the anti-tracking technologies examined above have much in common: both approaches may provide users with the highest level of privacy protection but at the cost of nonpersonalized services.

PIR refers to cryptography-based methods that enable a user to privately retrieve the contents of a database, indexed by a memory address sent by the user, in the sense that it is not feasible for the database provider to ascertain which of the entries was retrieved [82, 83]. In the context of Web search, PIR protocols allow a user to look up information in an online database without letting the database provider know the search query or response. A simple way to provide this functionality is as follows: the database provider submits a copy of the entire database to the user so that they can look up the information themselves. This is known as trivial download. The field of PIR is aimed at transferring less data while still preserving user privacy.

The first PIR protocol [66] traces back to 1995. Said protocol allowed users to privately retrieve records from a series of replicated copies of a database. In this scheme, each of the servers storing a copy of that database could not learn any information about the items retrieved by the user; this was, however, at the expense of a large amount of communication. In the current information systems, the implementation of this solution is impractical; normally these systems make use of a database stored on a single server. Despite these shortcomings, this initial work triggered numerous and important contributions to the field.

An alternative to this protocol was [67], which proposed the first single-server approach in 1997. As in many subsequent PIR protocols, the main problem with this alternative is that it requires the participation of the server itself. In other words, the single-server approach implicitly assumes that the database provider will have

some incentives to help users protect their protect. In practice, this is an unrealistic assumption.

Although the literature of PIR is particularly rich and extensive, the mechanisms proposed so far have several major limitations. First, considering the inherent operation of these protocols, we may conclude that personalization is unfeasible. Since the database provider does not know neither the queries nor the corresponding answers, users cannot be profiled by the provider. And secondly, there are several disadvantages that preclude the practical deployment of these cryptographic methods: PIR protocols require the provider's cooperation, are limited to a certain extent to query-response functions in the form of a finite lookup table of precomputed answers, and are burdened with a significant computational overhead. A comprehensive and detailed discussion of PIR protocols appears in [84].

Next, we quickly explore some other mechanisms relying on cryptographic techniques. An approach to conceal users interests in recommendation systems is [85, 86], which propose a method that enables a community of users to calculate a public aggregate of their profiles without revealing them on an individual basis. In particular, the authors use a homomorphic encryption scheme and a P2P communication protocol for the recommender to perform this calculation. Once the aggregated profile is computed, the system sends it to users, who finally use local computation to obtain personalized recommendations. This proposal prevents the system or any external attacker from ascertaining the individual user profiles. However, its main handicap is assuming that an acceptable number of users is online and willing to participate in the protocol. In line with this, [87] uses a variant of Pailliers' homomorphic cryptosystem which improves the efficiency in the communication protocol. Another solution [88] presents an algorithm aimed at providing more efficiency by using the scalar product protocol.

TTP-based Mechanisms

A conceptually-simple approach to protect user privacy consists in a TTP acting as an intermediary or *anonymizer* between the user and the untrusted personalized information system. In this scenario, the system cannot know the user ID, but merely

the identity of the TTP itself involved in the communication. One of the deficiencies of this approach is that personalized services cannot be provided, as the TTP forwards user data, e.g., queries, tags or ratings, of multiple users on their behalf.

As a solution to this problem, the TTP may act as a *pseudonymizer* by supplying a pseudonym ID' to the service provider, but only the TTP knows the correspondence between the pseudonym ID' and the actual user ID. A convenient twist to this approach is the use of digital credentials [16–18] granted by a trusted authority, namely digital content proving that a user has sufficient privileges to carry out a particular transaction without completely revealing their identity. The main advantage is that the TTP need not be online at the time of service access to allow users to access a service with a certain degree of anonymity.

Unfortunately, none of these approaches prevent the service provider from profiling a user and inferring their real identity. In its simplest form, reidentification is possible due to the personally identifiable information often included in user-generated data such as Web search queries or tags. However, even though no identifying information is included, an observed user profile might be so uncommon that the attacker could narrow their focus to concentrate on a tractable list of potential identities and eventually unveil the actual user ID. Another example that illustrates why pseudonyms are insufficient to protect both anonymity and privacy is described as follows. Suppose that an observer has access to certain behavioral patterns of online activity associated with a user, who occasionally discloses their ID, possibly during interactions not involving sensitive data. The same user could attempt to hide under a pseudonym ID' to exchange information of confidential nature. Nevertheless, if the user exhibited similar behavioral patterns, the unlinkability between ID and ID' could be compromised through these similar patterns. In this case, any past profiling inferences carried out for the pseudonym ID' would be linked to the actual user ID.

In addition to these vulnerabilities, we would like to note that a collusion of the TTP, the network operator or some entity involved in the communication could definitely jeopardize user privacy. Moreover, all TTP-based solutions require that users shift their trust from the personalized information system to another party, possibly capable of collecting user data from different applications, which finally might

facilitate user profiling via cross-referencing inferences. In the end, traffic bottlenecks are a potential issue with TTP solutions.

We have shown that anonymizers, pseudonymizers and digital credentials are TTP-based approaches that may be used as an alternative to hide users' identities from an untrusted service provider. In the remainder of this subsection, we shall explore a particularly rich class of PETs that also rely on trusted entities, but whose fundamental aim is to conceal the correspondence between users exchanging messages. In the scenario of personalized information systems, ACSs may contribute to protect user privacy against the intermediary entities enabling the communications between systems providers and users. As we shall see next, the majority of these systems build on the assumptions of the trusted model defined in Sec. 2.3.1. Only those systems consisting in a network of mixes may be classified into our semi-trusted model.

As commented at the beginning of Sec. 2.3.2, the inherent operation of the Internet poses serious privacy concerns. This is because users' IP addresses are attached to every message sent through the network. Clearly, the use of encryption techniques is not enough to mitigate such privacy risks. Hiding the content of messages hinders adversaries in their efforts to learn the information users exchange, but does not prevent those adversaries from unveiling who is communicating with whom, when, or how frequently. Motivated by this, the first high-latency ACS, Chaum's *mix* [10], appeared.

Fundamentally, a mix is a system that takes a number of input messages, and outputs them in such a way that it is infeasible to link an output to its corresponding input with certainty. In order to achieve this goal, the mix changes the appearance (by encrypting and padding messages) and the flow of messages (by delaying and reordering them). Specifically, users wishing to submit messages to other peers encrypt the intended recipients' addresses by using public key cryptography and send these messages to the mix. The mix collects a number of these encrypted messages and stores them in its internal memory. Afterwards, these messages are decrypted and the information about senders is removed. In a last stage, when the number of

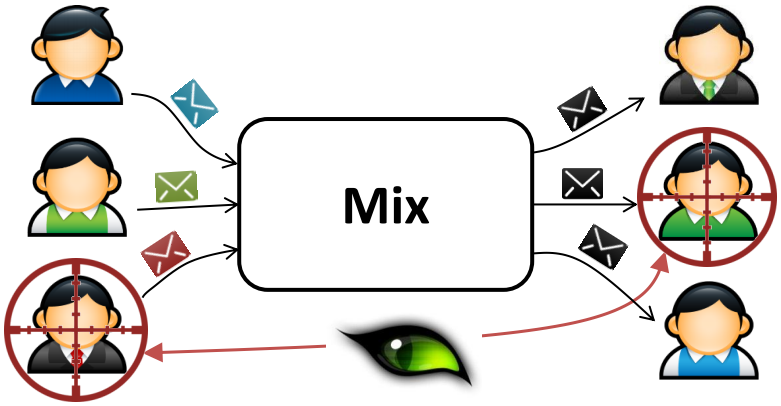


Figure 2.3: Many of the current ACSs are built upon the idea of Chaum's mix. Essentially, a mix can be seen as a black box that forwards messages in such a way that prevents an adversary from linking an outgoing message to its corresponding input message.

messages kept reaches a certain threshold, the mix forwards *all* these messages to their recipients in a random order.

In the literature, this process of collecting, storing and forwarding messages when a condition is satisfied is normally referred to as a *round*. An important group of mixes called *pool* mixes operate on this basis. Depending on the *flushing* condition, we may distinguish different types of pool mixes. Possibly, the most relevant form of pool mixes are *threshold* pool mixes [6], where the condition is imposed on the number of messages stored, as in the case of Chaum's mixes. The main difference is that threshold pool mixes do not flush all messages in each round, but keep some of them. Clearly, this strategy degrades the usability of the system: any incoming message can be stored in the mix for an arbitrarily long period of time. But these systems, in principle, achieve a better anonymity protection since they increase the set of possible incoming messages linkable to an outgoing target message to include all those messages that entered the mix before this target message was flushed.

Another important group of pool mixes outputs messages based on time [7]. Essentially, these *timed* mixes forward all messages kept in the memory every fixed interval of time called timeout. The major advantage of these mixes is that the delay experienced by messages is upper bounded, in contrast to the case of threshold pool mixes. The flip side is that the unlinkability between incoming and outgoing messages may be seriously compromised when the number of messages arriving in that interval

of time is small. Motivated by this, some of the current mix designs implement a combination of the strategies based on threshold and those based on time. Namely, these systems flush messages when a timeout expires, provided that the number of messages stored meets a threshold [8].

An alternative to pool mixes are the mixes based on the concept of *stop-and-go*, known as *continuous* mixes [9]. Specifically, this approach abandons the idea of rounds and gives the user the possibility of specifying the time that their messages will be stored in the mix before being submitted, for example, to a personalized information system. To this end, for each message to be sent the sender selects a random delay from an exponential distribution. This information is then attached to the message, which is encrypted with the mix's public key and then sent to the mix. Once the mix decrypts the message, the mix keeps it for the time specified by the user and then forwards it to its intended recipient.

The use of networks of mixes has also been thoroughly studied in the literature. The main reason to route over multiple mixes is to limit the trust that is placed on each single mix. This alternative is therefore in line with the semi-trusted model contemplated in Sec. 2.3.1. In order to trace messages, an adversary must ideally compromise all the mixes along the path. Depending on the network topology, we may classify the existent approaches into *cascade mixes*, *free-route networks* and *restricted-route networks*. The application of cascade mixes was already suggested by Chaum in his original work [10]. Fundamentally, this approach contemplates the concatenation of mixes to distribute trust. In contrast to this approach where messages are routed through a fixed path, free-route networks recommend that users choose random paths to route their own messages [11]. In the end, restricted-route networks consider the case where every mix in the network is connected to a reduced number of neighboring mixes [12].

An ACS that does not delay or reorder messages, which may be thus loosely regarded as a low-latency alternative to mixes, is *onion routing* [13, 14]. Such alternative approach is based on connections, rather than individual messages, but the net effect is that traffic is routed through a network of nodes in order to enhance anonymity, similarly to the scenario of cascade mixes. When a user wishes to send

a message, they submit it first to one of these nodes. Then, the node encrypts the message in a layered fashion and chooses the intermediate nodes to reach the recipient. Afterwards, each of these intermediate nodes peels off a layer of encryption and forwards the resulting message to the next node in the route. In the end, the last node delivers the message to the recipient. Considering how the system works, we may conclude that the functionality of the nodes essentially boils down to relaying messages. Clearly, this is in contrast to the case of mix systems, where messages are also delayed. Further, we would like to mention the second-generation version of onion routing, Tor [15], which has been available to Internet users since 2002. Despite being an improvement on onion routing, Tor nodes do not delay messages either, rendering the system susceptible to traffic analysis based on timing comparisons.

User Collaboration

In this subsection we examine those approaches where users collaborate to enhance their privacy. All these approaches may be understood under the semi-trusted model described in Sec. 2.3.1.

An archetypical example of user collaboration is the Crowds protocol [68]. This protocol is particularly helpful to minimize requirements for infrastructure and trusted intermediaries such as pseudonymizers, or to simply provide an additional layer of anonymity. In the Crowds protocol, a group of users collaborate to submit their messages to a Web server, from whose standpoint they wish to remain completely anonymous. In simple terms, the protocol works as follows. When sending a message, a user flips a biased coin to decide whether to submit it directly to the recipient, or to send it to another user, who will then repeat the randomized decision.

Crowds provides anonymity from the perspective of not only the final recipient, but also the intermediate nodes. Therefore, trust assumptions are essentially limited to fulfillment of the protocol. The original proposal suggests adding an initial forwarding step, which substantially increases the uncertainty of the first sender from the point of view of the final receiver, at the cost of an additional hop. As in most ACSs, Crowds enhances user anonymity but at the expense of traffic overhead and delay.

Closely inspired by Crowds, [54] proposes a protocol that enables users to report traffic violations anonymously in vehicular ad hoc networks. This protocol differs from the original Crowds in that, first, it does take into account transmission losses, and secondly, it is specifically conceived for multi-hop vehicular networks, rather than for wired networks. Also in the case of lossy networks, [55] provides a mathematical model of a Crowds-like protocol for anonymous communications. The authors establish quantifiable metrics of anonymity and quality of service, and characterize the trade-off between them.

Another protocol for enhancing privacy in communications, also relying on user collaboration and message forwarding, is [70]. The objective of the cited work is to hide the relationship between user identities and query contents even from the intended recipient, an information provider. The main difference with respect to the Crowds protocol is that instead of resorting to probabilistic routing with uncertain path length, it proposes adding a few forged queries.

In the context of personalized Web search, [89] proposes a P2P protocol to safeguard the privacy of users querying the Web search engine. The protocol follows the same philosophy of Crowds but leverages on social networks for grouping users with similar interests. Another approach exploiting user collaboration is [90], which suggests that two or more users exchange a portion of their queries before submitting them, in order to obfuscate their respective interest profiles versus the network operator or external observers. The idea of query profile obfuscation through multiple user collaboration has also been investigated from a game-theoretic perspective [91].

In LBSs, users submit queries along with the location to which these queries refer. An example would be the query "Where is the nearest Italian restaurant?", together with the geographic coordinates of the user's current location. In this scenario, [69] proposes a P2P spatial cloaking algorithm whereby users send their queries to an untrusted LBS provider without disclosing their precise location. The authors propose using the k-anonymity requirement [23, 24], a popular privacy criterion that we shall review later in Sec. 2.4.1. Accordingly, when a user wishes to submit a query to the provider, first they must find a group of $k-1$ neighboring peers willing to

collaborate. Once the group is formed, the originator of the query computes a geographical region including all users belonging to the group. After that, the user in question selects uniformly at random one of the members of the group. Ultimately, the originator sends both the query and the coordinates of that region to the selected user, which in turn is responsible for forwarding this information to the LBS provider on their behalf.

In the context of recommendation systems, some approaches suggest that users' private information be stored in a distributed way, in order to mitigate the potential privacy risks derived from the fact this information is kept in a single repository. One of these approaches is PocketLens [92], basically a collaborative-filtering algorithm specifically designed to be deployed to a P2P scenario. The proposed algorithm enables users to decide which private information should be exchanged with other users of the P2P community. In addition, the authors provide several architectures for the problem of locating neighbors. Closely in line with this alternative, [93] assumes a pure decentralized P2P scenario and proposes the use of several perturbative strategies. Namely, this scheme recommends replacing the actual ratings with (i) fixed, predefined values, (ii) uniformly distributed random values and (iii) with values drawn from a bell-curve distribution imitating the distribution of the population's ratings. In essence, this scheme could be regarded as a combination of the approaches in [92] and [94].

Data Perturbation

An alternative to hinder an attacker in its efforts to precisely profile users consists in perturbing the information they explicitly or implicitly disclose when communicating with a personalized information system. The submission of false data, together with the user's genuine data, is an illustrative example of data-perturbative mechanism. In this kind of mechanisms, the perturbation itself typically takes place on the user side. This means that users need not trust any external entity such as the recommender, the ISP or their neighboring peers. Obviously, this does not signify that data perturbation cannot be used in combination with other TTP-based approaches or mechanisms relying on user collaboration. It is rather the opposite—depending on the trust model

assumed by users, this class of PETs can be synergically combined with any of the approaches examined in Sec. 2.3.2. In any case, data-perturbative techniques come at the cost of system functionality and data utility, which poses a trade-off between these aspects and privacy protection.

An interesting approach to provide a distorted version of a user's profile of interests is query forgery. The underlying idea boils down to accompanying original queries or query keywords with bogus ones. By adopting this data-perturbative strategy, users prevent privacy attackers from profiling them accurately based on their queries, without having to trust neither the service provider nor the network operator, but clearly at the cost of traffic overhead. In other words, inherent to query forgery is the existence of a trade-off between privacy and additional traffic. Precisely, [95] studies how to optimize the introduction of forged queries in the setting of information retrieval.

Other alternatives relying on the principle of query forgery are [96–101], which propose a system for private Web browsing called PRAW. The purpose of this system is to preserve the privacy of a group of users sharing an access point to the Web while surfing the Internet. In order to enhance user privacy, the authors propose hiding the actual user profile by generating fake transactions, i.e., accesses to a Web page to hinder eavesdroppers in their efforts to profile the group. The PRAW system assumes that users are identified, i.e., they are logged in a Web site. However, the generation of false transactions prevents privacy attackers from the exact inference of user profiles.

The idea behind [102] is the same as in the PRAW system—the authors come up with the injection of false queries. In particular, they suggest a model working as a black box, switching between real queries and false queries. The proposed model operates as follows: it sends a real query with a certain probability, and a dummy query with the complement of that probability. The actual status of the switch and the probability of switching are assumed to be invisible or unknown to the attacker. The authors justify this assumption by arguing that this information is only available on the user side.

A software implementation of query forgery is the Web browser add-on TrackMeNot [103]. This popular add-on makes use of several strategies for generating and submitting false queries. Basically, it exploits RSS feeds and other sources of information to extract keywords, which are then used to generate false queries. The add-on gives users the option to choose how to forward such queries. In particular, a user may send bursts of bogus queries, thus mimicking the way people search, or may submit them at predefined intervals of time. Despite the strategies users have at their disposal, TrackMeNot is vulnerable to a number of attacks that leverage on the semantics of these false queries as well as timing information, to distinguish them from the genuine queries [104].

GooPIR [105] is another proposal aimed at obfuscating query profiles. Implemented as a software program [(r)], this approach enables users to conceal their search keywords by adding some false keywords. To illustrate how this approach works, consider a user wishing to submit the keyword "depression" to Google and willing to send it together with two false keywords. Based on this information, GooPIR would check the popularity of the original keyword and find that "iPhone" and "elections" have a similar frequency of use. Then, instead of submitting each of these three keywords at different time intervals, this approach would send them in a batch. The proposed strategy certainly thwarts attacks based on timing. However, its main limitation is that it cannot prevent an attacker from combining several of these batches, establishing correlations between keywords, and eventually inferring the user's real interest [106]. As an example, suppose that the user's next query is "prozac" and that GooPIR recommends submitting it together with the keywords "shirt" and "eclipse". In this case, one could easily deduce that the user is interested in health-related issues.

Another form of perturbation [107] consists in hiding certain categories of interests. In this work, user profiles are organized in a hierarchy of categories in such a way that lower-levels categories are regarded as more specific than those at higher levels. Based on this user-profile model, the idea is to disclose only those parts of the user profile corresponding to high-level interests.

[(r)] `http://unescoprivacychair.urv.cat/goopir.php`

In the case of perturbative methods for recommendation systems, [94] proposes that users add random values to their ratings and then submit these perturbed ratings to the recommender. After receiving these ratings, the system executes an algorithm and sends the users some information that allows them to compute the prediction. When the number of participating users is sufficiently large, the authors find that user privacy is protected to a certain extent and the system reaches a decent level of accuracy. However, even though a user disguises all their ratings, it is evident that the items themselves may uncover sensitive information. Simply put, the mere fact of showing interest in a certain item may be more revealing than the rating assigned to that item. For instance, a user rating a book called "How to Overcome Depression" indicates a clear interest in depression, regardless of the score assigned to this book. Apart from this critique, other works [108,109] stress that the use of *randomized* data distortion techniques might not be able to preserve privacy.

In line with these two latter works, [110] applies the same perturbative technique to collaborative-filtering algorithms based on singular-value decomposition. More specifically, the authors focus on the impact that their technique has on privacy. For this purpose, they use the privacy metric proposed by [111], which is essentially equivalent to differential entropy, and conduct some experiments with data sets from Movielens [(s)] and Jester [(t)]. The results show the trade-off curve between accuracy in recommendations and privacy. In particular, they measure accuracy as the mean absolute error between the predicted values from the original ratings and the predictions obtained from the perturbed ratings.

2.4 Privacy Metrics

In this section we review the state of the art in privacy metrics. We proceed by exploring, first, those metrics used in the application fields of SDC, ACSs and LBSs; and secondly, we examine those privacy measures specifically intended for personalized information systems.

[(s)] `http://movielens.umn.edu`

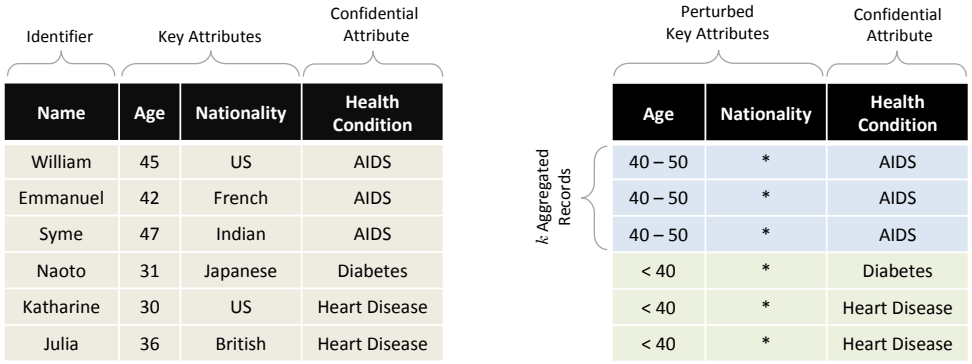
[(t)] `http://eigentaste.berkeley.edu/user/index.php`

2.4.1 Statistical Disclosure Control

Traditionally, institutes and governmental statistical agencies have systematically gathered information about individual respondents, either people or companies, with the aim of distributing this information to the research community [112]. Commonly, statistical agencies make this information public by releasing a *microdata set*, essentially a database table whose records carry data concerning said respondents. While these databases may be extremely useful for researchers, it is fundamental that their publication not compromise the respondents' privacy in the sense of revealing information about specific individuals. With this purpose, considerable effort has been devoted to the development of privacy-protecting mechanisms to be applied to the microdata sets before their release. SDC [113] is, precisely, the research area that deals with the inherent trade-off between protecting the privacy of the respondents and ensuring that those data are still useful for researchers.

Usually, a microdata set contains a set of attributes that may be classified into *identifiers*, *key attributes* or *quasi-identifiers*, or *confidential attributes*. First, identifiers allow to unequivocally identify individuals. It would be the case of social security numbers or full names, which would be removed before the publication of the microdata set. Secondly, key attributes are those attributes that, in combination, may be linked with external information to reidentify the respondents to whom the records in the microdata set refer. Examples include job, address, age, gender, height and weight. Last but not least, the microdata set contains *confidential attributes* with sensitive information on the respondent, such as salary, religion, political affiliation or health condition.

With the aim of protecting the privacy of the individuals appearing in a microdata set and, at the same time, preserving the usefulness of those data, the SDC community has proposed a wide range of mechanisms [53, 71–74]. In essence, these mechanisms rely upon some form of perturbation that permits enhancing privacy to a certain extent, at the cost of losing some of the data utility with respect to the unperturbed version. In order to assess the effectiveness of such mechanisms, numerous privacy metrics have been investigated.



(a) Original data (b) Perturbed data

Figure 2.4: We apply generalization and suppression to the key attributes age and nationality respectively, in such a manner that the requirement of 3-anonymity is satisfied. The upshot of this perturbation is that each tuple of key attributes in the released table (b) is shared by at least 3 records. This means that an attacker who knows the key-attribute values of a particular respondent cannot ascertain the record of this respondent beyond a subgroup of 3 records in the original table (a) and in any public database with identifier attributes.

Probably, the best-known privacy metric is *k-anonymity* [23, 24], which is the requirement that each tuple of key-attribute values be shared by at least k records in the database. This condition may be achieved through the mechanisms of generalization and suppression, as illustrated by the example depicted in Fig. 2.4, where age and nationality are regarded as key attributes, and health condition as a confidential attribute. Rather than making the original table available, we publish a k-anonymous version containing aggregated records, in the sense that all key-attribute values within each group are replaced by a common representative tuple. As a result, a record cannot be unambiguously linked to the corresponding record in the original table or, more generally, to any public database containing identifier attributes. Consequently, k-anonymity is said to protect microdata against *linking attacks*.

Unfortunately, while this criterion prevents identity disclosure, it may fail against the disclosure of the confidential attribute. Precisely, the definition of this privacy criterion establishes that complete reidentification is unfeasible within a group of records sharing the same tuple of perturbed key-attribute values. However, if the records in the group also share a common value of a confidential attribute, the association between an individual linkable to the group of perturbed key attributes and the

corresponding confidential attribute remains disclosed. More specifically, consider the example depicted in Fig. 2.4 and suppose that a privacy attacker knows Emmanuel's key-attribute values. If the attacker learned that Emmanuel is included in the released table, then the attacker might conclude that the individual in question suffers from AIDS even though such attacker is not able to ascertain which record belongs to this individual. This is known as *homogeneity attack*. Now suppose that the adversary strives to infer the confidential-attribute value of Naoto, who belongs to a group in which the distribution of this confidential-attribute value is not completely homogeneous. Even in this case, the adversary could exploit the fact that the Japanese have a low incidence of heart disease and, hence, it could be deduced that this individual is more likely to have diabetes. Such attack is known as *background-knowledge attack*.

Despite these two attacks, the main issue with k-anonymity as a privacy criterion is its vulnerability against the exploitation of the difference between the prior distribution of confidential data in the entire population, and the posterior conditional distribution of a group given the observed, perturbed key attributes. For example, imagine that the proportion of respondents with heart disease is much higher than that in the overall data set. This is normally referred to as a *skewness attack*.

All these vulnerabilities motivated the appearance of enhanced privacy criteria, some of which we proceed to sketch briefly. A restriction of k-anonymity called *p-sensitive* k-anonymity was presented in [25, 26]. In addition to the k-anonymity requirement, it is required that there be at least p different values for each confidential attribute within the group of records sharing the same tuple of perturbed key-attribute values. Clearly, large values of p may lead to huge data utility loss. A slight generalization called *l-diversity* [27, 28] was defined with the same purpose of enhancing k-anonymity. The difference with respect to p-sensitivity is that the group of records must contain at least l "well-represented" values for each confidential attribute. Depending on the definition of well-represented, l-diversity can reduce to p-sensitive k-anonymity or be more restrictive. Concretely, a microdata is said to meet the entropy l-diversity requirement if, for each group of records with the same tuple of perturbed key-attribute values, the entropy of the distribution of the confidential-attribute value within the group is at least $\log l$. We would like to stress

that neither of these enhancements succeeds in completely removing the vulnerability of k-anonymity against skewness attacks. Further, they are still susceptible to *similarity attacks*, in the sense that while confidential-attribute values within a cluster of aggregated records might be p-sensitive or l-diverse, they might also very well be semantically similar. For example, consider the confidential-attribute values to be lung cancer, prostate cancer or bladder cancer, compared to other, noncancerous diseases.

In an attempt to overcome all these deficiencies, *t-closeness* [29] was proposed. A perturbed microdata set satisfies t-closeness if for each group sharing a common tuple of perturbed key-attribute values, some measure of distance between the posterior distribution of the confidential attributes in the group and the prior distribution of the overall population does not exceed a threshold t. As argued in [114], to the extent to which the within-group distribution of confidential attributes resembles the distribution of those attributes for the entire dataset, skewness attacks will be thwarted. In addition, since the within-group distribution of confidential attributes mimics the distribution of those attributes over the entire dataset, no semantic similarity can occur within a group that does not occur in the entire dataset. The main limitation of the original t-closeness work [29], however, is that no computational procedure to reach t-closeness was specified.

An information-theoretic privacy criterion, inspired by t-closeness, was proposed in [32]. In the latter work, privacy risk is defined as the conditional KL divergence between the posterior and the prior distributions, a measure that may be regarded as an average-case version of t-closeness. Particularly, this average privacy risk is shown to be equal to the mutual information between the confidential attributes and the observed, perturbed key attributes. A related criterion named *δ-disclosure* is proposed in [30], a worst-case version that measures the maximum absolute log ratio between the prior and the posterior distributions. Lastly, [31] analyzes privacy for interactive databases, where a randomized perturbation rule is applied to a true answer to a query, before returning it to the user. Consider two databases that differ only by one record, but are subject to a common perturbation rule. Conceptually, the randomized perturbation rule is said to satisfy the *ϵ-differential privacy* criterion if

the two corresponding probability distributions of the perturbed answers are similar, according to a certain inequality.

2.4.2 Anonymous-Communication Systems

In the literature of ACSs, many proposals focus on measuring the extent to which these systems provide anonymity guarantees. A key point is that the degree of anonymity achieved by these systems depends on the capabilities of the adversary, and often anonymity metrics are tailored to the corresponding assumptions. A complete study on adversary models for these systems may be found in [115]. Next, we review the most relevant anonymity metrics in the field of anonymous communications.

In the important case of the mix systems examined in Sec. 2.3.2, [9] defined the *anonymity set* of users as the set of possible senders of a given message, or recipients, in the sense that the likelihood of them fulfilling the role in question is nonzero. A simple measure of anonymity was proposed by [33], namely the logarithm of the number of users involved in the communication, that is, the Hartley entropy of the anonymity set. The main drawback of this metric is that it does not contemplate the probabilistic information that an adversary may obtain about users when observing the system. In other words, this approach ignores the fact that certain users may be more likely to be the senders of a particular message.

Several approaches have considered the use of information-theoretic quantities to evaluate ACSs. The most significant are those proposed in [34, 35], in which the degree of anonymity observable by an adversary is measured essentially as the Shannon entropy of the probability distribution of possible senders of a given message. A well-known interpretation of Shannon's entropy refers to the game of 20 questions, in which one player must guess what the other is thinking through a series of yes/no questions, as quickly as possible. Informally, Shannon's entropy is a lower bound on —and often good approximation to the minimum of— the average number of binary questions regarding the nature of possible outcomes of an event, to determine which one in fact has come to pass, intelligently exploiting their known probabilities. The use of entropy as a measure of privacy, however, is by no means new. As a matter of fact, Shannon's work in the fifties introduced the concept of *equivocation* as the

conditional entropy of a private message given an observed cryptogram [116], later used in the formulation of the problem of the wiretap channel [117, 118] as a measure of confidentiality.

Still in the case of information-theoretic measures, [36] formalizes the notion of unlinkability by using Shannon's entropy. By contrast, [37, 38] argue that a worst-case metric should be considered instead of Shannon's entropy, since the latter contemplates an average case. The authors refer to this worst-case metric as *local anonymity*, essentially equivalent to min-entropy, and concordantly define the *source hiding* property as the requirement that no sender probability exceed a given threshold. Another approach [39] proposes a method for quantifying the property of *relationship anonymity*, as defined in [119]. More specifically, the authors make use of Shannon's entropy and min-entropy for measuring this property. Similarly, [40] evaluates Shannon's entropy, min-entropy and Hartley's entropy as anonymity metrics, and proposes then to use Rényi's entropy, which may be regarded as a generalization of those three metrics.

Besides Hartley's entropy, other possibilistic —rather than probabilistic— approaches include [41–43]. According to these metrics, subjects are considered anonymous if an adversary cannot determine their actions with absolute certainty. Further, [44] proposes a combinatorial anonymity metric that counts the number of possible one-to-one correspondences between a set of senders and a set of receivers, by means of the permanent of the matrix of adjacencies of the associated bipartite graph, consistent with message timing observations ruling out some of the permutations. It must be stressed that probability distributions weighting such possibilities are not considered, or from a mathematically equivalent perspective, that those probabilities are considered equally likely. Another difference with respect to most metrics based on probabilities is that this metric is directly defined on a group of consistent matchings between senders and receivers, rather than defined on the set of senders or receivers corresponding to one given message. Some limitations and extensions of this approach may be found in [120].

2.4.3 Personalized Information Systems

As discussed in Sec. 2.1, personalized information systems rely on some form of profiling to provide information tailored to users' preferences. Said otherwise, personalization comes at the risk of profiling. The literature of privacy metrics in this particular scenario typically measures user privacy based on the profile constructed by an attacker. Potential privacy attackers include the systems themselves but also any other entity capable of eavesdropping the information users reveal to such systems. As we shall see next, most of the proposed metrics quantify user privacy according to two profiles. The former is the profile capturing the genuine interests of a user, and the latter the profile observed by the attacker. In principle, the observed profile does not need to coincide with the original one. This may be as a result of adopting any of the PETs reviewed in Sec. 2.3.2, or even simpler, due to cookies being disabled for a period of time.

In the context of personalized Web search, [96] proposes PRAW, a system aimed at preserving the privacy of a group of users sharing an access point to the Web. The cited work and its successive improvements [97–101] suggest perturbing the actual user profile by generating fake transactions, that is, accesses to Web pages. In the PRAW system, user profiles are modeled as weighted vectors of queries, and privacy is computed as the similarity between the genuine profile and that observed from the outside. More specifically, the authors use the cosine measure to capture the similarity between both profiles. They assume, accordingly, that the lower the cosine similarity value between these two profiles, the higher the privacy level attained by such perturbation strategy.

Similarly to those works, [121] proposes to measure privacy as a generic function of both the actual profile and the profile observed by a recommender. The authors acknowledge that this function may, in principle, be different for each user, as users may perceive privacy risks differently. Their metric is justified in the same way as in the PRAW system. That is, it is assumed that the more those profiles differ, the higher the privacy protection. Then, a weighted version of the Euclidean distance is given as a particular instantiation of the generic function.

In the literature we also find examples of privacy criteria based on information-theoretic quantities. For example, [95] measures privacy risk as the relative entropy between the user's query distribution and the population's. In the context of personalized Web search, [102] identifies two privacy breaches when submitting search queries. The former refers to the disclosure of identifying information, e.g., asking Google Maps [u] how to get from your home to a restaurant. The latter refers to private information inferred indirectly from such queries, e.g., estimating the probability of suffering from a disease based on searches for medical assistance. The authors propose the injection of false queries to counter the latter kind of privacy breach, and measure privacy as the mutual information between the real queries X and the observed ones Y. Accordingly, when $\mathrm{I}(X;Y)$ is zero, the observed profile does not leak any information about the actual profile, and perfect privacy protection is attained.

Still in the scenario of personalized Web search, [89] defines a privacy criterion called *profile exposure level*. This criterion uses the mutual information between the genuine queries of a given user and the queries submitted to the search engines, including the genuine ones and those forwarded by this user on behalf of their neighbors. Specifically, user privacy is measured as the quotient between the mutual information and the Shannon entropy of the distribution of original queries. In the end, the authors justify their metric by interpreting it as an amount of uncertainty reduction.

Another information-theoretic privacy criterion is [107]. In this approach, user profiles are represented essentially as normalized histograms of queries. The profile categories are organized hierarchically so that the higher-level interests are more general than those at the lower levels. According to this representation, the authors define user privacy based on two parameters, *minDetail* and *expRatio*. The former parameter is a threshold that is used to filter out those components of the profile where the user has shown little interest in. The latter is the Shannon entropy of the filtered profile, a quantity that is taken as the level of privacy achieved. Finally, other approaches using Shannon's entropy as privacy criterion include [90, 91].

[u] `http://maps.google.com`

Part I

Privacy Metrics

Chapter 3

Measuring Privacy as an Attacker's Estimation Error

3.1 Introduction

The widespread use of information and communication technologies to conduct all kinds of activities has in recent years raised privacy concerns. There is a broad diversity of applications with a potential privacy impact, from social networking platforms to e-commerce or mobile phone applications.

At the same time, a variety of PETs have emerged to support the provision of new services and functionalities while mitigating potential privacy threats. The privacy concerns arising in different applications are diverse and so are the corresponding privacy-enhanced solutions that address these concerns. Similarly, a wide range of privacy metrics have been proposed in the literature to evaluate the level of protection offered by PETs. However, most of these metrics are specific to concrete systems and adversary models and are difficult to generalize or translate to other contexts. Therefore, a better understanding of the relationships between the different privacy metrics would enable a more grounded and systematic approach to measuring privacy, and would assist system designers in selecting the most appropriate metric for a given application.

In this chapter we propose a theoretical framework for privacy-preserving systems, endowed with a general definition of privacy in terms of the estimation error incurred by an attacker who aims to disclose the private information that the system is designed to conceal. Further, we show that the most widely used privacy metrics, such as k-anonymity, l-diversity, t-closeness, ϵ-differential privacy, as well as information-theoretic metrics such as Shannon's entropy, min-entropy, or mutual information, may be construed as particular cases of the estimation error. In a nutshell, our framework permits interpreting and comparing a number of well-known metrics under a common perspective.

The importance of privacy metrics, accompanied with utility metrics, lies in the fact that they provide a quantitative means of comparing the suitability of two or more privacy-enhancing mechanisms, in terms of the privacy-utility trade-off posed. Ultimately, such metrics enable us to systematically build privacy-aware information systems by formulating design decisions as optimization problems, solvable theoretically or numerically, capitalizing on a rich variety of mature ideas and powerful techniques from the wide field of optimization engineering.

In our interpretations of state-of-the-art privacy metrics as particular cases of the estimation error, we illustrate how the general framework can be instantiated in three very different areas of application, namely SDC, anonymous communications and LBSs.

In SDC, a great effort has been devoted to the investigation of privacy metrics. Sec. 2.4.1 already mentioned that the best-known metric is *k-anonymity*, which was first proposed in [23, 24]. In an attempt to address the weaknesses of this proposal, various extensions and enhancements were introduced later in [25,27,29–32]. While all these proposals have contributed to some extent to the understanding of the privacy requirements of this field, the SDC research community would undoubtedly benefit from the existence of a rule that could help them decide which privacy metric is the most suitable for a particular application.

In anonymous communications, one of the goals is to conceal who talks to whom against an adversary who observes the inputs and outputs of the communication channel. In Sec. 2.4.2 we introduced mixes as a fundamental component of anonymous

communications. In essence, mixes are systems that encrypt, pad, delay and reorder messages so that it is not possible to correlate their inputs and outputs. Among these four strategies, delaying messages causes the most noticeable impact on the usability of the system. However, such strategy allows at the same time for stronger levels of privacy protection. In other words, there is a trade-off between anonymity (privacy) and delay (utility), and the only way to tackle the problem of designing mix systems in an optimal trade-off sense, is to be equipped with quantifiable measures of both anonymity and utility.

In the end, we approach the particularly rich, important example of LBSs, where users submit queries along with the location to which those queries refer. In this scenario, a wide range of approaches have been proposed, many of them based on an intelligent perturbation of the user coordinates submitted to the provider [122]. Basically, users may contact an *untrusted* LBS provider directly, perturbing their location information so as to hinder providers in their efforts to compromise user privacy in terms of location, although clearly not in terms of query contents and activity, and at the cost of an inaccurate answer. In short, this approach presents again the inherent trade-off between data utility and privacy common to any perturbative privacy mechanism.

The connection between state-of-the-art privacy metrics and information theory, and the mathematical unification of these metrics as an attacker's estimation error presented in this chapter shed new light on the understanding of those metrics and their suitability when it comes to applying them to specific scenarios. We also hope to illustrate the riveting intersection between the fields of information privacy and information theory, in an attempt towards bridging the gap between the respective communities. Moreover, the fact that our metric boils down to an estimation error opens the possibility of applying notions and results from the mature, vast field of estimation theory [123].

The work presented in this chapter was published in [48].

Chapter Outline

The rest of this chapter is organized as follows. Sec. 3.2 introduces some background on Bayes decision theory (BDT). Then, Sec. 3.3 describes our notation, terminology and adversary model, and afterwards presents our measure of privacy. Secs. 3.4 and 3.6 are devoted to the classification of several privacy metrics, showing the relationships with our proposal and the correspondence with assumptions on the attacker's strategy. While the former section approaches this from a theoretical perspective, the latter illustrates the applicability of our framework to help system designers choose the appropriate metrics, without having to delve into the mathematical details. Sec. 3.5 provides two numerical examples that illustrate our formulation and the measurement of privacy as an attacker's estimation error. Finally, conclusions are drawn in Sec. 3.7.

3.2 Background on Bayes Decision Theory

In this section, we shall introduce some elementary concepts for those readers who are not familiar with BDT.

BDT is a statistical method that, fundamentally, uses a probabilistic model to analyze the making of decisions on uncertainties and the costs associated with those decisions [124, 125]. In general, Bayes decision principles may be formulated in the following terms. Consider the uncertainty refers to an *unknown* parameter modeled by an r.v. X. In decision-theoretic terminology, this is also known as *state of nature*. Let Y be another r.v. modeling an *observation* or measurement on the state of nature. Suppose that, given a particular observation y, we are required to make a decision on the unknown. Let $\hat{x}$ denote the estimator of X, that is, the rule that provides a decision or estimate $\hat{x}(y)$ for every possible observation y. Clearly, any decision will be accompanied by a cost. This is captured by the *loss function* $d\colon (x, \hat{x}) \mapsto d(x, \hat{x})$, which measures how costly the decision $\hat{x} = \hat{x}(y)$ will be when the unknown is x. However, since the actual loss incurred by a decision cannot be calculated with absolute certainty at the time the decision is made, BDT contemplates the average loss associated with this decision. Concretely, the *Bayes conditional risk* for

an estimator $\hat{x}$ is defined in the discrete case as

$$\mathcal{R}(y) = \mathrm{E}[d(X, \hat{x}(y))|y] = \sum_x p_{X|Y}(x|y)\, d(x, \hat{x}(y)),$$

where the expectation is taken over the *posterior* probability distribution $p_{X|Y}$. According to this, the *Bayes risk* associated with that estimator is defined as the average of the Bayes conditional risk over all possible observations y, that is,

$$\mathcal{R} = \mathrm{E}\,\mathrm{E}[d(X, \hat{x}(Y))|Y] = \sum_{x,y} p_{X\,Y}(x, y)\, d(x, \hat{x}(y)),$$

where the expectation is additionally taken over the probability distribution of Y. Based on this definition, an estimator is called *Bayes estimator* or *Bayes decision rule*, if it minimizes the Bayes risk among all possible estimators. It turns out that this optimal estimator is precisely

$$\hat{x}_{\mathrm{Bayes}}(y) = \arg\min_{\hat{x}} \mathrm{E}[d(X, \hat{x})|y],$$

for all y; i.e., the Bayes estimator is the one that minimizes the Bayes conditional risk for every observation.

Once some of the basic elements in Bayes analysis have been examined, we would like to establish a connection between maximum a posteriori (MAP) estimator and Bayes estimator. With this aim, first recall that a MAP estimator, as the name implies, is the estimator that maximizes the posterior distribution. Now consider the loss function d to be the Hamming distance between x and $\hat{x}$. The Hamming distance, which we introduced in Sec. 2.2, is in fact an indicator function. But recall that the expectation of an indicator r.v. is the probability of the event it is based on. Mathematically,

$$\mathrm{E}[d_{\mathrm{Hamming}}(X, \hat{x})|y] = \mathrm{P}\{X \neq \hat{x}|y\},$$

and consequently,

$$\hat{x}_{\mathrm{MAP}}(y) = \arg\min_{\hat{x}} \mathrm{P}\{X \neq \hat{x}|y\} = \arg\max_{\hat{x}} \mathrm{P}\{X = \hat{x}|y\}. \tag{3.1}$$

In conclusion, Bayes and MAP estimators coincide when the loss function is Hamming distance.

3.3 Measuring Privacy as an Attacker's Estimation Error

This section presents a general framework that lays the foundation for the establishment of a unified measurement of privacy. However, it is not until Sec. 3.4 where we shall show that a number of privacy criteria may be regarded as particular cases of our proposal. Previously, Sec. 3.3.1 introduces our notation. Next, Sec. 3.3.2 describes the adversary model. In Sec. 3.3.3 we present our privacy metric, and finally, in Sec. 3.3.4, we illustrate the proposed formulation with a simple but insightful example.

3.3.1 Mathematical Assumptions and Notation

In this section we provide the notation that we shall use throughout this chapter. To this end, we first introduce the key actors of the proposed framework:

- a *user*, who wishes to protect their privacy;
- a (trusted) *system*, to which each user entrusts their private data for its protection; the unique purpose of this entity is to guarantee the privacy of the user, and with this aim, the system may use any privacy-preserving mechanism at its disposal;
- and an *attacker*, who strives to disclose private information about this user.

To clarify the elements involved in our framework, consider a conceptually-simple approach to anonymous Web browsing, consisting in a TTP acting as an intermediary between Internet users and Web servers. From the perspective of our model, the users would be those subscribed to the anonymous proxy; the system would be this proxy; and the attackers those servers that attempt to compromise users' privacy from their Web browsing activity.

In the following, the term r.v. is used with full generality to include categorical or numerical data, vectors, tuples or sequences of mixed components, but for mathematical simplicity we shall henceforth assume that all r.v.'s in this chapter have finite alphabets. The variables that constitute our framework are described as follows.

- The *attacker's unknown* or *uncertainty* is denoted by the r.v. X, which models the private information about a user that the attacker wishes to ascertain.

- The *system's input* is represented by the r.v. X' and refers to user's data required by the system to make a decision.

- The *system's decision* is modeled by the r.v. Y' and denotes disclosed information, perhaps part of X', or a perturbation.

- The *attacker's input* is denoted by the r.v. Y and captures any evidence or measurement the attacker has about the unknown. As its name indicates, this variable models the information that serves as input for the adversary to ascertain X. In some cases, Y may be directly the information revealed by the system, i.e., $Y = Y'$. That is, the only information available to the attacker is exactly that disclosed by the system. In other circumstances, the attacker may observe a perturbed version of Y', maybe together with background knowledge about the unknown. In such cases, we have $Y \neq Y'$. Since the attacker's input is, in fact, the information *observed* by the attacker, directly from the system or indirectly from other sources, throughout this work we shall use the terms *attacker's input* and *attacker's observation* indistinguishably to refer to the variable Y.

- The *attacker's decision* is modeled by the r.v. $\hat{X}$ and represents the attacker's estimate of X from Y.

Table 3.1: Simplified representation of our notation.

	Unknown	**Input**	**Decision**
Attacker	X	Y	$\hat{X}$
System	-	X'	Y'

In order to clarify this notation, we provide an example in which the above variables are put in the context of SDC. In this scenario, the data publisher plays the role of the system. Concretely, X may represent identifying or confidential-attribute

values the attacker endeavors to ascertain with regard to an individual appearing in a released table. The individuals contained in this table are what we call users. The system's input becomes now the key-attribute values that the publisher has about the individuals. On the other hand, Y' is the perturbed version of those values, which jointly with the (unperturbed) confidential-attribute values, constitute the released table. Furthermore, the attacker's input consists of the released table and, possibly, background knowledge the privacy attacker may have. In the end, the attacker's decision is the estimate of X. All this information is shown in Table 3.2.

Similarly, now we specify the variables of our framework in the special case of a mix. Under this scenario, the mix represents the system, whose objective is to hide the correspondence between the incoming and outgoing messages. Precisely, the attacker's uncertainty is this correspondence. The system's input and system's decision are the arrival and departure times of the messages, respectively. On the other hand, the information available to the attacker, i.e., the attacker's observation Y, consists of X', Y' and the design parameters of the mix. Finally, $\hat{X}$ is the attacker's decision on the correspondence between the messages. This is depicted in Fig. 3.1 and summarized in Table 3.3.

Table 3.2: Description of the variables used in our notation in the special case of SDC.

	Unknown	**Input**	**Decision**
Attacker	identifier or confidential attributes	perturbed table, possibly with background knowledge	estimate of identifier or confidential attributes
System	-	key attributes	perturbed key attributes

3.3.2 Adversary Model

The consideration of a framework that encompasses a variety of privacy criteria necessarily requires the formalization of the attacker's model. In this spirit, we now proceed to present the parameters that characterize this model.

Firstly, we shall contemplate an adversary model in which the attacker uses a Bayes (best) decision rule. Conceptually, this corresponds to the estimation made by an attacker who uses optimally the available information, as we formally argued

in Sec. 3.2. Namely, for every possible decision of the system resulting in an observation y, the attacker will make a Bayes decision $\hat{x}(y)$ on X. With regard to this attacker's decision rule, we would like to remark the fact that, whereas it is a deterministic estimator, the system's decision is assumed to be a *randomized* perturbation rule given by $p_{Y'|X'}$. As a consequence of this, it is clear that the system does not leak any private information when deciding Y', provided that Y' and X' are statistically independent.

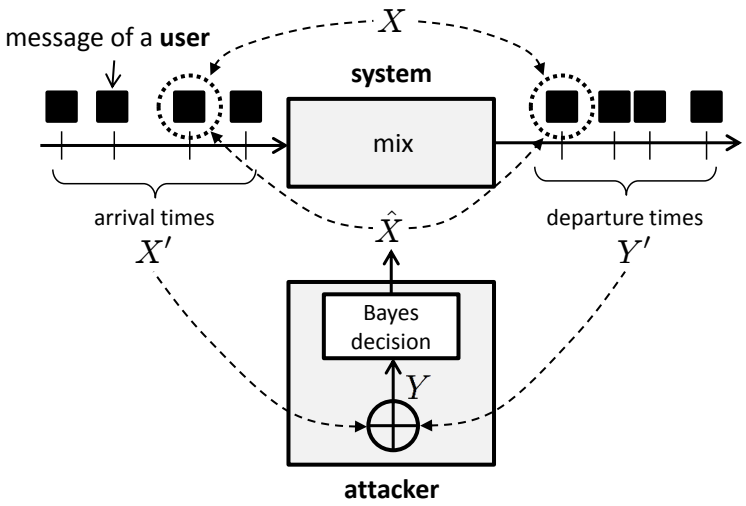


Figure 3.1: Our framework is put in the context of mixes.

Secondly, as explained in Sec. 3.2, we shall require to evaluate the cost of each decision made by the attacker. For this purpose, we consider the *attacker's distortion function* $d_{\mathrm{A}}\colon (x, \hat{x}) \mapsto d_{\mathrm{A}}(x, \hat{x})$, which measures the degree of dissatisfaction that the attacker experiences when $X = x$ and $\hat{X} = \hat{x}(y)$. Similarly, we contemplate the *system's distortion function* $d_{\mathrm{S}}\colon (x', y') \mapsto d_{\mathrm{S}}(x', y')$, which reflects the extent to which the system, and therefore the user, is discontent when $Y' = y'$ and $X' = x'$.

A crucial distinction in the type of attacker's distortion function d_{A} considered will be whether it captures a sort of geometry over the symbols of the alphabet, or not. The most evident example of distortion function that does not take into account this geometry is the Hamming function, which we already introduced at the end of Sec. 2.2. Concretely, this binary metric just indicates whether x and $\hat{x}$ coincide, and provides no more information about the discrepancy between them. On the

Table 3.3: Description of the variables used in our notation in the special case of mixes.

	Unknown	**Input**	**Decision**
Attacker	correspondence between incoming and outgoing messages	arrival and departure times of the messages, mix design parameters and maybe background knowledge	estimate of correspondence between incoming and outgoing messages
System	-	arrival times of the messages	departure times of the messages

other hand, the squared error loss $d_{\mathrm{A}}(x, \hat{x}) = (x - \hat{x})^2$ and the absolute error loss $d_{\mathrm{A}}(x, \hat{x}) = |x - \hat{x}|$ are just two commonly-used examples of distortion functions that do rely or induce a certain geometry.

3.3.3 Privacy-Metric Definition

Bearing in mind the above considerations, and consistently with Sec. 3.2, we define *conditional privacy* as

$$\mathcal{P}(y) = \mathrm{E}[d_{\mathrm{A}}(X, \hat{x}(y))|y], \tag{3.2}$$

which is the estimation error incurred by the attacker, conditioned on the observation y. Based on this definition, we contemplate two possible measures of privacy. In particular, we define *worst-case privacy* as

$$\mathcal{P}_{\min} = \min_{y} \mathcal{P}(y). \tag{3.3}$$

On the other hand, we define *average privacy* as

$$\mathcal{P}_{\mathrm{avg}} = \mathrm{E}\,\mathcal{P}(Y) = \mathrm{E}\,d_{\mathrm{A}}(X, \hat{x}(Y)), \tag{3.4}$$

which is the average of the conditional privacy over all possible observations y.

In order to measure the utility loss caused by the perturbation of the original data, we define the *average distortion* as

$$\mathcal{D} = \mathrm{E}\,d_{\mathrm{S}}(X', Y'). \tag{3.5}$$

According to these definitions, a privacy-protecting system and an attacker would adopt the following strategies. Namely, the system would select the decision rule $p_{Y'|X'}$ that maximizes either the average privacy or the worst-case privacy, while not allowing the average distortion to exceed a certain threshold. On the other hand, the

attacker would choose the Bayes estimator, which would lead to the minimization of *both* measures of privacy. The reason behind this is that the Bayes estimator also minimizes the conditional privacy, as stated in Sec. 3.2.

In light of the definitions above, the functions d_{S} and d_{A} clearly give us a measure of distortion and privacy, respectively. In the former case, distortion is measured from the *system*'s point of view, whereas in the latter case, privacy is quantified from the standpoint of the *adversary*. Despite the focus given, one could contemplate an alternative definition of d_{A} so that both functions are defined from the system's perspective. For example, we could define an alternative privacy function measuring the degree of *satisfaction* experienced by the system when $X = x$ and $\hat{X} = \hat{x}$. It turns out that the theoretical analysis presented in Sec. 3.4 could be readily adapted to this case. However, we have preferred to emphasize the role of the adversary and thus consider the perception that they have about their own error when estimating the unknown.

In this line, we would also like to remark that a privacy risk $\mathcal{R}$ in lieu of $\mathcal{P}$ could be defined for $-d_{\mathrm{A}}(x, \hat{x}(y))$ instead of $d_{\mathrm{A}}(x, \hat{x}(y))$. An analogous argument justifies the use of utility instead of distortion.

Last but not least, we would also like to note that, in the special case when the unknown variable X models the identity of a user, our measure of privacy may be regarded, in fact, as a measure of anonymity.

3.3.4 Example

Next, we present a simple example that sheds some light on the formulation introduced in the previous sections.

For the sake of simplicity, consider $X' = X$, that is, the system's input is the confidential information that needs to be protected. Suppose that X is a binary r.v. with $\mathrm{P}\{X = 0\} = \mathrm{P}\{X = 1\} = 1/2$. In order to hinder privacy attackers in their efforts to ascertain X, for each possible outcome x, the system will disclose a perturbed version y'. Namely, with probability p the system will decide to reveal the complementary value of x, whereas with probability $1 - p$ no perturbation will be applied, i.e., $y' = x$. Note that, in this example, the system's decision rule is

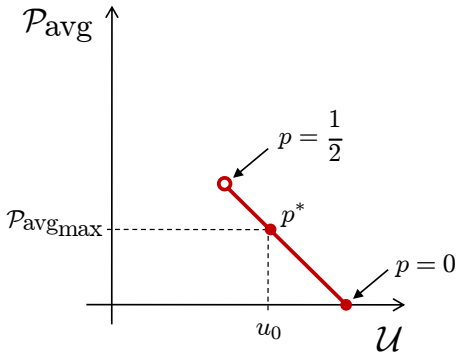


Figure 3.2: Representation of the trade-off curve between privacy and utility for the example provided in Sec. 3.3.4.

completely determined by p, for which we conveniently impose the condition $0 \leqslant p < 1/2$.

At this point, we shall assume that the attacker only has access to the disclosed information Y', and therefore the attacker's input Y boils down to it. We anticipate that, throughout this work, this supposition will be usual. In addition, we shall consider the attacker's distortion function to be the Hamming distance. However, as commented in Sec. 3.2, this implies that the Bayes estimator matches the MAP estimator. According to this observation, it is easy to demonstrate that the attacker's best decision is $\hat{X} = Y$. Therefore, the average privacy (3.4) becomes

$$\mathcal{P}_{\text{avg}} = \mathrm{P}\{X \neq \hat{X}\} = \mathrm{P}\{X \neq Y\} = \mathrm{P}\{X \neq Y'\} = p.$$

On the other hand, if we suppose that the system's distortion function is also the Hamming distance, from (3.5), it follows that

$$\mathcal{D} = \mathrm{P}\{X' \neq Y'\} = \mathrm{P}\{X \neq Y'\} = p.$$

Based on these two results, we now proceed to describe the strategy that the system would follow. To this end, we define the *average utility* $\mathcal{U}$ as $1 - \mathcal{D}$. According to this, the system would strive to maximize the average privacy with respect to p, subject to the constraint $\mathcal{U} \geqslant u_0$. Fig. 3.2 illustrates this simple optimization problem by showing the trade-off curve between privacy and utility. In this example, it is straightforward to verify that the optimal value of average privacy is $\mathcal{P}_{\text{avg}_{\text{max}}} = 1 - u_0$, for $1/2 < u_0 \leqslant 1$.

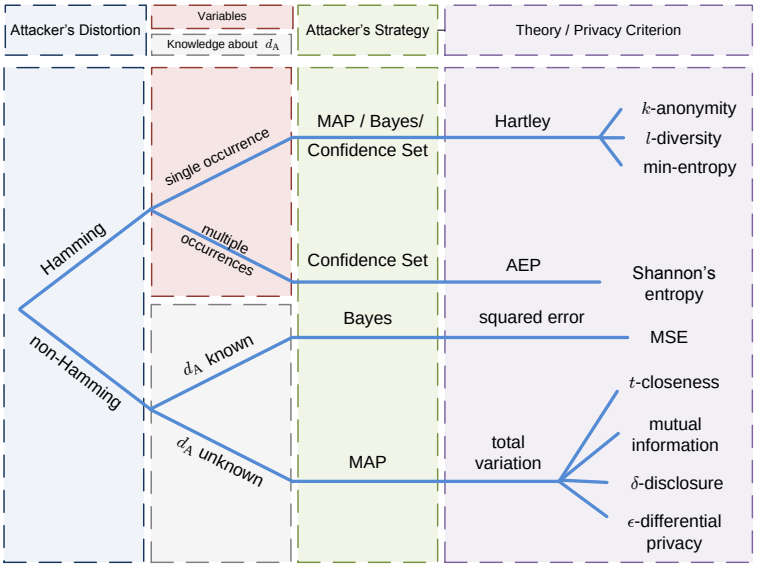


Figure 3.3: The arguments that lead to the interpretation of several privacy metrics as particular cases of our definition of privacy are conceptually organized in the above points. As can be observed, these arguments clearly depend on the attacker's distortion function, namely on the geometry of this function (Hamming or non-Hamming) and on the knowledge the system has about it, i.e., it is known or unknown to the system. Other parameters include the nature of the variables of our framework and, obviously, the attacker's strategy.

3.4 Theoretical Analysis

In this section we shall interpret several well-known privacy criteria as particular cases of our more general definition of privacy. Specifically, we shall show that many of the metrics examined in Chapter 2 are bijectively related to an estimation error and thus equivalent to our privacy measure—using a metric or a bijection of this metric is essentially the same, both in terms of comparison and optimization.

The arguments behind the interpretations of these metrics as a particularization of our criterion are based on numerous concepts from the fields of information theory, probability theory and BDT. For a comprehensive exposition of these arguments, the underlying assumptions and concepts will be expounded in a systematic manner, following the points sketched in Fig. 3.3. As mentioned in Sec. 3.3.2 and illustrated by the first branch of the tree depicted in this figure, our starting point makes the significant distinction between attacker's distortion measures based on the Hamming distance and the rest, according to whether we wish to capture a certain, gradual

measure of distance between alphabet values beyond sheer symbol equality. It is important to recall from Sec. 3.2 that in the case of a Hamming distortion measure, expected distortion boils down to probability of error, yielding a different class of estimation problems.

Bearing in mind the above remark, in Sec. 3.4.1 we shall contemplate the case when the attacker's distortion function is the Hamming distance, whereas in Sec. 3.4.2 we shall deal with the more general case in which d_{A} can be any other distortion function. In the special case of Hamming distance, we consider two alternatives for the variables in Table 3.1: single-occurrence and multiple-occurrence data. The former case considers the variables to be tuples of a small number of components, and the latter case assumes that these variables are sequences of data. In the scenario of single-occurrence data, we shall establish a connection between Hartley's entropy and our privacy metric, which will allow us to interpret k-anonymity, l-diversity and min-entropy criteria as particular cases of our framework. The arguments that will enable us to justify this connection stem from MAP estimation, BDT and the concept of confidence set. On the other hand, when we consider multiple-occurrence data, we shall use the asymptotic equipartition property (AEP) to argue that the Shannon entropy, as a measure of privacy, is a characterization of the cardinality of a high-confidence set of sequences.

In the more general case in which the attacker's distortion function is not the Hamming distance, we shall explore two possible scenarios. On the one hand, we shall consider the case where this function is known to the system. Under the assumption of a Bayes attacker's strategy, we shall use BDT to justify the system's best decision rule. On the other hand, we shall contemplate the case in which the attacker's distortion function is unknown to the system. Specifically, this scenario will allow us to connect our framework to several privacy criteria through the concept of total variation, provided that the attacker uses MAP estimation.

3.4.1 Hamming Distortion

In this section, we shall analyze the special case when the attacker's distortion function is the Hamming distance. In addition, we shall contemplate two cases for the variables of our framework: single-occurrence and multiple-occurrence data.

Single Occurrence

This section considers the scenario in which the variables defined in Sec. 3.3.1 are tuples of a relatively small number of components, including both categorical and numerical data, defined on a finite alphabet. In order to establish a connection between some of the most popular privacy metrics and our criterion, first we shall introduce the concept of confidence set and briefly recall a riveting generalization of Shannon's entropy.

Consider an r.v. X taking on values in the alphabet $\mathcal{X}$. A *confidence set* $\mathscr{F}$ with confidence p is defined as a subset of $\mathcal{X}$ such that $\mathrm{P}\{X \in \mathscr{F}\} = p$. In the case of continuous-valued random scalars, confidence sets commonly take the form of intervals. In these terms, it is clear that a privacy attacker aimed at ascertaining X will benefit the most from those confidence sets whose cardinality is reduced substantially with respect to the original alphabet size, with high confidence. To connect the concept of confidence set to our interpretation of privacy as an attacker's estimation error, consider an attacker model where the attacker only takes into account the shape of the PMF of the unknown X to identify a confidence set $\mathscr{F}$ for some desired confidence p, and beyond that, assumes all the included members equally relevant. This last assumption may be interpreted as an investigation on a tractable list of potential identities, carried out in parallel. MAP estimation within that set, considering it uniformly distributed, leads to an estimation error of $1 - \frac{1}{|\mathscr{F}|}$, that is, a bijection of its cardinality.

In our interpretations, we further use the Rényi entropy, a family of functionals widely used in information theory as a measure of uncertainty. Recall from Sec. 2.2

that Rényi's entropy of order α is defined as

$$\mathrm{H}_\alpha(X) = \frac{1}{1-\alpha} \log \sum_{i=1}^{n} p_X(x_i)^\alpha,$$

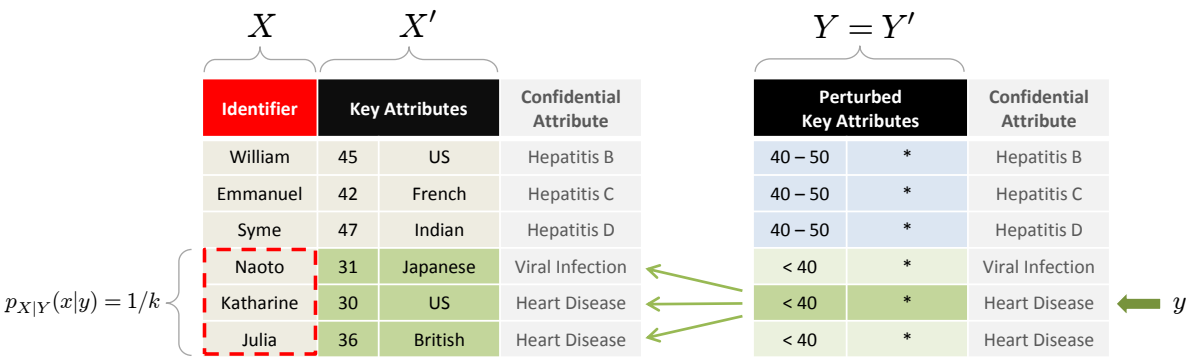
where p_X is the PMF of an r.v. X that takes on values in the alphabet $\mathcal{X} = \{x_1, \ldots, x_n\}$. Recall also that, in the special case when $\alpha = 0$, Rényi's entropy boils down to Hartley's entropy. Note that, when $p_X(x) > 0$ for all $x \in \mathcal{X}$, the Hartley entropy becomes $\mathrm{H}_0(X) = \log n$. Under this assumption, the Hartley entropy can be understood as a confidence set with $p = 100\%$. Lastly, $\mathrm{H}_1(X)$ and $\mathrm{H}_\infty(X)$ denote the Shannon entropy and min-entropy of the r.v. X, respectively.

We shall shortly interpret min-entropy, Shannon's entropy and Hartley's entropy within our general framework of privacy as an attacker estimation error, when Hamming distance is used as a distortion measure, first for single occurrences of a target information, and later for multiple occurrences. For now, we could loosely consider an attacker striving to ascertain the outcome of the finite-alphabet r.v. X, and the effect of the dispersion of its PMF on such task. Conceptually, we could then regard these three types of entropies simply as worst-case, average-case and best-case measurements of privacy, respectively, on account of the fact that

$$\mathrm{H}_\infty(X) \leqslant \mathrm{H}_1(X) \leqslant \mathrm{H}_0(X), \tag{3.6}$$

with equality if, and only if, X is uniformly distributed. More specifically, the min-entropy $\mathrm{H}_\infty(X)$ is the minimum of the *surprisal* or *self-information* $-\log p_X(x_i)$, whereas the Shannon entropy $\mathrm{H}_1(X)$ is a weighted average of such logarithms, and finally, the Hartley entropy $\mathrm{H}_0(X)$ optimistically measures the cardinality of the entire set of possible values of X regardless of their likelihood.

Once we have put the Hartley, Shannon and min entropies in the context of our framework, now we go on to describe a scenario that will allow us to relate our privacy metric to an extensively-used criterion. Specifically, we focus on the important case of SDC, where the data publisher plays the system's role. In this scenario, a data publisher wishes to release a microdata set and, before distributing it, the publisher applies some algorithm [25,27,29–32] to enforce the k-anonymity requirement [23,24]. As mentioned in Sec. 2.4.1, the objective of a linking attack is to unveil the identity



(a) Original data (b) Perturbed data

Figure 3.4: A data publisher plans to release a 3-anonymized microdata set. To this end, the publisher must enforce that, for a given tuple of key-attribute values in (b), the probability of ascertaining the identifier value of the corresponding record in (a) must be at most 1/3.

of the individuals appearing in a released table by linking the records in this table to any public data set including identifiers. Since k-anonymity is aimed at protecting the data against this attack, in our scenario the attacker's unknown X becomes the user identity. The other variables shown in Table 3.2 are as follows: X' are the key-attribute values, Y' are the perturbed key-attribute values, the attacker's observation Y is assumed to be Y', and finally, $\hat{X}$ is an estimate of the identity of a user. Although we consider $Y = Y'$, bear in mind that our interpretation of k-anonymity as an estimation error implicitly assumes that the adversary has access to any public database containing identifier attributes. Fig. 3.4 illustrates our notation.

In order to protect the data set from identity disclosure, the algorithm must ensure that, for any observation y consisting in a tuple of perturbed key-attribute values in the released table, the identifier value of the corresponding record in the original table cannot be ascertained beyond a subgroup of at least k records. As we shall see next, this requirement will be reflected mathematically by assuming that the probability distribution $p_{X|Y}(\cdot|y)$ of the identifier value, conditioned on the observation y, is the uniform distribution on a set of at least k individuals.

That said, our adversary model contemplates an attacker who uses a MAP estimator, which, as shown in Sec. 3.2, is equivalent to the Bayes estimator. Under this

model, given an observation y, the conditional privacy (3.2) becomes

$$\mathcal{P}(y) = \mathrm{P}\{X \neq \hat{x}(y)|y\} = 1 - \max_x p_{X|Y}(x|y), \tag{3.7}$$

which precisely is the MAP error ε_{MAP}, conditioned on that observation y; in terms of min-entropy, we may recast our metric as

$$\mathcal{P}(y) = \varepsilon_{\text{MAP}} = 1 - 2^{-\mathrm{H}_\infty(X|y)},$$

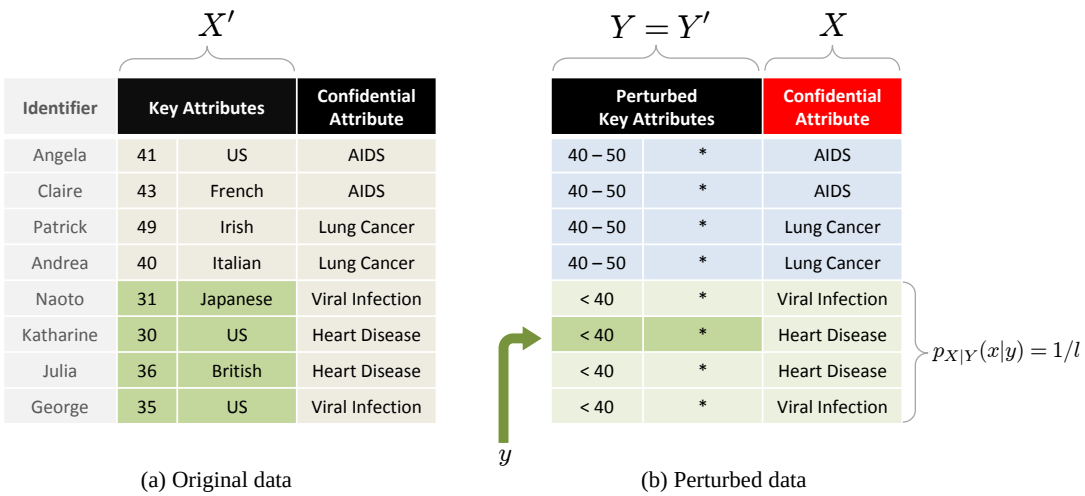
which shows that the concept of min-entropy is intimately related to MAP decoding. If we finally apply the aforementioned uniformity condition of $p_{X|Y}(\cdot|y)$, and assume that this PMF is the uniform distribution on a group of exactly k individuals, that is, $u_i = 1/k$ for all $i = 1, \ldots, k$, then

$$\mathcal{P}(y) = 1 - 1/k = 1 - 2^{-\mathrm{H}_0(X|y)},$$

which expresses the conditional privacy in terms of Hartley's entropy. In a nutshell, the k-anonymity criterion may be interpreted as a special case of our privacy measure, determined by this Rényi's entropy.

After examining this first interpretation, next we shall explore an enhancement of k-anonymity. As argued in Sec. 3.2, this criterion does not protect against confidential attribute disclosure. In an effort to address this limitation, several privacy metrics were proposed. In the remainder of this section, we shall focus on one of these approaches. In particular, we shall consider the l-diversity metric [27], which builds on the k-anonymity principle and aims at overcoming the attribute disclosure problem.

As commented in Sec. 2.4.1, a microdata set satisfies l-diversity if, for each group of records sharing a tuple of key-attribute values in the perturbed table, there are at least l "well-represented" values for each confidential attribute. In our new scenario, a data publisher, still playing the system's role, applies an algorithm on the microdata set to enforce this requirement. Since the aim of this criterion is to protect the data against attribute disclosure, we consider that the attacker's unknown X refers to the confidential attribute. The other variables remain the same as in our previous interpretation. Note, however, that we abandon the assumption that the attacker has access to any public database with identifiers—the adversary is not aimed at linking



(a) Original data (b) Perturbed data

Figure 3.5: In this example, the 2-diversity principle is applied to a microdata set. In order to meet this requirement, we assume that, for each group of records with the same tuple of perturbed key-attribute values, the probability distribution of the confidential-attribute value in (b) is the uniform distribution on a set of at least 2 values.

records between tables, but ascertaining the confidential-attribute value of a given record in the released table.

Having said that, we shall make the assumption that the l-diversity requirement is met by enforcing that, for a given tuple y of perturbed key-attribute values, the probability distribution $p_{X|Y}(\cdot|y)$ of the confidential attribute within the group of records sharing this tuple is the uniform distribution on a set of at least l values. This is depicted in Fig. 3.5. Note that this assumption entails that the data fulfill both the distinct and entropy l-diversity principles described in Sec. 2.4.1. Lastly, we shall suppose again that the attacker uses MAP estimator.

As mentioned before, under the premise of a MAP attacker, our measure of conditional privacy boils down to the MAP error (3.7). If we also apply the assumption above about the uniformity of $p_{X|Y}(\cdot|y)$, and suppose that this distribution is uniform on a group of l individuals, then the conditional privacy yields

$$\mathcal{P}(y) = 1 - 1/l = 1 - 2^{-\mathrm{H}_0(X|y)},$$

which expresses our privacy metric again in terms of Hartley's entropy. In short, the l-diversity criterion lends itself to be interpreted as a particular case of our more general privacy measure.

Multiple Occurrences

In this section, we shall consider the case when the variables shown in Table 3.1 are sequences of categorical and numerical data but in a finite alphabet. Recall from Sec. 2.2 that we use the notation X^k to denote a sequence $X_1, \ldots, X_k$.

The special case that we contemplate now could perfectly model the scenario in which a user interacts with an LBS provider, through an intermediate system protecting the user's location privacy. In this scenario, a user would submit queries along with their locations to the trusted system. An example would be the query "Where is the nearest parking garage?", accompanied by the geographic coordinates of the user's current location. As many approaches suggest in the literature of private LBSs, the system would perturb the user coordinates and submit them to the LBS provider. Concordantly, we may choose Euclidean distance as the natural attacker's distortion measure. Alternatively, if the attacker's interest lies in whether the user is at home, at work, shopping for groceries or at the movies, in order to profile their behavior, or more simply, whether the user is at a given sensitive location or not, then the appropriate model for the location space becomes discrete, and Hamming distance is more suited.

In this context, the consideration of sequences of discrete r.v.'s in our notation makes sense. Specifically, an attacker would endeavor to ascertain the sequence X^k of k unknown locations visited by the user, from the sequence Y'^k of k perturbed locations that the system would submit to the LBS. Put differently, the attacker's unknown would be the location data the user conveys to the system, i.e., $X^k = X'^k$, and the information available to the adversary the perturbed version of this data, that is, $Y^k = Y'^k$.

Having motivated the case of sequences of data, in this section we shall establish a connection between our metric and Shannon's entropy as a measure of privacy. But in order to emphasize this connection, first we briefly recall one of the pillars of information theory: the AEP [76], which derives from the weak law of large numbers and results in important consequences in this field.

Consider a sequence X^k of k independent, identically distributed (i.i.d.) r.v.'s, drawn according to p_X, with alphabet size n. Loosely speaking, the AEP states that

among all possible n^k sequences, there exists a *typical subset* $\mathscr{T}_\epsilon^k$ of sequences almost certain to occur. More precisely, for any $\epsilon > 0$, there exists a k sufficiently large such that $\mathrm{P}\{\mathscr{T}_\epsilon^k\} > 1 - \epsilon$, and $|\mathscr{T}_\epsilon^k| \leqslant 2^{k(\mathrm{H}_1(X)+\epsilon)}$. A similar argument called joint AEP [76] also holds for the i.i.d. sequences (X^k, Y^k) of length k drawn according to $\prod_{i=1}^k p_{XY}(x_i, y_i)$. Another information-theoretic result is related to those sequences x^k that are jointly typical with a given typical sequence y^k. Namely, the set of all these sequences x^k is referred to as the *conditionally typical set* $\mathscr{T}_\epsilon^{X^k|y^k}$ and satisfies, on the one hand, that $\mathrm{P}\{\mathscr{T}_\epsilon^{X^k|y^k}\} > 1 - \epsilon$ for large k, and on the other, that its cardinality is bounded by Shannon's conditional entropy, $|\mathscr{T}_\epsilon^k| \leqslant 2^{k(\mathrm{H}_1(X|Y)+\epsilon)}$. Further, it turns out that these conditionally typical sequences are equally likely, with probability $2^{-k\,\mathrm{H}_1(X|Y)}$, approximately in the exponent. While the most likely sequence may in fact *not* belong to the typical set, the set of typical sequences encompasses a sufficiently large number of sequences that amount to a probability arbitrarily close to certainty.

Next, we proceed to interpret, under the perspective of our framework, the Shannon entropy as a measure of privacy. To this end, consider the scenario in which a privacy attacker observes a typical Y^k and strives to estimate the unknown X^k. Conveniently, we assume $X^k = X'^k$ and $Y^k = Y'^k$, which models the LBS example described before, provided that the attacker ignores any spatial-temporal constraint. In other words, we model a scenario without memory and hence suppose that (X_i, Y_i) are i.i.d. drawn according to p_{XY}. We would like to stress that the consideration of this simplified model is just for the purpose of providing a simple, clear example that illustrates the application of our framework. Having said this, in the terms above we may regard $\mathscr{T}_\epsilon^{X^k|y^k}$ as a set of arbitrarily high confidence with cardinality $2^{k\,\mathrm{H}_1(X|Y)}$, approximately in the exponent.

The upshot is that the Shannon (conditional) entropy of an unknown r.v. (given an observed r.v.) is an approximate measure of the size of a high-confidence set, measure suitable for attacker models based on the estimation of sequences, rather than individual samples. Moreover, within this confidence set, sequences are equally likely, approximately in the exponent, concordantly with the interpretation of confidence-set cardinality as a measure of privacy made in Sec. 3.4.1 on single occurrences. Even

though for simplicity our argument focused on memoryless sequences, the Shannon-McMillan-Breiman theorem is a generalization of the AEP to stationary ergodic sequences, in terms of entropy rates [126].

3.4.2 Non-Hamming Distortion

This section investigates the complementary case described in Sec. 3.4 in which the attacker's distortion function is not the Hamming distance. Particularly, in this section we turn our attention to the scenario of SDC, and contemplate two possible alternatives regarding the system's knowledge on the function d_{A}—first, when this function is known to the data publisher, and secondly, when it is unknown. Under the former assumption, the system would definitely use BDT to find the decision rule $p_{Y'|X'}$ which maximizes either the worst-case privacy (3.3) or the average privacy (3.4), and satisfies a constraint on average distortion. The latter assumption, however, describes a more general and realistic scenario. The remainder of this subsection precisely interprets several privacy criteria under this assumption. The only piece of information which is though known to the publisher is $d_{\max} = \max_{x,\hat{x}} d_{\mathrm{A}}(x, \hat{x})$, that is, the maximum value attained by said function.

Bearing in mind the above consideration, in our new scenario a privacy attacker endeavors to guess the confidential-attribute value of a particular respondent in the released table. Initially, the attacker has a prior belief given by p_X, that is, the distribution of that confidential-attribute value in the whole table. Later, the attacker observes that the user belongs to a group of records sharing a tuple of perturbed key-attribute values y, which is supposed to coincide with the system's decision y'. Based on this observation, the attacker updates their prior belief and obtains the posterior distribution $p_{X|Y}(\cdot|y)$. This situation is illustrated in Fig. 3.6. A fundamental question that arises in this context is how much privacy the released table leaks as a result of that observation. In the remainder of this section, we elaborate on this question and provide an upper bound on the reduction in privacy incurred by the disclosure of that information.

Total Variation and t-Closeness

For notational simplicity, we occasionally rename the posterior and the prior distributions $p_{X|Y}(\cdot|y)$ and p_X simply with the symbols p and q, respectively, but bear in mind that p is a PMF of x parametrized by y. In addition, we shall assume that the attacker adopts a MAP strategy. More precisely, $\hat{x}_p$ and $\hat{x}_q$ will denote the attacker's estimate when using the distributions p and q. Under these assumptions, the *reduction* (prior minus posterior) in conditional privacy can be expressed as

$$\begin{aligned}\Delta\mathcal{P}(y) &= \mathrm{E}_p\, d_{\mathrm{A}}(X, \hat{x}_q) - \mathrm{E}_p\, d_{\mathrm{A}}(X, \hat{x}_p) \\ &= \mathrm{E}_p\, d_{\mathrm{A}}(X, \hat{x}_q) - \mathrm{E}_q\, d_{\mathrm{A}}(X, \hat{x}_q) + \mathrm{E}_q\, d_{\mathrm{A}}(X, \hat{x}_q) \\ &\quad - \mathrm{E}_q\, d_{\mathrm{A}}(X, \hat{x}_p) + \mathrm{E}_q\, d_{\mathrm{A}}(X, \hat{x}_p) - \mathrm{E}_p\, d_{\mathrm{A}}(X, \hat{x}_p),\end{aligned}$$

where E_p and E_q denotes that the expectation is taken over the posterior and the prior distributions, respectively, as PMFs of x.

In this expression, the first two terms can be upper bounded by $d_{\max} \sum_x |p_x - q_x|$, since $\sum_x (p_x - q_x) \leqslant \sum_x |p_x - q_x|$. Clearly, this same bound applies to the last two terms. On the other hand, the remaining terms $\mathrm{E}_q\, d_{\mathrm{A}}(X, \hat{x}_q) - \mathrm{E}_q\, d_{\mathrm{A}}(X, \hat{x}_p)$ are upper bounded by 0, since the error incurred by $\hat{x}_q$ is smaller than or equal to that of $\hat{x}_p$. In the end, we obtain that

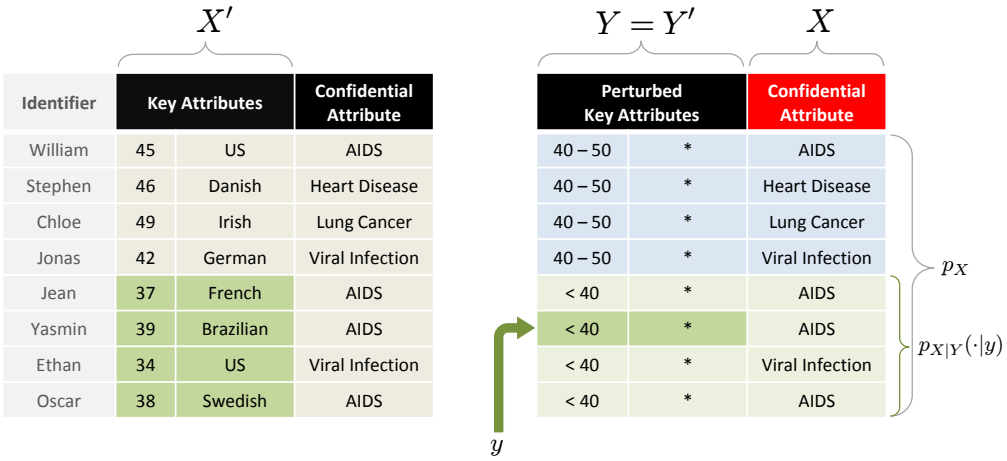
$$\Delta\mathcal{P}(y) \leqslant 2\, d_{\max} \sum_x |p_x - q_x| \,.$$

At this point, we shall briefly review the concept of *total variation*. For this purpose, consider P and Q to be two PMFs over $\mathcal{X}$. In probability theory, the total variation distance between P and Q is

$$\mathrm{TV}(P \,\|\, Q) = \tfrac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)| \,.$$

Furthermore, recall that, in information theory, *Pinsker's inequality* relates the total variation distance with the KL divergence. Particularly, $\mathrm{TV}(P \,\|\, Q) \leqslant \frac{\sqrt{2}}{2} \sqrt{\mathrm{D}(P \,\|\, Q)}$. Having stated this result, now the total variation distance permits writing the upper bound on $\Delta\mathcal{P}(y)$ in terms of the KL divergence:

$$\Delta\mathcal{P}(y) \leqslant 4\, d_{\max} \mathrm{TV}(p \,\|\, q) \leqslant 2\sqrt{2}\, d_{\max} \sqrt{\mathrm{D}(p \,\|\, q)},$$



(a) Original data (b) Perturbed data

Figure 3.6: At first, an attacker believes that the probability that a user appearing in (b) suffer from AIDS is 1/2. However, after observing that the user's record is one of the last four records, this probability becomes 3/4.

where the last inequality follows from Pinsker's inequality. Returning to the notation of prior and posterior distributions,

$$\begin{aligned}\Delta\mathcal{P}(y) &\leqslant 4\, d_{\max} \mathrm{TV}(p_{X|Y}(\cdot|y) \,\|\, p_X) \\ &\leqslant 2\sqrt{2}\, d_{\max} \sqrt{\mathrm{D}(p_{X|Y}(\cdot|y) \,\|\, p_X)}. \end{aligned} \tag{3.8}$$

This upper bound allows to establish a connection between our privacy criterion and t-closeness [29]. The latter criterion boils down to defining a maximum discrepancy between the posterior and prior distributions,

$$t = \max_y \mathrm{D}(p_{X|Y}(\cdot|y) \,\|\, p_X).$$

Under this definition and on account of (3.8),

$$\Delta\mathcal{P}(y) \leqslant 2\sqrt{2}\, d_{\max} \sqrt{t}.$$

Therefore, t-closeness is essentially equivalent to bounding the decrease in conditional privacy.

On a different note, we would like to make a comment on an issue of a purely technical nature. Clearly, in light of inequality (3.8), the minimization of either the total variation distance or the KL divergence leads to the minimization of an upper bound on $\Delta\mathcal{P}(y)$. However, the fact that the KL divergence imposes a worse upper

bound suggests us considering it when the resulting mathematical model be more tractable than the one built upon the total variation distance.

Mutual Information and Rate-Distortion Theory

The privacy criterion proposed in [32], called *(average) privacy risk* $\mathcal{R}$, is the average-case version of t-closeness. Formally, $\mathcal{R}$ is a conditional KL divergence, the average discrepancy between the posterior and the prior distributions, which turns out to coincide with the mutual information between the confidential data X and the observation Y:

$$\begin{aligned}\mathcal{R} &= \mathrm{E}_Y \, \mathrm{D}(p_{X|Y}(\cdot|Y) \,\|\, p_X) \\ &= \mathrm{E}_Y \, \mathrm{E}_{X|Y} \left[\log \frac{p_{X|Y}(X|Y)}{p_X(X)} \middle| Y \right] \\ &= \mathrm{E} \log \frac{p_{X|Y}(X|Y)}{p_X(X)} \\ &= \mathrm{I}(X;Y).\end{aligned}$$

Directly from their definition, $\mathcal{R} \leqslant t$, meaning that t-closeness is a stricter measure of privacy *risk*. Because the KL divergence is itself an average, $\mathcal{R}$ is clearly an average-case privacy criterion, but t-closeness is technically a maximum of an expectation, a hybrid between average case and worst case. The next subsection will comment on a third, purely worst-case criterion.

Further, we conveniently rewrite inequality (3.8) as

$$\frac{1}{8\, d_{\max}^2} \Delta\mathcal{P}(y)^2 \leqslant \mathrm{D}(p_{X|Y}(\cdot|y) \,\|\, p_X).$$

By averaging over all possible observation y, the right-hand side of this inequality becomes the privacy risk $\mathcal{R}$, which we showed to be equal to the mutual information. This leads to a bound on the privacy reduction in terms of mutual information,

$$\frac{1}{8\, d_{\max}^2} \mathrm{E} \left[\Delta\mathcal{P}(Y)^2 \right] \leqslant \mathrm{I}(X;Y).$$

Based on this observation, it is clear that the minimization of the mutual information contributes to the minimization of an upper bound on $\Delta\mathcal{P}(y)$. With this in mind, we now consider the more general scenario in which Y' and Y need not necessarily

coincide, and contemplate the case of a data publisher. Concretely, from the perspective of a publisher, we would choose a randomized perturbation rule $p_{Y'|X'}$ with the aim of minimizing the mutual information between X and Y, and consequently protecting user privacy. Evidently, the publisher would also need to guarantee the utility of the data to a certain extent, and thus impose a constraint on the average distortion. In conclusion, the data publisher would strive to solve the optimization problem

$$\min_{\substack{p_{Y'|X'} \\ \mathrm{E}\, d_{\mathrm{U}}(X',Y') \leqslant \mathcal{D}}} \mathrm{I}(X;Y), \tag{3.9}$$

which surprisingly bears a strong resemblance with the rate-distortion problem in the field of information theory. Specifically, the above optimization problem is a generalization of a well-known, extensively studied information-theoretic problem with more than half a century of maturity. Namely, the problem of lossy compression of source data with a distortion criterion, first proposed by Shannon in 1959 [127].

The importance of this lies in the fact that some of the information-theoretic results and methods for the rate-distortion problem can be extended to the problem (3.9). For example, in the special case when $X = X'$ and $Y = Y'$, our more general problem boils down to Shannon's rate-distortion and, interestingly, can be computed with the Blahut-Arimoto algorithm [76].

Bear in mind that the very same metric, or conceptually equivalent variations thereof, may in fact be interpreted under different perspectives. Recall, for instance, that mutual information is the difference between an unconditional entropy and a conditional entropy, effectively the posterior uncertainty modeled simply by the Shannon entropy, normalized with respect to its prior correspondence. Under this perspective, mutual information might also be connected to the branch of the tree in Fig. 3.3 leading to Shannon's entropy.

δ-Disclosure and Differential Privacy

Finally, we quickly remark on the connection of δ-disclosure and ϵ-differential privacy with our theoretical framework. *δ-disclosure* [30] is an even stricter privacy criterion than t-closeness, and hence much stricter than that average privacy risk $\mathcal{R}$ or mutual

information, discussed in the previous subsection. The definition of δ-disclosure may be rewritten in terms of our notation as

$$\delta = \max_{x,y} \left| \log \frac{p_{X|Y}(x|y)}{p_X(x)} \right|,$$

and understood as a worst-case privacy criterion. In fact,

$$\mathcal{R} \leqslant t \leqslant \delta.$$

We mentioned in the background section that [31] analyzes the case of the randomized perturbation Y of a true answer X to a query in a PIR system, before returning it to the user. Consider two databases d and d' that differ only by one record, but are subject to a common perturbation rule $p_{Y|X}$, and let p_Y and p'_Y be the two probability distributions of perturbed answers induced. After a slight manipulation of the definition given in the work cited, but faithfully to its spirit, we may say that a randomized perturbation rule provides ϵ-differential privacy when

$$\epsilon = \max_{y,d,d'} \log \frac{p_Y(y)}{p'_Y(y)}.$$

Even though it is clear that this formulation does not quite match the problem in terms of prior and posterior distributions described thus far, this manipulation enables us to still establish a loose relation with δ-disclosure, in the sense that the latter privacy criterion is a slightly stricter measure of discrepancy between PMFs, also based on a maximum (absolute) log ratio. We note, however, that although there is a formal similarity between the metrics, there are substantial differences between them in terms of their assumptions, objectives, models, and privacy guarantees.

3.5 Numerical Example

This section provides two simple albeit insightful examples that illustrate the measurement of privacy as an attacker's estimation error. Specifically, we quantify the level of privacy provided, first, by a privacy-enhancing mechanism that perturbs location information in the scenario of LBS, and secondly, by an anonymous-communication protocol largely based on Crowds [68].

3.5.1 Data Perturbation in Location-Based Services

Our first example contemplates a user who wishes to access an LBS provider. For instance, this could be the case of a user who wants to find the closest Italian restaurant to their current location. For this purpose, the user would inevitably have to submit their GPS coordinates to the (untrusted) provider. To avoid revealing their exact location, however, the user itself could perturb their location information by adding, for example, Gaussian noise. Alternatively, we could consider a user delegating this task to a (trusted) intermediary entity, as described in Sec. 3.4.1. In any case, data perturbation would enhance user privacy in terms of location, although clearly at the cost of data utility. Simply put, data-perturbative methods present the inherent trade-off between data utility and privacy.

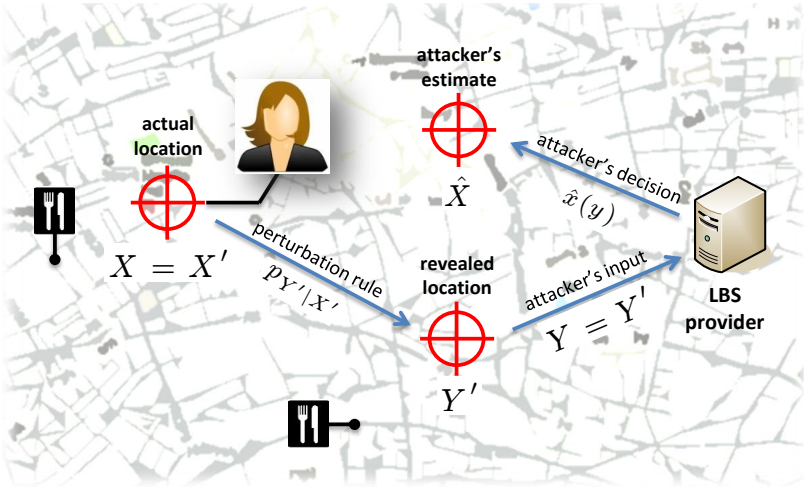


Figure 3.7: A user looking for a nearby Italian restaurant accesses an LBS provider. The user decides to perturb their actual location before querying the provider. In doing so, the user hinders the provider itself and any attacker capable of capturing their query, in their efforts to compromise user privacy in terms of location. In this example, we contemplate that the user is solely responsible for protecting their private data. In terms of our notation, this allows us to regard the user as the system. Notice that the user's actual location is, on the one hand, the attacker's unknown, and on the other, the information that the user (system) takes as input to generate the location that will be finally revealed. Thus we conclude that $X = X'$. Then, according to some randomized perturbation rule $p_{Y'|X'}$, the user discloses, for each location data x', a perturbed version y'. This perturbed location is submitted to the provider, which only has access to this information, i.e., $Y = Y'$. Lastly, based on this revealed information, the attacker uses a Bayes estimator $\hat{x}(y)$ to ascertain the user's actual location X.

Under the former strategy, and in accordance with the notation defined in Sec. 3.3.1, the user becomes the system—it is the user who is responsible for protecting their location data. Playing the role of the system, the user decides then to perturb their location data X on an individual basis for each query. In other words, we do not contemplate the case of sequences of data X^k, as Sec. 3.4.1 does.

A key element of our framework is the attacker's distortion function. In our example we assume the squared error between the actual location x and the attacker's estimate $\hat{x}$, that is, $d_{\mathrm{A}}(x, \hat{x}) = \|x - \hat{x}\|^2$. Unlike Hamming distance, note that the squared error does quantify how much the estimate differs from the unknown. As for the other variables of our model, we contemplate that the attacker's input Y is directly the location data perturbed by the user, Y', as illustrated in Fig. 3.7. Put differently, the attacker, assumed to be the service provider, has no more information than that disclosed by the user. Under all these assumptions, the average privacy (3.4) is

$$\mathcal{P}_{\mathrm{avg}} = \mathrm{E}[\|X - \hat{X}\|^2],$$

that is, the mean squared error (MSE).

As a final remark, we would like to connect our privacy criterion with a metric specifically conceived for the LBS scenario at hand [128]. In this cited work, the authors propose a framework that contemplates different aspects of the adversary model, captured by means of what they call *certainty*, *accuracy* and *correctness*. The information to be protected by a trusted intermediary system are traces modeling the locations visited by users over a period of time. The system accomplishes this task by hiding certain locations, reducing the accuracy of such locations or adding noise. As a result, the attacker observes a perturbed version of the traces and, together with certain mobility profiles of these users, attempts to deduce some information of interest X about the actual traces. In terms of our notation, the observed trajectories and the mobility patterns constitute the attacker's observation Y.

More accurately, given a particular observation y, the attacker strives to calculate the posterior distribution $p_{X|Y}$. However, since the adversary may have a limited number of resources, they may have to content themselves with an estimate $\hat{p}_{X|Y}$. The authors then use Shannon's entropy to measure the *uncertainty* of X, and define

accuracy as the discrepancy between $p_{X|Y}$ and $\hat{p}_{X|Y}$. Finally, they refer to location privacy as *correctness* and measure it as

$$\mathrm{E}_{\hat{p}_{X|Y}}[d_{\mathrm{S}}(X, x_t)|y],$$

where x_t is the true outcome of X, d_{S} a distance function specified by the system, and the expectation is taken over the estimate of the posterior distribution.

The most notable difference between [128] and the privacy criterion here proposed is that the former metric limits its scope to the specific scenario of LBSs; whereas in this thesis we attempt to provide a general overview. Besides, their proposal is a measure of privacy in an average-case sense. Another important distinction between the cited work and ours is that the former arrives to the conclusion that entropy and k-anonymity are not appropriate metrics for quantifying privacy in the context of LBS. Here, on the other hand, we do *not* argue against the use of entropy, k-anonymity and any of the other privacy metrics examined in Sec. 3.4. In fact, we regard these metrics as particular cases of the attacker's estimation error under certain assumptions on the adversary model, the attacker's strategy and a number of different considerations explored in that section.

3.5.2 Crowds-like Protocol for Anonymous Communications

In Chapter 2 we mentioned Chaum's mixes as a building block to implement anonymous communications networks. A different approach to communication anonymity is based on collaborative, P2P architectures. An example of collaborative approach is Crowds [68], in which users form a "crowd" to provide anonymity for each other.

In Crowds, a user who wants to browse a Web site forwards the request to another member of his crowd chosen uniformly at random. This crowd member decides with probability p to send the request to the Web site, and with probability $1 - p$ to send it to another randomly chosen crowd member, who in turn repeats the process. For the purpose of illustration, we consider a variation of the Crowds protocol. The main difference with respect to the original Crowds is that we do not introduce a mandatory initial forwarding step. We note that this variation provides worse anonymity than the original protocol, while also reducing the cost (in terms of delay and bandwidth) with

respect to Crowds. Further, we assume that the users participating in the protocol are honest; i.e., we only consider the Web site receiving the request as possible adversary.

More formally, consider n users indexed by $i = 1, \ldots, n$, wishing to communicate with an untrusted server. In order to attain a certain degree of anonymity, each user submits the message directly to said server with probability $p \in (0, 1)$, and forwards it to any of the other users, including themselves, with probability $1 - p$. In the case of forwarding, the recipient performs exactly the same probabilistic decision until the message arrives at the server. Fig. 3.8 shows the operation of this protocol.

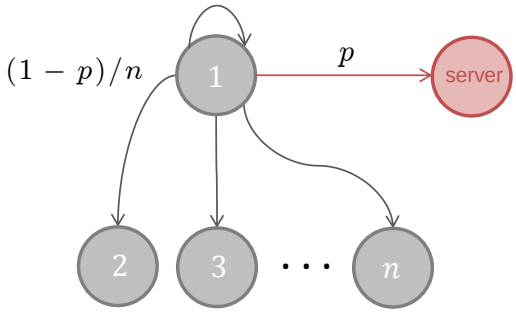


Figure 3.8: Anonymous-communication protocol inspired by Crowds. In our second numerical example, we contemplate a scenario where users send messages to a common, untrusted server, who aims at compromising sender anonymity. In response to this privacy threat, users decide to adhere to a modification of the Crowds protocol, whose operation is as follows: each user flips a biased coin and depending on the outcome chooses to submit the message to the server or else to another user, who is asked to perform the same process. The probability that a user forward the message to the server is denoted by p, whereas the probability of sending it to any other peer, including themselves, is $(1 - p)/n$.

In our protocol, we assume that the server attempts to guess the identity of the author of a given message, represented by the r.v. X, knowing only the user who last forwarded it, represented by the r.v. Y, consistently with the notation defined in Sec. 3.3.1. The other variables of our framework are as follows. Since the set of users involved in the protocol collaborate to frustrate the efforts of the server, they are in fact the system. The information that then serves as input to this system is simply the identity of the user who initiates the forwarding protocol, X. That is, the attacker's uncertainty and the system's input coincide, $X' = X$. Then again, the assumption that the server just knows the last sender in the forwarding chain leads to $Y = Y'$.

Under this model, and under the assumption of a uniform message-generation rate, that is, $p_X(x) = 1/n$ for all x, it can be proven that the conditional PMF of X given $Y = y$ is

$$p_{X|Y}(x|y) = \begin{cases} p + (1-p)/n & , \quad x = y \\ (1-p)/n & , \quad x \neq y \end{cases} . \tag{3.10}$$

Fig. 3.9 shows this conditional probability in the particular case when $x = 1$, i.e., the probability that the originator of a message be user 1, conditioned to the observation that the last sender is user y. Note that, because of the symmetry of our model, it would be straightforward to derive a PMF analogous to the one plotted in this figure, but for other originators of the message, namely $x = 2, \ldots, n$.

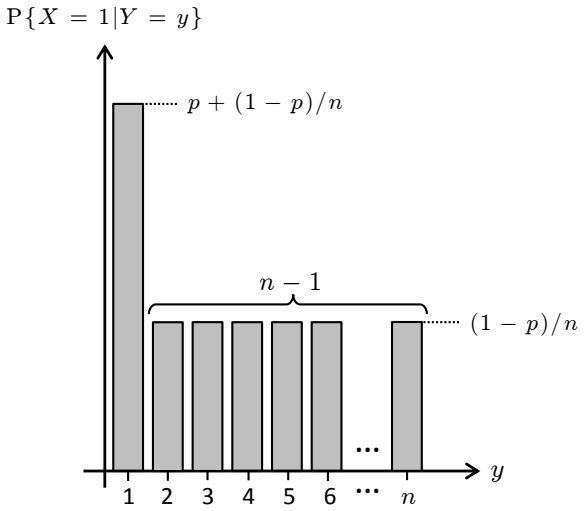


Figure 3.9: Probability that the original sender of a given message be the user 1, conditioned to the observation that the last sender in the forwarding path is user y. From this figure, we observe the PMF attains its maximum value when this last sender is precisely the user 1.

That said, assume that the attacker chooses Hamming distance as distortion function. Under this assumption, the conditional privacy (3.2) yields

$$\mathcal{P}(y) = \mathrm{P}\{X \neq \hat{x}(y)|y\},$$

that is, the MAP error conditioned on the observation y. Because Hamming distance implies, by virtue of (3.1), that Bayes estimation is equivalent to MAP estimation, it follows that the attacker's (best) decision rule is $\hat{x}(y) = y$. Leveraging on this

observation, we obtain that the privacy level provided by this variant of Crowds is

$$\mathcal{P}(y) = \varepsilon_{\text{MAP}} = 1 - \text{P}\{X = y|y\} = (1-p)(1-1/n),$$

from which it follows an entirely expected result—the lower the probability p of forwarding a message directly to the server, the higher the privacy provided by the protocol, but the higher the delay in the delivery of said message.

In the following, we consider the measurement of the privacy protection offered by this protocol, in terms of the three Rényi's entropies introduced in Sec. 3.4.1, namely the min-entropy $\text{H}_\infty(X|y)$, the Shannon entropy $\text{H}_1(X|y)$ and the Hartley entropy $\text{H}_0(X|y)$ of the r.v. X, modeling the actual sender of a given message (the privacy attacker's target), given the observation of the user who last forwarded it, y. Specifically, we connect the interpretations described in Sec. 3.4.1 to the example at hand.

But first we would like to recall from Sec. 3.4.1 that $\text{H}_\infty(X|y)$, $\text{H}_1(X|y)$ and $\text{H}_0(X|y)$ may be considered, from the point of view of the user, as a worst-case, average-case and best-case measurements of privacy, respectively, in the sense that

$$\text{H}_\infty(X|y) \leqslant \text{H}_1(X|y) \leqslant \text{H}_0(X|y),$$

owing to (3.6), with equality if and only if the conditional PMF of X given $Y = y$ is uniform. Note that a worst-case privacy metric from the point of view of the user is a best-case measure from the standpoint of the attacker and vice versa. Revisiting the interpretations given in that section, recall that the min-entropy $\text{H}_\infty(X|y)$ is directly connected with the maximum probability, in our case $\max_{x_i} p_{X|Y}(x_i|y) = p+(1-p)/n$, on account of (3.10). More concretely, and in the context of our example, min-entropy reflects the model in which a privacy attacker makes a single guess of the originator of a message, specifically the most likely one, which corresponds to $x = y$.

At the other extreme, the Hartley entropy $\text{H}_0(X|y)$ is a possibilistic rather than probabilistic measure, as it corresponds to the assumption that a privacy attacker would not content themselves with discarding all but the most likely sender, but consider instead all possible users. More accurately, measuring privacy as a Hartley's

entropy essentially boils down to the cardinality of the set of all possible originators of a message, namely $\mathrm{H}_0(X|y) = \log n$.

On a middle ground lies Shannon's entropy, which was interpreted in Sec. 3.4.1 by means of the AEP, specifically in terms of the cardinality of the set of typical sequences of i.i.d. samples of an r.v. Put in the context of our Crowds-like protocol, however, Shannon's entropy may be deemed as an average-case metric that considers the entire PMF of X given $Y = y$, and not merely its maximum value or its support set.

3.6 Guide for Designers of SDC and ACSs

The purpose of this section is to show the applicability of our framework to those designers of SDC and ACSs who, wishing to quantify the level of protection offered by their systems, do not want to delve into the mathematical details set forth in Sec. 3.4. In order to assist such designers in the selection of the privacy metric most appropriate for their requirements, this section revises the application scenarios of SDC and anonymous communications, and classifies some of the metrics used in these fields in terms of worst case, average case and best case, from the perspective of the user.

Before proceeding any further, we would like to briefly recall the distinction precisely between worst-case, average-case and best-case measurements of privacy. To this end, consider the scenario of ACSs in general and mixes in particular. In this specific scenario, the knowledge of the privacy attacker may be modeled by a probability distribution on the possible senders of a given message. A clear example of best-case privacy metric is Hartley's entropy, which measures the degree of anonymity attained by the mere cardinality of the set of candidate senders, or equivalently, by the logarithm of such cardinality. Loosely speaking, Hartley's entropy may be regarded as a best-case metric from the point of view of users (worst for adversaries), in the sense that it represents a privacy attacker's thorough effort in considering any and all possibilities, regardless of their likelihood. In the special case of threshold

pool mixes, however, the set of candidate output messages for a given input may be infinite, rendering Hartley's entropy inappropriate.

On the opposite extreme, min-entropy may be understood as the MAP estimation error where the attacker simply guesses the most likely outcome. This information-theoretic quantity may be construed as a worst-case metric, in the sense that the attacker is concerned with the most vulnerable statistical link between senders and messages. Finally, Shannon's entropy takes into account the underlying probability distribution in its entirety, between the extremes posed by the previous two metrics, yielding a quantity bounded according to (3.6). For this reason, one may think of it as an average-case metric.

Next, we elaborate on the distinction between Hamming and non-Hamming distortion functions, between whether these functions are known or unknown to the system, and finally between single and multiple-occurrence data. The reason is that the understanding of these concepts is fundamental for a system designer who, following the arguments sketched in Fig. 3.3, wants to choose the suitable metrics for their field of application. With this purpose, next we illustrate these concepts by means of a couple of simple albeit insightful examples.

The first consideration a system designer should take into account when applying our framework refers to the geometry of the attacker's distortion function d_{A}, namely whether it is a *Hamming* or a *non-Hamming* function. To illustrate this key point, consider a set of users in a social network. A Hamming function taking as inputs the users u_1 and u_2 would model an attacker who contemplates *only* their identities when comparing them, and ignores any other information such as the relationship between them within the social network, their profile similarity or their common interests. Another adversary, however, could represent said network by a graph, modeling users and relationships among them as nodes and edges, respectively. Leveraging on this graph, the attacker could use a non-Hamming function to compute the number of hops separating these two users and, accordingly, lead to the conclusion that they are, for example, close friends since $d_{\mathrm{A}}(u_1, u_2) = 1$.

The second consideration builds on the assumption of a non-Hamming attacker's distortion function. Under this premise, we contemplate two possible cases—when

the function is *known* to the system and when *not*. The former case is illustrated, for instance, in the context of LBSs—in this application scenario, an adversary will probably use the Euclidean distance to measure how their estimated location differs from the user's actual location. The latter case, i.e., when the measure of distortion used by the attacker is unknown to the system, would undoubtedly model a more general and realistic scenario. As an example of this case, consider a system perturbing the queries that a user wants to submit to a database, and an attacker wishing to ascertain the actual queries of this user. Suppose that these queries are one-word queries and that the perturbation mechanism replaces them with synonyms or semantically-similar words. Under these assumptions, our attacker could opt for a non-Hamming distortion function and measure the distance between the actual query and the estimate as the number of edges in a given ontology graph. Although the system could be aware of this fact, the specific ontology used by the attacker could not be available to the system, and consequently the distortion function would remain unknown.

Our last consideration is related to the nature of the variables of our framework, summarized in Table 3.1. Specifically, we contemplate two possible cases—*single* and *multiple-occurrence* data. The former case considers such variables to be tuples of a small number of components, and the latter assumes that these variables are sequences of data. An LBS attacker who observes the disclosed, possibly perturbed location of a user and makes a single guess about their actual location is an example of single-occurrence data. To illustrate the case of multi-occurrence data, consider a set of users exchanging messages through a mix system. Recall that such systems delay and reorder messages with the aim of concealing who is communicating with whom. Among the multiple attacks these systems are vulnerable to, the statistical disclosure attack [129] is a good example for our purposes of illustration, since it assumes an adversary who observes a large number or *sequence* of messages coming out of the mix, with the aim of tracing back their originators.

Having examined these key aspects of our framework, now we turn our attention, first, to the application scenario of SDC, and secondly, to the case of ACSs. In the former scenario, a data publisher aims at protecting the privacy of the individuals

appearing in a microdata set. Depending on the privacy requirements, the publisher may want to prevent an attacker from ascertaining the confidential-attribute value of any respondent in the released table. Under this requirement, t-closeness and mutual information appear as acceptable measures of privacy, since both criteria protect against confidential *attribute disclosure*. Recall that the assumptions on which they are based are a prior belief about the value of the confidential attribute in the table, and a posterior belief of said value given by the observation that the user belongs to a particular group of this table. Building on these premises, t-closeness may be regarded as a *worst-case* measurement of privacy, in the sense that it identifies the group of users whose distribution of the confidential attribute deviates the most from the distribution of this same attribute in the entire table. Recall that a worst-case measurement of privacy from the user's perspective is, in fact, a best-case measure from the attacker's point of view and vice versa.

Although t-closeness overcomes the similarity and skewness attacks mentioned in Sec. 2.4.1, its main deficiency is that no computational procedure has been given to enforce said criterion. An alternative is the mutual information between the confidential attributes and the observation, an average-case version of t-closeness that leads to a looser measure of privacy risk. In any of these two metrics, it is assumed the more general case in which the attacker's distortion function is not the Hamming distance. Specifically, this assumption models an adversary who does not content themselves with finding out whether the estimate and the unknown match, but wishes to quantify how much they diverge.

Another distinct privacy requirement is that of *identity disclosure*, whereby a publisher wishes to protect the released table against a linking attack. In this attack, the adversary's aim is to uncover the identity of the individuals in the released table by linking the records in this table to a public data set including identifier attributes. Under this requirement and under the assumption that the attacker regards each respondent within a particular group as equally likely, k-anonymity may be deemed as a *best-case* measure of privacy, determined by Hartley's entropy.

In the scenario of ACSs, there exists a wide variety of approaches. Among them, a popular anonymous-communication protocol is Crowds. Although in this section

Table 3.4: Guide for designers of SDC and ACSs. This table classifies several privacy metrics depending, first, on whether they are regarded as worst-case, average-case and best-case measures, and secondly on their application domain.

	Worst case	**Average case**	**Best case**
SDC	t-closeness	mutual information	k-anonymity
ACSs	min-entropy	Shannon's entropy	Hartley's entropy

we limit the discussion of the privacy provided by such systems to a variant of this protocol, we would like to stress that the conclusions drawn here may be extended to other anonymous systems. Having said this, recall that in the original Crowds protocol, a system designer makes available to users a collaborative protocol that helps them enhance the anonymity of the messages sent to a common, untrusted Web server. The design parameters are the number of users participating in the protocol and the probability of forwarding a message directly to the server.

In our variant of this protocol, however, we contemplate an attacker who strives to guess the identity of the sender of a given message, based on the knowledge of the last user in the forwarding path. Under this adversary model, we may regard min-entropy, Shannon's entropy or Hartley's entropy as particular cases of our measure of privacy, depending on the specific strategy of the attacker. For example, under an adversary who uses MAP estimation and, accordingly, opts for the last sender, min-entropy may be interpreted as a worst-case privacy metric. Alternatively, we may assume an attacker who takes into account the entire probability distribution of possible senders, and not only the most likely candidate. In this case, Shannon's entropy may be deemed as an average-case measure. Finally, suppose an attacker who thoroughly examines all potential originators of the message without considering their likelihood. Under this assumption, Hartley's entropy may be regarded as a best-case measurement of privacy. The discussion in this section is summarized in Table 3.4.

3.7 Conclusion

Numerous privacy metrics have been proposed in the literature. Most of these metrics have been conceived for specific applications, adversary models, and privacy threats,

and thus are difficult to generalize. Even for specific applications, we often find that various privacy metrics are available. For example, to measure the anonymity provided by anonymous-communication networks, several flavors of entropy (Shannon, Hartley, min-entropy) can be found in the literature, while no guidelines exist that explain the relationship between the different proposals, or provide an understanding of how to interpret or put in context the results provided by each of them. Also, these proposals fail to justify the choice, often simply neglecting alternatives, say min-entropy or any Rényi's entropy.

In the scenario of SDC, a variety of approaches attempt to capture, to a greater or lesser degree, the private information leaked as a result of the dissemination of microdata sets. In this spirit, k-anonymity is possibly the best-known privacy measure, mainly due to its mathematical tractability. Later, numerous extensions and enhancements were introduced with the aim of overcoming its limitations. While all these metrics have provided further insight into our understanding of privacy, the research community would benefit from a framework embracing those metrics and making it possible to compare them, and to evaluate any privacy-protecting mechanism by the same yardstick.

In this chapter, we propose a unifying view to choose and justify privacy measures in a more systematic manner. Our approach starts with the definition and modeling of the variables of a general framework. Then, we proceed with a mathematical formulation of privacy, which essentially emerges from BDT. Specifically, we define privacy as the estimation error incurred by an attacker. We first propose what we refer to as conditional privacy, meaning that our measure is conditioned on an attacker's particular observation. Accordingly, we define the terms of average privacy and worst-case privacy.

The formulation is then investigated theoretically. Namely, we interpret a number of well-known privacy criteria as particular cases of our more general metric. The arguments behind these justifications are based on fundamental results related to the fields of information theory, probability theory and BDT. More accurately, we interpret our privacy criterion as k-anonymity and l-diversity principles by connecting

them to Rényi's entropy and MAP estimation. Under certain assumptions, a conditional version of the AEP allows us to interpret Shannon's entropy as an arbitrarily high confidence set. Then, the total variation distance and Pinsker's inequality justify t-closeness requirement and the criterion proposed in [32] as particular instances of our measure of privacy. In the course of this interpretation, we find that our formulation bears a strong resemblance with the rate-distortion problem in information theory.

After our theoretical analysis, we provide some guidelines for those systems designers of SDC and ACSs who do not wish to delve into mathematical details. A couple of simple albeit insightful examples are also presented. Our first example quantifies the level of privacy provided by a privacy-enhancing mechanism that perturbs location information in the scenario of LBS. Under certain assumptions on the adversary model, our measure of privacy becomes the MSE. Then we turn our attention to the scenario of ACSs and measure the degree of anonymity achieved by a modification of the collaborative protocol Crowds. We contemplate different strategies for the attacker and, accordingly, interpret min-entropy, Shannon's entropy and Hartley's entropy as worst-case, average-case and best-case privacy metrics.

The establishment of connections between privacy metrics and concepts from the field of information theory, and the formulation of these metrics as estimation errors cast light on the understanding of the privacy properties associated with those metrics and the evaluation of their applicability to specific applications. With this work, we also show the riveting interplay between the field of information privacy on the one hand, and on the other the fields of information theory and stochastic estimation, while bridging the gap between the respective communities.

In closing, we hope that this unified perspective of privacy metrics, drawing upon the principles of information theory and Bayesian estimation, is a helpful, illustrative step towards the systematic modeling of privacy-preserving information systems.

Chapter 4

Measuring the Privacy of User Profiles

4.1 Introduction

In Chapters 2 and 3, we established the critical importance of quantifying privacy in order to assess, compare, improve and optimize privacy-protecting technologies. The main contribution presented in Chapter 3 was precisely the definition of a general framework where privacy was measured as an attacker's estimation error. The applicability of our framework was demonstrated in the scenarios of SDC, ACSs and LBS. In application scenarios involving user profiles, as it is the case of personalized information systems, there are several proposals specifically conceived for measuring privacy. The problem, however, is that these approaches are not appropriately justified and are defined in an ad hoc manner for a few specific applications.

This chapter approaches the fundamental problem of proposing quantitative measures of the privacy of user profiles. We tackle the issue by providing a thorough justification of KL divergence and Shannon's entropy as measures of anonymity and privacy. Our justification relies on fundamental principles from information theory and statistics, thereby drawing intriguing links between said fields and information privacy.

We consider two adversary models. The first model assumes an attacker aimed at targeting users who deviate from the average profile of interests; and the second one contemplates an attacker whose objective is to classify a given user into a predefined group of users. Under the former model, the use of divergence and entropy as measures of anonymity is justified by elaborating on Jaynes' rationale behind entropy-maximization methods and the method of types. Under the latter adversary model, a riveting argument in favor of divergence as privacy criterion stems from hypothesis testing and large deviation theory. The adversary model as well as the metrics defined here will serve as a reference for the next chapters.

The results presented in this chapter are an extension of [46, 50, 130].

Chapter Outline

The rest of this chapter is organized as follows. Sec. 4.2 delves into the technical literature of profiling and reviews some fundamental concepts related to it. Sec. 4.3 defines the adversary model used throughout this work. This includes the definition of an abstract model for representing user interests, our assumptions about the scenario, and the specification of concrete objectives for the adversary. The use of divergence and entropy as privacy and anonymity measures is justified in Secs. 4.4 and 4.5. Afterwards, Sec. 4.6 establishes a connection between our privacy criteria and other proposals for measuring user privacy in the context of personalized information systems. Finally, conclusions are drawn in Sec. 4.7.

4.2 User Profiling

In Sec. 2.1.2 we illustrated the privacy risks inherent in personalized information systems and emphasized the increasing pervasiveness of personalization technologies. As shown in that section, this kind of technologies appear in a variety of applications including personalized Web search and browsing, multimedia recommendation systems, collaborative tagging or personalized news. In all these applications, the ability to *profile* users is the cornerstone to provide a personalized service.

Profiling, however, is not only present in personalized information systems, but also plays a prominent role in a wide range of scenarios. As a matter of fact, before computers became a part of people's daily lives, detective and criminal investigators constructed profiles of their offenders, psychiatrists built behavioral profiles of people with some personality disorder, marketing researchers elaborated profiles of potential clients, and recruiting companies profiled candidates for particular job vacancies [131]. Currently, such types of profiles are no longer handmade, and profiling spans many other disciplines, from forensic medicine to immigration policy, from supply chain management to actuarial consultancy [132].

In the coming subsections, we shall dive into the technical literature of profiling to examine some fundamental concepts in the field. In the end, we shall recall a widely accepted definition of this term. The purpose of all this is to comprehend the meaning of profiling from a broad perspective, not limited to the context of personalization, so that we can define an adversary model consistent with the literature of profiling.

4.2.1 Construction and Application of Profiles

Profiling practices and technologies are characterized by the use of algorithms that collect and analyze data over a period of time; their ultimate objective is to acquire knowledge in the form of statistical patterns or correlations between data [133]. When those patterns are employed to identify and represent people, they are called *profiles* [132]. In the context of profiling, a profile may refer either to a person or to a group of people. For the sake of simplicity, in our scenario of personalized information systems we only contemplate the case of a single person.

In the literature, there exist several models that describe the technical process of profiling, namely the semiotic model knowledge discovery in databases [134] and de facto industry standard cross-industry standard process for data mining, CRISP-DM. In essence, these models characterize profiling as a process that consists of a number of phases. For instance, [134] defines profiling as an adaptive and dynamic process where data are collected, prepared, mined and finally applied. Although the phases in each model differ in the degree of sophistication, both models reduce the process of

profiling to the construction of profiles, i.e., data collection, preparation and mining, and the subsequent application of those profiles to people.

4.2.2 Individual and Group Profiling

Profiling can then be viewed as a type of knowledge that *identifies* and represents people by means of the construction and application of profiles. The technical literature of profiling [131, 132] attributes two meanings to the term *identify*:

- the discovery of the individual characteristics of a person, also referred to as *individuation*;
- and the *categorization* of a person as a specific type of person.

In other words, and according to the cited works, profiling refers both to the discrimination of one person *from* all other persons, and to the identification of a person *as part of* a certain group of persons. The application or usage of profiles to identify people in the sense of individuation or categorization motivates the distinction between *individual* and *group* profiling.

Individual profiling is frequently used in the information systems that motivate this thesis. Personalized information systems aim to ascertain the unique interests and preferences of users, that is, their mission is to discover what distinguishes a particular user from the general population of users. At the same time, personalization and many other technologies also capitalize on group profiling. Typically, these technologies take advantage of the fact that a user's profile may coincide with another profile built from a sheer volume of data belonging to a number of other people. In this latter kind of profiling, profiles are applied to persons whose data were not used to generate those profiles.

In the case of group profiling, there exists an important distinction between the groups of people that profiles may represent. In particular, a group profile may refer either to an existing *community* of people that consider themselves as a group, or to a *category* of people that do not necessarily constitute a community but share certain characteristics. An example of community could be a political party or a religious

organization, while the group of Internet users that regularly query databases with medical information could be deemed as a category.

4.2.3 Definition of Profiling

As the literature recognizes, profiling may seem to refer, in the first instance, to concepts of a rather different nature but connected to each other in important ways. After exploring such concepts, now we recall a widely accepted definition of profiling [131, 132, 135, 136]. Quoting [131, 132], the term profiling is defined as

- the process of constructing profiles that identify and represent either a person or a group of persons,
- and/or the application of profiles with the aim of
  - individuating a person,
  - or categorizing a person as a member of a specific group of persons.

The above definition illustrates the connection between the concepts of individual and group profiling on the one hand, and on the other, the construction and application of profiles. As we shall see later in Sec. 4.3, the consideration of these concepts will be key in the definition of our adversary model. In that section, the assumptions about the privacy attacker will be consistent with the profiling practices and the terminology reviewed here.

4.3 Adversary Model

In Secs. 2.1.3 and 2.1.4 we stressed the need for privacy metrics as the only way to evaluate, compare and design privacy-protecting mechanisms. When measuring the level of privacy provided by a PET, however, it is essential to specify the concrete assumptions about the adversary, that is, its capabilities, properties or powers, as well as the scenario where this attacker operates; this is known as the *adversary model*. The importance of such a model lies in the fact that the level of privacy provided is

measured with respect to it. In other words, if the assumptions change, so does the metric.

The objective of this section is precisely to specify these assumptions. In Sec. 4.3.1 we describe the particularities of the scenario of personalized information systems considered, and identify the potential privacy attackers contemplated in this scenario. In the next subsections we analyze two additional, key aspects of our adversary model. Specifically, Sec. 4.3.2 defines the user-profile model, that is, the model used by the attacker to represent user interests and preferences. And afterwards, Sec. 4.3.3 examines an essential element of the adversary model, the objective of profiling itself. As we shall see later, the objectives considered for the attacker will be in line with the technical literature of profiling that we briefly reviewed in Sec. 4.2.

4.3.1 Scenario

In the use case described in Sec. 2.1.2, a company illegitimately gained a competitive advantage from monitoring certain professional and personal activities of employees and other individuals. The example given in that section highlighted the serious privacy threats posed by personalized information systems.

These systems allow users to tag Web pages, post comments or rate information items of any type, that is, they enable users to take a series of actions from which these users expect to obtain some sort of benefit. The scenario considered in this chapter assumes that users are *identified* from the standpoint of such systems. This does not necessarily mean that personalized information systems have users' real names or other personally identifying information; it only implies that users' actions are monitored so that their profiles can be constructed and personalization can be provided. User identification, in that sense, could be achieved, for example, by using HTTP cookies. We would like to note that if users were neither logged in nor willing to be tracked, they would definitely not receive any personalized service.

Clearly not all the actions taken by a user are equally sensitive. Further, the sensitivity of such actions is context-dependent and subject to user perception. Resorting to the example given in Sec. 2.1.2, tagging the Web page `http://occupywallst.org` with "OWS" could be considered as a sensitive action for the protagonist of the story,

Jane Doe, given her aspirations for promotion. However, if this same tag was to be posted by another user, it might not have any impact at all.

That individual tag would not lead an attacker to draw far-reaching conclusions about her actual interests or political leaning and, in principle, the privacy of our protagonist would not be seriously compromised. Tagging that Web page could be regarded as expressing some sympathy for the Occupy Wall Street movement, but would not be interpreted as if she had a deep interest in the topic. However, if numerous tags were posted in this same direction, information providers could dispel their initial doubts about her concerns, and obtain a precise snapshot of her real interests, i.e., they could be able to build her profile and maybe conclude she is an activist.

The construction of this profile is essential to enable personalization, but at the very same time it raises serious privacy risks with regard to social sorting or segmentation [137]. In this work we are concerned about the risk of profiling, which goes hand in hand with the risk of reidentification[(a)]. For this reason, in our scenario of personalized information systems we assume that the set of potential privacy attackers encompasses any entity capable of *profiling* users based on the information they disclose when interacting with such systems. Clearly, this set includes the information systems themselves, which may have personally identifying information about users, and also comprises any attacker able to intercept the communications between users and systems. Besides, since the information conveyed (e.g., ratings, tags, comments or posts) is often publicly available to other users of those systems, any entity able to collect this information is also taken into consideration in our adversary model. Table 4.1 summarizes the assumptions about the scenario considered here.

[(a)] In the scenario considered in this work, we assume that users are identified by personalized information systems. By "identified" we simply mean that users are tracked by those systems. However, several actions taken by a certain user may eventually lead these systems, or any entity intercepting the communications between the user and such systems, to find out the user's real identity, provided that it has not been voluntarily given by the user. We refer to this as reidentification.

4.3.2 User-Profile Model

In the motivating scenario of this work, a user submits queries to a Web search engine, clicks on news links in a personalized news recommendation system, and assigns tags to resources on the Web, all according to their profile of interests. The information revealed, i.e., queries, news clicked and tags, allows those systems to extract a profile of interests or *user profile*, which is fundamental in the provision of personalized services.

In the context of personalized information systems, user profiles are frequently modeled as histograms. For example, collaborative tagging systems commonly represent profiles by using tag clouds, which, in essence, may be regarded as histograms. Recall that a tag cloud is a visual depiction in which tags are weighted according to their frequency of use. Those two possible representations for user profiles, tag clouds and histograms, are, in fact, simultaneously used in popular tagging systems such as BibSonomy [(b)], CiteULike [(c)], Delicious, LibraryThing [(d)] and SlideShare [(e)].

In the scenario of personalized recommendation systems, we also find examples of profiles modeled as histograms, especially in content-based recommenders [138] such as IMDb, Jinni [(f)] and Last.fm [(g)]. Of particular interest is the case of Google News, where news are classified into a predefined set of topic categories; and accordingly, users are modeled by their distribution of clicks on news, i.e., as histograms of relative frequencies of clicks within that set of categories [139]. In this same spirit, recent privacy-protecting approaches in the scenario of recommendation systems also propose using histograms of absolute frequencies for modeling user profiles [140, 141].

Motivated by all these examples and inspired by other works in the field [89, 90, 95, 102, 107, 142], in this chapter we justify and interpret a privacy criterion under the assumption that user profiles are modeled as PMFs, that is, as histograms of relative frequencies of user data (e.g., queries, clicks, tags and ratings) within a set

[(b)] `http://www.bibsonomy.org`
[(c)] `http://www.citeulike.org`
[(d)] `http://www.librarything.com`
[(e)] `http://slideshare.net`
[(f)] `http://www.jinni.com`
[(g)] `http://www.last.fm`

of categories of interest. Our user-profile model is, therefore, much in line with the representations used in numerous tagging systems and personalized recommendation systems. In addition to its extensive use, we would also like to emphasize its mathematical tractability. Other user-profile models include semantic networks, weighted concepts and association rules [143]. Fig. 4.1 shows an example of the user profile representation assumed in this work. This example could perfectly resemble the case of Jane Doe described in Sec. 2.1.2. Fig. 4.2, on the other hand, depicts the profile of a user as shown in Movielens.

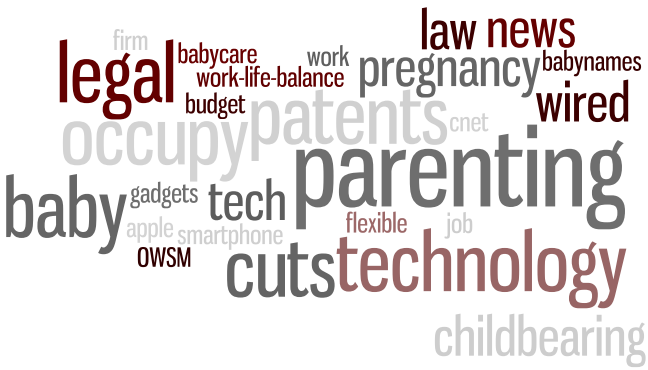


Figure 4.1: User profile modeled as a tag cloud in a collaborative tagging system. The tags posted by users are frequently depicted as tag clouds, not only in those tagging systems, but also in multimedia recommendation systems such as Jinni.

An important ingredient of our profile model are the categories of interests employed to represent user preferences. In tagging systems these categories are usually the tags themselves, and profiles are just a counter of the number of times each tag has been posted. The main drawbacks of such profiles are that, first, they become untractable when tagging activity is significant, and secondly, they do not allow easy inspection of user interests. The categorization of user data may help in both regards. A coarser representation of those data could make it easier to have a quick overview of said preferences. Consider, as an example, a user posting the tags "nyfw" and "jen kao". Rather than using this information to model their interests, it could be more convenient to have a higher level of abstraction that enables the attacker to conclude, directly from the observation of their profile, that the user is interested in fashion. The granularity level used to represent user preferences certainly will depend

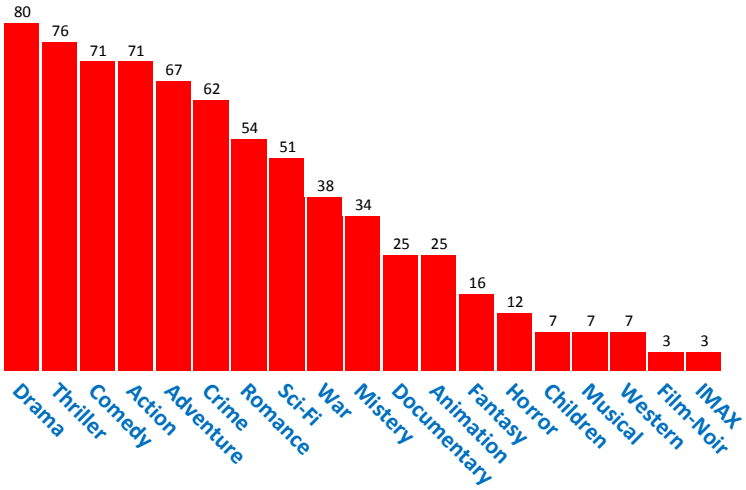


Figure 4.2: User profile modeled as a histogram of absolute frequencies of ratings within a set of predefined movie genres. Many personalized information systems use this kind of representation, or slight variations of this idea, to model user interests.

on the attacker's capabilities. For instance, a rudimentary attacker will possibly have to content themselves with a histogram of raw data such as tags or search queries. A more sophisticated attacker, on the other hand, could cluster these tags into hierarchical tag categories. In a nutshell, the categorization of user data is an element to be considered in the definition of our user-profile model and hence in the adversary model.

Actual and Apparent Profiles

In view of the assumptions described in 4.3.2, our privacy attacker boils down to an entity that aims to profile users by representing their interests in the form of normalized histograms, on the basis of a given categorization. To achieve this aim, the attacker may exploit any explicit and implicit information that users communicate to information systems. To mitigate the risk of profiling, naturally users may adopt any privacy-protecting mechanism.

Among the different approaches in the literature, this thesis focuses on those mechanisms based on data perturbation. As mentioned in Sec. 2.3.1, the key strengths of data perturbation are its simplicity in terms of infrastructure requirements and its

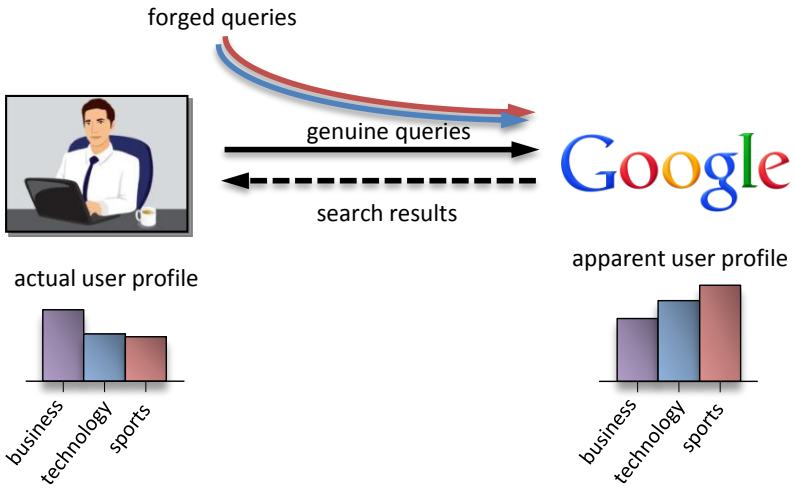


Figure 4.3: Query forgery in personalized Web search. A user submits false queries, accompanied with genuine queries, to perturb his *actual* profile of interests. By adopting query forgery, the adversary, possibly the service provider itself, observes a distorted version of his profile. We refer to this profile as the *apparent* user profile.

strong privacy guarantees, as users need not trust the information provider, nor the network operator nor other peers.

Under the assumption of an untrusted model, and as a response to the privacy threats described in Sec. 2.1.2, users therefore contemplate the possibility of unveiling only some pieces of their private data, or slightly perturbed versions of it. In doing so, users gain some privacy, although at the cost of certain loss in usability. Users may consider, for example, the elimination of some sensitive tags or comments, and the submission of false ratings and search queries. As a result of this, the attacker observes a perturbed version of the genuine profile, also in the form of a relative histogram, which does not reflect the actual interests of the user. In short, the attacker believes that the observed behavior characterizes the actual user's profile. Thereafter, we shall refer to these two profiles as the *actual user profile* and the *apparent user profile*. Fig. 4.3 shows an example of such profiles.

4.3.3 Attacker's Objective

In Sec. 4.2 we analyzed various concepts related to profiling. Specifically, we showed that profiling is defined based on the concepts of construction and application of profiles on the one hand, and individuation and classification on the other. In this

section we connect our adversary model with those concepts. Our aim is to define a model in line with the literature of profiling.

Recall that, in this work, the set of potential privacy attackers comprises any entity aimed at profiling users of personalized information systems. Bearing in mind the terminology reviewed in Sec. 4.2, we shall refer to our adversary strictly as an entity which constructs and applies profiles to identify and represent those users. Recall from Sec. 4.2.2 that the term "identify" refers to either individuation or classification, and that it has nothing to do with learning the real identity of a user or other personally identifying information. Actually, in our scenario of personalized information systems we contemplate that users may be completely identified by such systems.

In addition to define of our privacy attacker in those more technical terms, our adversary model also captures the objective of profiling itself. In particular, we consider the two forms of profiling described in Sec. 4.2.2, i.e., individual and group profiling, and integrate them into our model as concrete objectives for the adversary. These two objectives are interpreted as follows:

- On the one hand, we may consider the attacker strives to target users who deviate from the average profile of interests. In accordance with Sec. 4.2.2, we refer to this objective as *individuation*, meaning that the adversary aims at discriminating a given user from the whole population of users, or said otherwise, wishes to learn what distinguishes that user from the other users.

- On the other hand, we may assume that the attacker's goal is to classify a user into a predefined group of users. To conduct this *classification*, the attacker contrasts the user's profile with the profile representative of a particular group.

These two objectives, together with the assumptions about the scenario and the user profile representation, constitute the adversary model upon which our privacy metric builds. Table 4.1 provides a summary of our adversary model. In Secs. 4.4 and 4.5, we shall justify KL divergence and entropy as privacy criteria. This justification will rely on two adversary models differing only in the attacker's

Table 4.1: Main conceptual highlights of the adversary model assumed in this work.

What scenario is assumed?	We consider those information systems that provide users with profile-based personalization. In this scenario we assume that users are *identified* by such information systems. Accordingly, we contemplate users who may provide these systems with their names or other personally identifying information during the registration process. However, we also consider the possibility that those users use pseudonyms, or are not logged into the system. In any case, the only requirement is that users are disposed to be tracked by the personalized information system they wish to interact with. Otherwise, personalized services cannot be provided.
Who can be the privacy attacker?	Any entity able to *profile* users is taken into account. This includes service providers and any entity capable of eavesdropping users' data, e.g., ISPs, proxies, switches, routers, firewalls, users of the same local area network, system administrators and so on. Further, we also contemplate any other entity which can collect publicly available users' data.
How does the attacker model user interests?	User profiles are modeled as *histograms of relative frequencies* of user data across a predefined set of categories of interest. The *categorization* of those data plays a fundamental role in the modeling of user interests.
What is the attacker after when profiling users?	We contemplate two possible objectives for an attacker: *individuation* and *classification*. The former objective reflects an attacker wishing to target peculiar users, while the latter objective is associated with an adversary aimed at identifying a given user as a member of a specific group of users.

objective. Depending on the objective chosen, we shall regard those information-theoretic quantities as measures of *privacy risk* against individuation, or as measures of *privacy gain* against classification.

4.4 Privacy Metric against Individuation

Next, we shall proceed with our first interpretation of KL divergence and Shannon's entropy as a privacy criterion. Both in this section and in Sec. 4.5, the information-theoretic arguments and justifications in favor of our metric will be expounded in a systematic manner, following the points sketched in Fig. 4.4. Henceforth, we shall use the notation $\mathrm{H}(X)$ instead of $\mathrm{H}_1(X)$ to refer to the Shannon entropy of an r.v. X.

In the section at hand, we shall interpret divergence and entropy under the assumptions of the adversary model defined in Sec. 4.3, in the special case when the attacker's objective is to individuate a user in the sense of discriminating this user from all other users; this interpretation corresponds to the first branch of the tree in Fig. 4.4, which we term *individuation*. For that purpose, we shall adopt the perspective of Jaynes' celebrated *rationale on entropy maximization methods* [144], which is

based on the *method of types* [76, §11], a powerful technique in large deviation theory whose fundamental results we also explore in this section.

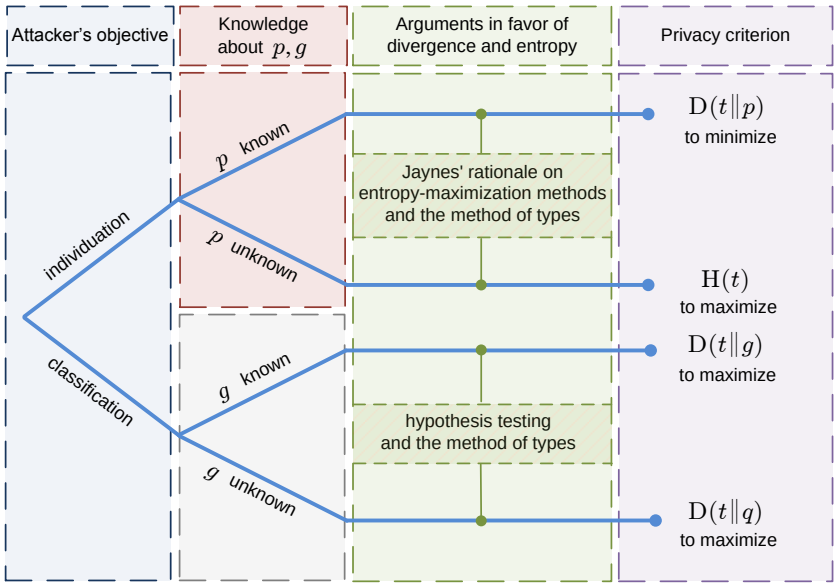


Figure 4.4: Summary of our interpretations of KL divergence and Shannon's entropy as measures of privacy. This figure illustrates, at a conceptual level, the assumptions upon which our privacy criterion builds. First, we follow Jaynes' rationale behind entropy-maximization methods to justify divergence and entropy when the attacker's goal is to *individuate* users. The knowledge of the population's distribution p determines whether the metric to be used is divergence or entropy. Secondly, when the attacker aims at *classifying* a user as a member of a particular group, our arguments in favor of divergence stem from hypothesis testing and the method of types. In the special case when the group profile g is unknown to the user, they may wish to maximize the divergence between the actual profile q and the perturbed, observed profile t, in order to avoid being classified as they actually are.

The first part of this section, Sec. 4.4.1, tackles an important question. Suppose we are faced with a problem, formulated in terms of a model, in which a probability distribution plays a major role. In the event this distribution is unknown, we wish to assume a feasible candidate. What is the most likely probability distribution? In other words, what is the "probability of a probability" distribution? We shall see that a widespread answer to this question relies on choosing the distribution *maximizing the Shannon entropy*, or, if a reference distribution is available, the distribution *minimizing the KL divergence* with respect to it, commonly subject to feasibility constraints determined by the specific application at hand.

Our review of the maximum-entropy method is crucial because it is unfortunately not always known in the privacy community. As we shall see in the last part of this section, Sec. 4.4.2, the key idea is to model a user profile as a probability distribution, as considered in Sec. 4.3.2, apply the maximum-entropy method to measure the likelihood of a user profile either as its entropy or as its divergence with respect to the population's average profile, and finally take that likelihood as a measure of anonymity.

4.4.1 Rationale behind the Maximum-Entropy Method

A wide variety of models across diverse fields have been explained on the basis of the intriguing principle of entropy maximization. A classical example in physics is the Maxwell-Boltzmann probability distribution $p(v)$ of particle velocities V in a gas [145,146] of known temperature. It turns out that $p(v)$ is precisely the probability distribution maximizing the entropy, subject to a constraint on the temperature, equivalent to a constraint on the average kinetic energy, in turn equivalent to a constraint on $\mathrm{E}\, V^2$. Another well-known example, in the field of electrical engineering, of the application of the maximum-entropy method, is Burg's spectral estimation method [147]. In this method, the power spectral density of a signal is regarded as a probability distribution of power across frequency, only partly known. Burg suggested filling in the unknown portion of the power spectral density by choosing that maximizing the entropy, constrained on the partial knowledge available. More concretely, in the discrete case, when the constraints consist in a given range of the cross-correlation function, up to a time shift k, the solution turns out to be a k^{th} order Gauss-Markov process [76]. A third and more recent example, this time in the field of natural language processing, is the use of log-linear models, which arise as the solution to constrained maximum-entropy problems [148] in computational linguistics.

Having motivated the maximum-entropy method, we are ready to proceed to describe Jaynes' attempt to justify, or at least interpret it, by reviewing the method of types of large deviation theory, a beautiful area lying at the intersection of statistics and information theory. Let $X_1, \ldots, X_k$ be a sequence of k i.i.d. drawings of an r.v. uniformly distributed in the alphabet $\{1, \ldots, n\}$. Let k_i be the number of times

symbol $i = 1, \ldots, n$ appears in a sequence of outcomes $x_1, \ldots, x_k$, thus $k = \sum_i k_i$. The *type* t of a sequence of outcomes is the relative proportion of occurrences of each symbol, that is, the *empirical distribution* $t = \left(\frac{k_1}{k}, \ldots, \frac{k_n}{k}\right)$, not necessarily uniform. In other words, consider tossing an n-sided fair dice k times, and seeing exactly k_i times face i. In [144], Jaynes points out that

$$\mathrm{H}(t) = \mathrm{H}\left(\frac{k_1}{k}, \ldots, \frac{k_n}{k}\right) \simeq \frac{1}{k} \log \frac{k!}{k_1! \cdots k_n!} \quad \text{for } k \gg 1.$$

Loosely speaking, for large k, the size of a *type class*, that is, the number of possible outcomes for a given type t (permutations with repeated elements), is approximately $2^{k\,\mathrm{H}(t)}$ in the exponent. The fundamental rationale in [144] for selecting the type t with maximum entropy $\mathrm{H}(t)$ lies in the approximate equivalence between entropy maximization and the maximization of the number of possible outcomes corresponding to a type. In a way, this justifies the infamous *principle of insufficient reason*, according to which, one may expect an approximately equal relative frequency $k_i/k = 1/n$ for each symbol i, as the uniform distribution maximizes the entropy. The principle of entropy maximization is extended to include constraints also in [144].

Obviously, since all possible permutations count equally, the argument only works for uniformly distributed drawings, which is somewhat circular. A more general argument [76, §11], albeit entirely analogous, starts with a prior knowledge of an arbitrary PMF p, not necessarily uniform, of such samples $X_1, \ldots, X_k$. Because the empirical distribution or type T of an i.i.d. drawing is itself an r.v., we may define its PMF $p_T(t) = \mathrm{P}\{T = t\}$; formally, the PMF of a random PMF. Using indicator r.v.'s, it is straightforward to confirm the intuition that $\mathrm{E}\,T = p$. The general argument in question leads to approximating the probability $p_T(t)$ of a type class, a fractional measure of its size, in terms of its relative entropy, specifically $2^{-k\,\mathrm{D}(t\,\|\,p)}$ in the exponent, i.e.,

$$\mathrm{D}(t\,\|\,p) \simeq -\frac{1}{k} \log p_T(t) \quad \text{for } k \gg 1,$$

which encompasses the special case of entropy, by virtue of (2.1). Roughly speaking, the likelihood of the empirical distribution t exponentially decreases with its KL divergence with respect to the average, reference distribution p.

In conclusion, the most likely PMF t is that minimizing its divergence with respect to the reference distribution p. In the special case of uniform $p = u$, this is equivalent to maximizing the entropy, on account of (2.1), possibly subject to constraints on t that reflect its partial knowledge or a restricted set of feasible choices.

4.4.2 Measuring the Privacy of User Profiles

We proceed to justify, or at least interpret, KL divergence and Shannon's entropy as measures of the privacy of a user profile. Before we dive in, we must stress that the use of entropy as a measure of privacy traces back to Shannon's work in the fifties [116]. More recent studies [34,149] rescue the suitable applicability of the concept of entropy as a measure of privacy, by proposing to measure the degree of anonymity observable by an attacker as the entropy of the probability distribution of possible senders of a given message. Sec. 2.4.2 provides further details on this.

In the context of this work, an intuitive justification in favor of entropy maximization is that it boils down to making the apparent user profile as uniform as possible, thereby hiding a user's particular bias towards certain categories of interest. But a much richer argumentation stems from Jaynes' rationale behind entropy-maximization methods [144, 150], more generally understood under the beautiful perspective of the method of types and large deviation theory [76, §11], which we motivated and reviewed in the previous subsection.

Under Jaynes' rationale on entropy-maximization methods, the entropy of an apparent user profile, modeled by a relative frequency histogram of categorized user data (e.g., queries, ratings or tags), may be regarded as a measure of privacy, or perhaps more accurately, anonymity. The leading idea is that the method of types from information theory establishes an approximate monotonic relationship between the likelihood of a PMF in a stochastic system and its entropy. Loosely speaking and in our context, the higher the entropy of a profile, the more likely it is, and the more users behave according to it. Under this interpretation, entropy is a measure of anonymity, *not* in the sense that the user's identity remains unknown, but only in the sense that higher likelihood of an apparent profile, believed by an external observer to be the actual profile, makes that profile more common, hopefully helping the user go

unnoticed, less interesting to an attacker whose objective is to target peculiar users. This is, of course, in the absence of a probability distribution model for the PMFs, viewed abstractly as r.v.'s themselves; if available, that distribution of profiles would be the measure of anonymity to be used, in the same sense of user-profile density regarded above.

If an aggregated histogram of the population were available as a reference profile, the extension of Jaynes' argument to relative entropy would also give an acceptable measure of anonymity. Recall from Sec. 2.2 that KL divergence is a measure of discrepancy between probability distributions, which includes Shannon's entropy as the special case when the reference distribution is uniform. Conceptually, a lower KL divergence hides discrepancies with respect to a reference profile, say the population's, and there also exists a monotonic relationship between the likelihood of a distribution and its divergence with respect to the reference distribution of choice, which enables us to deem KL divergence as a measure of anonymity in a sense entirely analogous to the above mentioned.

Under this interpretation, the KL divergence is therefore interpreted as an (inverse) indicator of the commonness of similar profiles in said population. As such, we should hasten to stress that the KL divergence is a measure of anonymity rather than privacy, in the sense that the obfuscated information is the uniqueness of the profile behind the online activity, rather than the actual profile itself. Indeed, a profile of interests already matching the population's would not require perturbation.

In conclusion, our justification of entropy and divergence as measures of anonymity builds upon these two ideas:

- *user-profile density* may be regarded as a measure of *anonymity*.
- The probabilistic model describing the distribution of profiles is frequently unknown to users. In the absence of this model, Jaynes' rationale allows us to interpret *Shannon's entropy* and *KL divergence* as measures of *user-profile density*.

Fig. 4.5 illustrates these ideas by means of a simple but insightful example. The figure in question shows a distribution of profiles in the probability simplex, in the

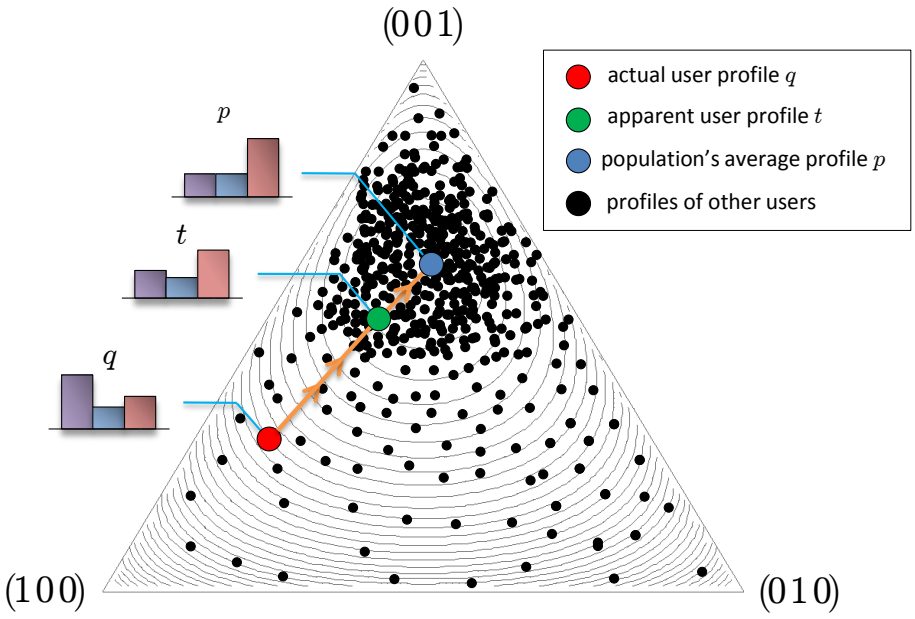


Figure 4.5: A privacy attacker aims at distinguishing a particular user among the population of users. Under Jaynes' rationale, KL divergence may be regarded as a measure of user-profile density. It is important to stress that the distribution of profiles here depicted is unknown to this particular user; only the average is known. Accordingly, the user adopts some perturbative strategy whereby the observed profile t gets close, in terms of divergence, to the average population's distribution p. As a result, the apparent profile becomes more common, getting lost in the crowd, and thus thwarting the attacker's intention.

case when profiles are modeled across $n = 3$ categories of interest, e.g., business, technology and sports. Note, however, that the justification provided in this section presumes that this information is not at the disposal of users. If available, users would certainly use it as a measure of anonymity. In this figure, we also represent the actual profile of a particular user, their apparent profile, and the average population's profile. Besides, we plot the contours of the divergence between a point in the simplex and the reference distribution p, that is, $\mathrm{D}(\cdot\|p)$. Bear in mind Jaynes' rationale, this particular user perturbs their actual profile in such a way that the resulting profile approaches, in terms of KL divergence, the population's profile. In doing so, the apparent profile gets lost in the crowd, thus hindering privacy attackers in their efforts to distinguish this user from other users.

Last but not least, we would like to emphasize that, under the assumptions this justification relies on, i.e., an adversary aimed at discriminating a given user from

the population of users, KL divergence is, in fact, a measure of privacy *risk* or, more accurately, anonymity *loss*. This contrasts with the interpretation given in Sec. 4.5, where the assumption of an attacker operating as a classifier leads us to consider KL divergence as a measure of privacy *gain*.

4.5 Privacy Metric against Classification

In Sec. 4.4, we interpreted KL divergence and Shannon's entropy as privacy criteria, under the assumption that the attacker attempted to target users who deviated from the average profile. In this section, we justify our metric under the premise that the attacker strives to classify a particular user into a predefined group. Put differently, the attacker's objective boils down to a classification problem. The justification provided in this section corresponds to the branch called *classification* in the tree of Fig. 4.4.

The use of KL divergence as a classifier is justified by its extensive application in the fields of speech and image recognition, machine learning, data mining, and in information security as well [151–157]. In recommender systems, we also find numerous examples where KL divergence is used to classify users with similar characteristics [158–160]. In this application scenario, divergence is a popular similarity measure for comparing users and items. A more elaborated justification in favor of KL divergence as a classifier, however, stems from hypothesis testing [76, §11] and the method of types of large deviation theory. In the following, we shall interpret our privacy metric as false positives and negatives when an attacker applies a binary hypothesis test to find out whether a sequence of observed data (e.g., ratings, tags or queries) belongs to a predefined group of users or not.

Let H be a binary r.v. representing two possible hypothesis about the distribution of an r.v. X. Precisely, $H = 1$ with probability θ and $H = 2$ with probability $1 - \theta$, and X conditioned on H has PMF g when $H = 1$ and g' when $H = 2$. Let $(X_j)_{j=1}^{k}$ be k i.i.d. drawings of this reference r.v. X and let t denote the type or empirical distribution of a k-tuple of their observed values $(x_j)_{j=1}^{k}$. Recall that the MAP estimate of a finite-alphabet r.v. is its most likely value. Also, recall from

Sec. 2.2 that $\mathrm{H}(p\|q)$ denotes the cross entropy between two distributions p and q over the same alphabet. It can be shown [76] that:

(i) The log-likelihood

$$-\frac{1}{k}\log \mathrm{P}\left\{ (X_j)_{j=1}^k = (x_j)_{j=1}^k \middle| H \right\} = \begin{cases} \mathrm{H}(t\|g)\,, & \text{if } H=1. \\ \mathrm{H}(t\|g'), & \text{if } H=2. \end{cases}$$

(ii) The MAP estimate $\hat{H}_{\mathrm{MAP}}$ of the hypothesis H from the observed sequence $(X_j)_j$ is determined by the Neyman-Pearson criterion, namely $\hat{H}_{\mathrm{MAP}} = 1$ if, and only if,

$$\mathrm{D}(t\|g) \leqslant \mathrm{D}(t\|g') + \gamma, \tag{4.1}$$

with $\gamma = \frac{1}{k}\log\frac{\theta}{1-\theta}$, and $\hat{H}_{\mathrm{MAP}} = 2$ otherwise.

Even if the prior probability θ is unknown or if the hypothesis is not modeled as an r.v., for any $\gamma \in \mathbb{R}$, criterion (ii) still optimizes the trade-off between the probabilities of false positives and false negatives, in the sense that one of these errors is minimized for a fixed value of the other. In short, γ parametrizes the trade-off curve in the error plane.

Our interpretation contemplates the scenario where an attacker knows, or is able to estimate, the distribution g representing a group into which a given user does *not* want to be categorized. The attacker observes then a sequence of k i.i.d. data (e.g., tags) generated by this user. Based on the type t of this sequence, which we regard as the user's apparent profile, the adversary attempts to ascertain whether said user is a member of that group. More accurately, the attacker considers the *hypothesis testing* between two alternatives, namely whether the data have been drawn according to g, hypothesis $\mathscr{H}_1$, or g', hypothesis $\mathscr{H}_2$, where g' may represent the complement of the sensitive group at hand, or any other group. In this interpretation we assume that the profiles belonging to a group are concentrated mainly around the representative distribution of that group.

Define the *acceptance region* $\mathscr{A}_k$ as the set of sequences of observed data over which the attacker decides to accept $\mathscr{H}_1$. Concordantly, consider the following two probabilities of decision error:

(a) the probability of a false negative $\alpha_k = g(\bar{\mathscr{A}_k})$, defined as the probability of accepting $\mathscr{H}_2$ when $\mathscr{H}_1$ is true,

(b) and the probability of a false positive $\beta_k = g'(\mathscr{A}_k)$, defined as the probability of accepting $\mathscr{H}_1$ when $\mathscr{H}_2$ is true.

Above, $\bar{\mathscr{A}_k}$ denotes the complement of $\mathscr{A}_k$. $g(\mathscr{A}_k)$, for example, represents the probability of all data sequences in $\mathscr{A}_k$, i.i.d. according to g, and similarly for $g'(\bar{\mathscr{A}_k})$. Hence, α_k is the probability that the attacker mistakenly classifies the user as not belonging to the group, and β_k the probability of the attacker incorrectly assuming that the user does belong to it.

According to the preliminaries in this section, an intelligent attacker would perform a Neyman-Pearson test (4.1) to infer whether the user belongs in fact to the group, in an optimal fashion, that is, minimizing the classification error α_k for a given error β_k, or vice versa. In the event that a suitable representation g' of the alternative group is unavailable, or that a simpler approach is deemed preferable, the user shall strive to counter such an intelligent attacker by merely maximizing the discrepancy $\mathrm{D}(t\|g)$ between the observed profile t and the representation g of the sensitive group to avoid.

Fig. 4.6 provides an example that illustrates our justification of divergence as a measure of privacy against classification. Particularly, this figure plots a distribution model for profiles in the simplex of probability, under the assumption that user profiles are represented across $n = 3$ categories of interest, exactly as in Fig. 4.5. We also depict the actual profile q of a particular user, their apparent profile t and the profile g representative of a group into which this user does not want to be classified. The contours correspond to the divergence $\mathrm{D}(\cdot\|g)$ between a point in the simplex and the group profile g. The figure in question also shows the region of the simplex that leads the attacker to classify a user as belonging to this particular group.

Last but not least, we would like to stress that the justifications provided in this section are clearly under the premise that the user knows the distribution g. An alternative to the absence of this information is assuming $g = q$, that is, considering the user as the group into which they do not want to be classified. Building on this assumption, the user's strategy consists in maximizing $\mathrm{D}(t\|q)$. Conceptually,

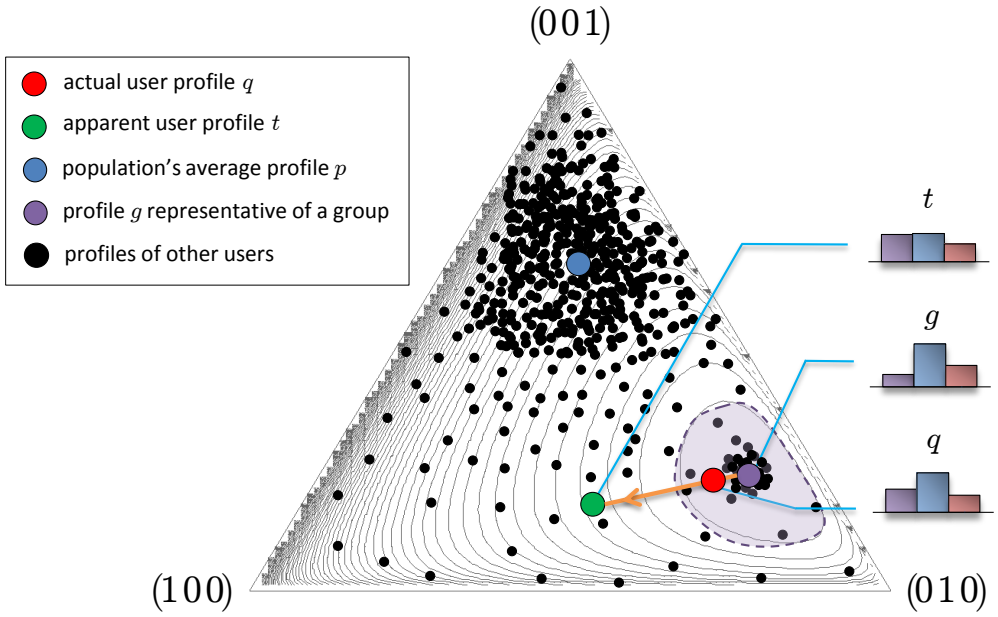


Figure 4.6: A user distorts their actual profile q to counter an attacker who strives to classify this user as belonging to a particular group. Under our interpretation of divergence through hypothesis testing, the probability of being classified as a member of that group decreases as the observed profile t moves away, in terms of divergence, from the profile g representative of said group.

this reflects the situation in which a user does not want the perturbed, observed profile resemble their actual profile. As we shall see in the next section, Sec. 4.6, the resulting privacy metric, i.e., the divergence between the apparent user profile and the actual user profile, is much in line with other criteria in the literature that suggest quantifying privacy by using some measure of similarity between these two profiles. Fig. 4.4 illustrates the assumptions about the adversary model and the information-theoretic arguments that we have followed to justify and interpret KL divergence and Shannon's entropy as privacy criteria.

4.6 Connection with Other Privacy Metrics

The aim of this section is twofold. First, we shall link KL divergence and Shannon's entropy to the privacy metrics for user profiles examined in Sec. 2.4.3. Secondly, we shall interpret both information-theoretic quantities as an attacker's estimation error, thus tying the more general privacy criterion defined in Chapter 3 to the metrics proposed in this chapter.

In Sec. 2.4.3, we showed that most of the criteria for quantifying the privacy of user profiles reduce to functions that take as inputs the actual user profile q and the apparent user profile t. A simple classification consists in grouping these criteria into similarity-based privacy measures and uncertainty-based privacy metrics. The cosine similarity [96–101] and the weighted Euclidean distance [121] fall into the former category. The latter category includes the Shannon entropy of the apparent profile [90, 91, 107] and the mutual information between the distributions q and t [89, 102].

The arguments provided to justify both types of privacy metrics are frequently presented as follows. In the case of similarity-based measures, it is assumed that the greater the disparity between the profiles q and t, the lower the privacy risk. In the case of uncertainty-based metrics, the justification consists merely in noting that entropy is a measure of uncertainty and mutual information is a measure of the reduction in uncertainty. While there is some intuition behind these criteria, the fact is that they lack a rigorous justification, accompanied by solid and convincing arguments. Besides, these metrics are often not defined in terms of an adversary model that contemplates assumptions such as the attacker's capabilities or objectives. Ultimately, they are conceived specifically for assessing the effectiveness of concrete privacy-preserving mechanisms.

In this chapter we propose KL divergence and Shannon's entropy as privacy metrics, and justify and interpret them by leveraging on fundamental principles from information theory and statistics. Particularly, under an adversary who aims at individuating users, we show that divergence and entropy may be regarded as measures of user-profile density, or profile likelihood, and thus anonymity. Under an attacker whose objective is to classify users, we interpret divergence as a measurement of privacy risk. Although our criteria and the state-of-the-art privacy metrics certainly build upon different assumptions, we may establish a connection between the adversary models considered in this work and those of the aforementioned metrics. Specifically, we may interpret the similarity-based metrics in the special case when the attacker's goal is to classify a given user. If the distribution of the group into which this user does not want to be classified is unavailable to them, the adversary

Table 4.2: Relationship among the state-of-the-art metrics for user profiles, our adversary models and the privacy criteria proposed in this work.

Attacker's objective	Knowledge about p, g	Proposed criteria	Related metrics
individuation	p known	$\mathrm{D}(t \,\|\, p)$	-
	p unknown	$\mathrm{H}(t)$	[90, 91, 107]
classification	g known	$\mathrm{D}(t \,\|\, g)$	-
	g unknown	$\mathrm{D}(t \,\|\, q)$	[89, 96–102, 121]

model defined in Sec. 4.3.3 would clearly fit with the assumptions of the similarity-based criteria. We note that this adversary model could also be valid for the mutual information between the actual and apparent profiles. Similarly, we may explain the uncertainty-based metrics from the perspective of an attacker who wishes to target singular users, and under the assumption that the population's distribution is unknown to these users. Table 4.2 summarizes this discussion.

In the remainder of this section we shall link the metrics proposed in this chapter, which are specific for measuring the privacy of user profiles, to the more general privacy measure defined in Chapter 3. To this end, consider a medical search engine (e.g., PubMed [(h)]) playing the role of an attacker. Our particular attacker is assumed to have a database table including identifiers and user profiles associated with those identifiers. The identifiers correspond to users registered with the search engine. The profiles are supposed to be constructed by exploiting any explicit or implicit information about users. For example, PubMed collects the words searched, the pages visited, the data and time of these visits and the user's Internet address. Fig. 4.7 (a) shows this database table.

Suppose, at some point, that a registered user wishes to submit some queries and prefers to do it without being logged in. The user thinks this may provide them with a certain level of anonymity, although clearly at the cost of nonpersonalized search results. If that level of protection were considered insufficient, the user could in principle adopt a PET such as the submission of false queries. Having said this,

[(h)] `http://www.ncbi.nlm.nih.gov/pubmed`

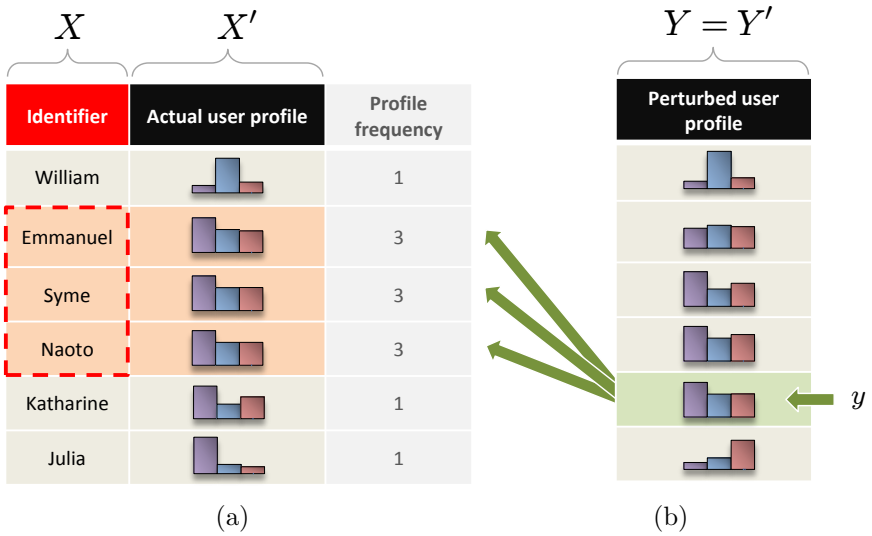


Figure 4.7: An attacker attempts to disclose the identity of a user based on the observed profile y (b). The user is registered with the search engine but uses the system without being logged in. We suppose that the attacker has profiled every registered user on the basis of their search history. All these profiles are stored in a database (a). To identify the user at hand, the adversary strives to find a match between the observed profile and all profiles X' stored in this database. When the frequency of the observed profile within the database is large enough, the attacker is likely to fail in its bid to ascertain X.

in this example we assume the adversary knows that the user is registered with the search engine and consequently attempts to ascertain their identity.

At this point we make a slight digression to put the variables of the framework of Sec. 3.3.1 in the context of the example at hand. For this, recall that the attacker's unknown X is private data about a user which the adversary endeavors to unveil. In this case, X becomes the identity of the user who wants to submit queries anonymously. The system's input X' is the information that serves as input for the system to make a decision. In our scenario we assume that the user plays the role of the system, i.e., they are the solely responsible for protecting their privacy. Accordingly, the variable X' represents the actual user profile and the system's decision Y' is directly the perturbed version of this profile. On the other hand, we suppose that the attacker's observation reduces to the perturbed profile, that is, $Y = Y'$. Note, however, that the adversary has the actual profile of the user in question, but does not know which of the profiles stored in the database belongs to this particular user; the

profile of this user was constructed while they were logged in. Finally, the attacker's decision $\hat{X}$ is the estimate of the user's identity.

Having contextualized the variables of the framework described in Chapter 3, now we proceed to interpret Shannon's entropy as an attacker's estimation error. Given a sequence of observed queries, modeled in our case as the type Y of these queries, the attacker makes a decision $\hat{X}$ on the identity of the originator X of said sequence. This decision is made by comparing the observed profile with all profiles stored in the search engine's database. Recall that Jaynes' rationale allows us to regard the Shannon entropy of a type as a measure of its probability, and therefore of its anonymity. Bearing this in mind, it follows that the higher the entropy of the observed type, the greater the number of profiles in the database that may be linked to this type, and consequently the greater the attacker's error in estimating X. If the population's distribution p were available to the user, an entirely analogous argument could be used to justify KL divergence as an estimation error. Fig. 4.7 illustrates the example above.

4.7 Conclusion

Numerous PETs have been proposed to mitigate the privacy risk inherent in personalized information systems. Unfortunately, these technologies have not yet gained wide adoption, mainly because, frequently, their effectiveness as well as their penalties in terms of utility remain unclear. In this context, privacy metrics, together with utility metrics, help pave the way for their adoption, as the only manner to evaluate, compare, improve and optimize them.

The literature of privacy metrics in personalized information systems is still in its infancy. There exist several criteria for measuring the privacy of user profiles but these are merely ad hoc proposals for specific applications and, what is more important, they are not duly justified.

To the best of our knowledge, our work is the first to rigorously justify a measure of the privacy of user profiles. The proposed metric is KL divergence, an information-theoretic quantity that we interpret under two distinct adversary models. First, we

consider an attacker who strives to target users who deviate from the average profile of interests; and secondly, we contemplate an attacker whose objective is to classify a given user into a predefined group of users.

For the former model, the use of KL divergence is justified by elaborating on Jaynes' rationale behind entropy-maximization methods and the method of types of large deviation theory. Under this interpretation, divergence is a measure of privacy risk, or more accurately, anonymity loss. In essence, our justification builds on three main principles. First, we model the profile of a user as a type or empirical distribution. Secondly, through Jaynes' rationale, the KL divergence between the user's profile and the population's may be deemed as a measure of the probability of the former profile. And thirdly, we consider that the probability of a profile may be a suitable measure of its anonymity. Only under this interpretation, the uniform profile is of particular interest since entropy may be justified as anonymity criterion in a sense entirely analogous to that of divergence.

For the latter adversary model, our privacy criterion is supported by its extensive use in fields such as speech and image recognition, machine learning, data mining and information security. But a richer argument stems from hypothesis testing and the method of types, which enable us to interpret KL divergence as false positives and negatives. Under this perspective, divergence is a measure of privacy gain.

Lastly, we show that the KL divergence and Shannon's entropy may be viewed, in fact, as an attacker's estimation error. This allows us to link the former information-theoretic quantities, which are specific to user profiles, to the more general privacy definition proposed in Chapter 3.

In a nutshell, in this chapter we have accomplished the following goals:

- Jaynes' rationale for entropy maximization has been applied to other scientific areas —for example spectral estimation— to both justify and interpret a variety of models and algorithms —continuing with the example of spectral estimation, Burg's method. The application of the same celebrated rationale and its extension to relative entropy, now to the field of information privacy, is one of the novel, exciting contributions of this chapter. An analogous application is made for the method of hypothesis testing in the field of statistics.

- By doing so, we argue in favor of the use of KL divergence and Shannon's entropy to measure profile privacy, along with conceptual insight into their information-theoretic, statistical meaning.
- Further, we introduce completely new models tying up the notions of profiling and profile classification with information theory.
- The value of these contributions stems from the fact that drawing a connection between information theory and information privacy, at the level of privacy metrics in mathematical modeling of privacy-enhancing mechanisms, opens the door for further application of powerful, mature concepts from the former field to the latter, and transitively, fields related to the former such as data compression and convex optimization, as we illustrate here with concrete examples.

Part II

Data-Perturbative Mechanisms and Privacy-Utility Trade-Off

Chapter 5

Tag Suppression in the Semantic Web

5.1 Introduction

The World Wide Web constitutes the largest repository of information in the world. Since its invention in the nineties, the form in which information is organized has evolved substantially. At the beginning, Web content was classified in directories belonging to different areas of interest, manually maintained by experts. These directories provided users with accurate information, but as the Web grew they rapidly became unmanageable.

Although they are still available, they have been progressively dominated by the current search engines based on Web crawlers, which explore new or updated content in a methodic, automatic manner. However, even though search engines are able to index a large amount of Web content, they may come back with irrelevant results or fail when terms are not explicitly included in Web pages. A query containing the keyword *accommodation*, for instance, would not retrieve pages with terms such as *hotel* or *apartment* not including that keyword.

A wide range of personalization technologies has been gradually integrating into these crawler-based search engines. Drawing upon profiling techniques, this new

generation of search engines is able to capture users' interests and provide them with information tailored to their preferences. Although these systems may come back with more accurate search results, they experience the same problem as the predecessor search engines since they retrieve Web content based on term matching.

A new form of conceiving the Web, called the *semantic Web* [57], has emerged to address this issue. The semantic Web, envisioned by Tim Berners-Lee, is expected to provide Web content with a conceptual structure so that information can be interpreted by machines. For this to become a reality, the semantic Web requires to explicitly associate meaning with resources on the Web. A widely spread manner to accomplish this is by means of *semantic tagging*.

One of the major benefits of associating concepts with Web pages is that machines will start to gain some level of understanding of information expressed in natural language, thus helping humans deal with information overload. When the semantic Web goes live, intelligent software agents will be able to automatically book flights for us, update our medical records at our request and provide us with personalized answers to particular queries, without the hassle of exhaustive literal searches across myriads of disorganized data [161].

In this scenario where information is processed on a conceptual basis, personalization will definitely overcome the one-size-fits-all paradigm and provide individually optimized access to information. In a nutshell, the semantic Web paints the most appropriate environment for personalized information systems [162]. In the meantime, we can enjoy some instances, although limited in scope, of this new conception of the Web, namely the *collaborative tagging systems* that have proliferated over the last years. Some examples include BibSonomy, CiteULike, Delicious and StumbleUpon [(a)], where users add short, usually one-word descriptions to resources they find on the Web.

Tagging systems are therefore the basis for the complete development of personalized information systems. Currently, numerous recommendation systems are incorporating tagging services to enhance the quality of their recommendations. For

[(a)] `http://www.stumbleupon.com`

example, the movie recommendation system Movielens began as a traditional recommender, using the ratings submitted by users as the source of information to generate personalized recommendations. But recently it included collaborative tagging techniques to enrich user profiling. In parallel, other systems that started as pure tagging systems are now offering recommendation services to their users [163]. Examples of these services include suggesting Web resources similar to those tagged previously, or recommending users to a target user, given the fact that they have similar tag-based profiles. In short, tagging is synonymous with personalization and vice versa.

Despite the many advantages the semantic Web is bringing to the Web community, the continuous tagging activity prompts serious privacy concerns. The tags submitted by users to semantic-Web servers could be used not only by these servers but also by any privacy attacker capable of collecting this information, to extract an accurate representation of user interests or *user profiles* [164, 165], leading these attackers to infer sensitive information such as health-related issues, political leaning or income level. This could be the case of the tagging systems mentioned above and many other applications where tags are used to build user profiles, normally in the form of some kind of histogram or tag cloud.

In this chapter we investigate a privacy-enhancing mechanism that has the purpose of hindering privacy attackers in their efforts to profile users on the basis of the tags they specify. In our approach, users inevitably reveal their personal preferences when tagging resources on the Web. To avoid being accurately profiled, though, they may wish to refrain from tagging some of those resources. In doing so, users protect their privacy to a certain degree without having to trust the semantic-Web server nor any other external entity. However, this is at the cost of some processing overhead and, what is more important, the semantic loss incurred by suppressing tags; since tagging is a means of classifying resources based on their content, those affected by suppression could not, for example, be retrieved by a user searching for the tags they have lost. Put another way, tag suppression poses a trade-off between privacy on the one hand, and on the other the semantic functionality enabled by tagging, which we also refer to as data utility in accordance with Sec. 2.1.4.

Our first contribution is an architecture that describes, at a high level, the components of a possible implementation of the tag-suppression technique. Our approach, which relies on the assumptions of the untrusted model detailed in Sec. 2.3.1, would be implemented as a software application installed on the user's computer. The purpose of this architecture is to assist users with the elimination of tags, in the presence of an attacker whose aim is to individuate these users.

The theoretical analysis of the inherent trade-off between privacy and data utility is our second contribution. Specifically, we present a mathematical formulation of optimal tag suppression in the semantic Web. We measure privacy as Shannon's entropy of the user's tag distribution after the suppression of certain tags, a privacy metric that we thoroughly justified in Chapter 4. Accordingly, we formulate and solve an optimization problem modeling the privacy-utility trade-off.

In addition, we experimentally evaluate the extent to which our technique contributes to privacy protection in a real-world tagging application. Namely, we apply tag suppression to BibSonomy, a popular tagging system for sharing bookmarks and publications, and show, in a series of experiments, how our approach enables its users to enhance their privacy. The work presented in this chapter builds on the adversary model proposed in Chapter 4.

A major portion of this chapter was published in [47, 51].

Chapter Outline

The rest of this chapter is organized as follows. Sec. 5.2 describes our privacy-enhancing mechanism and compares it to other approaches in the literature. Sec. 5.3 specifies the privacy metric used in this chapter and the properties of our adversary model. Sec. 5.4 presents the building blocks of the architecture and Sec. 5.5 introduces a formulation of the optimal trade-off between privacy and data utility. Sec. 5.6 presents a detailed theoretical analysis of the optimization problem characterizing the privacy-utility trade-off. In addition, this section shows a simple but insightful example that illustrates the formulation and theoretical analysis of the previous sections. Sec. 5.7 provides an experimental evaluation of our technique in BibSonomy. Conclusions are drawn in Sec. 5.8.

5.2 Privacy-Enhancing Mechanism

In the offline world, it is possible to eliminate those data that might compromise our privacy, those which have become useless or that we just want to get rid of. It would suffice, for example, to erase the information stored on a hard drive or to physically destroy the storage device where data are kept. In the online world, however, the user loses control of their data as they are managed and stored by other parties, e.g., cloud, Web and e-mail servers and other information systems. Consequently, it is not that easy to delete user-generated data such as comments posted on blogs, tags submitted to Web content or queries sent to Web search engines [(b)]. Even though these data could be removed from one of the above systems, it would be difficult to ensure that the data have been completely eliminated, given the ease with which data can be copied, distributed and shared with other parties. In other words, it is likely that our online data remain stored somewhere on the Internet for a long time.

In this situation, prevention is better than cure. That is, it would be desirable that certain data be eliminated on the user side, before these data be disseminated through the Internet and become an issue. The mechanism proposed in this chapter follows this philosophy. Particularly, our PET builds on the assumptions of the untrusted model defined in Sec. 2.3.1, where users mistrust any external entity and therefore strive to reveal as little private information as possible. Put differently, since users just trust themselves, privacy protection takes places on their side.

More specifically, *tag suppression* is a data-perturbative mechanism that has the purpose of preventing privacy attackers from accurately profiling users on the basis of the tags they specify. Conceptually, our approach protects user privacy to a certain extent, by dropping those tags that make a user profile show bias towards certain categories of interest. From a practical perspective, our tag-suppression technique is conceived to be implemented as a software application running on the users' local machine. The software implementation is responsible, on the one hand, for warning the user when their privacy is being compromised, and on the other, for helping them

[(b)] Search engines routinely record users' IP addresses, their search terms and the time when these searches are made. For example, Google and Yahoo! store all this information for a period of 9 and 18 months respectively [166].

decide which tags should be eliminated and which should not. Consequently, our approach guarantees user privacy to a certain degree without having to trust an external entity, but at the cost of some local processing overhead and, more importantly, the semantic loss incurred by suppressing tags.

5.2.1 Tag Suppression vs. Other Privacy-Protecting Techniques

In this section we compare tag suppression with other mechanisms that may help users protect their privacy in the scenario of personalized information systems. Our comparison is based on the classification used in Sec. 2.3.2, where we divided the state of the art in PETs into five categories, namely basic anti-tracking mechanisms, cryptography-based methods, TTP-based solutions, technologies relying on user collaboration and data-perturbative strategies. In this comparative analysis we shall resort to the trust models described in Sec. 2.3.1.

A naive approach to provide anonymous tagging would be using some of the anti-tracking mechanisms explored in Sec. 2.3.2. Disabling HTTP cookies or resorting to more sophisticated anti-tracking software (e.g., DoNotTrackMe [(c)] and Ghostery [(d)]) may hinder privacy attackers in their efforts to track users and thus profile them. On the one hand, this type of solutions may provide the highest level of privacy protection against an adversary wishing to profile users. But on the other, since personalized information systems cannot build user profiles, personalization is not possible. An alternative would be rejecting third-party cookies and accepting only those cookies issued by the personalized information system in question. The problem with this approach is that users would not be protected against this information system. Lastly, the fact that nearly all tagging systems require that users be logged in dismisses these anti-tracking mechanisms as a practical solution for the scenario of resource tagging.

The cryptography-based methods upon which PIR builds allow a user to retrieve an information item from a database without the owner of the database learning which particular item has been retrieved. The application of these methods to the specific scenario at hand, i.e., resource tagging, would not be straightforward as the

[(c)] `https://www.abine.com/dntdetail.php`
[(d)] `http://www.ghostery.com`

database should store the tags associated with the retrieved items. In other scenarios such as personalized Web search, PIR protocols are not an appropriate approach—the database owner is unable to ascertain the items users are interested in, and consequently profiling and therefore personalization are unfeasible. From the perspective of a user who wants to protect their privacy against a personalized Web search engine, PIR protocols are comparable, in terms of privacy and personalization, to the encryption of the search queries submitted by this user. Although PIR protocols do not require that users trust the database owner, they assume that the latter will collaborate in the execution of such protocols. Other important limitations that impede the direct application of these cryptographic mechanisms to personalized information systems are discussed in Sec. 2.3.2.

Another approach to provide anonymous tagging consists in a TTP forwarding users' tags to a personalized information system on their behalf. In adopting this simple strategy, the system does not know the user ID, but only the identity of the trusted entity. The problem with this strategy, however, is that personalized services cannot be provided since the information provider sees the TTP as a single user. An alternative is to use a TTP as a pseudonymizer. That is, the trusted entity gives each user a pseudonym; and each time a user wishes to post a tag, the TTP sends this tag, together with their pseudonym, to the service provider. While this alternative enables personalization, it does not prevent the provider from profiling users and eventually reidentify them [(e)]. Besides, all solutions relying on trusted entities require that users shift their trust from the service provider to these entities, possibly capable of collecting tags from different systems, which ultimately might facilitate user profiling via cross-referencing. However, even though users may be willing to assume such a trusted model, those entities may fail in the protection of user data. The most clear example is the AOL search data scandal [78] in 2006. More recent cases include Sony's security breach [167] and Evernote's [168].

[(e)] Secs. 2.3.2 and 4.6 elaborate on the reasons why pseudonyms may fail to protect anonymity and privacy.

Another class of TTP-based approaches are ACSs [6–15]. In the context of semantic tagging, anonymous communications may protect user privacy against the intermediary entities enabling the communications between tagging systems and their users. As we described in Sec. 2.3.2, routing messages through mix systems makes it more difficult for an attacker to track these messages. But while mixes may provide unlinkability to a certain extent, this is at the cost of delaying messages, which affects the usability of these systems and hence imposes a cost on them. In other words, mix systems pose a trade-off between anonymity and utility. In addition to this trade-off, other drawbacks are the deployment of infrastructure and, more importantly, the assumption that users are disposed to trust mixes. However, even though those systems were completely trustful, they could not prevent the recipient of those messages, i.e., the tagging system, from profiling users. Finally, we would like to highlight those systems relying on the principle of onion routing [13–15], which do not delay or reorder messages. Exactly as in mix networks, here trust is distributed among the onion routing nodes that collaborate in the forwarding protocol. These systems reduce the delay inherent in mixes, but suffer from the same limitations in terms of infrastructure and privacy protection.

There exist a myriad of alternatives based on user collaboration. One of the most popular is *Crowds* [68], which contemplates that a group of users wanting to browse the Web will collaborate to submit their requests. With this purpose, a user who decides to send a request to a Web server, selects first a member of the group at random and then forwards the request to it. When this member receives the request, it flips a biased coin to determine whether to send the request to another member or to submit it to the Web server. This process is repeated until the request is finally relayed to the intended destination. As a result, the Web server and any of the members forwarding the request cannot ascertain the identity of the true sender, that is, the member who initiated the request.

The protocol described above builds on the assumptions of the semi-trusted model defined in Sec. 2.3.1. This approach does not require the use of a TTP, but there are still several shortcomings that hinder its applicability to tagging systems, and more

generally, to personalized information systems. First, personalization is effective provided that all members of the group have similar profiles. Secondly, Crowds assumes that a number of users will participate in the protocol. However, even though it was possible, this solution could not protect user privacy against the collusion of all participants. Finally, another important drawback is the additional traffic intrinsic to this forwarding mechanism.

The above-mentioned shortcomings are, in fact, present in most of the PETs that leverage on user collaboration [69, 70, 90, 91]. An attempt to overcome these deficiencies is [89], which proposes a variation of the original Crowds protocol. The operation of this approach is essentially the same as in Crowds. The main difference is that users of this protocol are "friends" in some social network. This feature allows the protocol to create groups of users with similar interests, which makes this protocol suitable to be deployed in the scenario of personalized information systems. This proposal, however, does not overcome the drawbacks of the original Crowds protocol, in terms of traffic overhead and trustworthiness.

Unlike the traditional privacy-protecting mechanisms relying on access control policies, which determine whether the access to certain private data is granted or denied, data perturbation does not only contemplate these states "granted" and "denied", but also any other possibility between them. For example, the disclosure of certain parts of those data, or a slight perturbation of this private information. This kind of strategies permits users to preserve their privacy to a certain degree, although at the cost of certain loss in data utility. Further, unlike other approaches to protect user privacy, data perturbation may take place on the user side, without having to trust other entities. In other words, data perturbation is in line with the assumptions of our untrusted model.

In the scenario of personalized Web search, a widely used method to perturb user profiles consists in accompanying original queries or query keywords with false ones [95–103]. This conceptually-simple approach prevents privacy attackers from profiling users accurately based on their query history, but certainly at the expense of traffic overhead or redundancy. The main problem with query forgery is in the

generation of those false queries, since they should be indistinguishable from the genuine ones [104, 169, 170].

We could consider the application of this strategy to the scenario of tagging systems. A possible implementation of this *tag forgery* could be as follows: a user wishing to tag the Web page `www.mentalhelp.net` with "depression" could use the tag "sports" instead, to conceal their interest for this resource. On the one hand, adding random tags may distort the actual profile of interests, which provides this user with a certain level of privacy. But on the other, this strategy may have a far greater impact on semantic functionality than suppression does, since resources are assigned tags that do not describe, in principle, the actual content of such resources. In other words, the use of this technique is wholly inappropriate in collaborative tagging applications, where tags have the primary purpose of constructing meaning.

We would like to stress that the fact that forgery is not suitable for the tagging scenario does not mean that its applicability is limited to the context of Web search previously mentioned. Actually, forgery has shown to be appropriate for other applications such as PIR and recommendation systems. Precisely, later in Chapter 7 we propose the simultaneous use of forgery and suppression as a promising approach to privacy enhancement in personalized recommendation systems.

Another form of tag perturbation consists in replacing (specific) user tags with (general) tag categories. In conceptual terms, and resorting to the example above, the user would use the tag "health" instead of "depression". In this manner, the user would hide, to a certain extent, their genuine interest in that resource, but clearly at the cost of some vagueness or inaccuracy in the description of that Web page.

Among these approaches, we consider tag suppression as a suitable strategy for the enhancement of user privacy in the scenario of tagging systems, not only because of its simplicity in terms of implementation costs, but also because of its lower impact on semantic functionality. Lastly, we would like to emphasize the synergic effect of our approach in combination with other strategies based on data perturbation.

To sum up, the proposed technique appears as a simple approach in terms of infrastructure requirements, as users need not trust an external entity, the network

operator nor other users. Our PET, which contributes to the principle of *data minimization* [(f)], enables users to protect their privacy against the collusion of any passive attackers, but at the cost of semantic loss incurred by suppressing tags. Precisely, this privacy-utility trade-off also appears in ACSs and collaborative approaches. In these two cases, the degradation in utility is the delay introduced by mixes and the traffic overhead incurred by a forwarding strategy, respectively. Table 5.1 summarizes the major conclusions of this section.

Finally, despite the fact that the proposed strategy and other privacy-protecting mechanisms (such as those based on TTP or user collaboration) rely upon different assumptions, we would like to emphasize that these alternatives are not mutually exclusive and, more importantly, that users could benefit from the synergy of our approach and other systems building on the trusted or semi-trusted models. As a matter of fact, there are examples in the literature in which techniques assuming an untrusted model may complement TTP-based approaches perfectly. One example of this could be the use of dummy messages in combination with the traditional mix networks proposed in [10].

[(f)] According to [171], the data-minimization principle means that a data controller, e.g., the tagging server, should restrict the collection of personal data to what is strictly necessary to achieve its purpose. Also, it implies that the controller should store the data only for as long as is necessary to fulfil the purpose for which the information was collected.

Table 5.1: Comparison between our privacy-enhancing technique and other approaches that may contribute to privacy protection in the scenario of personalized information systems.

Approaches	**Underlying mechanism**	**Trust model**	**Disadvantages**
PIR [66, 67]	cryptographic methods	untrusted	◦ no personalization, ◦ database owner must collaborate, ◦ computational overhead.
anonymizer, pseudonymizer, digital credentials [16–18]	TTP	trusted	◦ users must trust an external entity, ◦ vulnerable to collusion attacks, ◦ traffic bottlenecks.
mix-based systems [6–10, 13–15]	TTP	trusted	◦ delay experienced by messages, ◦ users must trust an external entity, ◦ vulnerable to collusion attacks, ◦ infrastructure requirements.
Crowds and other P2P protocols [68, 68–70]	user collaboration	semi-trusted	◦ numerous users must collaborate, ◦ vulnerable to collusion attacks, ◦ traffic overhead.
tag suppression	data perturbation	untrusted	◦ semantic loss incurred by suppressing tags.

5.3 Adversary Model and Privacy Metric

In this chapter we assume the adversary model described in Chapter 4. Next we shall briefly review the main features of this model, and put it in the particular context of tagging systems. Afterwards, we shall specify the privacy metric used to evaluate our tag-suppression technique.

In this scenario, we assume that users are logged into the tagging system. It is only in this case that the system can profile users based on the tags they post and therefore can provide them with personalized services. Accordingly, our set of potential privacy attackers include, first, the tagging system, and secondly, the ISP and any networking infrastructure capable of capturing the tags submitted by users. Further, as tags are often publicly available to other users of the tagging system, we consider any other entity able to collect this information.

On the other hand, we suppose that the attacker models user profiles as *histograms* of relative frequencies of tags within a predefined set of categories of interest. As mentioned in Chapter 4, histograms, or equivalently, tags clouds, are the two models used by tagging systems to represent users' tagging activity.

According to this user profile representation, we suppose that the privacy attacker observes a perturbed version of this profile (i.e., the apparent profile), resulting from the suppression of certain tags. Based on this observation, we assume that the attacker is unable to discern whether the user is adhered to our tag-suppression technique or not, and thus this attacker cannot estimate the user's tag-suppression rate. We believe that this is a realistic assumption since, as we shall see later in Sec. 5.4, the proposed tag suppression strategy is conceived to be implemented as a software program running on the user's local machine. We would like to emphasize that this assumption must not be interpreted as *security through obscurity*, a principle that capitalizes on secrecy of design or implementation to provide security. Our adversary model does not pretend to hide the way tag suppression operates, but merely the fact that it is being used, an information that is only available on the user's side. This assumption is in line with other works [95, 102] that build on our untrusted model.

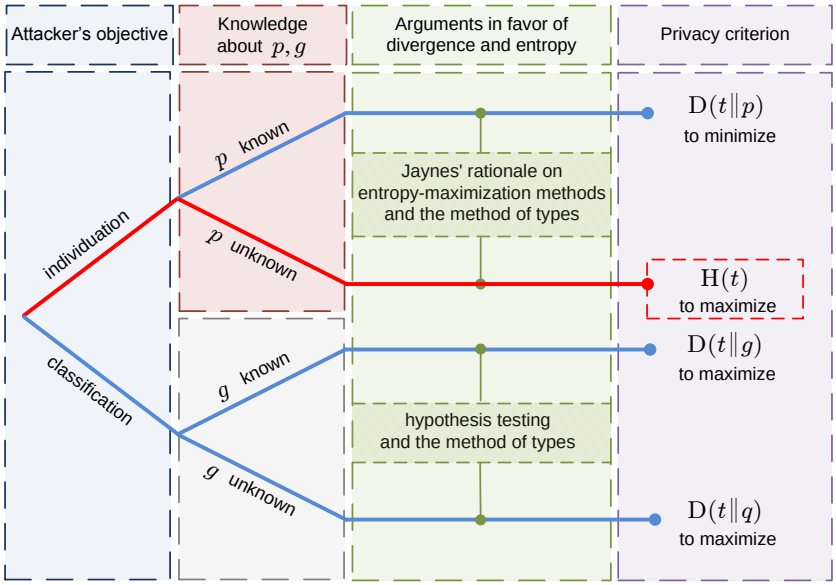


Figure 5.1: This figure corresponds to Fig. 4.4. Here we highlight in red the assumptions about the adversary model considered in this chapter. We assume, on the one hand, that our attacker wishes to individuate users, and on other that the population's tag distribution is not available to those users. Under these assumptions, the Shannon entropy of a user profile may be considered as a measure of anonymity gain.

The adversary model defined in Chapter 4 also contemplates the ultimate goal of profiling. In this chapter we assume that the attacker aims to *individuate* users, that is, its objective is to capture users whose interests deviate from the average profile. Under this assumption and according to the arguments presented in Chapter 4, Shannon's entropy and KL divergence are measures of the privacy of a user profile, or more precisely, of its anonymity. Since the population's tag distribution is frequently not at the disposal of users of tagging systems, we choose the entropy of the apparent profile as privacy metric. Recall that, under Jaynes' rationale behind entropy-maximization methods, the entropy of profile is a measure of its probability. In particular, the higher the entropy of a profile, the more likely it is, and the larger the number of users who have this profile and thus behave similarly.

Another interpretation of entropy stems from the observation that a privacy attacker will have actually gained some information about a user whenever their interests are significantly concentrated on a subset of categories. In other words, a user without any apparent interest in any category hides their preferences from an

attacker. Fig. 5.1 illustrates the assumptions about the adversary model considered in this section.

5.4 Architecture

In this section, we describe the functional components of a possible implementation of our tag-suppression technique. The proposed architecture helps users decide which tags should be suppressed and which should not. The fundamental purpose of our PET, and therefore of this architecture, is to hinder privacy attackers in their efforts to individuate users.

As anticipated in Sec. 5.2, our approach is devised to be implemented as a software application installed on the user's computer, for example, in the form of a Web browser add-on. Our architecture builds on the untrusted model defined in Sec. 2.3.1, which implies that users need not trust any external entity to protect their private data. We only assume, however, that users trust this software, in terms of the data it collects and its execution, exactly as they trust their own Web browser.

As we shall detail later, our approach triggers an alarm when user privacy is at risk. Afterwards, it recommends users which particular tags should be avoided in order to cope with such a threat. We would like to underline, though, that it is the user who has the last word, as they may decide to follow this recommendation or not. For this reason, we may view our approach as a recommendation system. Next, we describe some general characteristics of the architecture, and subsequently examine its internal components at a functional level.

Recall from Sec. 5.3 that we assume a passive attacker capable of ascertaining the tags posted by users of a tagging system. Our adversary can therefore be the system storing the tags posted by these users, or any attacker able to capture this information. In addition, we may contemplate the definition of the profile of a user tagging across several systems. In this case, we may also suppose that an attacker has the ability to link several profiles across different tagging applications. For the sake of simplicity, in this section we consider a user interacting with a single system. Our

architecture, however, could be easily extended to the more general case in which a user tagging activity spans a number of systems.

The proposed architecture gives some high-level specifications on how the profile of a user could be locally obtained by a software application implementing our technique. Our approach makes two assumptions about this user profile.

- First, when no perturbation is applied, we suppose that the profile computed on the user's side coincides with the profile built by the attacker. In other words, the profiling techniques used by the software application and those employed by the attacker lead to the same user profile. This means that the software and the adversary use the same predefined set of categories of interest and the same categorization algorithm, so that any tag posted by the user is classified into the same category by both the software and the adversary. We believe this is plausible assumption as long as the categorization process relies on a set of standard and widespread categories of interest.

- Secondly, as in any personalized recommendation system, our approach needs the user profile to start making recommendations about whether to eliminate a particular tag or not. Simply put, we contemplate a training phase before the proposed architecture starts working. Because an attacker might learn about the user's actual profile during this phase, we consider, as an alternative, that the user explicitly expresses their interests.

  In addition, the core component of our approach, the *suppression strategy generator*, assumes that the user profile remains stable over a long period of time. If the user does not explicitly declare their profile, we suppose that this steady-state condition is achieved after the training phase, once the user has tagged a sufficiently large number of items. This assumption is in line with the so-called *long-term* profiles which, in contrast to the *short-term* profiles, capture interests that are not subject to frequent changes [143]. We acknowledge, however, that a practical implementation of our technique should take into account that the user's tagging interests may vary significantly with time.

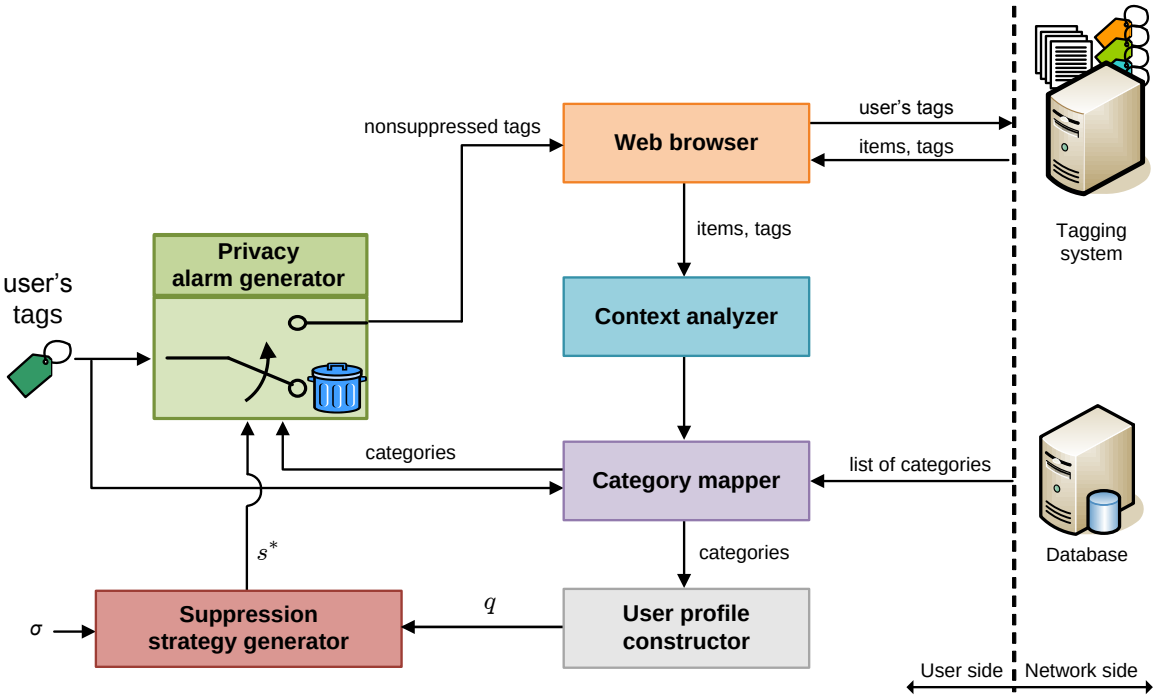


Figure 5.2: Internal components of the proposed architecture.

Fig. 5.2 depicts the proposed architecture, which consists of a number of modules, each of them performing a specific task. From a general perspective, this figure shows a user interacting with a tagging system, essentially an entity that stores information items (e.g., music, videos and Web pages) and tags associated with these items. Next, we provide a functional description of the modules of this architecture.

Web browser. This module is essentially responsible for the communication with the tagging system. Upon request of the user, it downloads information about the items the user wants to tag, as well as the tags posted by other users of the system. Afterwards, the retrieved data are delivered to the *context analyzer*, which processes all this information. Last but not least, the Web browser is also in charge of submitting the tags proposed by the user to the tagging system.

Context analyzer. This module is aimed to process the information retrieved by the Web browser. The purpose is to help the *category mapper* module to decide which category of the user profile should be updated. Said processing could be done by using the vector space model [172], as normally done in information retrieval,

to represent Web pages as tuples containing their most representative terms. For example, the term frequency-inverse document frequency (TF-IDF) could be applied to calculate the weights of each term appearing in the Web page that includes the item to be tagged. Later, the context analyzer could take a number of the most weighted terms of the tuple, and send them to the *category mapper* module. The selection of these terms could be done according to these two possible alternatives: a user could choose either a fixed number of terms, or those terms with weights above a threshold. This selection poses a compromise between accuracy and computational overhead, regardless the alternative chosen. The higher the resulting number of terms, the higher the accuracy in the categorization of the tag, but the higher the computational processing performed by the category mapper.

Category mapper. This component maps the tags submitted by the user into a predefined set of categories. This set of categories could be obtained by querying databases with this kind of information. For example, the Open Directory Project [(g)] could be used for this end. In some cases, these categories are provided by the tagging system, as it happens in YouTube. The categorization process performed by this module uses both the tag proposed by the user and the contextual information given by the context analyzer. The resulting categories are delivered to the modules *user profile constructor* and *privacy alarm generator*.

User profile constructor. It is responsible for the estimation of the user profile. Specifically, this module receives the categories corresponding to the tags submitted by the user and, accordingly, updates their profile. As mentioned before, our architecture assumes that, when estimating the histogram, the relative frequencies of activity are sufficiently stable once the user has posted a significant number of tags. An aspect that a real implementation of this module should consider is the initialization of the profile. An alternative could be initializing this profile to zero [142]. Another approach building on the principle of maximum entropy would use the uniform distribution instead.

We would like to emphasize that this module is active even when the user explicitly declares their profile. Since the profile specified by the user may not be an accurate

[(g)] `http://www.dmoz.org`

reflection of their online behavior, our architecture may decide, after the training phase, to replace it with the profile implicitly inferred from their tagging activity.

Suppression strategy generator. This module is the core of the architecture as it is directly responsible for the user privacy. First, this component is provided with the user profile and a *tag suppression* rate σ, which is a parameter reflecting the proportion of tags that the user is willing to suppress. Next, this module computes the optimal tuple of suppressing tags s^*, which contains information about the tags that should be suppressed. In particular, the component s_i^* is the percentage of tags that our architecture suggests eliminating in the category i. Finally, this tuple is given to the *privacy alarm generator* module. Later in Sec. 5.5, we provide a more detailed specification of this module by using a formulation of the trade-off between privacy and tag-suppression rate, which will enable us to compute the tuple s^*.

Privacy alarm generator. The functionality of this module is to warn the user when their privacy is being compromised. When the user submits a tag, this module waits for the category mapper block to send the category corresponding to that tag. Let i be the index of this category. The module afterwards receives the tuple s^* and proceeds as follows. With probability s_i^*, a privacy alarm is generated to warn the user. If the alarm is triggered, it is the user who must decide whether to eliminate the tag or not. Otherwise, our approach is not aware of any privacy threat and then sends the tag to the Web browser.

Having examined each individual component, we shall next describe how our approach would operate. For this, we may consider the case of a collaborative bookmarking system (e.g., Delicious), where users essentially tag Web pages. Fig. 5.3 illustrates this case. At the training phase, the user would browse the Web and submit tags to those pages of their interest (Fig. 5.3(a,b)). The contextual information derived by the context analyzer would be used to transform those tags into categories, and thus to construct the user profile.

The user profile would be used to calculate the tuple s^*. Then, at a certain point, the user could receive a privacy alarm when trying to submit a tag that would contribute to make the user profile significantly different from the uniform profile. If this was the case, the user would have to decide whether to eliminate the tag or not.

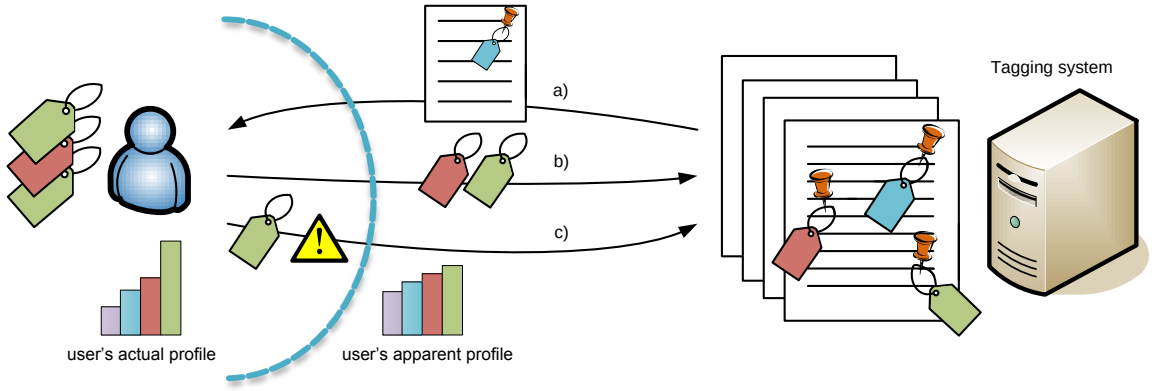


Figure 5.3: A user retrieves a Web page and the tags submitted by the other users from a server (a). Later, the user submits their own tags to that server (b). Afterwards, the user receives a privacy alarm when trying to submit a new tag (c).

Finally, if this tag was suppressed, the user's apparent profile would diverge from the actual user profile (Fig. 5.3(c)).

5.5 Trade-Off between Privacy and Tag-Suppression Rate

In this section we present a formulation that will enable us to specify the main block of the architecture proposed in Sec. 5.4, namely the suppression strategy generator.

We model the tags posted by a user as r.v.'s taking on values on a common finite alphabet of categories or topics, namely the set $\{1, \ldots, n\}$ for some integer $n \geqslant 2$. In our mathematical model, we assume these r.v.'s are i.i.d. This assumption allows us to describe user profiles by means of the PMF according to which such r.v.'s are distributed, which leads to an equivalent representation than that used in tagging systems. Accordingly, we define q as the probability distribution of the tags of a particular *user* and $\sigma \in [0, 1)$ as a *tag suppression* rate, which is the ratio of suppressed tags to total tags that the user is willing to eliminate. Concordantly, we define the user's *apparent* tag distribution t as $\frac{q-s}{1-\sigma}$ for some suppression strategy $s = (s_1, \ldots, s_n)$ satisfying $0 \leqslant s_i \leqslant q_i$ and $\sum s_i = \sigma$ for $i = 1, \ldots, n$. Conceptually, the user's apparent tag distribution may be interpreted as the result of, on the one hand, the suppression of certain tags from the actual user profile, that is, $q - s$, and one the other, the subsequent normalization by $\frac{1}{1-\sigma}$ so that $\sum_i t_i = 1$. The

information about which tags should be suppressed is encoded in the tag suppression strategy s. Namely, the component s_i is the relative frequency of tags that our mechanism suggests eliminating in the category i.

Based on the assumptions made in Sec. 5.3, we use Shannon's entropy [76] to quantify user privacy. Specifically, our privacy metric is the entropy of the user's apparent tag distribution t, according to Sec. 4.4.1, a measure of the probability of that distribution. Furthermore, the tag-suppression rate is our simplified measure of any loss in semantic functionality or data utility due to suppression. Consistently with both measures, we define the *privacy-suppression* function

$$\mathcal{P}(\sigma) = \max_{\substack{s \\ 0 \leqslant s_i \leqslant q_i, \\ \sum s_i = \sigma}} \mathrm{H}\left(\frac{q - s}{1 - \sigma}\right), \tag{5.1}$$

which characterizes the optimal trade-off between privacy and data utility, and formally expresses the intuitive reasoning behind tag suppression: the higher the tag-suppression rate σ, the higher the entropy of the apparent distribution, the likelihood of this distribution and thus the user privacy. We would like to stress that, in the context of this formulation, tag suppression does not attempt to hide the user's actual, genuine profile of interests, but the fact that such profile can make this user unique. In other words, we aim to perturb the user's actual profile so that their interests are more common, thus hindering an attacker in its efforts to individuate the user. Accordingly, if a user had a profile $q = u$, perturbation would not be needed, yet the exact profile would be disclosed.

For simplicity, we shall use natural logarithms throughout this chapter and refer to $\log_e$ as ln, particularly because all bases produce equivalent optimization objectives.

5.6 Theoretical Analysis

In this section, we shall analyze the fundamental properties of the privacy-suppression function (5.1) defined in Sec. 5.5, and present a closed-form solution to the maximization problem. Our theoretical analysis only considers the case when all

given probabilities are strictly positive:

$$q_i > 0 \text{ for all } i = 1, \ldots, n. \tag{5.2}$$

This assumption will be properly justified in Sec. 5.6.2. We shall suppose further, now without loss of generality, that

$$q_1 \leqslant \cdots \leqslant q_n. \tag{5.3}$$

Before proceeding with the mathematical analysis, it is immediate from the definition of the privacy-suppression function that its initial value is $\mathcal{P}(0) = \mathrm{H}(q)$. The behavior of $\mathcal{P}(\sigma)$ for $0 < \sigma < 1$ is characterized by the theorems presented in this section. The notation used throughout this section is summarized in Table 5.2.

5.6.1 Monotonicity and Quasiconcavity

Our first theoretical characterization, namely Lemma 5.1, investigates two elementary properties of the privacy-utility trade-off. The lemma in question shows that the trade-off is nondecreasing and quasiconcave. The importance of these two properties is that they confirm the evidence that an optimal tag suppression strategy will never lead to a degradation in privacy protection. In other words, an increase in the tag-suppression rate does not lower the entropy of the apparent profile.

Theorem 5.1. *The privacy-suppression function $\mathcal{P}(\sigma)$ is nondecreasing and quasiconcave.*

Proof: First, let $0 \leqslant \sigma < \sigma' \leqslant 1$. Based on the solution s to the maximization problem corresponding to $\mathcal{P}(\sigma)$, consider the tag suppression strategy s' given by the equation

$$\frac{q - s'}{1 - \sigma'} = \frac{q - s}{1 - \sigma}.$$

The feasibility of s' may be checked, on the one hand, by observing that the constraints $0 \leqslant s'_i \leqslant q_i$ are equivalent to $0 \leqslant \frac{q_i - s'_i}{1-\sigma'} \leqslant \frac{q_i}{1-\sigma'}$ for $i = 1, \ldots, n$. According to the implicit definition of s', we may rewrite these constraints as $0 \leqslant \frac{q_i - s_i}{1-\sigma} \leqslant \frac{q_i}{1-\sigma'}$. Given that s is feasible, the left-hand inequality is satisfied. The right-hand inequality is

also verified by simply noting that $\frac{q_i}{1-\sigma} < \frac{q_i}{1-\sigma'}$. On the other hand, it is immediate to check that $\sum_i s'_i = \sigma$.

Once we have confirmed that s' is feasible, we now turn to prove the first part of the lemma. Since the feasibility of s' does not necessarily imply that s' is a maximizer of the problem corresponding to $\mathcal{P}(\sigma')$, it follows that $\mathcal{P}(\sigma') \geqslant \mathrm{H}\left(\frac{q-s'}{1-\sigma'}\right) = \mathcal{P}(\sigma)$, and consequently, that the privacy-suppression function is nondecreasing.

Finally, the quasiconcavity of the privacy-suppression function is directly proved by the fact that $\mathcal{P}(\sigma)$ is a nondecreasing function of σ. ■

The quasiconcavity of the privacy-suppression function (5.1) guarantees its continuity on the interior of its domain, namely $(0, 1)$, but it is fairly straightforward to verify, directly from the definition of $\mathcal{P}(\sigma)$ and under the positivity assumption (5.2), that continuity also holds at the interval endpoint 0.

5.6.2 Critical Suppression

The following theorem will confirm the intuition that there must exist a tag-suppression rate beyond which *critical privacy* is achievable, in the sense that the privacy-suppression function attains its maximum value, that is, $\mathcal{P}(\sigma) = \ln n$. Precisely, this *critical suppression* is

$$\sigma_{\mathrm{crit}} = 1 - n \min_i q_i = 1 - n\, q_1,$$

according to the labeling assumption (5.3). From the above, it is interesting to note that σ_{crit} becomes worse (closer to one) with worse (smaller) ratio $\frac{q_1}{u_1} = n\, q_1$.

Theorem 5.2 (Critical Suppression)**.** *Let u be the uniform distribution on $\{1, \ldots, n\}$, that is, $u_i = 1/n$. For all $\sigma \in [0, 1)$, if $\sigma \geqslant \sigma_{\mathrm{crit}}$, then $\mathcal{P}(\sigma) = \mathrm{H}(u) = \ln n$. In addition, the optimal tag suppression strategy is $s^* = q - u\,(1-\sigma)$, for which the user's apparent distribution and the uniform's match. Conversely, if $\sigma < \sigma_{\mathrm{crit}}$, then $\mathcal{P}(\sigma) < \ln n$.*

Proof: We consider only the nontrivial case when $q \neq u$, which implies that $q_1 < 1/n$ and, consequently, $\sigma_{\mathrm{crit}} > 0$. To confirm this implication, assume $q \neq u$ and suppose now that $q_1 \geqslant 1/n$. Taking into account the labeling assumption (5.3) and the fact that q is a probability distribution in the sense that $\sum_i q_i = 1$, we arrive at the contradiction that q must be the uniform distribution. Given that $q_1 < 1/n$, it

Table 5.2: Description of the variables used in our notation.

Symbol	Description
n	number of categories of interest into which tags are classified
q	the *actual* user profile is the genuine profile of interests
σ	the *tag-suppression rate* is the percentage of tags that the user is willing to suppress
s	a *suppression strategy* is an n-tuple with the percentage of tags that the user should eliminate in each category
t	the *apparent* user profile is the perturbed profile, as observed from the outside, resulting from the elimination of certain tags
u	uniform profile across the n tag categories
$\mathrm{H}(t)$	*user privacy* is measured as the Shannon entropy of the apparent user profile
$\mathcal{P}(\sigma)$	*privacy-suppression function* modeling the trade-off between privacy and utility, the latter being measured as the tag-suppression rate
σ_{crit}	the *critical suppression* is the suppression rate beyond which the privacy-suppression function attains its maximum value or critical privacy $\mathcal{P}_{\mathrm{crit}}$

immediately follows that $\sigma_{\mathrm{crit}} > 0$. The converse, that is, $\sigma_{\mathrm{crit}} > 0$ implies $q \neq u$, is easily checked by noting that when $q_1 < 1/n$, q cannot be, by definition, the uniform distribution. On the other hand, the positivity assumption (5.2) ensures that $\sigma_{\mathrm{crit}} < 1$.

Once we have determined the interval of values in which σ_{crit} is defined, we now proceed to confirm the feasibility of s^*. It is clear from its form that $\sum_i s_i^* = \sigma$, thus it suffices to verify that $0 \leqslant s_i^* \leqslant q_i$. First, observe that the right-hand inequality is satisfied for all i as $\sigma < 1$. Secondly, note that requiring that $s_i^* = q_i - \frac{1}{n}(1-\sigma) \geqslant 0$ for all i is equivalent to $\sigma \geqslant 1 - n\, q_i$, and finally to

$$\sigma \geqslant \max_i\, 1 - n\, q_i = 1 - n\, \min_i q_i,$$

as assumed in the theorem. Interestingly, observe that the expression for the critical suppression is independent of the privacy criterion assumed. To complete the first part of the proof, it is immediate to check that the proposed s^* maximizes the user privacy, since the uniform distribution maximizes entropy.

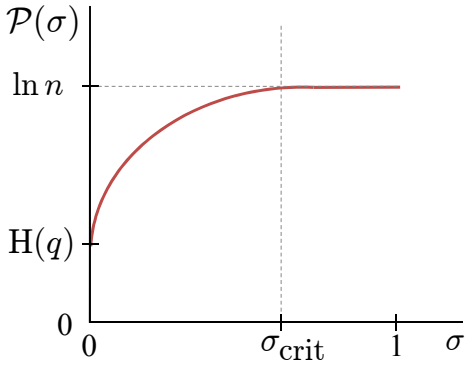


Figure 5.4: Conceptual plot of the privacy-suppression function.

Now it remains to prove that $\mathcal{P}(\sigma) < \ln n$ when $\sigma < \sigma_{\mathrm{crit}}$. To this end, recall that the KL divergence between the user's apparent distribution and the uniform distribution may be written as

$$\mathrm{D}(t\|u) = \sum_i t_i \ln \frac{t_i}{u_i} = \ln n - \mathrm{H}(t),$$

as argued in Sec. 2.2. But the information inequality [76] asserts that $\mathrm{D}(t\|u) \geqslant 0$, with equality if, and only if, $t = u$ for all i. Hence, when $\sigma < \sigma_{\mathrm{crit}}$, the solution t to the optimization problem corresponding to $\mathcal{P}(\sigma)$ satisfies that $t \neq u$, and therefore $\mathcal{P}(\sigma) = \mathrm{H}(t) < \ln n$. ■

After routine manipulation, we may write the optimal solution at exactly the critical suppression as

$$s_i^* = q_i - q_1,$$

equal to zero if, and only if, $q = u$. Owing to the fact that we are dealing with relative rather than absolute frequencies, it is not surprising that $s_1^* = 0$ at $\sigma = \sigma_{\mathrm{crit}}$. More generally, by virtue of the labeling assumption (5.3), we observe that only the first components of s^* may vanish. Fig. 5.4 conceptually illustrates the results derived from Lemma 5.1 and Theorem 5.2.

Before proceeding further with our theoretical analysis, we would like to remark that our assumption about the strictly positiveness of q is conveniently made, albeit not without loss of generality, to guarantee that the critical privacy $\mathcal{P}_{\mathrm{crit}}$ is attained for a suppression $\sigma < 1$, as proved in Theorem 5.2.

5.6.3 Closed-Form Solution

Our last theorem, Theorem 5.4, will provide a closed-form solution to the maximization problem involved in the definition of the privacy-suppression function (5.1). This solution will be obtained from a resource allocation lemma, namely Lemma 5.3, which addresses an extension of the usual water filling problem. Even though Lemma 5.3 provides a parametric-form solution, fortunately, we shall be able to proceed towards an explicit closed-form solution, albeit piecewise.

More specifically, this lemma considers the allocation of resources $x_1, \ldots, x_n$ minimizing the sum $\sum_i f_i(x_i)$ of convex cost functions on the individual resources. Resources are assumed to be nonnegative, upper bounded by positive thresholds b_i, and to amount to a total of $\sum_i x_i = \theta$, for some $\theta > 0$. The well-known water-filling problem [75, §5.5] may be regarded as a special case when resources are not upper bounded and $f_i(x_i) = -\ln(\alpha_i + x_i)$, for $\alpha_i > 0$.

Lemma 5.3 (Resource Allocation)**.** *For all $i = 1, \ldots, n$, let $f_i : [0, b_i] \to \mathbb{R}$ be twice differentiable on $[0, b_i)$, with $f_i'' > 0$, and hence strictly convex. Additionally, assume that $\lim_{x_i \to b_i^-} f_i'(x_i) = \infty$. Because $f_i'' > 0$, f_i' is strictly increasing, and, interpreted as a function from $[0, b_i)$ to $f_i'([0, b_i))$, invertible. Denote the inverse by $f_i'^{-1}$. Consider the following optimization problem in the variables $x_1, \ldots, x_n$:*

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{n} f_i(x_i) \\ &\text{subject to} && 0 \leqslant x_i \leqslant b_i, \text{ for all } i, \\ &&& \text{and } \sum_{i=1}^{n} x_i = \theta, \text{ for some } \theta > 0. \end{aligned}$$

(i) *The solution to the problem exists, is unique and of the form $x_i^* = \max\{0, f_i'^{-1}(\nu)\}$, for some $\nu \in \mathbb{R}$ such that $\sum_i x_i^* = \theta$.*

(ii) *Suppose further, albeit without loss of generality, that $f_n'(0) \leqslant \cdots \leqslant f_1'(0)$. Then, either $f_i'(0) < \nu \leqslant f_{i-1}'(0)$ for $i = 2, \ldots, n$, or $f_i'(0) < \nu$ for $i = 1$, and*

for the corresponding index i,

$$x_j^* = \begin{cases} f_j'^{-1}(\nu), & j = i, \dots, n \\ 0 \qquad\quad , & j = 1, \dots, i-1 \end{cases},$$

and

$$\sum_{j=1}^{n} x_j^* = \sum_{j=i}^{n} f_j'^{-1}(\nu) = \theta.$$

Proof: The existence and uniqueness of the solution is a consequence of the fact that we minimize a strictly convex function over a compact set. Systematic application of the Karush-Kuhn-Tucker (KKT) conditions [75] leads to the Lagrangian cost

$$\mathcal{L} = \sum f_i(x_i) - \sum \lambda_i x_i + \sum \mu_i (x_i - b_i) - \nu \left(\sum x_i - \theta \right),$$

which must satisfy $\frac{\partial \mathcal{L}}{\partial x_i} = 0$, and finally to the conditions

$$\begin{aligned} 0 \leqslant x_i \leqslant b_i, \sum x_i = \theta \quad & \text{(primal feasibility)}, \\ \lambda_i, \mu_i \geqslant 0 \quad & \text{(dual feasibility)}, \\ \lambda_i x_i = 0, \mu_i (x_i - b_i) = 0 \quad & \text{(complementary slackness)}, \\ f_i'(x_i) - \lambda_i + \mu_i - \nu = 0 \quad & \text{(dual optimality)}. \end{aligned}$$

Since $\lim_{x_i \to b_i^-} f_i'(x_i) = \infty$, it follows from the dual optimality condition that $x_i < b_i$. But then, the complementary slackness condition implies that $\mu_i = 0$, and consequently, we may rewrite the dual optimality condition as $f_i'(x_i) = \lambda_i + \nu$. By eliminating the slack variables λ_i, we finally obtain the simplified condition $f_i'(x_i) \geqslant \nu$. In addition, observe that since $f_i'(x_i) = \lambda_i + \nu$, the complementary slackness condition implies that $(f_i'(x_i) - \nu)\, x_i = 0$. In short, we may rewrite the dual optimality and the complementary slackness conditions equivalently as

$$\begin{aligned} f_i'(x_i) \geqslant \nu \quad & \text{(dual optimality)}, \\ (f_i'(x_i) - \nu)\, x_i = 0 \quad & \text{(complementary slackness)}. \end{aligned}$$

Now, we proceed to directly solve these equations. To this end, recall that, since $f_i'' > 0$, f_i' is strictly increasing. Consider, first, the case when $f_i'(0) \geqslant \nu$, or equivalently, $f_i'^{-1}(\nu) \leqslant 0$. Suppose that $x_i > 0$, so that by complementary slackness,

$f_i'(x_i) = \nu \leqslant f_i'(0)$, contradicting the fact that f_i' is strictly increasing. Consequently, $x_i = 0$.

Consider now the opposite case, that is, when $f_i'(0) < \nu$, or equivalently $f_i'^{-1}(\nu) > 0$. In this case, the only conclusion consistent with the dual optimality condition is $x_i > 0$. But then, it follows from the complementary slackness condition that $f_i'(x_i) = \nu$, or equivalently, $x_i = f_i'^{-1}(\nu)$. This could be interpreted as a Pareto equilibrium. Specifically, for all positive resource $x_i > 0$, the marginal ratios of improvements $f_i'(x_i)$ must all be the same. Otherwise, minor allocation adjustments on the resources could improve the overall objective. In summary,

$$x_i = \max\{0, f_i'^{-1}(\nu)\},$$

which proves claim (i) in the lemma.

In order to verify (ii), observe that whenever $\nu \leqslant f_{i-1}'(0) \leqslant \cdots \leqslant f_1'(0)$ holds for some $i = 2, \ldots, n$, then $f_{i-1}'^{-1}(\nu), \ldots, f_1'^{-1}(\nu) \leqslant 0$, and thus $x_{i-1} = \cdots = x_1 = 0$. Note that the index $i = n + 1$ is not permitted, since the zero solution, that is, $x_i = 0$ for all $i = 1, \ldots, n$, contradicts the primal feasibility condition $\sum_i x_i = \theta$. ■

Next, we shall provide a closed-form solution for the privacy-suppression function. However, before presenting the theorem in question, we shall introduce some notation. Let $\bar{Q}_i = \sum_{j=i+1}^{n} q_j$ denote the complementary cumulative distribution function. In addition, define

$$\sigma_i = \bar{Q}_i - q_i(n - i),$$

for $i = 1, \ldots, n$, and, conveniently, define $\sigma_0 = 1$. Note that $\sigma_n = 0$, that $\sigma_1 = 1 - n\, q_1 = \sigma_{\text{crit}}$, and consistently with Theorem 5.2, the solution in this theorem at $\sigma = \sigma_{\text{crit}}$ becomes $\frac{q_j - s_j^*}{1-\sigma} = \frac{1}{n}$, for $j = 1, \ldots, n$. Further, define

$$\tilde{q} = \left(q_1, \ldots, q_{i-1}, \tfrac{\bar{Q}_{i-1}}{n-i+1}, \ldots, \tfrac{\bar{Q}_{i-1}}{n-i+1}\right),$$
$$\tilde{s} = \left(0, \ldots, 0, \tfrac{\sigma}{n-i+1}, \ldots, \tfrac{\sigma}{n-i+1}\right),$$

a distribution in the probability simplex in $\mathbb{R}^n$, and an n-tuple representing a tag suppression strategy, respectively.

Theorem 5.4. *For any $i = 2, \ldots, n$, $\sigma_i \leqslant \sigma_{i-1}$, with equality if, and only if, $q_i = q_{i-1}$. For any $i = 1, \ldots, n$ and any $\sigma \in [\sigma_i, \sigma_{i-1}]$, the optimal suppression strategy is*

$$s_j^* = \begin{cases} 0 & , \quad j = 1, \ldots, i-1 \\ q_j - \frac{\bar{Q}_{i-1} - \sigma}{n-i+1}, & j = i, \ldots, n \end{cases},$$

and, consequently, the corresponding optimal user's apparent tag distribution is

$$t_j^* = \begin{cases} \frac{q_j}{1-\sigma} & , \quad j = 1, \ldots, i-1 \\ \frac{\bar{Q}_{i-1} - \sigma}{(1-\sigma)(n-i+1)}, & j = i, \ldots, n \end{cases}$$

Accordingly, the corresponding, maximum entropy yields the privacy-suppression function

$$\mathcal{P}(\sigma) = \mathrm{H}\left(\frac{\tilde{q} - \tilde{s}}{1 - \sigma}\right).$$

Proof: From the definition of σ_i and under the labeling assumption (5.3), it is immediate to check the monotonicity of these suppression thresholds.

Now, we proceed to prove the rest of the theorem for the nontrivial case $\sigma \in (0, 1)$. Using the definition of entropy, we may write the objective function in the (original) optimization problem (5.1) as $-\mathrm{H}(t) = \sum_i t_i \ln t_i$, with $t_i = \frac{q_i - s_i}{1-\sigma}$, since the maximization of entropy is equivalent to the minimization of negative entropy. Recall that s is optimal for the original problem if, and only if, s is optimal for the scaled problem. After this convenient, straightforward transformation, the objective function exposes the structure of the privacy-suppression optimization problem as a special case of the resource allocation lemma, Lemma 5.3. Specifically, the functions $f_i(s_i) = t_i \ln t_i$ of s_i are twice differentiable on $[0, q_i)$, and satisfy $f_i'' > 0$ and $\lim_{s_i \to q_i^-} f_i'(s_i) = \infty$. Further, the equality constraint in (5.1) becomes $\sum_i s_i = \sigma$. In this special case, $f_i'(s_i) = -\frac{1}{1-\sigma}\left(\ln \frac{q_i - s_i}{1-\sigma} + 1\right)$ and

$$f_i'^{-1}(\nu) = q_i - (1 - \sigma)\, e^{-(1-\sigma)\nu - 1},$$

the solution for s_i when $s_i > 0$.

The labeling assumption (5.3) is equivalent to the assumption that $f_n'(0) \leqslant \cdots \leqslant f_1'(0)$ in the lemma, since $f_i'(0) = -\frac{1}{1-\sigma}\left(\ln \frac{q_i}{1-\sigma} + 1\right)$ is a strictly decreasing function

of q_i. From the second part of the lemma,

$$\sigma = \sum_{j=i}^{n} {f'_j}^{-1}(\nu) = \bar{Q}_{i-1} - (n-i+1)(1-\sigma)\, e^{-(1-\sigma)\nu - 1},$$

and hence,

$$\nu = -\frac{1}{1-\sigma}\left(\ln \frac{\bar{Q}_{i-1} - \sigma}{(1-\sigma)(n-i+1)} + 1\right).$$

Now it suffices to substitute ν into $f'_i(\nu)$ in order to obtain the expression for the nonzero optimal suppression strategy s_j in the theorem. The optimal user's apparent tag distribution t is easily derived from this expression.

Next, we shall confirm the interval of values of σ in which it is defined. To this end, observe that the condition $f'_i(0) < \nu$ in the lemma, is equivalent to

$$-\frac{1}{1-\sigma}\left(\ln \frac{q_i}{1-\sigma} + 1\right) < -\frac{1}{1-\sigma}\left(\ln \frac{\bar{Q}_{i-1} - \sigma}{(1-\sigma)(n-i+1)} + 1\right),$$

and finally, after routine algebraic manipulation, to

$$\sigma > \bar{Q}_i - q_i(n-i).$$

We could proceed to carry out an analogous analysis on the upper bound condition $\nu \leqslant f'_{i-1}(0)$ of the lemma to find out the interval of values of σ in which the solution is defined. However, we note that, because a unique solution will exist for each σ, the intervals resulting from imposing $f'_i(0) < \nu \leqslant f'_{i-1}(0)$ must be contiguous and nonoverlapping, hence, of the form $(\sigma_i, \sigma_{i-1}]$. Further, since $\mathcal{P}(\sigma)$ is continuous on $[0, 1)$, we may write the intervals as $[\sigma_i, \sigma_{i-1}]$ in lieu of $(\sigma_i, \sigma_{i-1}]$.

To complete the proof, we shall express the privacy-suppression function in terms of the optimal user's apparent tag distribution, that is, $\mathcal{P}(\sigma) = -\sum_{j=1}^{n} t_j \ln t_j$. We split the sum into two parts, namely,

$$-\sum_{j=1}^{i-1} \frac{q_j}{1-\sigma} \ln \frac{q_j}{1-\sigma} - \sum_{j=i}^{n} \frac{\bar{Q}_{i-1} - \sigma}{(1-\sigma)(n-i+1)} \ln \frac{\bar{Q}_{i-1} - \sigma}{(1-\sigma)(n-i+1)},$$

where we observe that the terms in the second sum do not depend on j. From this expression, it is straightforward to identify the terms of $\mathcal{P}(\sigma)$ as the entropy of the

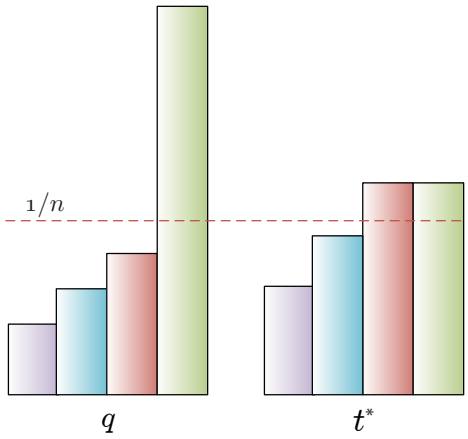


Figure 5.5: A user's tag distribution q and their corresponding apparent tag distribution t^* after an optimal suppression of tags.

distribution

$$\left(\frac{q_1}{1-\sigma},\ldots,\frac{q_{i-1}}{1-\sigma},\ \frac{\bar{Q}_{i-1}-\sigma}{(1-\sigma)(n-i+1)},\ldots,\frac{\bar{Q}_{i-1}-\sigma}{(1-\sigma)(n-i+1)}\right),$$

precisely the distribution $\frac{\tilde{q}-\tilde{s}}{1-\sigma}$, given at the end of the theorem. ■

The optimal tag suppression strategy in Theorem 5.4 is interpreted as follows. On the one hand, only tags corresponding to the categories $j = i,\ldots,n$ are suppressed. This is not surprising because, precisely, these are the categories with the highest probabilities, or roughly speaking, with probabilities furthest away from the uniform distribution. On the other, the optimal user's apparent tag distribution within those categories does not depend on j, and hence they all have the same probability. Further, consistently with the fact that we are dealing with relative frequencies, the components of the apparent distribution belonging to the categories $j = 1,\ldots,i-1$ are obtained by normalizing the genuine user distribution. Fig. 5.5 captures this intuitive analysis by illustrating a simple example with $n = 4$ categories. Namely, this figure shows a user with an actual profile q who is willing to accept a tag-suppression rate $\sigma \in [\sigma_3, \sigma_2]$, causing that a privacy attacker observe an optimal user's apparent profile t^* significantly different from q, specially in those categories with the highest ratio $\frac{q_j}{u_j} = \frac{q_j}{1/n}$.

A number of conclusions can be drawn from the results obtained in this last theorem. The following two sections will be focused on the analysis of the behavior of the privacy-suppression function at low suppression rates and high privacy.

5.6.4 Low-Suppression Case

This section investigates the privacy-suppression $\mathcal{P}(\sigma)$ in the case when $\sigma \simeq 0$.

Proposition 5.5 (Low Suppression)**.** *In the nontrivial case when $q \neq u$, there exists a positive integer i with suppression thresholds satisfying $0 = \sigma_n = \cdots = \sigma_i < \sigma_{i-1}$. For all $\sigma \in [0, \sigma_{i-1}]$, the optimal tag suppression strategy s^* contains $n-i+1$ nonzero components, and the slope of the privacy-suppression function at the origin is $\mathcal{P}'(0) = \mathrm{H}(q) + \ln q_n$.*

Proof: The hypothesis $q \neq u$ implies that $n > 1$, and the existence of a positive integer i enabling us to rewrite the labeling assumption (5.3) as

$$q_1 \leqslant \cdots \leqslant q_{i-1} < q_i = \cdots = q_n,$$

and to express q_j as $\frac{\bar{Q}_{i-1}}{n-i+1}$, for $j = i, \ldots, n$. On account of Theorem 5.4,

$$0 = \sigma_n = \cdots = \sigma_i < \sigma_{i-1} \leqslant \cdots \leqslant \sigma_1,$$

and for all $\sigma \in [0, \sigma_{i-1}]$, we have that

$$\mathcal{P}(\sigma) = \mathrm{H}\left(\frac{\tilde{q} - \tilde{s}}{1 - \sigma}\right).$$

It is routine to check that

$$\mathcal{P}'(0) = -\sum_{j=1}^{i-1} q_j \ln q_j - \sum_{j=i}^{n} q_j \ln \tfrac{\bar{Q}_{i-1}}{n-i+1} + \ln \tfrac{\bar{Q}_{i-1}}{n-i+1} = -\sum_{j=1}^{n} q_j \ln q_j + \ln q_n,$$

where the last equality follows from the fact that $q_i = \cdots = q_n$, as shown at the beginning of this proof. ■

Now we define the *relative increment factor*

$$\delta = \frac{\mathcal{P}'(0)}{\mathcal{P}(0)} = 1 + \frac{\ln q_n}{\mathrm{H}(q)}.$$

The results from Proposition 5.5 allows us to approximate the privacy-suppression function at $\sigma \simeq 0$ as

$$\mathcal{P}(\sigma) \simeq \mathrm{H}(q) + \sigma\,(\mathrm{H}(q) + \ln q_n)$$

or, in terms of the relative increment,

$$\frac{\mathcal{P}(\sigma) - \mathrm{H}(q)}{\mathrm{H}(q)} \simeq \delta\,\sigma. \tag{5.4}$$

In conceptual terms, q_n characterizes the privacy gain at low suppression rates, together with $\mathrm{H}(q)$, in contrast to the fact that the ratio $\frac{q_1}{1/n}$ determines σ_{crit}, the minimum suppression rate for which the critical privacy is achievable, as defined in Sec. 5.6.2. We mentioned in that section that $q_1 < 1/n$ in the nontrivial case when $q \neq u$. An entirely analogous argument shows that $q_n \geqslant 1/n$, with equality if, and only if, $q = u$, since the opposite, that is, $q_i < 1/n$, leads to a contradiction. This result allows us to conclude that $\delta < 1$, unless $q = u$, for which, unsurprisingly, δ becomes zero. In other words, the relative privacy gain (5.4) is lower than the suppression introduced. Namely, the privacy increment at low suppression rates becomes less noticeable with smaller q_n, for a fixed $\mathrm{H}(q)$.

5.6.5 High-Privacy Case

Next, we shall analyze the case when $\sigma \simeq \sigma_{\mathrm{crit}}$ and consequently the privacy-suppression function attains its maximum value. To this end, consider the index $i = 2$ just to check that, whenever $\sigma \in [\sigma_2, \sigma_{\mathrm{crit}}]$, for $q \neq u$,

$$\mathcal{P}(\sigma) = \mathrm{H}\left(\frac{\left(q_1, \frac{1-q_1}{n-1}, \ldots, \frac{1-q_1}{n-1}\right) - \left(0, \frac{\sigma}{n-1}, \ldots, \frac{\sigma}{n-1}\right)}{1-\sigma}\right) < \ln n.$$

In addition, we are implicitly assuming that $q_1 \neq q_2$, so that, by virtue of Theorem 5.4, $\sigma_2 < \sigma_{\mathrm{crit}}$. Consequently, we skip an empty interval and may express the privacy-suppression function as

$$\mathcal{P}(\sigma) = -\frac{q_1}{1-\sigma} \ln \frac{q_1}{1-\sigma} - \frac{1-q_1-\sigma}{1-\sigma} \ln \frac{1-q_1-\sigma}{(1-\sigma)(n-1)}.$$

From this expression, it is routine to conclude that $\mathcal{P}'(\sigma_{\text{crit}}) = 0$ and $\mathcal{P}''(\sigma_{\text{crit}}) = -\frac{1}{q_1^2 n^2 (n-1)}$, and finally,

$$\mathcal{P}(\sigma) \simeq \ln n + \frac{1}{2}\mathcal{P}''(\sigma_{\text{crit}})(\sigma - \sigma_{\text{crit}})^2.$$

We would like to remark that the fact that $\mathcal{P}(\sigma)$ admits a quadratic approximation for $\sigma \simeq \sigma_{\text{crit}}$, with $\mathcal{P}'(\sigma_{\text{crit}}) = 0$, may be determined directly from the fundamental properties of Fisher information [76]. Recall that for a family of distributions f_θ indexed by a scalar parameter θ, $\mathrm{D}(f_\theta \| f_{\theta'}) \simeq \frac{1}{2}\,\mathrm{I}(\theta')(\theta' - \theta)^2$, where $\mathrm{I}(\theta') = \mathrm{E}\left(\frac{\partial}{\partial \theta'} \ln f_{\theta'}\right)^2$ is Fisher information. Denote by $t^*_\sigma = \frac{q - s^*}{1-\sigma}$ the family of optimal apparent tag distributions, indexed by the suppression rate. Theorem 5.2 guarantees that $t^*_{\sigma_{\text{crit}}} = u$, thus we may write $\mathcal{P}(\sigma) = \mathrm{H}(t^*_\sigma) = \ln n - \mathrm{D}(t^*_\sigma \| t^*_{\sigma_{\text{crit}}})$. Under this formulation, it is clear that the Fisher information associated with the suppression rate is $\mathrm{I}(\sigma_{\text{crit}}) = -\mathcal{P}''(\sigma_{\text{crit}})$.

Lastly, we would like to note that the observation at the end of Sec. 5.6.2 that $s^*_1 = 0$ at $\sigma = \sigma_{\text{crit}}$ is consistent with the fact that σ_{crit} is the endpoint of the interval corresponding to the solution for s^* with $n - 1$ nonzero components in Theorem 5.4.

5.6.6 Numerical Example

In this section, we show various numerical results for a simple but insightful example that attempts to illustrate the formulation and the theoretical analysis presented in Secs. 5.5 and 5.6. The evaluation of our privacy-enhancing mechanism in a real-world application is presented later in Sec. 5.7.

In this practical example, we shall consider three categories and assume that the user's distribution is $q = (0.100, 0.200, 0.700)$, thus fulfilling both the positivity and the labeling assumptions (5.2,5.3). On account of Theorem 5.4, the suppression thresholds are $\sigma_3 = 0$, $\sigma_2 = 0.500$ and $\sigma_1 = \sigma_{\text{crit}} = 0.700$. In addition, the initial privacy value is $\mathcal{P}(0) \simeq 0.8018$, which is the privacy level achieved by a user who is not willing to accept the suppression of any tag. Furthermore, Sec. 5.6.4 and 5.6.5 allow us to characterize the behavior of the privacy-suppression function for $\sigma = 0$ and $\sigma = \sigma_{\text{crit}}$. Concretely, the first and second order approximations are determined

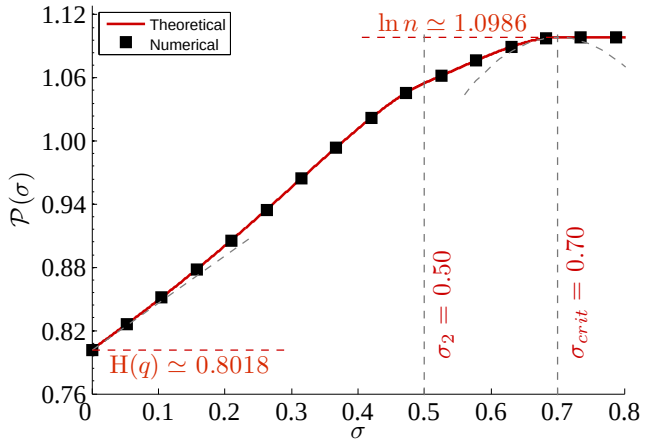


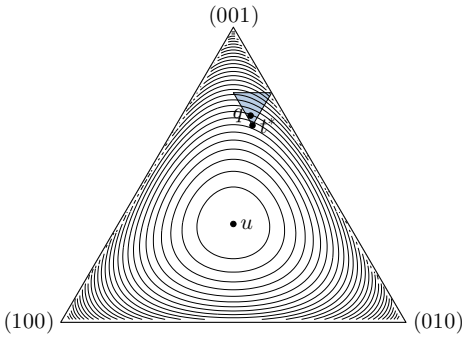
Figure 5.6: Optimal trade-off curve between privacy and suppression, and the corresponding approximations and suppression thresholds for $q = (0.100, 0.200, 0.700)$.

by the quantities $\mathcal{P}'(0) \simeq 0.4451$ and $\mathcal{P}''(\sigma_{\text{crit}}) \simeq -5.56$. All these results are captured in Fig. 5.6, where the privacy-suppression function $\mathcal{P}(\sigma)$ is represented. The optimization problem involved in the definition of this function has been computed theoretically, by simply applying Theorem 5.4, and numerically [(h)].

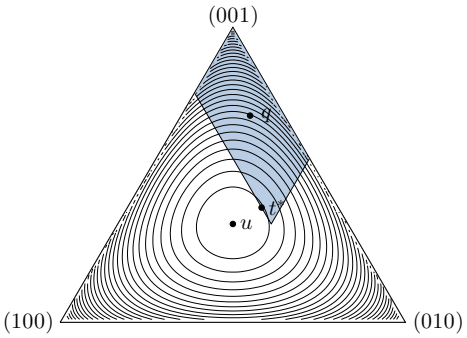
After observing the behavior of the optimal trade-off curve between privacy and suppression, now we turn to examine the optimal apparent tag distribution for a set of suppression rates. To this end, the user's distribution q, the optimal apparent distribution t^* and the uniform distribution u are represented in the probability simplices shown in Fig. 5.7. In addition, the contours of the entropy $\mathrm{H}(\,\cdot\,)$ of a distribution in the simplex are depicted. More interestingly, this figure also shows the region, highlighted in dark blue, which corresponds to all the possible apparent tag distributions, not necessarily optimal, for a given suppression rate. Namely, this feasible region results from the intersection of the set $\left\{ t = \frac{q-s}{1-\sigma} \middle| \, 0 \leqslant s_i \leqslant q_i, \sum_i s_i = \sigma \right\}$, and the probability simplex.

We now turn our attention to Fig. 5.7(a), where a suppression $\sigma \in [\sigma_3, \sigma_2]$ has been selected to check that, according to the notation of Theorem 5.4, s^* has $n - i + 1 = 1$ nonzero components. Geometrically, this places the solution t^*, not entirely unexpectedly, at one vertex in the feasible region. In addition, observe that a

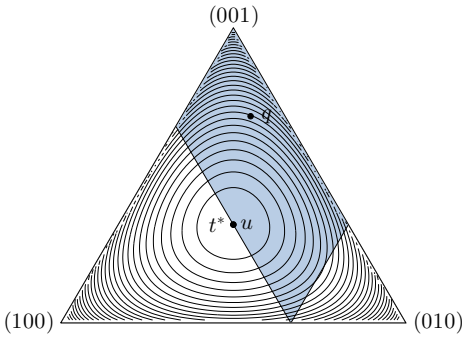
[(h)] The numerical method chosen is the interior-point optimization algorithm [75] implemented by the Matlab R2012b function `fmincon`.



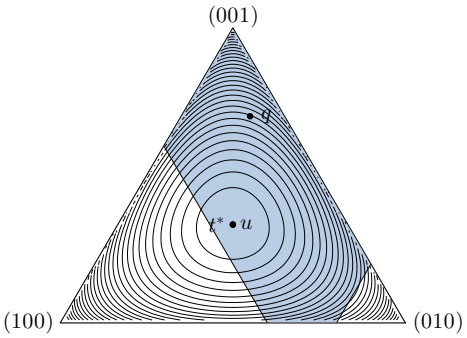
(a) $\sigma = 0.100$, $\sigma/\sigma_{\text{crit}} \simeq 0.143$, $\mathcal{P}(\sigma) \simeq 0.8487$, $\mathcal{P}(\sigma)/\mathcal{P}(0) \simeq 1.0585$, $s^* = (0.000, 0.000, 0.100)$, $t^* \simeq (0.111, 0.222, 0.667)$.



(b) $\sigma = 0.550$, $\sigma/\sigma_{\text{crit}} \simeq 0.786$, $\mathcal{P}(\sigma) \simeq 1.0688$, $\mathcal{P}(\sigma)/\mathcal{P}(0) \simeq 1.3330$, $s^* = (0.000, 0.025, 0.525)$, $t^* \simeq (0.222, 0.389, 0.389)$.



(c) $\sigma = 0.700$, $\sigma/\sigma_{\text{crit}} = 1$, $\mathcal{P}(\sigma) \simeq 1.0986$, $\mathcal{P}(\sigma)/\mathcal{P}(0) \simeq 1.3702$, $s^* = (0.000, 0.100, 0.600)$, $t^* = u$.



(d) $\sigma = 0.750$, $\sigma/\sigma_{\text{crit}} \simeq 1.071$, $\mathcal{P}(\sigma) \simeq 1.0986$, $\mathcal{P}(\sigma)/\mathcal{P}(0) \simeq 1.3702$, $s^* \simeq (0.017, 0.117, 0.617)$, $t^* = u$.

Figure 5.7: Probability simplices showing u, q and t^* for several interesting values of σ.

suppression of 10 % increases the user privacy to a 5.8 % of the original privacy $\mathrm{H}(q)$. This confirms an interesting result obtained in Sec. 5.6.4, where we concluded that the relative increment factor δ for low-suppression rates was lower than the suppression introduced. In Fig. 5.7(b) the suppression rate is on the interval $[\sigma_2, \sigma_{\text{crit}}]$, leading to an optimal suppression strategy s^* with $n - i + 1 = 2$ nonzero components. In this case, the solution t^* is placed on one edge of the feasible region. Additionally, note that a suppression of 55 % increments the user privacy to a 33 % of its original value. The case in which $\sigma = \sigma_{\text{crit}}$ and thus user privacy attains its maximum value is depicted in Fig. 5.7(c). When this happens, s^* still has $n - i + 1 = 2$ nonzero components. Precisely, note that $s_3^* = q_3 - q_1$ and $s_2^* = q_2 - q_1$, which perfectly agree

with the results obtained at the end of Sec. 5.6.2. Finally, the case when $\sigma > \sigma_{\text{crit}}$, which certainly does not make sense, is shown in Fig. 5.7(d). In this particular case, s^* has $n - i + 1 = 3$ nonzero components and t^* falls into the interior of the feasible region.

5.7 Experimental Analysis

In this section, we analyze the extent to which our technique enables users to enhance their privacy in a real-world tagging application. In this analysis we contemplate the impact that the suppression of tags has on the semantic functionality of this application, but tackle this in a simplified manner, by using a tractable measure of data utility, namely the tag-suppression rate. More sophisticated metrics of any loss in semantic functionality due to suppression will be explored later in Chapter 6.

We start, in Sec. 5.7.1, by examining the data set used to conduct the experimental evaluation. To make user profiles tractable, Sec. 5.7.2 describes a methodology for mapping tags into a small set of meaningful categories of interest. Finally, Sec. 5.7.3 presents the experimental results.

5.7.1 Data set

We applied the proposed technique to BibSonomy, a popular social bookmarking and publication-sharing system. In particular, we experimented with the data set retrieved by the Knowledge & Data Engineering Group at the University of Kassel [173]. The data set in question comprises those bookmarks and publications tagged by approximately two thousand users. The information is organized in the form triples (*username*, *resource*, *tag*), each one modeling the action of a user who associates a resource, being a bookmark or a publication, with a tag. Our data set contains 671 807 of these triples, which were posted from Jan. 1989 to Dec. 2007, and includes 1 921 users, 206 941 resources and 58 755 tags. It is worth mentioning that no preprocessing was done, although usernames were anonymized.

5.7.2 Tag Categorization

The representation of a user profile as a normalized histogram across these 58 755 tags is clearly an inappropriate approach for our experiments; not only because of the intractability of the profile, but also because it makes it difficult to have a quick overview of the user interests. For example, for users posting the tags "welfare", "Dubya" and "campaign" it would be preferable to have a higher level of abstraction that enables us to conclude, directly from the inspection of their profiles, that they are interested in politics. This level of abstraction is not only interesting for our experimental evaluation, but also it represents what an attacker would eventually do to capture user interests.

The categorization of tags therefore permits modeling user profiles in a tractable manner, on the basis of a reduced set of meaningful categories of interest, consistently with the representation assumed in Sec. 5.3. In this section we summarize the methodology that we followed to categorize the tags of our data set. This categorization process is described in more detail later in Chapter 6, where we focus on the more practical and experimental aspects of our tag-suppression technique.

To accomplish this categorization, first we carried out some preprocessing to discard those tags considered as spam. With this intention, we eliminated the tags with a number of characters over 26, which in our data set represented the 99^{th} percentile. Furthermore, we got rid of those triples without tags. As a result of this preprocessing, the number of triples became 665 052, and thus the number of users, resources and tags reduced to 1 916, 206 697, and 50 900, respectively.

In what follows, the categorization process can be roughly conceptualized in three steps. This process is in line with other works in the field [174, 175].

(i) Computation and recording of simultaneous occurrence of two tags under a common resource, in the form of a *co-occurrence matrix*. Tags may then be modeled as numeric vectors of co-occurrences, obtained as columns or rows within this matrix.

(ii) Definition of a quantitative measure of semantic dissimilarity, namely the *cosine distance*, between tag vectors, under the principle that similar tags should induce similar co-occurrence profiles.

(iii) Clustering of said tag vectors with the *Lloyd's algorithm*, replacing all tags within each cluster by a common representative tag, minimizing the average semantic distance just defined.

The application of the first step allowed us to obtain a matrix of co-occurrences c_{ij}. Then we filtered this matrix as we wanted to preserve just those tags with a sufficiently high level of co-occurrence. For this reason, we dropped those tags satisfying $\sum_j c_{ij} < \tau$, for a certain threshold τ. We chose $\tau = 100$ since we wished to retain at least 80% of the triples. After this filtering process the number of users, resources, tags and triples became 1 737, 190 478, 5 057 and 540 904, respectively.

Equipped with the cosine distance as a measure of dissimilarity, we proceeded to apply Lloyd's algorithm [(i)]. The application of this clustering algorithm enabled us to group the 5 057 tags into 5 categories, which gave us a granularity level sufficiently aggregated as to avoid having user profiles with many empty categories. Subsequently, the resulting categories were sorted in increasing order of popularity of their tags, with the aim of satisfying the labeling assumption (5.3). Although this classification does not necessarily imply that all user profiles meet this condition, in our experiments we shall ultimately rearrange the categories of each individual profile to fulfil it. Lastly, the tags in each category were ordered in decreasing order of proximity to the centroid [(j)].

In a last stage, and on account of the positivity assumption (5.2), we eliminated those users who did not tag across all categories. In addition, we dropped users with an activity level lower than 50 tags, since it would have been difficult to calculate a reliable estimate of their profiles with such a few tags. Accordingly, the number of users, resources and triples became 209, 144 904 and 447 203, respectively.

[(i)] Lloyd's algorithm [176], which is normally referred to as k-means in the computer science community, is a popular iterated algorithm for grouping data points into a set of k clusters. Sec. 6.5.2 provides further details on this algorithm.

[(j)] The complete results of this clustering are available to other researchers at `http://sites.google.com/site/javierparraarnau/publications`.

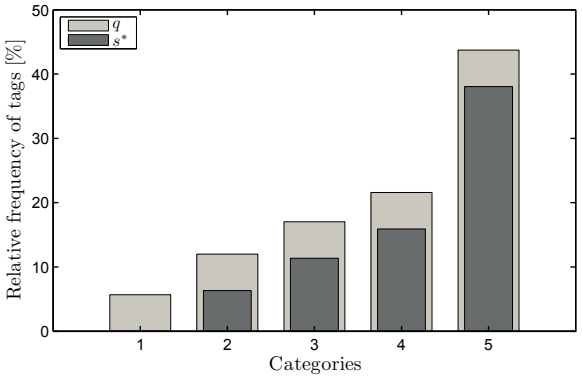


Figure 5.8: In this figure, we plot the actual user profile q of the particular user considered in Sec. 5.7.3, who posted a total of 1 075 tags across all categories. Additionally, we plot the optimal suppression strategy s^* for $\sigma = \sigma_{\text{crit}} \simeq 0.7163$, that is, the percentage of tags that the user should refrain from tagging in each category in order to achieve the uniform profile.

5.7.3 Results

In this section, we examine the extent to which our technique contributes to privacy preservation. For this purpose, first we explore how a particular user in our data set benefits from the application of an optimal tag suppression strategy; and secondly, we analyze the effect of this optimal suppression when the whole population of users enhance their privacy by using a common tag-suppression rate.

As detailed in previous sections, tag suppression requires that a user specify a rate indicating the fraction of tags they are disposed to eliminate. Based on this suppression rate and the user profile across the $n = 5$ categories obtained in Sec. 5.7.2, our approach solves the optimization problem (5.1). The result of this optimization is a suppression strategy s^*, that is, an n-tuple containing the percentage of tags that the user should eliminate in each category. In our first series of experiments, we select a particular user in our data set [(k)] and compute this suppression strategy in the special case when the user specifies $\sigma = \sigma_{\text{crit}}$. Both the actual profile of the user in question and the optimal strategy are plotted in Fig. 5.8, where it is shown one of the theoretical results obtained in Sec. 5.6.2, namely the fact that $s_i^* = q_i - q_1$ for any category i. In addition, Fig. 5.9 illustrates the optimal trade-off curve between privacy and suppression, which we calculated theoretically and numerically. The suppression

[(k)] This specific user is identified by the number `633` in [173].

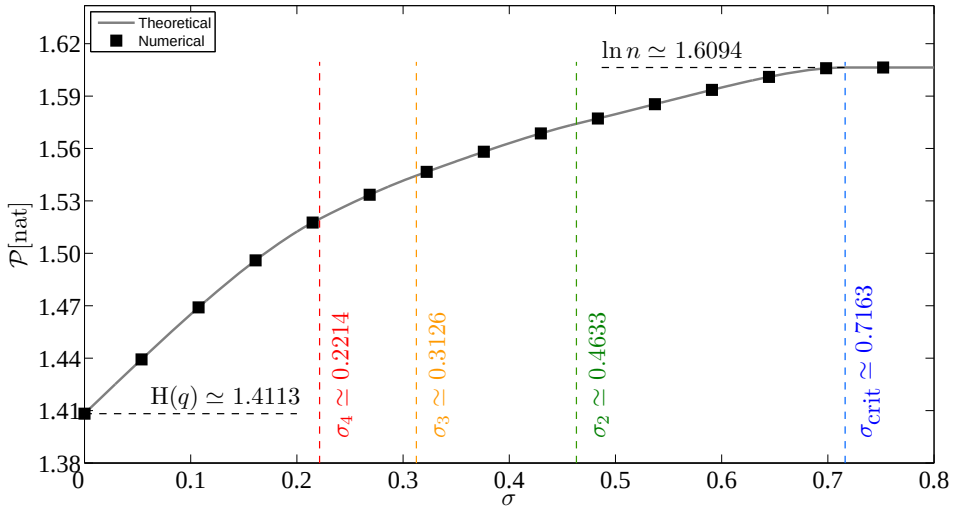
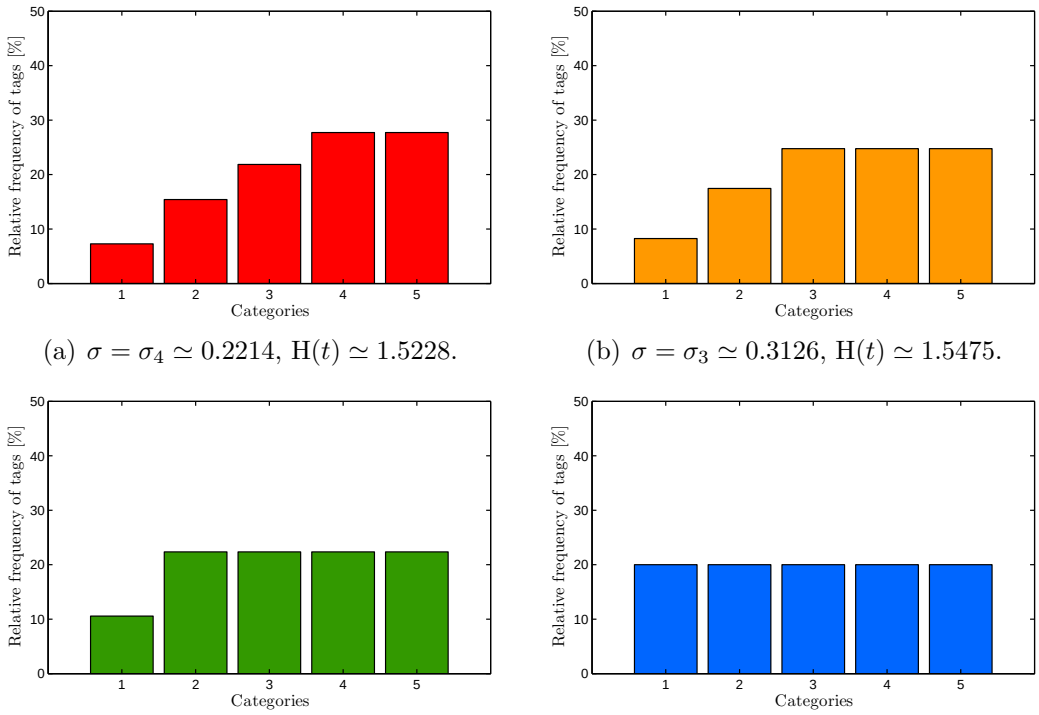


Figure 5.9: We plot the privacy-suppression function and the suppression thresholds for one particular user in our data set.

thresholds σ_i shown in this figure indicate the suppression rates beyond which the components $j = i, \ldots, n$ of the apparent profile t have the same probability. This effect is observed in Fig. 5.10, where we represent t precisely for these interesting values of σ.

The second set of experiments contemplates a scenario where all users apply our technique by using a common tag-suppression rate. Under this assumption, Fig. 5.12 shows the privacy protection achieved by these users in terms of percentile curves (10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th}) of relative privacy gain. Noteworthy is the fact that certain users obtain privacy gains between 100% and 235%, although, clearly, at the cost of high suppression rates. Another eye-opening finding is the distribution of the suppression thresholds σ_i plotted in Fig. 5.11. Recall that we also refer to σ_1 as the critical suppression σ_{crit}. Particularly, we observe that 86.6% of users have $\sigma_1 \in [0.9, 1)$, whereas the remaining percentage of users lie in the interval $[0.7, 0.9)$. In practice, this means that all users will require a high suppression rate for their profiles to become completely uniform. Although this might be certainly controversial, this is not a poor performance of our mechanism, but a consequence of the stringent privacy requirement imposed by such uniformity. As a matter of fact, the distributions of σ_i, for $i = 2, 3, 4$, indicate that the components t_j with $j = i, \ldots, n$ may be uniform at



(a) $\sigma = \sigma_4 \simeq 0.2214$, $\mathrm{H}(t) \simeq 1.5228$.

(b) $\sigma = \sigma_3 \simeq 0.3126$, $\mathrm{H}(t) \simeq 1.5475$.

(c) $\sigma = \sigma_2 \simeq 0.4633$, $\mathrm{H}(t) \simeq 1.5772$.

(d) $\sigma = \sigma_1 \simeq 0.7163$, $\mathrm{H}(t) \simeq 1.6094$.

Figure 5.10: We represent the apparent profile t of a particular user in the special case when the suppression rate coincides with the suppression thresholds σ_i, $i = 1, \ldots, 4$. Recall that t is the perturbed profile resulting from the elimination of tags and observed from the outside. At these interesting values of suppression, we observe how the components of t corresponding to the categories $j = i, \ldots, 5$ are balanced. In the end, when the critical suppression σ_1 is attained, t becomes u and $\mathrm{H}(t) = \ln 5 \simeq 1.6094$. The actual profile of this specific user is depicted in Fig. 5.8.

a significantly lower cost. For example, 32.5% of users have 3 out of 5 components evenly balanced for a suppression rate below 68%.

In closing, the results shown in this section illustrate how our mechanism perturbs the user profile observed from the outside and how this perturbation enables users to protect their privacy to a certain degree.

5.8 Conclusions

There exists a large number of proposals for privacy protection in the semantic Web. Within these approaches, tag suppression arises as a simple strategy in terms of

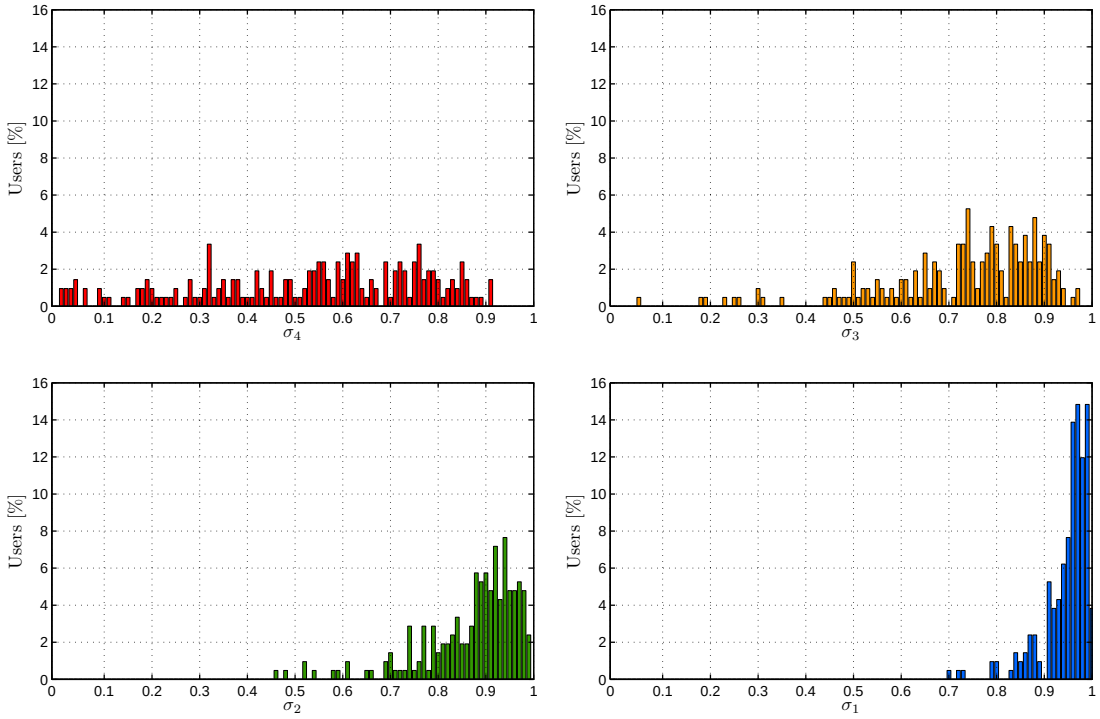


Figure 5.11: We plot the distribution of the suppression thresholds σ_i, for $i = 1, \ldots, 4$. In the special case when $\sigma \geqslant \sigma_1 = \sigma_{\text{crit}}$, the apparent user profile is the uniform profile across all categories.

infrastructure requirements, as users need not trust an external entity nor the network operator. The fact that the proposed strategy builds on the assumptions of the untrusted model does not prevent it from being used in combination with other mechanisms, e.g., those based on TTP or user collaboration. Our technique, in fact, may contribute to improve the effectiveness of these mechanisms. However, like other approaches in the literature, our data-perturbative approach comes at the cost of some processing overhead but more importantly at the expense of semantic loss incurred by suppressing tags. In other words, tag suppression poses an inherent trade-off between privacy on the one hand, and data utility on the other.

Our first contribution is an architecture that outlines how our tag-suppression technique could be implemented in practice. The proposed architecture helps users refrain from proposing certain tags in order to hinder attackers in their efforts to target peculiar users. The main component of our proposal is a module responsible for obtaining an optimal tag suppression strategy. Our approach uses this information

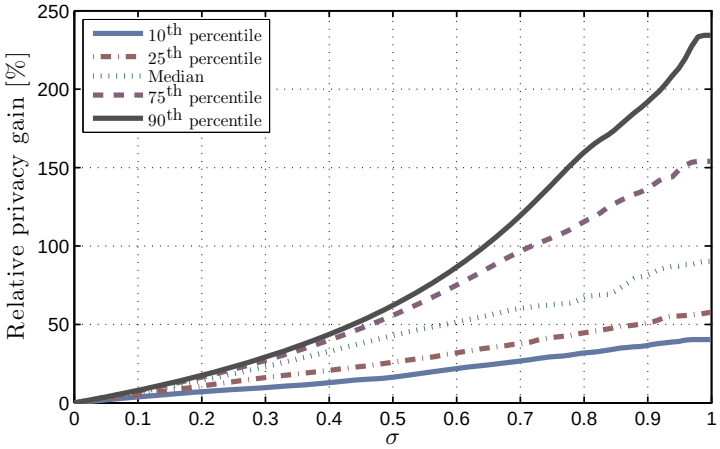


Figure 5.12: We consider the case when all users in our data set protect their privacy by using a common tag-suppression rate. Built on this premise, we then plot some percentiles curves of privacy gain against this common suppression rate.

to warn users when their privacy is being compromised and it is then for the users to decide whether to eliminate those tags or not.

Our second contribution is a systematic, mathematical approach to the problem of optimal tag suppression. On the one hand, we measure privacy as the entropy of the user's apparent tag distribution, after the suppression of some tags, and justify it under an adversary whose objective is to individuate users. On the other hand, we model any loss in semantic functionality as the tag-suppression rate, that is, the fraction of tags a user consents to eliminate. This simplified measure of data utility enables us to formulate the privacy-utility trade-off as a mathematically tractable optimization problem.

In our model, we represent user tags as r.v.'s taking on values on a common finite alphabet of categories or topics. This allows us to describe user profiles as PMFs, a representation that is frequently used in popular tagging systems such as BibSonomy, CiteULike and Delicious. The proposed model, however, is restricted to relative frequencies, relevant against content-based attacks, but does not deal with differences in the absolute frequencies, which certainly could be exploited by traffic analysis. Further, we assume that the adversary is unable to know whether a particular user is applying our PET. We consider this is a reasonable assumption as the elimination of tags takes place on the user side.

As a result of our theoretical analysis, we provide a closed-form solution for the optimal tag suppression strategy and a privacy-suppression function modeling the optimal trade-off curve. Our theoretical study first proves that the privacy-suppression function $\mathcal{P}(\sigma)$ is nondecreasing and quasiconcave. Subsequently, we show that, under the positivity assumption (5.2), there exists a critical suppression $\sigma_{\text{crit}} < 1$ beyond which the critical privacy is achievable. Specifically, this σ_{crit} only depends on the minimum ratio $\frac{q_j}{u_j}$ of probabilities between the user's tag distribution q and the uniform distribution u. More interestingly, for a given suppression σ the suppression of tags only affects the categories $j = i, \ldots, n$, precisely those with the highest probabilities among all categories. Not unexpectedly, the number of categories exposed to suppression, that is, $n - i + 1$, increases with σ. In the particular case when $\sigma = \sigma_{\text{crit}}$, only the category $i = 1$ remains unchanged. With regard to the optimal user's apparent distribution, the components of t^* corresponding to the categories $j = i, \ldots, n$ have the same probability, whereas the probability of the other components is obtained by normalizing the actual user distribution.

In addition, the privacy-suppression function is characterized at low suppression rates and at high privacy. Specifically, we present a first-order Taylor approximation for $\sigma \simeq 0$ in the nontrivial case when $q \neq u$, from which we conclude that q_n determines, together with the initial privacy value, the privacy gain at low suppression. Also, we prove that this privacy gain is lower than the suppression introduced. Besides, we provide a second-order approximation for $\sigma \simeq \sigma_{\text{crit}}$, assuming that probabilities q_j are strictly increasing. Finally, the fact that $\mathcal{P}'(\sigma)$ vanishes at $\sigma = \sigma_{\text{crit}}$ is regarded as a property of the Fisher information.

Our theoretical analysis is then illustrated with a simple but insightful example. But this is not until Sec. 5.7 where we provide an experimental evaluation of our privacy-enhancing mechanism. In that section, we consider the use of tag suppression in a real-world application and assess experimentally the extent to which our approach could help users protect their privacy. Our experiments also evaluate the impact that our PET would have on semantic functionality, but approach this in a simplified manner, by using the tag-suppression rate as a measure of utility. The next chapter, Chapter 6, is entirely devoted to investigate the impact on the semantic functionality

of an enhanced collaborative tagging application, by using a more sophisticated utility metric, in particular the percentages of tags that resources lose as a result of tag suppression.

Chapter 6

Privacy-Preserving Enhanced Collaborative Tagging

6.1 Introduction

Collaborative tagging is one of the most widespread and popular services available online. First provided by social bookmarking sites only—e.g., Delicious, Digg and StumbleUpon—, it is currently supported by nearly any type of social Web application, and it is used to annotate any kind of online and offline resources, such as Web pages, images, videos, movies, music, and even blog posts.

The main purpose of collaborative tagging is to classify resources based on user feedback, expressed in the form of free-text labels, i.e., tags. The novelty of such an approach to content or resource categorization has been seen, in recent years, as a challenging research topic, in part because collaborative tagging provides the basis for the semantic Web, a network that will connect online resources based on their meanings, and not only on their uniform resource identifiers [177].

Although these days collaborative tagging is mainly used to support tag-based resource discovery and browsing, it can also be exploited for other purposes. In Chapter 5 we mentioned that semantic tagging is intimately related to personalization and that many collaborative tagging systems have recently begun to offer personalized

services. In these systems, the user's tastes and interests are inferred *implicitly*, based on the tags they submit.

This implicit form of profile construction is actually the most used way to model user preferences. The problem with such approach, however, is that it requires users to interact frequently enough so that their profiles become an accurate reflection of their interests and the provider can then start to offer an effective personalized service. This is known as the *cold-start problem* [178], and a solution to this is for personalized information systems to allow users to *explicitly* express their preferences initially, for example, in the form of content-filtering rules or categories of interest.

While these two forms of user profile construction, i.e., implicit and explicit, are employed by current personalized information systems, no collaborative tagging service enables its members to explicitly specify preferences. In order to achieve this enhanced use, the current architecture of collaborative tagging services should be extended by including a *policy layer*. The aim of this layer would be to enforce user preferences, intentionally denoting resources on the basis of the set of tags associated with them, and, possibly, other parameters concerning their trustworthiness, e.g., the percentage of users who have added a given tag or the social relationships and characteristics of those users. This is a new research topic, and, to the best of our knowledge, the only work addressing this issue is reported in [179], where a multi-layer policy-based collaborative tagging system is described.

The incorporation of this policy layer would provide users with enhanced Web access functionalities like content filtering and discovery. The downside of these policy-based collaborative tagging services is that they may exacerbate the risk of privacy. First, because users explicitly communicate part of their interests. And secondly, because this explicit feedback does not preclude the collaborative tagging system from profiling users based on their tags. Ultimately, the combination of explicit and implicit data may lead this system to construct more precise profiles. In other words, besides the support to policy enforcement, enhanced collaborative tagging would require another layer addressing privacy protection.

Although the collection of end users' private information stored by social services, like Facebook, is now recognized as a privacy threat [180, 181], it is worth noting

that the public availability of user-generated data (as tags are) would allow even a rudimentary attacker to profile users. Further, the huge number of users using collaborative tagging services, and the fact that collaborative tagging is a service supported virtually by any social online application, increase the risk of cross referencing, thereby seriously compromising user privacy. Indeed, it could be possible to correlate the account of a user with other accounts they may have at different services, which would imply gaining far more accurate information about their profile.

Consequently, collaborative tagging would require the enforcement of mechanisms that enable users to protect their privacy by allowing them to hide certain user-generated contents (unless they desire otherwise), without making them useless for the purposes they have been provided in a given online service. This means that privacy-preserving mechanisms must not negatively affect the accuracy and effectiveness of the service, e.g., tag-based browsing, filtering, or personalization.

In this chapter we make a first contribution in this direction by proposing an architecture that incorporates two layers on support of enhanced and private collaborative tagging. More specifically, the proposed architecture consists of a bookmarking service and two additional services built on it. The former service enables users to specify policies both to block undesired Web content and to denote resources of interest. The latter implements tag suppression, a privacy-preserving technology that we investigated in Chapter 5.

The combination of these two services allows us then to broaden the functionality of collaborative tagging systems and, at the same time, to provide users with a mechanism to preserve their privacy when tagging. However, the fact that our PET comes at the cost of data utility poses a trade-off between privacy on the one hand, and on the other the effectiveness of the enhanced collaborative tagging services enabled by said policy layer. Our second and main contribution is an extensive performance evaluation of this architecture, showing its effectiveness in terms of privacy guarantees, data utility and filtering capabilities for two key scenarios, namely parental control and resource recommendation.

The results presented in this chapter are an extension of [45].

Chapter Outline

The remainder of this chapter is organized as follows. Sec. 6.2 examines the architecture of the proposed enhanced social tagging service. Sec. 6.3 describes how tag suppression fits into this architecture. Sec. 6.4 introduces the two reference scenarios on which our PET has been tested, whereas performance results are reported and discussed in Sec. 6.5. Sec. 6.6 concludes this chapter.

6.2 Overview of the Proposed Approach

As we discussed in Sec. 6.1, social bookmarking services are among the most used social services, and, thanks to their support to collaborative tagging, they can be currently considered as the most valuable knowledge acquisition tools, as far as online resources are concerned.

We also pointed out that collaborative tagging is not exploited to its full potential, since it is typically used just to support tag-based resource browsing and search, despite the fact that, collaborative tagging systems can be easily enhanced without modifying their core architecture, since they provide access to the collected information via APIs, which can be easily exploited by external applications. One of the reasons is that the size of the collected data sets is too big to allow the enforcement of even simple mechanisms, concerning, e.g., personalization, content filtering and quality assessment.

In addition, we commented that current collaborative tagging systems do not enable users to explicitly convey their preferences. As a matter of fact, the exploitation of explicit relationships and user preferences has been studied only in [179], where a multi-layer architecture is proposed integrating a basic social tagging service with trust relationships and user preferences. One of the notable characteristics of such framework is the support of a rule layer, which can be used to express and enforce user preferences. Such preferences are coded into policies explicitly specifying the set of trustworthy tags by denoting their creators in terms of their relationships and/or characteristics. Also, they state which action must be performed by the system when

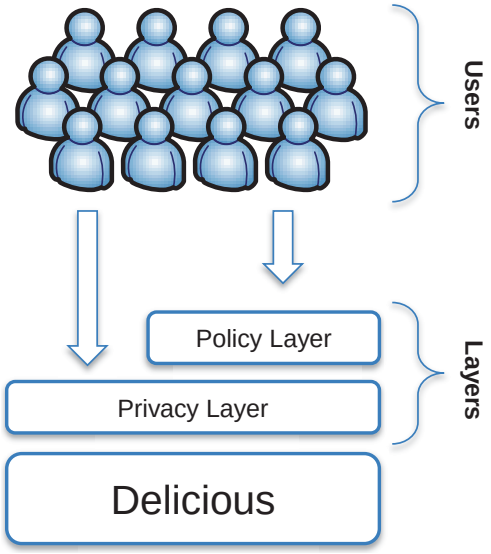


Figure 6.1: Architecture of the proposed enhanced social tagging service.

accessing a resource associated with a given set of tags (mark it as trustworthy or not, as un/safe, etc.).

Motivated by all this, in this chapter we describe an enhanced collaborative tagging system which consists of a "traditional" bookmarking service, such as Delicious, and two main additional services built on top of it (see Fig. 6.1). Such services address two main issues. The former allows end users to specify policies which can be used either to explicitly denote resources of interests or to enforce blocking conditions on the browsed data. The latter features *tag suppression*, a PET that has the purpose of hindering privacy attackers in their efforts to profile users. Such an architecture is a specific implementation of the multi-layer framework mentioned before, with the relevant difference that in [179] the privacy layer is missing. Lastly, we would also like to emphasize that our approach is not limited to the specific bookmarking application here contemplated, i.e., Delicious. That is, it could be built on top of any social bookmarking system.

But which is the purpose of combining a policy layer with a privacy layer? As discussed in Sec. 6.1, privacy is usually considered an issue for those social services which collect end users' sensible information (e.g., personal data, opinions, photos, and videos). Social bookmarking services do not fall in this category, since they

do not require the user to specify personal data (with the exception of the users' name and e-mail) and they do not collect user-generated contents. Due to this, social bookmarking services do not provide data protection mechanisms—even those available, e.g., in Facebook, which are not enough to prevent the disclosure of private data. As an example, Delicious allows registered users to flag a bookmark as public (default option) or private. When a user marks a bookmark as private, this bookmark and its associated tags are hidden to other users of Delicious. Note, however, that even if a user flags all their bookmarks and tags as private, Delicious still records this information.

Nevertheless, if tags were not sensible information per se, they could easily be exploited to infer users' personal information, such as personal interests, preferences and opinions. This is even easier when it is possible to statistically analyze huge collections of tags as those made publicly available by social bookmarking services, thus obtaining accurate tag-based user profiles. In this field, privacy-preserving techniques should guarantee privacy protection and, at the same time, the effectiveness of the services enabled by the policy layer.

The problem here is not only to find the correct trade-off between these two issues. In fact, since collaborative tagging is used to find/browse resources based on the associated tags, suppressing tags might decrease accuracy, and increase the number of false positives/negatives. Moreover, if tags are used for more sensible purposes, e.g., parental control and quality assessment, this might have even worse consequences. For these reasons, the support to privacy-preserving techniques is a key requirement when we come to enhanced policy-based uses of collaborative tagging. Actually, in such cases, users may tend to annotate resources by using tags which can be re-used for specific purposes—e.g., parental control. Such tags are then even more sensitive than the ones collected by traditional collaborative tagging services. Our aim is to verify whether and how tag suppression can be effectively applied in an enhanced collaborative tagging service such as the one illustrated in this chapter.

Next, we briefly review the assumptions upon which our tag-suppression technique builds, and then describe the reference scenarios we have used to carry out the experimental evaluation.

6.3 Tag Suppression at the Privacy Layer

In our scenario of enhanced collaborative tagging, users tag resources on the Web, e.g., music, pictures, videos or bookmarks, according to their personal preferences. Users therefore contribute to describe and classify those resources, but this is at the expense of revealing their profile of interests. In order to avoid being precisely profiled by the tagging system, or in general by any attacker able to collect the tags posted, users may adopt a privacy-protecting technology based on *data perturbation*.

The data-perturbative technology considered in this chapter is *tag suppression*, a conceptually-simple strategy that allows a user to refrain from tagging certain resources in such a manner that the profile resulting from this perturbation does not capture their interests so accurately. Our approach protects user privacy to a certain degree, but at the cost of the effectiveness of the enhanced collaborative tagging system.

In this chapter we assume exactly the same adversary model and privacy metric considered in Sec. 5.3. More specifically, we assume that users are logged into to the tagging system, that user profiles are modeled as PMFs, and that the attacker aims at individuating users in the sense regarded in Sec. 4.3.3. Further, we use the Shannon entropy of the apparent user profile as a measure of privacy, or more precisely, anonymity [(a)].

Our privacy layer therefore implements tag suppression. In practice, this means that said layer will be responsible for choosing a suppression strategy s so that t maximizes $\mathrm{H}(t)$ for a given σ. Formally speaking, its aim will be to solve the multiobjective optimization problem given by the *privacy-suppression* function (5.1),

$$\mathcal{P}(\sigma) = \max_{\substack{s \\ 0 \leqslant s_i \leqslant q_i, \\ \sum s_i = \sigma}} \mathrm{H}\left(\frac{q-s}{1-\sigma}\right),$$

which characterizes the optimal trade-off between privacy and tag-suppression rate. In Chapter 5 we found a closed-form solution to this problem, but the optimization

[(a)] Recall from Sec. 4.4 that Shannon's entropy of a profile may be regarded as an inverse measure of its uniqueness.

was carried out for suppression rate as a measure of utility, which made the problem mathematically tractable. In the remainder of this chapter, our objective is to assess the loss in semantic functionality and accuracy by using more elaborate and meaningful utility metrics. In particular, we shall evaluate the impact that tag suppression has on the enhanced collaborative tagging system described in Sec. 6.2, in terms of certain percentages regarding missing tags on bookmarks, on the one hand, and on the other in terms of false positives and negatives.

According to Sec. 5.3, our formulation is built upon the premise that the population's tag distribution p is unknown to users, which leads us to assume $p = u$. Under this assumption, entropy maximization is a special case of divergence minimization. Note, however, that if p was available to users, it would be preferable to use KL divergence as a measure of privacy risk. This is because divergence minimization may reduce the degradation in utility compared to entropy maximization, which strives to make the apparent profile close to the uniform distribution u, ignoring the fact that certain categories may be more popular than others.

In the end, we recall an important result from our theoretical analysis of the privacy-suppression function. Concretely, Sec. 5.6 showed that there exists a tag-suppression rate beyond which this function achieves its maximum value or *critical privacy* $\mathcal{P}_{\text{crit}}$. We referred to this rate as the *critical suppression* σ_{crit} and proved that

$$\sigma_{\text{crit}} = 1 - n \min_i q_i,$$

which implies that the critical suppression is never attained for $\sigma < 1$ provided that q has at least one zero component. The importance of this result lies in the fact that a user not tagging across all categories will not achieve an apparent user profile close to u for any suppression rate. Put differently, no suppression strategy fulfilling the constraints in (5.1) can lead to the uniform distribution whenever the genuine profile vanishes at some components. This fundamental property about our tag-suppression mechanism will be later used to justify some of the results shown in Sec. 6.5.

6.4 Reference Scenarios

As most PETs, tag suppression must address two main issues: protecting user privacy and granting that the perturbed data set can be effectively used. Specifically, we must verify whether the semantic loss incurred by tag suppression in order to protect private data can be acceptable. Clearly, the acceptable semantic loss threshold may highly depend on the purpose for which social bookmarking is used. Depending on it, we may require different levels of semantic accuracy, and we may have a higher or lower error tolerance.

As an example, we can figure out two different scenarios, which are both examples of enhanced uses of social bookmarking, and share the notion of "user-defined policy", i.e., a tag-based intentional definition of resource classes, explicitly expressed by users. Such classes, depending on the purpose for which policies are specified, may denote an assessment of the quality, safety or relevance of tagged resources. In the former scenario users specify policies in order to inform the bookmarking service about the resources they consider relevant. Based on them, the social bookmarking service regularly updates users, e.g., by using Web feeds, about the resources denoted by the policies. It can be considered as a subscription service which makes use of a recommendation system relying primarily on the explicit preferences expressed by users. Note that this is in contrast to the traditional recommenders, where preferences are inferred implicitly from users' past behavior. For example, content-based recommenders suggest those resources whose profiles are similar to the profile of a given user [182].

The latter scenario concerns parental control. Here policies denote which resources are un/safe. Whenever a user requests access to a resource, such policies are then used to determine whether access to that resource can be granted or should be denied. Note that the parental-control scenario has very low tolerance of false negatives; we refer to *false negatives* as those resources classified as safe, but that are actually unsafe. More precisely, in this scenario, granting access to an unsafe resource is not acceptable at all. By contrast, in the former scenario we can tolerate a higher threshold of false

negatives, since recommending a not relevant resource would not compromise the safety of users.

We introduce here the general definition of policy which can be applied to both scenarios.

Definition 6.1 (Policy)**.** A policy *pol* is a pair $(CC, sign)$, where: 1) CC is a conjunction of *category constraints* $(cc_1 \wedge \cdots \wedge cc_n)$, and 2) $sign \in \{+, -\}$. Each category constraint is a triple (c, op, θ), where c is a tag category, $\theta \in [0, 1]$, and op is a comparison operator.

A category constraint intentionally denotes the set of resources associated with a percentage of tags in the category c which is greater than (less than, equal to, etc., depending on op) the value denoted by θ. For example, category constraint $(c, >, 0.5)$ denotes those resources associated with a percentage of tags in category c which is greater than 50%. On the other hand, the semantics of the *sign* component depends on the scenario. More accurately, in the resource recommendation scenario it denotes whether the resources matching the category constraint CC are relevant $(+)$ or not $(-)$, whereas in the parental-control scenario it denotes whether they are safe $(+)$ or unsafe $(-)$.

Since the support for both positive and negative policies may raise conflicts (i.e., we may have a resource covered by both positive and negative policies), a conflict resolution mechanism must be enforced. The scientific literature provides several examples of approaches which can be adopted. A comprehensive survey on this topic is [183]. Here, for simplicity we adopt the one according to which negative policies are prevailing, since this approach is the one giving stronger guarantees with regard to the risk of accessing not appropriate contents. However, other conflict resolution policies can be easily adopted as well.

Next, we provide examples of policies for the reference scenarios introduced above. In the examples, we shall refer to some of the tag categories we have obtained from our experimental data set (see Sec. 6.5 for more details). For brevity, in this section we shall denote the relevant tag categories by $c_1, \ldots, c_n$. Also, for simplicity and clarity, in the examples we shall keep using the policy formal notation introduced in Definition 6.1. We would like to note, however, that such notation, describing how

policies are actually implemented in the system, is supposed to be made transparent in the front end both to improve usability and to help users specify policies reflecting as much as possible their intentions. Several strategies can be devised for this purpose, e.g., the use of textual labels instead of numeric values and comparison operators. Nevertheless, a discussion on such issues is out of the scope of this work.

Example 6.2 (Policies for resource recommendation)**.** Suppose that Carol (C) is interested in literature, but not in resources concerning science fiction. C realizes that the relevant tag categories are c_1 ("books") and c_2 ("literary criticism"), and she decides that the resources she is interested in are those associated with not less than 40% of the tags in either c_1 or c_2. In contrast, C finds out that the tag category which corresponds to the resources she is not interested in is c_3 ("science fiction, fantasy"), and she decides to discard all the resources associated with not less than 20% of the tags in c_3. Consequently, C specifies the following policies:

- $pol_1 = (\{(c_1, \geq, 0.4)\}, +)$,
- $pol_2 = (\{(c_2, \geq, 0.4)\}, +)$,
- $pol_3 = (\{(c_3, \geq, 0.2)\}, -)$.

Suppose now that there exists a resource R_1, which satisfies content constraints $(c_1, \geq, 0.4)$, $(c_2, \geq, 0.4)$, and $(c_3, \geq, 0.2)$. In such a case, we have a conflict, since all policies pol_1, pol_2, and pol_3 apply. According to our conflict resolution mechanism, policy pol_3 prevails over policies pol_1 and pol_2, since the latter are positive policies. Consequently, resource R_1 is marked as irrelevant to C.

Example 6.3 (Policies for parental control)**.** Suppose that Alice (A) would like to enable a Web filter for her son Bob (B) by granting him access only to contents specifically tailored for children. By checking the available tag categories, she realizes that the suitable one is c_4, "entertainment for children." She then decides that resources suitable to children are those associated with not less than 60% of the tags from category c_4. Moreover, just to be sure that no harmful content is accessed, she also would like to prevent access to "entertainment" resources which may include any content for adults. In order to achieve this, A specifies the following policies:

- $pol_4 = (\{(c_4, \geq, 0.6)\}, +)$;
- $pol_5 = (\{(c_5, \geq, 0.1)\}, -)$, where c_5 is the tag category corresponding to "entertainment for adults."

Suppose now that B requests access to a resource R_2, which satisfies both content constraints $(c_1, \geq, 0.6)$ and $(c_2, \geq, 0.1)$. In such a case, we have a conflict, since both policies pol_4 and pol_5 apply. According to our conflict resolution mechanism, policy pol_5 prevails over pol_4, since pol_5 is a negative policy. Consequently, Bob is denied access to resource R_2.

In the following section, we report the results of a series of experiments which have been carried on for the parental-control scenario. The reason of this choice is that such scenario is the most demanding as far as error tolerance is concerned. Therefore, if good results are obtained for this more demanding scenario they can be extended to the other one as well.

6.5 Experimental Analysis

In this section, we delve into the impact that tag suppression may have on the collaborative tagging system proposed in Sec. 6.2, which exploits the bookmarking application Delicious to provide enhanced services. With this aim, Sec. 6.5.1 first examines the data set that we used to conduct the experimental evaluation. To make user profiles tractable, Sec. 6.5.2 describes the methodology that we followed for mapping tags into a small set of meaningful categories of interest. Finally, Sec. 6.5.3 shows a comprehensive analysis of the degradation in data utility and accuracy, incurred by the application of our privacy-protecting technique.

6.5.1 Data Set

In our experiments, we used the Delicious data set retrieved by the Distributed Artificial Intelligence Laboratory (DAI-Labor), at Technische Universität Berlin [184]. This data set contains those bookmarks and tags marked as public by approximately 950 000 users. It consists of triples (*username*, *bookmark*, *tag*), each one representing

the action of a user associating a bookmark with a tag. The data set includes 420 millions of these triples, posted from Sep. 2003 to Dec. 2007.

The data set that we considered in our analysis is a subset of the entire data set described above. Concretely, we selected out a subset covering approximately one year and including 1 241 029 triples. We decided to choose this subset because, on the one hand, it spanned a significant period of time, and, on the other, it did not involve the processing of millions of triples that would overload our experiments. Our data subset therefore contains 9 588 users, 390 008 resources and 59 505 tags.

6.5.2 Tag Categorization and Methodology

As we commented in Sec. 5.7.2, modeling a user profile as a normalized histogram across these 59 505 tags would be certainly unfeasible from various practical perspectives, mainly concerning the unavailability of data to reliably, accurately measure interests across such fine-grained categorization, and, should the data be available, its overwhelming computational intractability. Further, in our experiments but also in data mining procedures, a coarser categorization makes it easier to have a quick overview of the user interests.

Motivated by this, we categorize the tags in our data set into a coarser representation with just a few high-level tag categories. We have followed the same methodology used in Sec. 5.7.2, where we clustered the tags of a data set from BibSonomy into 5 categories. Specifically, we have used Lloyd's algorithm [176] to group tags into 20 categories; and then, for each of those categories, we have clustered its tags into 10 subcategories. The next subsection provides a complete description of such methodology, which we sketched out in Sec. 5.7.2. Right after this, we describe the aforementioned Lloyd's algorithm.

Methodology

In Sec. 5.7.2 we summarized the tag categorization process in three steps, namely the computation of a co-occurrence matrix, the definition of a similarity metric between

tags, and the application of Lloyd's algorithm. Next, we proceed to describe these three steps in detail.

Exactly as in Chapter 5, we first filtered out those tags considered as spam. For this purpose, we collected some statistics about the number of characters contained by tags. After observing that 98% of tags had less than 23 characters, we dropped those tags with a number of characters over 22. In addition, we eliminated those posts with more than 50 tags, as they are usually spam [174]. Additionally, posts with no tags were not considered. After this simple preprocessing, the number of triples reduced to 1 149 895, and, consequently, the number of users, bookmarks and tags to 9 207, 349 658 and 54 024, respectively.

In a second stage, we aimed at identifying clusters or groups of semantically similar tags. As frequently done in the literature, we performed a clustering analysis based on the *co-occurrence* between tags, that is, the number of times each pair of tags simultaneously appears in a same bookmark. Specifically, we modeled the relationships among tags as a matrix of co-occurrences c_{ij}, where each entry with $i \neq j$ corresponds to the co-occurrence between tags i and j, and each entry in the diagonal is a self-occurrence, i.e., the absolute frequency of appearance of a tag. Note that, clearly, this is a symmetric matrix and that each row (column) describes one tag in terms of the semantic similarity to the other tags. Repeated tagging is taken into consideration. For example, if a given resource is tagged with tag i 10 times, and with tag j 5 times, we increase the self-occurrence counter c_{ii} of the first by 10, the self-occurrence counter c_{jj} of the second by 5, and the co-occurrence counter c_{ij} of these two tags by 5, ignoring the transposed position c_{ji} in a practical implementation of the procedure ($i < j$).

In an attempt to concentrate on the significant relationships among these tags, we eliminated those rows satisfying $\sum_j c_{ij} < \tau$, for a certain threshold τ. Similarly, we dropped those columns fulfilling an equivalent condition. In this regard, observe that, the higher the threshold, the lower the number of resulting tags, and thus the lower the number of triples containing those tags. Since we aimed at preserving at least 80% of the triples, and at the same time, we required the resulting tags to have a strong co-occurrence, we chose $\tau = 95$. In doing so, we obtained a reduced co-occurrence matrix

with dimension 5 999 tags. In conclusion, after this filtering process the number of triples, users and bookmarks became 985 273, 8 882, and 310 923, respectively.

Once we filtered the co-occurrence matrix, we proceeded to use a well-known clustering algorithm to create a two-level hierarchy of categories. But before applying this algorithm, we first required to specify a measure of similarity among tags. Recall that we modeled tags as rows and columns of a matrix, that is, vectors. As often done in the literature, we employed the cosine metric [185], a simple and robust measure of similarity between vectors. More precisely, two tags are represented numerically by column (or row) vectors x and y of the co-occurrence matrix, with 5 999 entries. Let $\bar{x} = x/\|x\|$ denote the Euclidean normalization of x, and similarly for y. The cosine distance is defined as

$$d(x, y) = 1 - \langle \bar{x}, \bar{y} \rangle = 1 - \frac{\langle x, y \rangle}{\|x\| \|y\|} = \|\bar{x} - \bar{y}\|^2;$$

strictly speaking, the square of their Euclidean distance, after normalization. Note that $d(x, y) = 0$ if, and only if, $\bar{x} = \bar{y}$, meaning that the normalized co-occurrence profile of tags x and y is identical, to be expected, approximately, for complete synonyms.

Equipped with this measure of dissimilarity, we applied Lloyd's algorithm [176], a popular iterated algorithm for grouping data points into a set of k clusters. As a result, we grouped the 5 999 tags into 20 categories. Afterwards, for each of those categories, we turned to apply the same algorithm to get 10 subcategories. The process yielded a total of 200 subcategories, which provided us with a granularity level thin enough as to define precise filtering policies, and sufficiently aggregated as to avoid noisy behaviors. The resulting categories were classified in decreasing order of popularity of their tags. Then, tags in each subcategory were sorted in decreasing order of proximity to the centroid. As an illustrating example, Fig. 6.2 represents two of the subcategories corresponding to the top-level tag category "entertainment". The complete results of our clustering, that is, the list of all tags belonging to each of the 200 subcategories, is directly downloadable at `http://hdl.handle.net/2117/16623`.

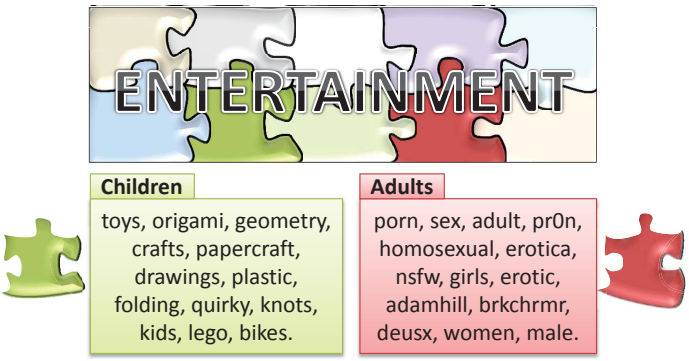


Figure 6.2: According to our hierarchical clustering, each category is composed of 10 subcategories. In this example, we represent two subcategories belonging to the "entertainment" category. In particular, we show the tags falling into the subcategory 62 "entertainment for children" and 68 "entertainment for adults", which are used in the specification of policies for the parental-control scenario described in Sec. 6.4. The two examples of subcategories shown here also illustrate a key result of the categorization process—tags in each subcategory are sorted in decreasing order of proximity to the centroid, which in practice means that those tags at the top of the list are the most representative tags of the subcategory they belong to.

As a result of our categorization process, the first tag in each subcategory, i.e., the closest tag to the centroid, is considered to be the most representative tag of its subcategory. This tag could be used as the reconstruction value of its corresponding subcategory. For example, when a user assigned the tag "nsfw" to a resource (see Fig. 6.2), we could replace automatically this tag with the tag "porn". An alternative would consist in replacing this tag with a descriptor that could be manually assigned to the subcategory the tag belongs to. For instance, instead of considering the tag "porn" as a reconstruction value, we could use the descriptor "entertainment for adults". In our experiments, we opted for the former approach—first, because of its straightforward implementation, and secondly, because it only affects how subcategories are named.

Lloyd's Algorithm

This subsection provides a brief description of Lloyd's algorithm [176], the clustering algorithm that we used in the previous subsection to categorize tags into high-level tag categories.

Assume that we are given a data set $\{x_1, \ldots, x_n\}$ composed of n points in $\mathbb{R}^d$, and that we wish to partition this data set into k disjoint, non-empty subsets or *clusters*.

Specifically, suppose that our aim is to find out how these points should be assigned to said clusters so as to minimize a measure of distortion $\mathcal{D}$. Intuitively, we may interpret a cluster as a set of data points where the distances among these points are relatively small, compared with the distances to those which do not belong to the cluster. Let c_j for $j = 1, \ldots, k$ be the centroids of such clusters. Define the indicator variables γ_{ij} for $j = 1, \ldots, k$, denoting whether a point x_i is assigned to the cluster j, that is,

$$\gamma_{ij} = \begin{cases} 1, & x_i \in \text{cluster } j \\ 0, & x_i \notin \text{cluster } j \end{cases} .$$

According to this notation, assume the following measure of distortion,

$$\mathcal{D} = \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} \|x_i - c_j\|^2,$$

i.e., the sum of squared Euclidean distances from each point to its assigned centroid. Under this assumption, Lloyd's algorithm is a heuristic for solving the aforementioned clustering problem, specifically for finding those values of $\{\gamma_{ij}\}$ and $\{c_j\}$ minimizing $\mathcal{D}$.

The algorithm in question starts with an initial set of k centroids. These centroids may be chosen simply at random from the data set. Other initialization methods are described in [186, 187]. After this initialization, Lloyd's algorithm follows an iterative procedure. At each iteration, it carries out these two steps:

- *Assignment of clusters.* In this first step, the algorithm holds the centroids fixed and finds those assignments $\{\gamma_{ij}\}$ that minimize $\mathcal{D}$. It can be shown that the optimal $\{\gamma_{ij}\}^*$ consists in assigning each point to the cluster with the nearest centroid in Euclidean distance.

- *Update of centroids.* In this second step, the algorithm holds the assignments fixed and minimizes distortion with respect to $\{c_j\}$. Similarly, it can be proved that the optimal centroids are

$$c_j = \frac{\sum_{i=1}^{n} \gamma_{ij} x_i}{\sum_{i=1}^{n} \gamma_{ij}},$$

that is, the average of the points assigned to their clusters.

The algorithm proceeds by alternating between these two steps until the assignments do not change. When this happens, the algorithm is said to converge.

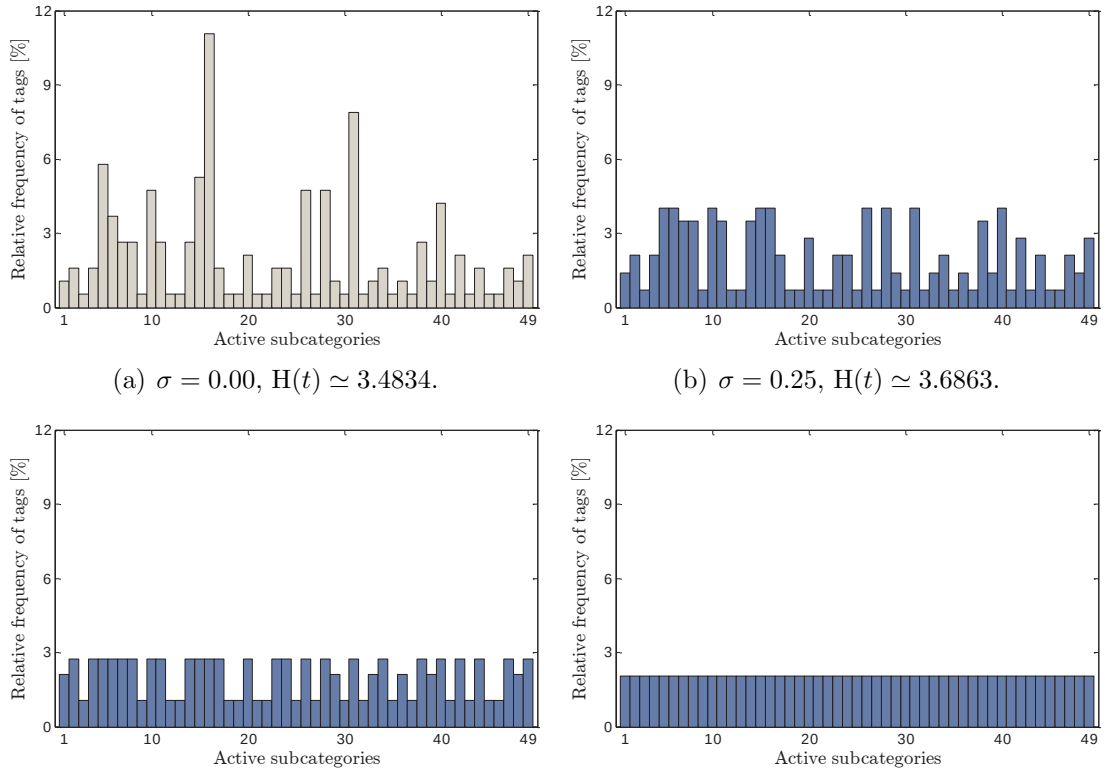
6.5.3 Results

This section presents a number of experimental results that will allow us to evaluate the proposed enhanced collaborative tagging service in terms of privacy protection, utility loss and filtering accuracy. Specifically, Sec. 6.5.3 analyzes the privacy gain as a result of the application of our tag-suppression technique, Sec. 6.5.3 evaluates the utility loss, whereas Sec. 6.5.3 provides insight into the loss in filtering accuracy for the parental-control scenario.

Privacy

In our architecture, a user specifies a suppression rate indicating the fraction of tags they are disposed to eliminate. Based on this suppression rate and the user profile across the $n = 200$ subcategories, our approach computes the optimal tag suppression strategy s directly from Theorem 5.4. Recall that s is an n-tuple containing the percentage of tags that a user should eliminate in each subcategory.

In Sec. 6.3 we mentioned that the critical suppression beyond which critical privacy is attained is given by $\sigma_{\text{crit}} = 1 - n \min_i q_i$. A consequence of this fact is that, in the case when a user does not tag across all subcategories, the critical privacy $\mathcal{P}_{\text{crit}} = \ln n$ is not achieved for any $\sigma < 1$. This is precisely what happens in our data set, that is, no user has tagged across all subcategories, which in practice means that these users will not get an apparent user profile close to u. However, without loss of generality we may consider the subset of subcategories that have been tagged by a particular user. Note that this is consistent with the theoretical analysis presented in Chapter 5, where we assumed that the components of q are strictly positive. We denote these categories as the *active* subcategories of that user, and the cardinality of this subset as n_{act}. In terms of these subcategories, we may assume the existence of an equivalent critical suppression σ'_{crit} and an equivalent critical privacy $\mathcal{P}'_{\text{crit}}$, in the sense that,



(a) $\sigma = 0.00$, $\mathrm{H}(t) \simeq 3.4834$.

(b) $\sigma = 0.25$, $\mathrm{H}(t) \simeq 3.6863$.

(c) $\sigma = 0.50$, $\mathrm{H}(t) \simeq 3.8123$.

(d) $\sigma = 0.75$, $\mathrm{H}(t) \simeq 3.8918$.

Figure 6.3: We represent the apparent profile of a particular user, that is, the perturbed profile resulting from the suppression of tags and observed from the outside. We only show the active subcategories of this profile, i.e., those subcategories tagged by the user. In this particular case, the user posted 190 tags belonging to 49 subcategories. As expected, we observe that as σ increases, t approaches u and $\mathrm{H}(t)$ tends to $\ln 49 \simeq 3.8918$. When there is no suppression, the apparent profile is plotted in gray to emphasize that this profile is actually the genuine profile. This is consistent with Fig. 6.5.

beyond this suppression rate, t becomes the uniform distribution across the active subcategories and $\mathcal{P}'_{\mathrm{crit}} = \mathrm{H}(t) = \ln n_{\mathrm{act}}$. This interesting property is illustrated in Fig. 6.3, where we plot the apparent profile of a specific user [(b)].

The figure in question shows the user's apparent profile just for the active subcategories. For convenience, we rearranged these subcategories and indexed them from 1 to 49. Clearly, when no suppression is applied, the apparent profile is in fact the actual user profile q. On the other hand, when $\sigma = 0.25$ we observe that the subcategories affected by suppression are those with a percentage of tags furthest away

[(b)] This particular user is identified by the string `674f779ba3b445937fd9876054a6e` in [184].

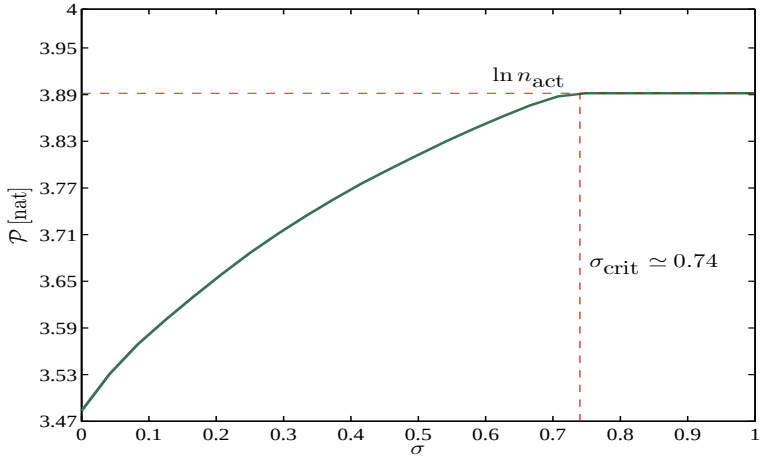


Figure 6.4: Our PET poses a trade-off between privacy and tag-suppression rate. This is illustrated here, where we plot function (5.1) for the particular user considered in Sec. 6.5.3. Further, we observe that when $\sigma \geqslant \sigma'_{\text{crit}} \approx 0.74$, the function achieves its maximum value $\mathcal{P}'_{\text{crit}}$, which is given by the number of active subcategories n_{act}.

from u. In the special case when the user consents to eliminate a fraction of tags $\sigma \geqslant \sigma'_{\text{crit}} \approx 0.74$, t becomes the uniform distribution across the active subcategories and hence $\mathrm{H}(t)$ attains its maximum value, $\ln 49$. This effect is also highlighted in Fig. 6.4, where we represent the trade-off between privacy and tag-suppression rate for this particular user. In short, the results shown in these two figures confirm the existence of an (equivalent) critical-suppression rate beyond which the privacy-suppression function achieves its maximum value. Also, we observe that the trade-off is concave.

In addition, we plot in Fig. 6.5 an example of suppression strategy in the case when $\sigma = \sigma'_{\text{crit}}$. In this figure, we superimpose the optimal suppression strategy on the genuine user profile q, in order to reflect the proportion of tags that the user should eliminate from each subcategory of q to become the uniform distribution.

Lastly, Fig. 6.6 shows the privacy protection that users of the proposed collaborative tagging application achieve as a result of the suppression of tags. More accurately, we consider the case when all users in our data set have adhered to tag suppression and use the same suppression rate. Under these assumptions, we plot the percentile curves (10^{th}, 50^{th} and 90^{th}) of relative privacy gain. We observe an important difference between these results and those obtained for BibSonomy in Sec. 5.7.3. As shown

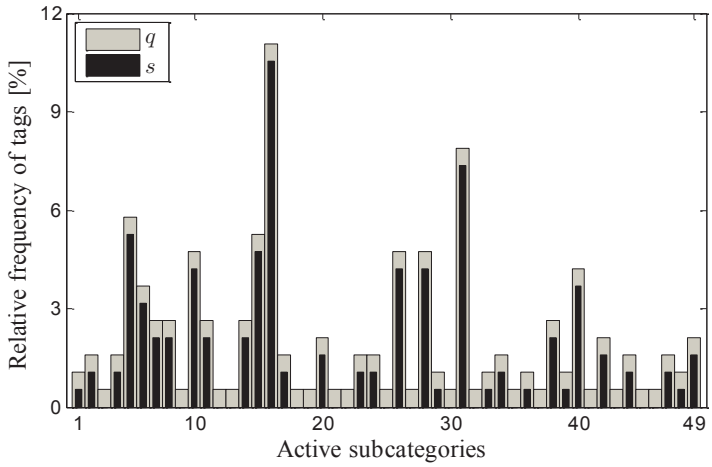


Figure 6.5: In this figure, we represent the genuine user profile q of the particular user considered in Sec. 6.5.3. In addition, we plot the suppression strategy s solving the optimization problem (5.1) in the special case when $\sigma = \sigma'_{\text{crit}} \approx 0.74$.

in Fig. 5.12, the relative privacy gain is much greater in this latter application. For example, in the limit when σ approaches 1, the 90^{th} percentile of relative privacy gain is five times greater than that for Delicious. The reason for this is due to the fact that we removed those users of BibSonomy who had not tagged across all categories and who did not have a significant tagging activity. This filtering has not been done in the case of Delicious, as we have preferred to evaluate our approach in a scenario with an important number of users [(c)].

To sum up, the experimental results presented in this section show how tag suppression contributes to privacy protection in our enhance collaborative tagging service.

Data Utility

As we have just seen, our approach helps users protect their privacy. Nevertheless, as in any perturbative mechanism, this protection comes at the expense of a loss in data utility. In this section, we assess quantitatively the degradation in data utility caused by our privacy-protecting mechanism.

In Chapter 5, we used a simplified measure of loss in data utility, the tag-suppression rate, which allowed us to formulate the optimal privacy-utility trade-off

[(c)] In Sec. 5.7.3 we used 209 users of BibSonomy to evaluate the privacy protection provided by tag suppression. The experiments conducted in Sec. 6.5.3 for Delicious involve 8 882 users.

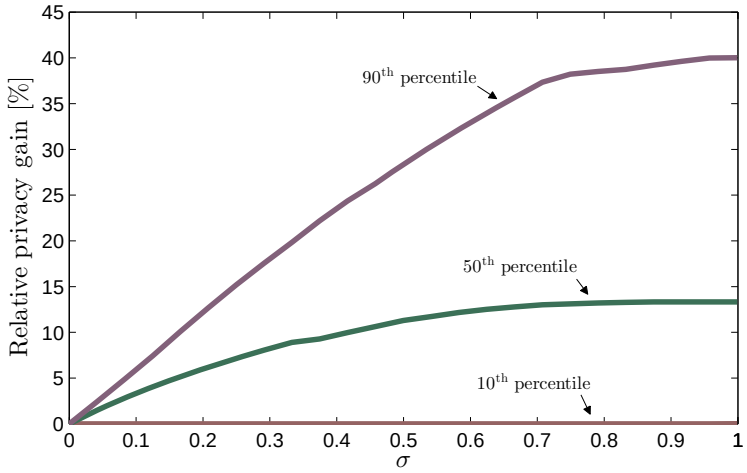


Figure 6.6: Percentiles curves of relative privacy gain in the special case when all users in our data set apply tag suppression and use the same suppression rate.

in a mathematically tractable manner. In this section, we evaluate the impact that suppression has on utility by considering a more sophisticated albeit computationally-feasible metric—the percentage of tags that each bookmark loses as a result of the elimination of tags. To highlight that tags make bookmarks meaningful, throughout this section we shall refer to this loss in data utility as *semantic loss*. Occasionally, we shall also refer to it as utility loss.

In our experiments, the set of tags that users assign to a particular bookmark is referred to as the *bookmark profile* and is modeled exactly as we do with user profiles, that is, as a normalized histogram of these tags across the $n = 200$ subcategories mentioned in Sec. 6.5.2. In addition to this characterization, we contemplate a fraction Σ of the user population suppressing tags with a common suppression rate, and assume that the remaining users do not eliminate their tags.

In order to calculate the utility loss experienced by every bookmark in our data set, we first computed the optimal suppression strategy for every user suppressing tags. Afterwards, the resulting suppression strategies were applied to the specific bookmarks tagged by these users. Next, we briefly describe how our tag suppression algorithm subtracts tags from these bookmarks.

Given a user and a tag-suppression rate, we use Theorem 5.4 to calculate s. Let α be the total number of tags posted by this user. Accordingly, αs_i is the absolute

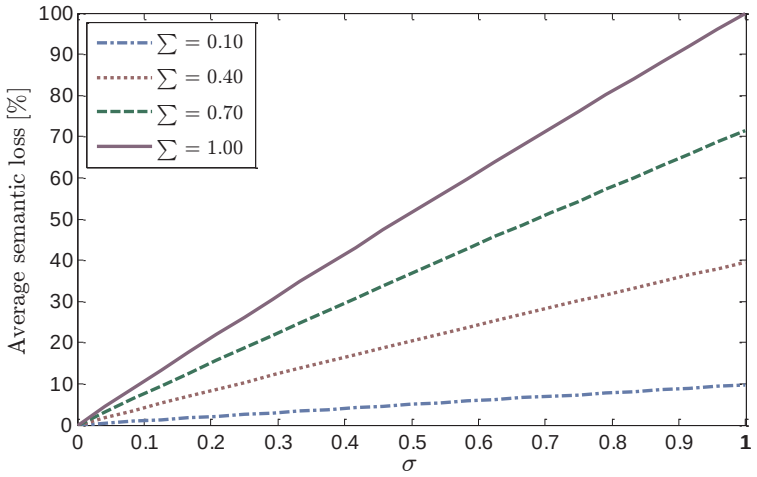
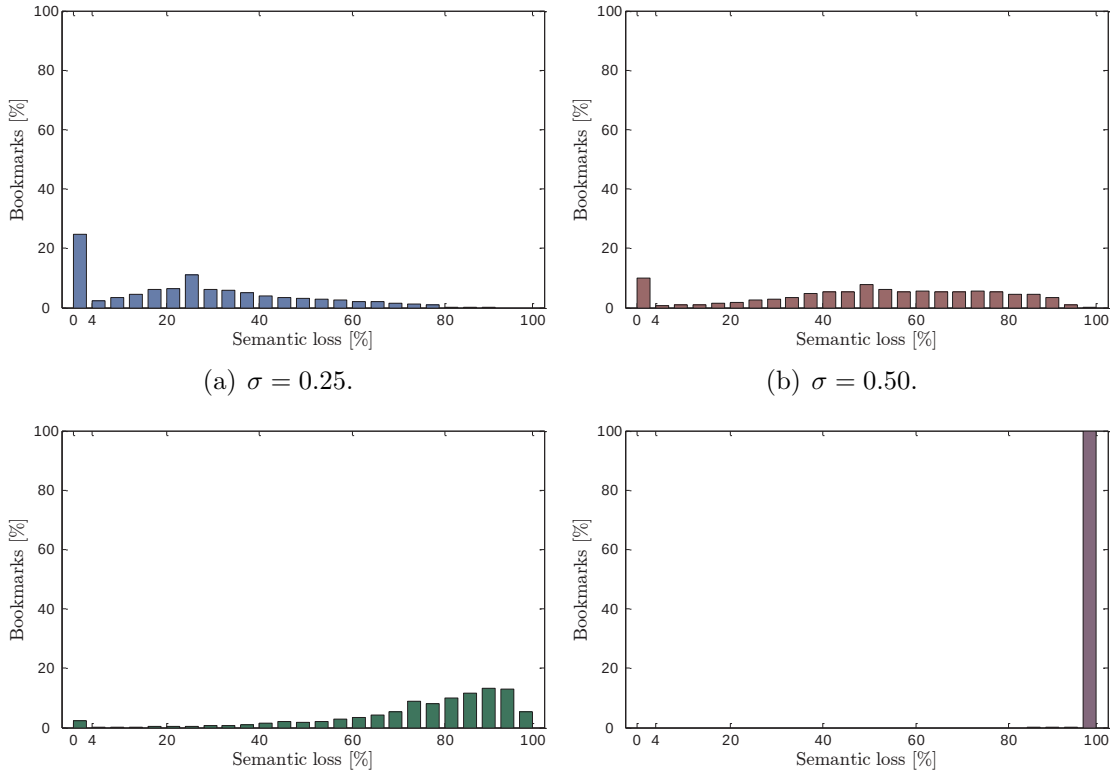


Figure 6.7: We plot the average loss in semantic functionality as a function of the tag-suppression rate. As expected, we observe that, regardless of the fraction of users eliminating tags, the semantic loss exhibits a linear behaviour with the suppression rate.

number of tags that the user should eliminate in the category i. Note that this number may not be an integer. Denote by $\mathcal{S}_i = \{b_1, \ldots, b_m\}$ the set of bookmarks to which the user assigned tags corresponding to the category i, and denote by $\beta_1, \ldots, \beta_m$ the number of tags that the user associated with each of these bookmarks. Then, for each $b_j \in \mathcal{S}_i$, our algorithm eliminates $\alpha s_i \frac{\beta_j}{\sum_{k=1}^{m} \beta_k}$ tags from the ith component of the histogram of absolute frequencies of this bookmark. This process is repeated for each category and for each user. To illustrate how it operates, consider a user who must eliminate $\alpha s_i = 1.5$ tags from one particular category. Suppose that the user's bookmarks belonging to this category are b_1, b_2 and b_3. Also, assume that the user assigned 1 tag to b_1, 1 tag to b_3, and 2 tags to b_2. According to all this, our algorithm would eliminate $\frac{1.5}{4} = 0.375$ tags from each b_1 and b_3, and $\frac{1.5}{2} = 0.750$ tags from b_2.

Having described how we computed the utility loss, next we show the results obtained in our experiments. Fig. 6.7 represents the semantic loss averaged for all bookmarks. Unsurprisingly, the results indicate that the average semantic loss is roughly linear with the common suppression rate. Specifically, we appreciate that such measure of utility is given approximately by the multiplication of σ and Σ. Fig. 6.8 provides more extensive results with regard to semantic loss, but in the form of histograms of relative frequencies. In particular, this figure depicts the percentage of bookmarks affected by a given semantic loss, for $\sigma = 0.25, 0.50, 0.75, 0.99$ and the



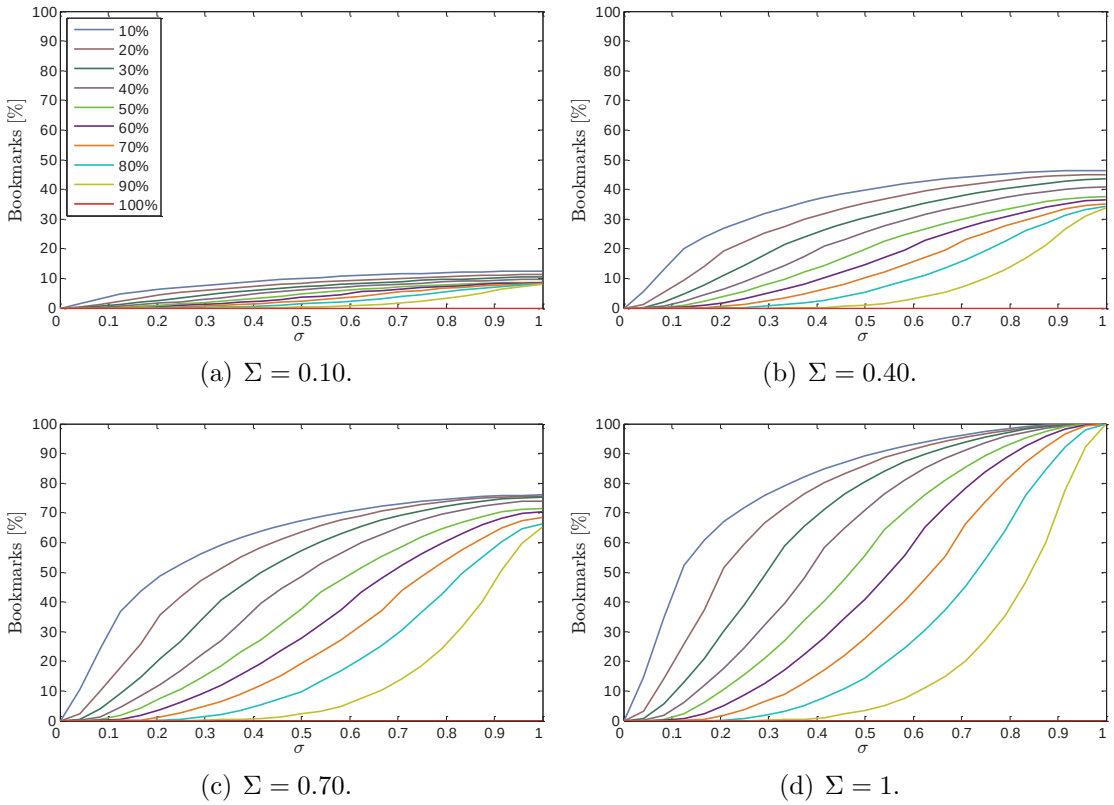
(a) $\sigma = 0.25$. (b) $\sigma = 0.50$.

(c) $\sigma = 0.75$. (d) $\sigma = 0.99$.

Figure 6.8: Loss in semantic functionality in the case when all users apply tag suppression.

worst-case scenario where all users are adhered to tag suppression, i.e., $\Sigma = 1$. For $\sigma = 0.25$, we observe that around 24% of resources experienced a reduction in their number of tags less than or equal to 4%. For a suppression rate of 0.75, we note that most of resources lost 68-100% of their tags. Not entirely unexpectedly, when users eliminated almost all their tags, we observe that nearly all bookmarks were affected by a semantic loss between 96% and 100%.

Additionally, Fig. 6.9 plots the curves of semantic loss for different values of Σ. More accurately, we depict a curve for the fraction of bookmarks with at least a 10% loss in the number of tags with respect to the case without suppression, and similarly for $20\%, 30\%, \ldots, 100\%$, where 100% refers to completely untagged bookmarks. For any Σ, we note that there is bijective relation between the semantic loss and the tag-suppression rate. Also, we see that, as σ increases, the evolution of the curves is



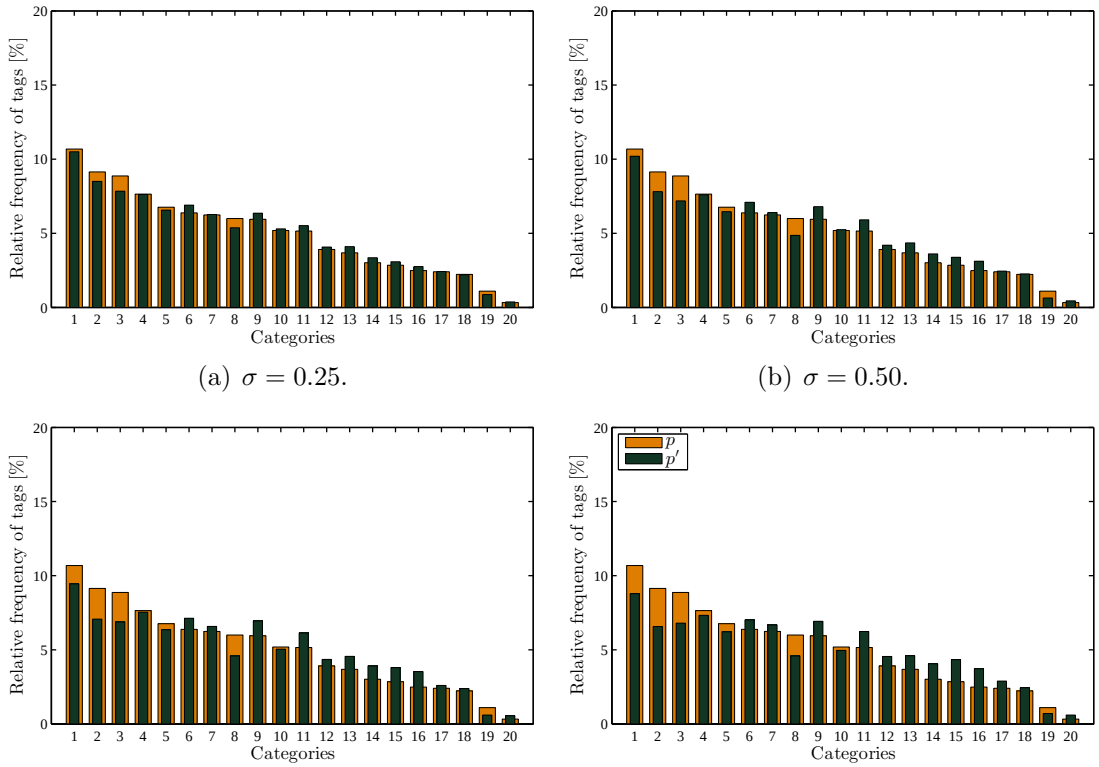
(a) $\Sigma = 0.10$. (b) $\Sigma = 0.40$.

(c) $\Sigma = 0.70$. (d) $\Sigma = 1$.

Figure 6.9: Curves of semantic loss showing the percentages of bookmarks that experienced, at least, a $10, 20\%, \ldots, 100\%$ loss in the number of tags, for distinct fractions of the population suppressing tags Σ. The 100% curve of semantic loss refers to those bookmarks that lost all their tags.

rather similar. For example, in the limit when σ approaches 1, we observe that the range of values taken by the percentage curves falls around Σ.

All these results have shown the degradation in data utility in terms of percentages of tags that bookmarks lose. Our next experimental results, on the other hand, show which categories of interest are primarily affected by suppression. In particular, Fig. 6.10 illustrates how tag suppression impacts on each of the 20 high-level categories found in Sec. 6.5.2. This figure represents the content profile of Delicious, which we computed as the aggregated profile of all bookmarks. We note that this profile corresponds to the population's tag distribution, resulting from the aggregation of the profiles of all users. This is the reason why we refer to this profile as p. The modified version of this histogram due to suppression is denoted by p'. Two remarks are in order. First, the categories most affected by suppression are those with the highest



(a) $\sigma = 0.25$. (b) $\sigma = 0.50$.

(c) $\sigma = 0.75$. (d) $\sigma = 0.99$.

Figure 6.10: The overall profile of Delicious is shown as the aggregated profile of all bookmarks across the categories obtained in Sec. 6.5.2. We denote this profile as p. In this figure, we consider the case when $\Sigma = 1$. As a consequence of suppression, the profile p results in the modified profile p'.

percentage of tags. This is the case of the first three categories, which, according to the categorization conducted in Sec. 6.5.2, seem to refer to "software", "Web" and "programming". Secondly, owing to the fact that we are dealing with relative rather than absolute frequencies, we observe an increase in the frequency of tags of those categories with the lowest percentage in p. This is what happens, for example, to the categories 15 and 16, which we refer to as "shopping and travel" and "film and television", respectively.

In summary, this section has examined the extent to which the application of tag suppression affects data utility, in terms of percentages of missing tags on bookmarks, depending on the fraction of users suppressing tags and a common suppression rate.

In addition, we have shown which content of the underlying bookmarking service is most affected by suppression.

Accuracy in Content Filtering

In this section, we quantitatively evaluate the degradation in the classification of Web content due to the suppression of tags. Specifically, this section measures the loss in accuracy in the parental-control scenario described in Sec. 6.4. Throughout this section, we shall resort to the example of Web filter referred to as "Example 2", which classifies resources on the Web into two states, "granted" or "denied".

Recall that our Web filter first retrieves the profile of the Web page to be accessed, which we model as a normalized histogram of tags across the set of subcategories described in Sec. 6.5.2, and secondly checks whether certain subcategories of this profile exceed a particular threshold. The subcategories of our example are "entertainment for children" and "entertainment for adults", identified, after the categorization process, as the subcategories 62 and 68, respectively [(d)]. The threshold values for these subcategories are $\theta_{62} = 60\%$ and $\theta_{68} = 10\%$. That said, suppose w is the profile of a Web page and that w_{62} and w_{68} are the components of this profile, corresponding to the aforementioned subcategories. According to Sec. 6.4, the operation of the filter is as follows: if $w_{68} < \theta_{68}$ and $w_{62} \geqslant \theta_{62}$, then that resource is classified as granted; otherwise the access to the Web page is denied.

Having reviewed how the parental-control filter works, in this series of experiments we shall assume that this filter is installed by default in the users' Web browser. In other words, we shall suppose that all users specify the same policies for parental control, which may describe a fairly realistic scenario, as most users do not change default settings [188]. Moreover, we shall assume that the filter works perfectly when tag suppression is not applied. When users skip tagging some resources, however, this filter may classify them incorrectly. In this regard, we shall refer to the *initial* state and the *final* state of a resource as the states before and after the suppression of tags, respectively.

[(d)] The list with the 200 subcategories resulting from our hierarchical clustering may be downloaded at `http://hdl.handle.net/2117/16623`.

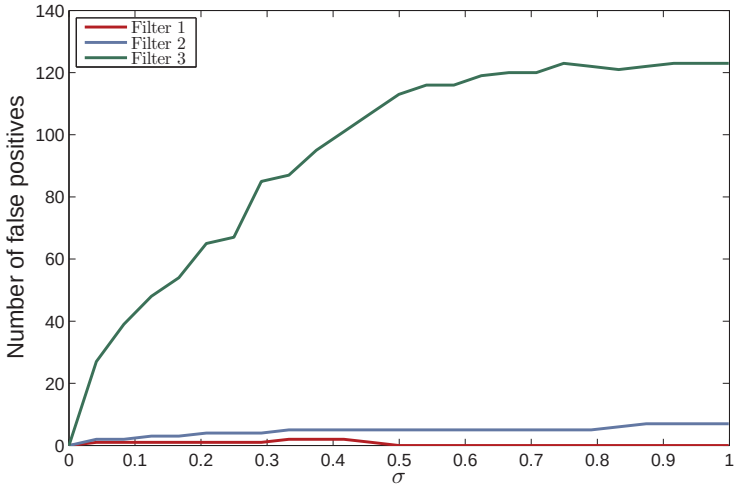


Figure 6.11: A false positive represents a resource that changes from the initial state granted to the final state denied, due to the suppression of tags. In this figure, we observe that the most permissive filter (filter 3) exhibits much more false positives than the other two filters.

In order to quantify the loss in the accuracy of this filter, we contemplate the following measures of utility: the number of false negatives and false positives, precision and recall. In our scenario, a *false negative* is defined as a resource that changes from the initial state denied to the final state granted, as a consequence of tag suppression. To illustrate this case, consider Alice enables our Web filter for her son Bob. Suppose that, at some point, Bob wishes to access a Web page with profile w, and components $w_{62} = 50\%$ and $w_{68} = 10\%$. According to the operation of the filter, the access to this resource would be blocked. Nevertheless, after the suppression of tags by other users, it could be possible that this Web page experienced a reduction in the percentage of tags such that $w_{68} < \theta_{68}$. Due to the fact that we are dealing with relative frequencies, this reduction could cause that $w_{62} \geqslant \theta_{62}$, and therefore Bob would be able to access said resource. Should this be the case, we would classify this Web page as a false negative.

Having described the case of a false negative, next we contemplate the other three possible combinations for the initial and final states. Specifically, we define a *true negative* as a resource whose access is granted before and after the suppression of tags. Similarly, a *false positive* denotes a resource passing from the initial state granted to

the final state denied. And finally, a *true positive* corresponds to a resource that is blocked before and after tag suppression.

Note that all bookmarks in our data set belong to one of these four cases—every resource is classified as denied or granted before (initial state) and after (final state) our technique is applied, which means they necessarily fall into one of the cases mentioned above. However, among these cases, false negatives are clearly the most sensitive in the scenario of parental control, as described in our example. On the other hand, false positives are less critical, even if important, since they represent resources that should be granted but are blocked due to tag suppression. Thus, false positives could be considered as an availability problem rather than a disclosure of potentially dangerous content.

We shall refer to *fn*, *tn*, *fp*, and *tp* as the number of false negatives, true negatives, false positives and true positives. According to this notation, *precision* may be defined as $\frac{tp}{tp+fp}$ and *recall* as $\frac{tp}{tp+fn}$. These two measures may be interpreted in probabilistic terms—precision may be regarded as the probability that a resource with final state denied has been classified correctly; and recall as the probability that a resource is classified correctly, given that its initial state is denied. Seen from another perspective, in the context of a medical test used to identify a disease, precision and recall are interpreted as follows. Let D be the event "the patient is ill" and T be the event "the test is positive". Precision is defined accordingly as $\mathrm{P}(D|T)$ and recall as $\mathrm{P}(T|D)$.

The experimental results are shown in Figs. 6.11, 6.12, 6.13 and 6.14, in the special case when all users eliminate tags, i.e., $\Sigma = 1$. In these figures, we test the Web filter described at the beginning of this section, specified more formally in Sec. 6.4. However, in order to enrich our analysis, we also include two slight variations of this (original) filter. Particularly, we contemplate different values for the thresholds θ_{62} and θ_{68}. Accordingly, in our experiments we refer to the original filter as *filter 2*. A more restrictive version of this filter is *filter 1*, whereas *filter 3* is more permissive. Next, we summarize the set of filters used in our evaluation:

- *filter 1*, with $\theta_{62} = 75\%$ and $\theta_{68} = 5\%$,
- *filter 2*, with $\theta_{62} = 60\%$ and $\theta_{68} = 10\%$,

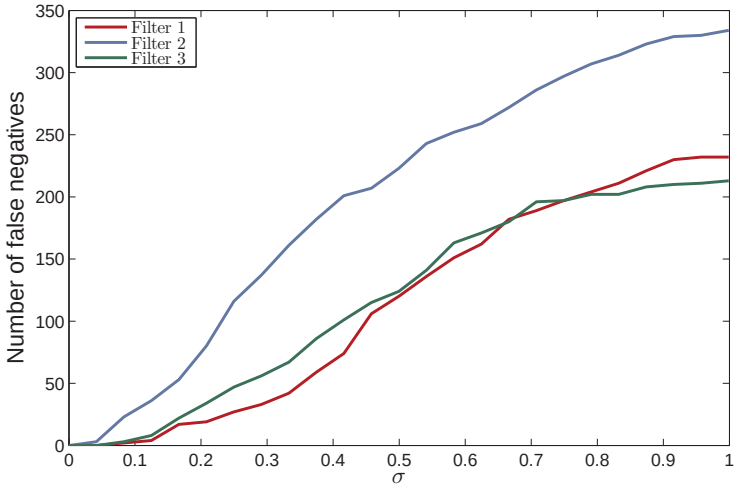


Figure 6.12: A false negative refers to a Web resource whose access is denied before tag suppression, but after the elimination of tags, the access to this resource is granted.

- *filter 3*, with $\theta_{62} = 45\%$ and $\theta_{68} = 15\%$.

Fig. 6.11 shows the number of false positives. As can be observed, the maximum number of cases is around 120 for the least restrictive filter. Since the total number of resources is 310 923, the number of false positives only represents 0.04% of all cases. The differences in terms of false positives between filter 3 on the one hand, and filters 1 and 2 on the other, are due to the nature of the resources granted by those filters. In particular, before the suppression of tags, 99% of the resources classified as granted by filter 1 have a distribution of tags such that *all* tags are concentrated on the subcategory "entertainment for children". In other words, the profile of each of those Web pages has only one positive component, namely the component 62. As a consequence of this fact, the profile of those resources will remain exactly the same no matter which suppression rate is applied. Recall that profiles are *relative* histograms of tags and that our suppression approach simply subtracts tags from positive components. Therefore, after the suppression of tags, almost none of those resources will be blocked and, consequently, they will not be considered as false positives. This is the reason why the number of false positives is so low in the case of filter 3.

The above reasoning also applies to filter 2, where, before tag suppression, 94% of the resources granted have a profile with 100% of their tags in the subcategory 62.

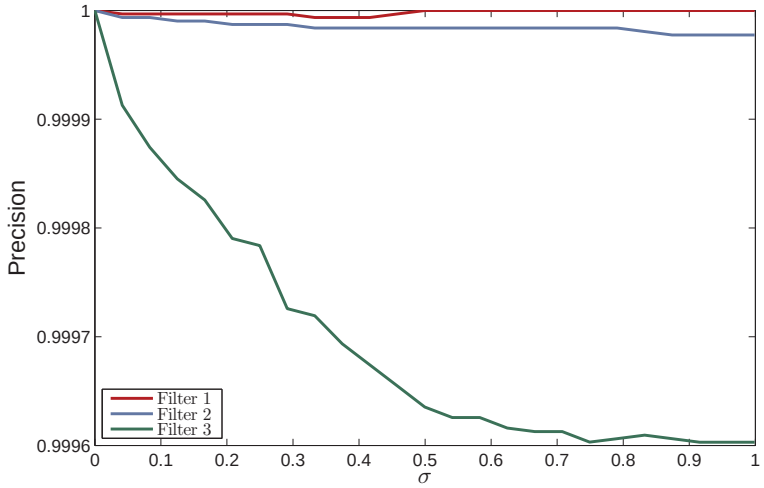


Figure 6.13: Precision may be interpreted as the probability that a resource has been classified correctly, given the fact that it was considered denied after the suppression of tags. Whilst at first glance it may seem that there is a great difference, in terms of precision, between filters 1 and 2 on the one hand, and filter 3 on the other, it should be noted that suppression has a negligible effect on the precision of any of the three filters.

But this is not the case of filter 3—this particular distribution of tags is only observed in 54% of the resources classified as granted. As a result, we notice a greater number of resources blocked after the suppression of tags, and therefore, a larger number of false positives, as shown in Fig. 6.11.

The number of false negatives is plotted in Fig. 6.12. Here we observe that the maximum number of cases is around 340, which accounts for 0.11% of all cases. In Fig. 6.13 we appreciate that precision is practically unaffected by the suppression of tags. The differences between filter 3 on the one hand, and filters 1 and 2 on the other, are essentially due to the larger number of false positives observed in filter 3, an effect that we examined above. Similarly, Fig. 6.14 shows that recall is reduced only by a 0.11% in the worst-case scenario, corresponding to filter 2.

In summary, these results indicate that tag suppression does not have a significant impact on the accuracy of a parental-control filter. Further, because the scenario of resource recommendation described in Sec. 6.4 is more tolerant to false negatives than the scenario analyzed in these experiments, we may extend the above results to the former scenario and then assert that our technique would have a similar impact on the accuracy of the recommendations.

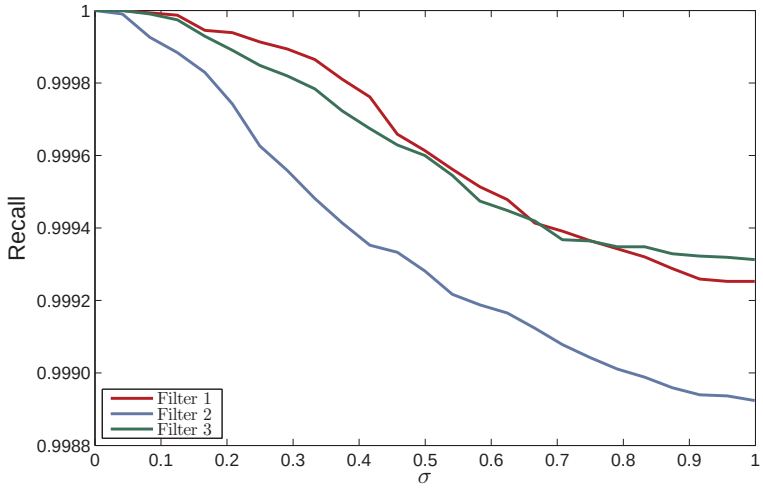


Figure 6.14: Recall lends itself to be interpreted in probabilistic terms. In particular, it may be regarded as the probability that a resource with initial state denied has been classified correctly. In this figure, we observe how tag suppression decreases this probability, but only to an insignificant extent.

In closing, we would like to emphasize the suitability of the tag-suppression rate as a measure of utility in our formulation of the trade-off; not only because the suppression rate allows us to model this trade-off as a mathematically tractable optimization problem, as we have shown in Chapter 5; but also because it is bijectively related to more elaborate utility metrics, as observed in Fig. 6.9 for the percentage of tags that bookmarks lose, and in Figs. 6.13 and 6.14 for precision and recall in a parental-control application.

6.6 Conclusions

Collaborative tagging is currently an extremely popular online service. Although it is basically used to support resource search and browsing, its potential is still to be exploited. Recently, some of the most popular collaborative tagging systems have come to understand the real value of tagging, and have started offering new services where personalization comes in.

At the heart of these services is the ability to profile users based on their tags. The flaw of this implicit form of user profile construction, however, is that the effectiveness of these services strongly depends on whether tagging systems have collected a large

amount of tags. This is the same problem arising in recommendation systems, i.e., the cold-start problem, and an alternative to this is that users *explicitly* provide their preferences. While this option is available in many personalized information systems, the fact is that it is not supported by any collaborative tagging system. Consequently, in order to exploit the potential of collaborative tagging, it would be necessary to extend the architecture of current tagging services to include a policy layer that supports the enforcement of user preferences.

On the other hand, as collaborative tagging has been gaining popularity, it has become more evident the need for privacy protection; not only because tags are sensitive information per se, but also because of the risk of cross-referencing. Besides, the fact that users can explicitly communicate their preferences to these enhanced collaborative tagging systems may facilitate the task of profiling. In a nutshell, collaborative tagging would also benefit from a layer helping users protect their privacy.

Motivated by all this, our first contribution is an architecture including two new layers on support of enhanced and private collaborative tagging. In particular, our architecture is composed of a bookmarking application and two additional services built on it. The former service is provided by a policy layer that permits users denoting resources of interests and specifying block conditions on the browsed data. The latter service is provided by the PET examined in Chapter 5, i.e., tag suppression.

Integrating these two layers enables us, first, to boost the services currently offered by collaborative tagging systems, and secondly, to thwart privacy attackers from profiling users. The problem of combining both services, however, is that the latter layer comes at the cost of data utility, which ultimately may impact the effectiveness of the enhanced collaborative tagging services enabled by the former layer. Our second and main contribution is precisely a thorough experimental analysis assessing the extent to which tag suppression, on the one hand, may contribute to privacy protection, and on the other it may negatively affect the functionality of two enhanced collaborative tagging services. Among other results, our empirical evaluation shows that tag suppression has a relatively small impact on the functionality of a parental-control application. We interpret this result as a consequence of the reduced number of active subcategories of bookmarks and our model of bookmark profile.

Chapter 7

Forgery and Suppression of Ratings in Recommendation Systems

7.1 Introduction

A personalized recommendation system [(a)] may be regarded as a type of information-filtering system that suggests information items users may be interested in. Examples of such systems include recommending music at Last.fm and Pandora Radio, movies by MovieLens and Netflix, and books and other products at Amazon.

As any personalized information system, recommenders capitalize on the creation of profiles to provide users with targeted information. On the one hand, such profiles may be *explicitly* declared by users. This is the case of the enhanced collaborative tagging application examined in Chapter 6. On the other hand, users' preferences may be *implicitly* inferred by the system based on their past activity and behavior. This is the most common form of profile construction, as typically users are reticent to voluntarily disclose their profile of interests.

In this latter kind of recommenders, a distinction is frequently made between *explicit* and *implicit* forms of data collection. The most popular form of explicit data collection is that users communicate their preferences by rating items. Such is

[(a)] For the sake of brevity, we shall often refer to personalized recommendation systems simply as recommendation systems or recommenders.

the case of many of the applications mentioned above, where users assign ratings to songs, movies or news they have already listened, watched or read. Other strategies to capture users' interests include asking them to sort a number of items by order of predilection, or suggesting that they mark the items they like. By contrast, recommendation systems may collect data from users without requiring them to rate information items. These practices comprise observing the items clicked by users in an online store, analyzing the time it takes users to examine an item, or simply keeping a record of the purchased items.

The prolonged collection of these personal data allows the system to build a profile of interests. With this invaluable source of information, the recommendation system applies some technique [163, 189] to generate a prediction of users' preferences for those items they have not yet considered. For example, Movielens and Digg use collaborative-filtering techniques to predict the rating that a user would give to a movie and to create a personalized list of recommended news, respectively.

Despite the many advantages recommendation systems are bringing to users, the information collected, processed and stored by these systems poses serious privacy risks. Such risks were carefully examined in Sec. 2.1.2 from a more general perspective, not limited to the particular case of recommenders. In response to the privacy concerns prompted by these information systems, it is not surprising that some users are reticent to reveal their interests. In fact, [190] reports that the 24% of Internet users surveyed provided false information in order to avoid giving private information to a Web site. Alternatively, another study [191] finds that 95% of the respondents refused, at some point, to provide personal information when requested by a Web site. In closing, these studies seem to indicate that submitting false information and refusing to give private information are strategies accepted by users concerned with their privacy.

In this chapter we approach the problem of protecting user privacy in those recommendation systems that profile users on the basis of the items they rate. Given the willingness of users to provide fake information and elude disclosing private data, we investigate a PET that simultaneously combines these two forms of data perturbation, namely the forgery and the suppression of ratings. Concordantly, in our scenario

a user rates those items they have an opinion on. But to prevent a privacy attacker from getting an accurate estimate of their profile, the user may want to refrain from rating some of those items and/or rate items that do not reflect their actual preferences. Our data-perturbative approach thus protects user privacy to a certain extent, and does not require the user to trust the recommendation system, nor the network operator nor any other external entity. The flip side, however, is that it comes at the cost of data utility, namely a degradation of the quality of the recommendation. In simple terms, the proposed PET poses a trade-off between privacy and utility.

The first contribution of this chapter is an architecture that describes the conceptual design and fundamental operational structure of a practical implementation of our PET. As in Chapter 5, our data-perturbative technique is intended to be implemented as a software-based service, e.g., a Web browser plug-in. The proposed architecture specifies how such software should operate. The ultimate aim of this architecture is to help users decide which ratings should be forged and which ones should be suppressed.

The theoretical analysis of the trade-off between the contrasting aspects of privacy and utility is the second and main contribution of this chapter. We tackle the issue in a systematic fashion, drawing upon the methodology of multiobjective optimization. Before proceeding, though, we adopt a quantifiable measure of user privacy—the KL divergence between the probability distribution of the user's items and the population's distribution, a criterion that we proposed in Chapter 4 and justified by leveraging on the rationale behind entropy-maximization methods. Equipped with a measure of both privacy and utility, we formulate an optimization problem modeling the trade-off between privacy on the one hand, and on the other forgery rate and suppression rate as utility metrics. Our extensive theoretical analysis finds a closed-form solution to the problem of optimal forgery and suppression of ratings, and characterizes the optimal trade-off surface between the aspects of privacy and utility.

Further, we provide an empirical evaluation of our data-perturbative approach. Specifically, we apply the forgery and the suppression of ratings to the popular movie recommendation system Movielens, and show how these two strategies may preserve

the privacy of its users. As we did in Chapter 5, the work presented here is also based on the adversary model defined in Chapter 4.

The work presented in this chapter is an extension of [49, 52, 192].

Chapter Outline

The remainder of this chapter is organized as follows. Sec. 7.2 introduces our PET. Sec. 7.3 defines the privacy criterion used in this chapter and specifies the assumptions about the adversary's capabilities. Sec. 7.4 presents an architecture describing a possible implementation of our privacy-protecting technology. Sec. 7.5 formulates the optimal trade-off between privacy and utility. Sec. 7.6 provides a theoretical analysis of the optimization problem characterizing this trade-off. Sec. 7.7 evaluates our privacy-protecting mechanism in a real recommendation system. Finally, conclusions are drawn in Sec. 7.8.

7.2 Privacy-Enhancing Mechanism

The privacy-enhancing mechanism investigated in this chapter combines two strategies based on data perturbation. On the one hand, the elimination of user data, a technique that we proposed and examined in depth in the context of semantic tagging; and on the other hand, the release of false information, a mechanism widely used not only in personalized information systems but also in anonymous communications and PIR.

In Sec. 5.2 we mentioned that refraining from sending sensitive, private information avoids potential privacy breaches and constitutes a great step forward in terms of the data-minimization principle. The *submission* of *false* data is a conceptually different approach to privacy protection. Actually, it can be viewed as the opposite strategy to suppression or the *retention* of user's *genuine* data. In Chapter 5 we discarded forgery as a privacy-enhancing mechanism for resource tagging, and deemed suppression a more fitting approach. The reason given was that forgery could lead to further degradation in semantic functionality, as tags aim at tying meaning up with resources; consider, for example, tagging a Web page about mental health with

the tag "car" [(b)]. In this chapter we rescue the suitable applicability of forgery to the scenario of recommendation systems, since ratings do not classify resources but merely assess their relevance.

The synergy of these two strategies, forgery and suppression, appears as a promising paradigm in the context of recommender systems. When users are adhered to this combined technique, they may submit ratings of items that do not reflect their actual preferences, and/or skip rating some items of their interest. This is what we refer to as the forgery and the suppression of ratings, respectively. Our PET thus enhances user privacy to a certain extent since the perturbed profile, as observed from the outside, no longer captures the precise and actual interests of the user in question [(c)]. In addition, the perturbative nature of our mechanism facilitates its implementation as a software program operating on the user's computer. This implies that users need not trust the recommendation system, nor the network operator nor any external entity.

In Sec. 5.2.1 we provided a thorough comparative analysis between tag suppression and the state of the art in PETs. The analysis carried out for suppression in that section can be directly extrapolated to the scenario of recommendation systems, in the more general case when suppression is combined with forgery. Next, we extend such analysis to include a couple of data-perturbative approaches specifically designed for the application of recommender systems.

In the context at hand, a common approach to privacy preservation consists in adding random values to ratings. An archetypical example is [94, 110]. In these works, the perturbation takes place on the users' side and affects only those items which users have previously rated. Once the ratings are modified, they are sent to the recommendation system, which calculates a weighted average of the ratings submitted

[(b)] In general, a particular data-perturbative mechanism will not be appropriate for all types of applications and contexts. There will be applications allowing only suppression, and others a combination of several mechanisms, for example. Essentially, this will depend on the type of information to be perturbed and the impact of such perturbation on system functionality and data utility.

[(c)] In fact, the purpose of our approach will not be to hide the actual profile of interests, but to make the perturbed profile more ordinary, less intriguing to an adversary who aims to target singular users. In other words, we shall assume the adversary model described in Sec. 4.3.3, where the attacker's objective is to individuate users. We shall explain it later in Sec. 7.3.

by all users. This information is then sent back to users, who ultimately use it to compute predictions about the unrated items. There are two important differences between this approach and ours. First, we send false ratings only to those items users have not rated yet. Secondly, and more importantly, concealing the actual ratings does not preclude a privacy attacker from profiling users on the basis of the items they rate. Put differently, the cited works overlook the fact that rating an item may be more sensitive than the particular score given.

7.3 Adversary Model and Privacy Metric

In this section we shall first specify our assumptions about the attacker. Based on these assumptions, we shall then define a privacy criterion which, later, will enable us to evaluate and subsequently optimize the privacy-enhancing mechanism proposed in Sec. 7.2. In essence, we shall assume the same adversary model described in Chapter 5. The main difference is that we shall suppose that users know or are able to estimate the population's item distribution. Next, we describe our assumptions about the adversary considered in this chapter.

We suppose that users are *identified* to the recommendation system. Recall from Sec. 4.3.1 that, by identified, we mean that users' activity is monitored by the system. This monitoring could be accomplished, for example, if users are logged into the recommender. On the other hand, our set of potential privacy attackers comprises any entity that can profile users based on their ratings. In other words, we contemplate an attacker who learns about the interests of users from the ratings they assign to information items. Our attacker may therefore be the recommender itself, but also the network operator and any passive eavesdropper.

In Sec. 4.3.2 we commented that user profiles are frequently modeled as histograms of user-generated data. In particular, we showed how numerous recommendation systems make use of this kind of representation. Examples of these systems include BibSonomy, Delicious, IMDb, Movielens and Pandora Radio. According to these examples, and consistently with the user-profile model considered in Chapter 5, we

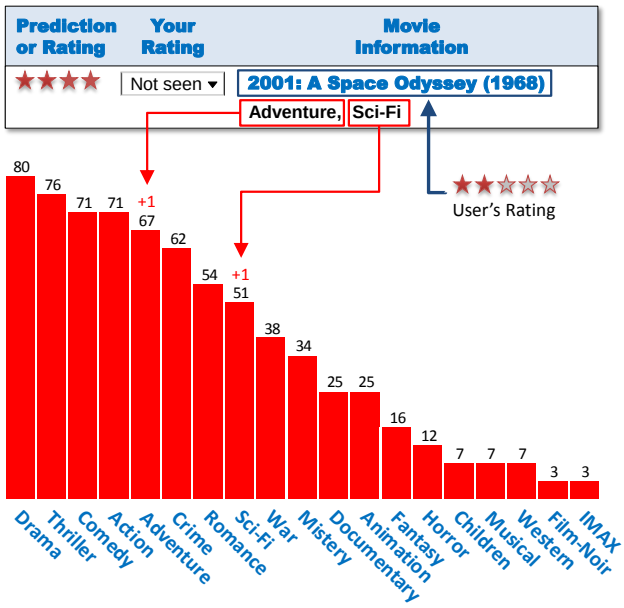


Figure 7.1: The profile of a user is modeled in Movielens as a histogram of absolute frequencies of ratings within a set of movie genres (bottom). Based on this profile, the recommender predicts the rating that the user would probably give to a movie (top). After having watched the movie, the user rates it and their profile is updated.

assume that our adversary models user interests by using histograms. More specifically, we consider a tractable model of user profile as a PMF, that is, a normalized histogram of ratings within a predefined set of categories of interest. We would like to remark that, under this model, user profiles do not capture the particular scores given to items, but what we consider to be more sensitive: the categories these items belong to. This corresponds to the case of Movielens, which we illustrate in Fig. 7.1. In this figure, we represent a user assigning two stars to a movie, meaning that they consider it to be "fairly bad". The recommender, however, updates their profile based only on the categories this movie belongs to.

In this chapter, users resort to our privacy-enhancing mechanism to prevent an attacker from constructing a precise characterization of their profiles. By adopting the forgery and the suppression of ratings, our attacker actually sees a perturbed version of the genuine profile of interests. In our adversary model, we assume that the attacker believes that the observed, perturbed profile is the actual one. Put another way, our adversary is unable to know if a particular user is using our mechanism. In Sec. 5.3

we argued that this assumption must not be considered as security through obscurity. The reason is that our data-perturbative mechanism is conceived to be implemented as a software program running on the user's local machine. As our approach operates on the user's side, it seems reasonable to assume that an external attacker cannot ascertain whether this software is running or not on the user's computer.

Unlike in previous chapters, here we contemplate that users are able to estimate the population's item distribution. In practice, this is an information that a software program could retrieve from the recommender. For example, Movielens provides users with the average rating assigned to each item, and IMDb shows the population's rating distribution of each item. However, if this information was not available at the recommender, an alternative might be querying other recommendation systems or using information services such as the Google Display Network Ad Planner [(d)]. This latter contains data about the distribution of user interests.

As in our tag-suppression mechanism, we suppose that the attacker's intention is to individuate users, that is, its goal is to find users who deviate from the average profile of interests. In view of the above, in this chapter we measure privacy risk, or more accurately, anonymity loss, as the KL divergence between the user's apparent item distribution and the population's item distribution. According to the arguments given in Sec. 4.4, our privacy criterion may be construed as a measure of the probability of the apparent profile. More precisely, the lower the discrepancy, in terms of KL divergence, between this profile and the population's distribution, the higher the likelihood of the apparent profile, and the greater the number of users who behave according to it. The assumptions made in this section are summarized in Fig. 7.2.

7.4 Architecture

In this section we define the major components of an architecture implementing our data-perturbative technique. The proposed architecture provides high-level functional aspects so that our PET can be implemented as a software tool installed on the user's

[(d)] `https://www.google.com/adplanner`

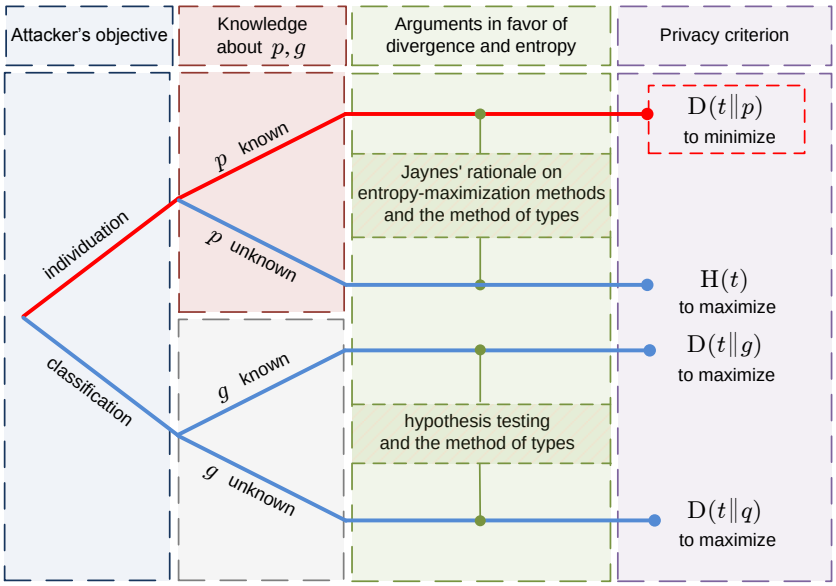


Figure 7.2: Summary of the main assumptions of our adversary model. Exactly as in Chapter 5, we contemplate an attacker who aims to individuate users. In this chapter, however, we assume that the population's item distribution is known to users. Under these assumptions, the KL divergence between the user's apparent profile and the population's distribution may be regarded as a measure of privacy, or more precisely, anonymity.

computer. We would like to stress that the description provided in this section does not pretend to serve as an exhaustive guide for programming such tool.

Our approach builds on the same assumptions than those of the architecture proposed in Sec. 5.4. For the sake of completeness, next we go through them briefly.

- First, we suppose that users trust the software implementation of our PET.
- Secondly, the proposed approach operates as a recommendation system, in the sense that it suggests which items should be forged and which ones should be suppressed. The software implementing our mechanism requires user permission to proceed with the forgery and the suppression of ratings.
- Also, we assume that the software implementation and the adversary use the same set of categories to model user interests. Further, we suppose that each information item is categorized in the same way by both the software and the attacker. In other words, when users neither forge nor eliminate ratings, the profile computed locally on their side matches the profile built by the attacker.

We believe this is a reasonable assumption as the categorization of items is frequently available to users of recommenders. This is the case, for example, of the movie recommendation systems IMDb, Jinni, Movielens and Netflix.

- In addition, our approach requires the user profile to start working. For this reason, we consider a training period before it can begin recommending which items should be forged and eliminated. The duration of this training phase will depend on the user's rating activity. Since the user's profile might be exposed during this phase, the user could alternatively declare their interests at the beginning. In this manner, ratings perturbation could be applied from the first moment. If it was the case, the declared profile would be replaced after the training phase by the profile estimated implicitly from tagging activity. The reason is that the former profile might not be a precise representation of the user's interests.

- Finally, we assume that, when estimating the user profile, the components of the relative histogram remain stable after the training phase. We recognize that this assumption may be an over-simplification, since user interests might vary considerably over time.

Having examined the assumptions about our architecture, next we make a distinction based on the user's knowledge about items. Hereafter we shall refer to the user's *known items* as those items they have an opinion on. In the case of the movie recommendation system IMDb, for example, the known items of a particular user would be those movies the user has already watched. Analogously, we shall refer to the user's *unknown items* as those items the user is not in the position to rate. For instance, this could be the case of a movie the user is not aware of or a movie the user has heard about, but has not watched yet.

This distinction allows us to specify more precisely the operation of our architecture. Namely, our approach makes use of the submission of ratings of unknown items and the suppression of ratings of known items. For the sake of brevity, we shall refer to these techniques simply as the forgery and suppression of ratings, respectively. Having said this, we would like to stress that the fact that forgery only applies to

unknown items is basically because users may be reluctant to assign false ratings to known items. Despite the above, our approach could also give the user the option to forge ratings of known items. However, for brevity, in this section we describe only the case where forgery applies just to unknown items. Next, we provide a high-level description of each of the components of the proposed architecture.

Communication manager. This module is in charge of interacting with the recommendation system. Specifically, it downloads information about the items the user finds when browsing the recommender's Web site. This information may include a description about the items, the ratings that other users assigned to them, and the categories of interest these items belong to. In Amazon, for instance, all this information is available to users. However, since this is not always the case, our approach incorporates modules intended to retrieve the population's ratings and to categorize all the items that the user explores.

On the other hand, this module receives the ratings of unknown items suggested by the *forgery alarm generator* and the ratings of known items sent by the *suppression alarm generator*. Afterwards, the module submits these ratings to the recommendation system.

Category extractor. This component is responsible for obtaining the categories the items belong to. To this end, the module uses the information provided by the communication manager. Should this information not be enough, the module will have to get additional data by searching the Web or by querying an information provider. Afterwards, the categorization of these items is carried out by using the vector space model and the TF-IDF weights, similarly as in Sec. 5.4. In a last stage, this module sends the items and their corresponding categories to the *known/unknown item classifier*.

Known/unknown item classifier. This module requires the active involvement of the user. Namely, it shows the user the items categorized by the category extractor module, and then asks the user to classify them as known or unknown. Evidently, this module will have previously checked whether these items have already been rated by the user. Should this be the case, the rated items would not be shown to the user, since these items would be classified as known items. For this purpose, the module keeps

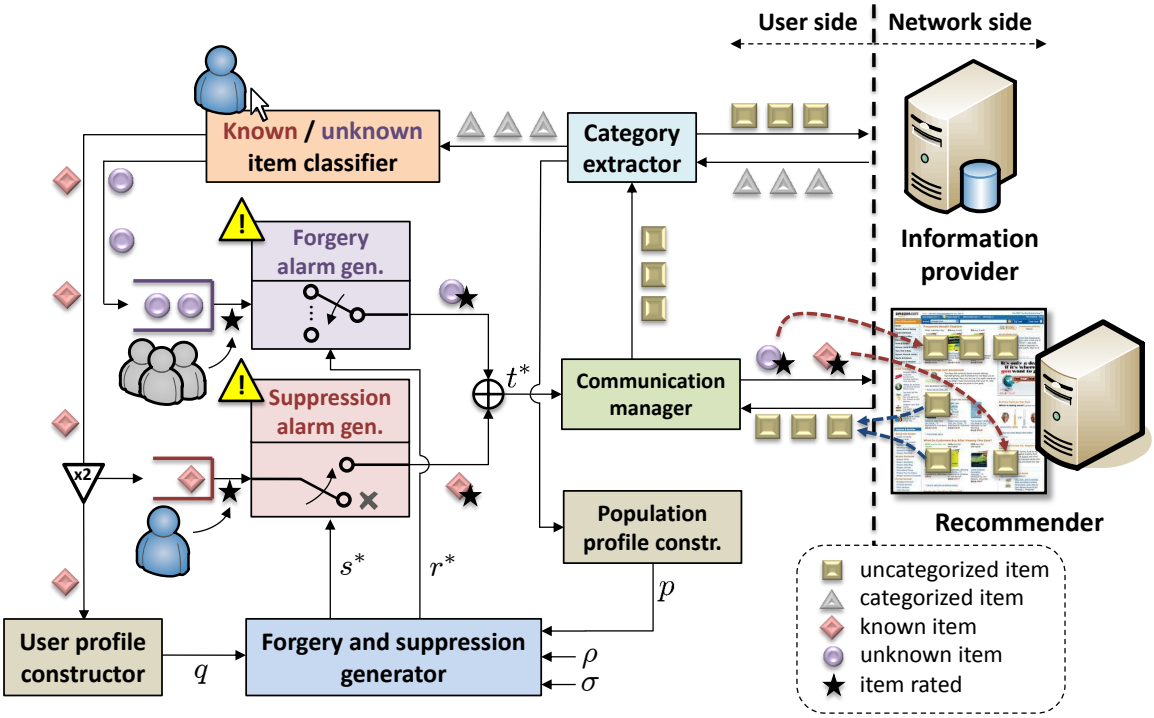


Figure 7.3: Block diagram of the proposed architecture.

a record of all the items that the user rates. Once these items have been classified as known or unknown, they are sent to the *forgery alarm generator* and the *suppression alarm generator*, respectively. In addition, the known items are submitted to the *user profile constructor*.

User profile constructor. This module is in charge of obtaining the user profile. To this end, the module is provided with the user's known items, i.e., those items capturing their preferences. Based on these items, it generates the user profile as described in Sec. 7.3. As mentioned at the beginning of this section, we assume that the relative frequencies of activity stabilize after the user has rated a large number of known items. Analogously to the architecture presented in Sec. 5.4, we also contemplate the possibility that the user explicitly specifies their profile to prevent an attacker from profile them during the training phase.

Population profile constructor. This module is responsible for the estimation of the population's item distribution. For this purpose, the module relies on the items

retrieved by the communication manager. As commented in Sec. 7.3, many recommendation system provides the categories their items belong to as well as detailed statistics about the ratings assigned by users. If this information were not enough to build the population's profile, this module could resort to other recommender or databases containing this kind of information.

Forgery and suppression generator. This block is the centerpiece of the architecture. First, the block is provided with the user profile and the population's distribution. In addition, the user specifies a forgery rate ρ and a suppression rate σ. The former is the fraction of ratings of unknown items that the user is willing to submit. The latter is the relative frequency of ratings of known items that the user is disposed to eliminate. Having specified these two rates, the module computes the optimal tuples of forgery r^* and suppression s^*, which indicate the fraction of ratings that should be forged and suppressed, respectively. More accurately, the component r_i^* is the percentage of ratings of unknown items that our architecture suggests submitting in the category $i \in \mathbb{N}^+$. The component s_i^* is defined analogously for suppression.

In the end, these two tuples are sent to the forgery alarm generator and the suppression alarm generator, respectively. Later in Sec. 7.5, we shall provide a more detailed specification of this module by using a formulation of the trade-off among privacy, forgery rate and suppression rate, which will enable us to compute the tuples r^* and s^*.

Suppression alarm generator. This module is responsible for warning the user when their privacy is at risk. Concretely, this module receives the tuple s^* and stores the known items provided by the known/unknown item classifier. These items are kept in an array. When the user decides to assign a rating to one of these items, the selected item is removed from such array. The user then rates this item, and the module proceeds as follows. Suppose that the item in question belongs to the category i. According to this, and exactly as in Sec. 5.4, the module generates an alarm with probability s_i^*. If the alarm is eventually triggered, the user must choose either to drop the rating or to send it to the recommender through the communication

manager module. If the alarm is not triggered, the rating is forwarded directly to the latter module.

Forgery alarm generator. Our approach also relies on the forgery of ratings. Specifically, this module selects, on the one hand, which unknown items should be forged, and on the other hand, which particular ratings should be assigned to these unknown items. With regard to the ratings to be given to the items, we follow a method similar to the one pointed out in [85]. Namely, our approach assigns each unknown item a random rating, drawn according to the distribution of the other users' ratings to that item. Alternatively, we could also contemplate the distribution of ratings of a user with similar preferences, or the distribution of ratings across all items. In order to obtain this information, the module will have to query information providers or explore other recommenders. In the case of Amazon, for example, this is not necessary since users are provided with the population's ratings.

In parallel, the module receives unknown items and stores them in an array. After getting the tuple r^*, the module proceeds as follows. Every time the user decides to assign a rating to a known item, regardless of whether this rating is finally submitted to the recommender or it is eliminated, the module chooses, at random, unknown items from the array. This selection is done according to the percentages specified by r^*. Specifically, the probability that an item corresponding to the category i be chosen is r_i^*. Once the module has chosen one item, our architecture encourages the user to submit it to the recommender. However, it is the user who finally decides whether to send this rating or not. If the user accepts the recommendation, then the rating is sent to the communication manager module, and the unknown item is removed from the array.

After having explored each of the modules of the architecture, now we shall describe how it would work. Initially, the user would browse the recommendation system's Web site and would find some items. In order for the user to obtain future recommendations from the system, they would have to rate some of those items. Before proceeding, though, our approach would retrieve information about the items and extract the categories they belong to. Afterwards, the user would be asked to

classify the items as known or unknown. The known items would allow the proposed architecture to build the user profile during the training phase.

After this phase, our approach would compute the tuples r^* and s^*. When trying to rate some of the known items, the user could receive two types of alarms. In particular, our architecture could suggest that the user refrain from rating some of those known items and could also recommend submitting a random rating to one or more of the unknown items. In any case, it would be up to the user to decide whether to eliminate and forge such ratings.

7.5 Trade-Off among Privacy, Forgery and Suppression

Our data-perturbative mechanism allows users to enhance their privacy to a certain extent, since the resulting profile, as observed from the outside, appears to be much more ordinary, and therefore less valuable to an attacker aimed at targeting singular users. The price to be paid, however, is a loss in data utility, in particular in the accuracy of the recommender's predictions.

Next, we present a formulation of the optimal trade-off between the contrasting aspects of privacy and utility. For the sake of tractability, we consider as utility metrics the forgery rate and the suppression rate. We would like to remark that the study of more sophisticated metrics of any loss in the accuracy of the recommendations due to ratings perturbation is an open problem.

The consideration of those two rates as simplified utility measures enables us to formulate the privacy-utility trade-off by means of a mathematically tractable model. Specifically, in this section we shall be able to formulate the problem of choosing a forgery strategy and a suppression strategy as a multiobjective optimization problem that takes into account privacy, forgery rate and suppression rate. As we shall show next, this formulation will enable us to go into the details of one of the functional blocks of the proposed architecture.

We begin by formalizing some of the concepts that we introduced in previous sections. Specifically, we model the known items of a user as a sequence of i.i.d. r.v.'s taking on values in a common finite alphabet of categories, in particular the set

$\{1, \ldots, n\}$ for some integer $n \geqslant 2$. Concordantly, we represent the profile of a user as the common PMF of such r.v.'s, $q = (q_1, \ldots, q_n)$, conceptually a histogram of relative frequencies of items within that set of categories.

When users adhere to the forgery and the suppression of ratings, they specify a *forgery rate* $\rho \in [0, \infty)$ and a *suppression rate* $\sigma \in [0, 1)$. The former is the ratio of forged ratings to total genuine ratings that a user consents to submit. The latter ratio is the fraction of genuine ratings that the user agrees to eliminate. Note that, in our approach, the number of false ratings submitted by the user can exceed the number of genuine ratings, that is, ρ can be greater than 1. Nevertheless, the number of suppressed ratings is always lower than the number of genuine ratings.

By forging and suppressing ratings, the *actual* profile of interests q is then perceived from the outside as the *apparent* PMF $t = \frac{q+r-s}{1+\rho-\sigma}$, according to a *forgery strategy* $r = (r_1, \ldots, r_n)$ and a *suppression strategy* $s = (s_1, \ldots, s_n)$. Such strategies represent the proportion of ratings that the user should forge and eliminate in each of the n categories. Naturally, these strategies must satisfy, on the one hand, that $r_i \geqslant 0$, $s_i \geqslant 0$ and $q_i + r_i - s_i \geqslant 0$ for $i = 1, \ldots, n$, and on the other, that $\sum_{i=1}^{n} r_i = \rho$ and $\sum_{i=1}^{n} s_i = \sigma$. In conclusion, the apparent profile is the result of the addition and the substraction of certain items to/from the actual profile, and the posterior normalization by $\frac{1}{1+\rho-\sigma}$ so that $\sum_{i=1}^{n} t_i = 1$.

According to the adversary model and privacy metric assumed in Sec. 7.3, we define *initial privacy risk* as the KL divergence [76] between the user's genuine profile and the population's item distribution p, that is,

$$\mathcal{R}_0 = \mathrm{D}(q \,\|\, p).$$

Similarly, we define *(final) privacy risk* $\mathcal{R}$ as the KL divergence between the user's apparent profile and the population's distribution,

$$\mathcal{R} = \mathrm{D}(t \,\|\, p) = \mathrm{D}\left(\frac{q + r - s}{1 + \rho - \sigma} \,\middle\|\, p\right).$$

Under the assumption that the population of users is large enough to neglect the impact of the choice of r and s on p, we define the *privacy-forgery-suppression* function

$$\mathcal{R}(\rho, \sigma) = \min_{\substack{r,s \\ r_i \geqslant 0,\, s_i \geqslant 0, \\ q_i + r_i - s_i \geqslant 0, \\ \sum r_i = \rho,\, \sum s_i = \sigma}} \mathrm{D}\left(\frac{q + r - s}{1 + \rho - \sigma} \,\middle\|\, p \right), \tag{7.1}$$

which characterizes the optimal trade-off among privacy, forgery rate and suppression rate, and allows us to formally specify the module *forgery and suppression generator* described in Sec. 7.4. More accurately, this functional block will be in charge of solving the optimization problem inherent in the definition of function (7.1).

Lastly, we would like to emphasize that, in our mathematical formulation, the KL divergence is more precisely regarded as a measure of anonymity, rather than privacy. Specifically, in Sec. 4.4 we interpreted the KL divergence as an indicator of the uniqueness of a profile within a population. Under this interpretation, the objective of our data-perturbative approach is not to conceal the user's actual profile, but to make the observed profile as common as possible. Table 7.1 summarizes the notation introduced in this section.

7.6 Theoretical Analysis

This section is entirely devoted to the theoretical analysis of the privacy-forgery-suppression function (7.1) defined in Sec. 7.5. In our attempt to characterize the trade-off among privacy risk, forgery rate and suppression rate, we shall present a closed-form solution to the optimization problem inherent in the definition of this function. Afterwards, we shall analyze some fundamental properties of said trade-off. For the sake of brevity, our theoretical analysis only contemplates the case when all given probabilities are strictly positive:

$$q_i, p_i > 0 \text{ for all } i = 1, \ldots, n. \tag{7.2}$$

Table 7.1: Description of the variables used in our notation.

Symbol	Description
n	number of interest categories into which information items are classified
q	the *actual* user profile is the genuine profile of interests
ρ, σ	the *rating-forgery rate* and the *rating-suppression rate* are the percentages of ratings the user is willing to forge and suppress, respectively
r, s	a *forgery strategy* and a *suppression strategy* are two n-tuples containing the percentage of ratings the user should forge and eliminate, respectively, in each category
t	the *apparent* user profile is the perturbed profile, resulting from the forgery and the suppression of certain ratings
p	*population*'s item distribution
$\mathrm{D}(t \,\|\, p)$	(final) *privacy risk* is measured as the KL divergence between the user's apparent distribution and the population's distribution
$\mathcal{R}(\rho, \sigma)$	function modeling the trade-off among privacy risk, forgery rate and suppression rate

The general case can easily be dealt with, occasionally via continuity arguments. Additionally, we suppose without loss of generality that

$$\frac{q_1}{p_1} \leqslant \cdots \leqslant \frac{q_n}{p_n}. \tag{7.3}$$

Before diving into the mathematical analysis, it is immediate from the definition of the privacy-forgery-suppression function that its initial value is $\mathcal{R}(0, 0) = \mathrm{D}(q \,\|\, p)$. The characterization of the optimal trade-off surface modeled by $\mathcal{R}(\rho, \sigma)$ at any other values of ρ and σ is the focus of this section.

7.6.1 Closed-Form Solution

Our first theorem, Theorem 7.3, will present a closed-form solution to the minimization problem involved in the definition of function (7.1). The solution will be derived from Lemma 7.1, which addresses a resource allocation problem. This a theoretical problem encountered in many fields, from load distribution and production planning

to communication networks, computer scheduling and portfolio selection [193]. Although this lemma provides a parametric-form solution, we shall be able to proceed towards an explicit closed-form solution, albeit piecewise.

Lemma 7.1 (Resource Allocation)**.** *For all $k = 1, \ldots, n$, let f_k be a real-valued function on $\{(x_k, y_k) \in \mathbb{R}^2 \colon \kappa_k + x_k - y_k \geqslant 0\}$, twice differentiable in the interior of its domain. Assume that $\frac{\partial f_k}{\partial x_k} = -\frac{\partial f_k}{\partial y_k}$, that $\frac{\partial^2 f_k}{\partial x_k^2} = \frac{\partial^2 f_k}{\partial y_k^2} > 0$ and that the Hessian $H(f_k)$ is positive semidefinite. Define $h_k = \frac{\partial f_k}{\partial x_k}$. Because $\frac{\partial h_k}{\partial x_k} > 0$ and $\frac{\partial h_k}{\partial y_k} < 0$, it follows that h_k is strictly increasing in x_k and strictly decreasing in y_k. Consequently, for a fixed y_k, $h_k(x_k, y_k)$ is an invertible function of x_k. Denote by h_k^{-1} the inverse of $h_k(x_k, 0)$. Suppose further that $h_k(x_k, y_k) = h_k(x_k - y_k, 0)$ and finally that $\lim_{x_k \downarrow y_k - \kappa_k} h_k(x_k, y_k) = -\infty$. Now consider the following optimization problem in the variables $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$:*

$$\begin{aligned} &\text{minimize} && \sum_{k=1}^{n} f_k(x_k, y_k) \\ &\text{subject to} && x_k,\, y_k \geqslant 0, \\ & && \kappa_k + x_k - y_k \geqslant 0 \text{ for } k = 1, \ldots, n, \\ & && \text{and } \sum_{k=1}^{n} x_k = \eta,\, \sum_{k=1}^{n} y_k = \theta \text{ for some } \eta, \theta \geqslant 0. \end{aligned}$$

(i) *The solution to the problem (x_k^*, y_k^*) depends on two real numbers ψ, ω that satisfy the equality constraints $\sum_k x_k^* = \eta$ and $\sum_k y_k^* = \theta$. The solution exists provided that $\psi \leqslant \omega$. If $\psi < \omega$, then the solution is unique and yields*

$$(x_k^*, y_k^*) = \left(\max\left\{0, h_k^{-1}(\psi)\right\}, \max\left\{0, -h_k^{-1}(\omega)\right\}\right).$$

If $\psi = \omega$, then there exists an infinite number of solutions of the form $(x_k^* + \alpha_k, y_k^* + \alpha_k)$ for all $\alpha_k \in \mathbb{R}_+$ meeting the two aforementioned equality constraints.

Without loss of generality, suppose that $h_1(0, 0) \leqslant \cdots \leqslant h_n(0, 0)$.

(ii) For $\psi < \omega$, consider the following cases:

(a) $h_i(0,0) < \psi \leqslant h_{i+1}(0,0)$ *for some* $i = 1, \ldots, j-1$ *and* $h_{j-1}(0,0) \leqslant \omega < h_j(0,0)$ *for some* $j = 2, \ldots, n$.

(b) $h_{j-1}(0,0) \leqslant \omega$ *for* $j = n+1$ *and, either* $h_i(0,0) < \psi \leqslant h_{i+1}(0,0)$ *for some* $i = 1, \ldots, n-1$ *or* $h_i(0,0) < \psi$ *for* $i = n$.

(c) $\psi \leqslant h_{i+1}(0,0)$ *for* $i = 0$ *and, either* $h_{j-1}(0,0) \leqslant \omega < h_j(0,0)$ *for some* $j = 2, \ldots, n$ *or* $\omega < h_j(0,0)$ *for* $j = 1$.

(d) $h_{j-1}(0,0) \leqslant \omega$ *for* $j = n+1$ *and* $\psi \leqslant h_{i+1}(0,0)$ *for* $i = 0$.

In each case, and for the corresponding indexes i *and* j,

$$x_k^* = \begin{cases} h_k^{-1}(\psi) & , \quad k = 1, \ldots, i \\ 0 & , \quad k = i+1, \ldots, n \end{cases},$$
$$y_k^* = \begin{cases} 0 & , \quad k = 1, \ldots, j-1 \\ -h_k^{-1}(\omega) & , \quad k = j, \ldots, n \end{cases}.$$

(iii) For $\psi = \omega$, *consider the following cases:*

(a) either $h_i(0,0) < \psi < h_j(0,0)$ *for some* $j = 2, \ldots, n$ *and* $i = j-1$, *or* $h_i(0,0) < \psi = h_{i+1}(0,0) = \cdots = h_{j-1}(0,0) < h_j(0,0)$ *for some* $i = 1, \ldots, j-2$ *and some* $j = 3, \ldots, n$.

(b) for $j = n+1$, *either* $h_i(0,0) < h_{i+1}(0,0) = \cdots = h_{j-1}(0,0) = \omega$ *for some* $i = 1, \ldots, j-2$ *or* $h_{j-1}(0,0) < \omega$ *with* $i = n$.

(c) for $i = 0$, *either* $\psi = h_{i+1}(0,0) = \cdots = h_{j-1}(0,0) < h_j(0,0)$ *for some* $j = 2, \ldots, n$ *or* $\psi < h_{i+1}(0,0)$ *with* $j = 1$.

In each case, and for the corresponding indexes i *and* j,

$$x_k^* = \begin{cases} h_k^{-1}(\psi) + \alpha_k & , \quad k = 1, \ldots, i \\ \alpha_k & , \quad k = i+1, \ldots, n \end{cases},$$
$$y_k^* = \begin{cases} \alpha_k & , \quad k = 1, \ldots, j-1 \\ -h_k^{-1}(\omega) + \alpha_k & , \quad k = j, \ldots, n \end{cases}.$$

Proof: The proof of statement (i) consists of two steps. In the first step, we show that the optimization problem stated in the lemma is convex; then we apply KKT

conditions to said problem, and finally reformulate these conditions into a reduced number of equations. The bulk of this proof comes later, in the second step, where we proceed to solve the system of equations for the two cases considered in the lemma, $\psi < \omega$ and $\psi = \omega$. Lastly, statements (ii) and (iii) follow from (i).

To see that the problem is convex, simply observe that the objective function is convex on account of $H(f_k) \succeq 0$, and that the inequality and equality constraint functions are affine. Since the objective and constraint functions are also differentiable and Slater's constraint qualification holds, KKT conditions are necessary and sufficient conditions for optimality [75]. Systematic application of these optimality conditions leads to the Lagrangian cost,

$$\begin{aligned}\mathcal{L} = \sum f_k(x_k, y_k) - \sum \lambda_k x_k - \sum \mu_k y_k \\ + \sum \nu_k(y_k - \kappa_k - x_k) - \psi\left(\sum x_k - \eta\right) + \omega\left(\sum y_k - \theta\right),\end{aligned}$$

and finally to the conditions

$$\begin{aligned}
&x_k \geqslant 0,\ y_k \geqslant 0,\ \kappa_k + x_k - y_k \geqslant 0, && \\
&\textstyle\sum x_k = \eta,\ \sum y_k = \theta, && \text{(primal feasibility)} \\
&\lambda_k \geqslant 0,\ \mu_k \geqslant 0,\ \nu_k \geqslant 0, && \text{(dual feasibility)} \\
&\lambda_k\, x_k = 0,\ \mu_k\, y_k = 0, && \\
&\nu_k\,(y_k - \kappa_k - x_k) = 0, && \text{(complementary slackness)} \\
&\tfrac{\partial \mathcal{L}}{\partial x_k} = h_k(x_k, y_k) - \lambda_k - \nu_k - \psi = 0, && \\
&\tfrac{\partial \mathcal{L}}{\partial y_k} = h_k(x_k, y_k) + \mu_k - \nu_k - \omega = 0, && \text{(dual optimality).}
\end{aligned}$$

Because $\lim_{x_k \downarrow y_k - \kappa_k} h_k(x_k, y_k) = -\infty$, it follows from the dual optimality conditions that $\kappa_k + x_k - y_k > 0$, which implies, by complementary slackness, that $\nu_k = 0$. Subsequently, we may rewrite the dual optimality conditions as $\lambda_k = h_k(x_k, y_k) - \psi$ and $\mu_k = \omega - h_k(x_k, y_k)$. By eliminating the slack variables λ_k, μ_k, we obtain the simplified conditions $h_k(x_k, y_k) \geqslant \psi$ and $h_k(x_k, y_k) \leqslant \omega$. Lastly, we substitute the above expressions of λ_k and μ_k into the complementary slackness conditions, so that we can formulate the dual optimality and complementary slackness conditions

equivalently as

$$h_k(x_k, y_k) \geqslant \psi, \tag{7.4}$$

$$h_k(x_k, y_k) \leqslant \omega, \tag{7.5}$$

$$(h_k(x_k, y_k) - \psi)\, x_k = 0, \tag{7.6}$$

$$(h_k(x_k, y_k) - \omega)\, y_k = 0. \tag{7.7}$$

In the following, we shall proceed to solve these equations which, together with the primal and dual feasibility conditions, are necessary and sufficient conditions for optimality. To this end, first note that, if $\psi > \omega$, then there exists no (x_k, y_k) that satisfies equations (7.4) and (7.5) at the same time, and consequently, as stated in part (i) of the lemma, there is no solution. Concordantly, next we shall study the case when $\psi < \omega$; afterwards we shall tackle the other case when $\psi = \omega$.

Before plunging into the analysis of the former case, recall that the function h_k is strictly increasing in x_k and strictly decreasing in y_k. Having said this, observe that, under the assumption $\psi < \omega$, the variables x_k and y_k cannot be positive simultaneously by virtue of equations (7.6) and (7.7). Bearing this in mind, consider these three possibilities for each k: $h_k(0,0) < \psi$, $\psi \leqslant h_k(0,0) \leqslant \omega$ and $\omega < h_k(0,0)$.

When $h_k(0,0) < \psi$, the only conclusion consistent with (7.4) and with the fact that h_k is strictly increasing in x_k is that $x_k > 0$. Since x_k must be positive, the complementary slackness condition (7.6) implies that $h_k(x_k, y_k) = \psi$ and, because of (7.7), that $y_k = 0$. As a result, x_k must satisfy $h_k(x_k, 0) = \psi$, or equivalently, $x_k = h_k^{-1}(\psi)$. Next, we show that the solution $(x_k, 0)$ is unique. For this purpose, suppose that $y_k > 0$ and, in consequence, that $x_k = 0$. It follows from (7.7), however, that $h_k(0, y_k) = \omega$, which contradicts the fact that h_k is a strictly decreasing function of y_k. In the end, we verify that $x_k = y_k = 0$ does not satisfy (7.4) and thus prove that $(x_k, y_k) = (h_k^{-1}(\psi), 0)$ is the unique minimizer of the objective function when $h_k(0,0) < \psi$.

Now consider the case when $\psi \leqslant h_k(0,0) \leqslant \omega$. First, suppose that $x_k > 0$, and therefore that $y_k = 0$. By complementary slackness, it follows that $h_k(x_k, 0) = \psi$, which is not consistent with the fact that h_k is strictly increasing in x_k. Consequently,

x_k cannot be positive. Secondly, assume that x_k is zero and y_k positive. Under this assumption, equation (7.7) implies that $h_k(0, y_k) = \omega$, a contradiction since h_k is a strictly decreasing function of y_k. Accordingly, y_k cannot be positive either. Finally, check that $x_k = y_k = 0$ satisfies the optimality conditions and hence it is the unique solution.

The last possibility corresponds to the case when $\omega < h_k(0, 0)$. Note that, in this case, the only conclusion consistent with (7.5) and with the fact that h_k is strictly decreasing in y_k is that $y_k > 0$. Thus, because of (7.7), y_k must satisfy $h_k(0, y_k) = \omega$. Recalling from the lemma that $h_k(x_k, y_k) = h_k(x_k - y_k, 0)$, we may express the condition $h_k(0, y_k) = \omega$ equivalently as $y_k = -h_k^{-1}(\omega)$. Lastly, we check that this solution is unique in the case under study. To this end, note that a solution such that $x_k > 0$ and $y_k = 0$ contradicts the fact that h_k is strictly increasing in x_k. As a result, x_k cannot be positive. Finally, we confirm that equation (7.5) does not hold for $x_k = y_k = 0$ and therefore prove that $(x_k, y_k) = (0, -h_k^{-1}(\omega))$ is the unique solution when $\omega < h_k(0, 0)$.

In summary, $x_k = h_k^{-1}(\psi)$ if $h_k(0, 0) < \psi$, or equivalently, $h_k^{-1}(\psi) > 0$; otherwise $x_k = 0$. Further, $y_k = -h_k^{-1}(\omega)$ if $h_k(0, 0) > \omega$, or equivalently, $h_k^{-1}(\omega) < 0$; otherwise $y_k = 0$. Accordingly, we may write the solution compactly as

$$(x_k, y_k) = (\max\left\{0, h_k^{-1}(\psi)\right\}, \max\left\{0, -h_k^{-1}(\omega)\right\}),$$

where ψ, ω must satisfy the primal equality constraints $\sum_k x_k = \eta$ and $\sum_k y_k = \theta$.

Having examined the case when $\psi < \omega$, next we proceed to solve the optimality conditions at hand for $\psi = \omega$. Observe that, in this new case, (7.4) and (7.5) transform into the equation

$$h_k(x_k, y_k) = \psi. \tag{7.8}$$

Moreover, note that any pair (x_k, y_k) satisfying (7.8) also meets the complementary slackness conditions (7.6) and (7.7). However, notice that this does not mean that all those pairs are optimal. To elaborate on this point, consider the following three possibilities for each k: $h_k(0, 0) < \psi$, $h_k(0, 0) = \psi$ and $\psi < h_k(0, 0)$.

In the case when $h_k(0,0) < \psi$, the only condition consistent with (7.8) and with the fact that h_k is strictly increasing in x_k is that $x_k > 0$. From the lemma, it is immediate that $\frac{\partial h_k}{\partial x_k} = -\frac{\partial h_k}{\partial y_k}$, which implies that x_k must also be greater than y_k. Hence, the set of solutions is

$$\{(x_k, y_k)\colon h_k(x_k, y_k) = \psi,\, x_k > y_k\},$$

where every pair in this set must also fulfill the primal equality conditions. Let x'_k satisfy $h_k(x'_k, 0) = \psi$, or equivalently, $x'_k = h_k^{-1}(\psi)$. Then, because $h_k(x'_k + \alpha_k, \alpha_k) = \psi$ for any $\alpha \geqslant 0$, this set may be recast equivalently as

$$\{(x_k, y_k)\colon x_k = x'_k + \alpha_k,\, y_k = \alpha_k\}.$$

For the two remaining cases, i.e., $h_k(0,0) = \psi$ and $\psi < h_k(0,0)$, the set of solutions is obtained in a completely analogous way as above. In the former case, the pairs (x_k, y_k) must satisfy $x_k = y_k$, and the set of solutions may be expressed as

$$\{(x_k, y_k)\colon x_k = \alpha_k,\, y_k = \alpha_k\}.$$

In the latter case, it follows that $y_k > x_k$ and, consequently, that the set of solutions is

$$\{(x_k, y_k)\colon x_k = \alpha_k,\, y_k = y'_k + \alpha_k\},$$

where y'_k must satisfy $h_k(0, y'_k) = \psi$.

To sum up, the case $\psi = \omega$ leads to the following solutions: $x_k = h_k^{-1}(\psi) + \alpha_k$ if $h_k(0,0) < \psi$, or equivalently, $h_k^{-1}(\psi) > 0$; otherwise $x_k = \alpha_k$. In addition, $y_k = -h_k^{-1}(\omega) + \alpha_k$ if $h_k(0,0) > \omega$, or equivalently, $h_k^{-1}(\omega) < 0$; otherwise $y_k = \alpha_k$. Accordingly, the solutions (x_k, y_k) yield

$$(\max\left\{0, h_k^{-1}(\psi)\right\} + \alpha_k, \max\left\{0, -h_k^{-1}(\omega)\right\} + \alpha_k), \tag{7.9}$$

for some ψ, ω and nonnegative sequence $\alpha_1, \ldots, \alpha_n$ such that $\sum_k x_k = \eta$ and $\sum_k y_k = \theta$. Note that, although $\psi = \omega$, we intentionally write ω instead of ψ to highlight that

the solutions for $\psi < \omega$ and for $\psi = \omega$ just differ in the term α_k, as we claimed in part (i) of the lemma.

To complete the proof of statement (i), it suffices to show that the number of solutions is infinite when $\psi = \omega$. To this end, simply observe that there exists an infinite number of sequences $\alpha_1, \ldots, \alpha_n$ such that

$$\sum_k x_k = \sum_k h_k^{-1}(\psi) + \sum_k \alpha_k = \eta \quad \text{and}$$

$$\sum_k y_k = -\sum_k h_k^{-1}(\psi) + \sum_k \alpha_k = \theta,$$

which results in an infinite number of solutions of the form given in (7.9).

Now we proceed to prove (ii), which is an immediate consequence of (i). For this purpose, observe that if $\psi \leqslant h_{i+1}(0,0) \leqslant \cdots \leqslant h_n(0,0)$ holds for some $i = 0, \ldots, n-1$, then $h_{i+1}^{-1}(\psi), \ldots, h_n^{-1}(\psi) \leqslant 0$, and accordingly $x_{i+1} = \cdots = x_n = 0$. Similarly, if $h_1(0,0) \leqslant \cdots \leqslant h_{j-1}(0,0) \leqslant \omega$ is satisfied for some $j = 2, \ldots, n+1$, then $h_1^{-1}(\omega), \ldots, h_{j-1}^{-1}(\omega) \geqslant 0$, and thus $y_1 = \cdots = y_{j-1} = 0$.

Note that the particular case when the index i ranges from 1 to $j-1$ and the index j goes from 2 to n is the case described in (ii) (a), which corresponds to $\eta, \theta > 0$. Further, observe that the case assumed in (ii) (b), i.e., when $j = n+1$, implies that $\theta = 0$. Here, the index i starts at 1, therefore excluding $\eta = 0$, and ends at n, including the possibility that $x_i > 0$ for all i. In part (ii) (c), we consider $i = 0$, which is equivalent to the condition $\eta = 0$. In this case, the index j starts at 1, permitting $y_j > 0$ for all j, and ends at n, avoiding $\theta = 0$. Finally, the case described in (ii) (d), namely when $j = n+1$ and $i = 0$, is precisely the trivial case $x = y = 0$.

To verify statement (iii), we proceed analogously by noting that if $\psi = h_{i+1}(0,0) = \cdots = h_{j-1}(0,0)$ holds for some $i = 1, \ldots, j-2$ and some $j = 3, \ldots, n$, then $h_{i+1}^{-1}(\psi) = \cdots = h_{j-1}^{-1}(\psi) = 0$, and consequently $x_k = y_k = \alpha_k$ for $k = i+1, \ldots, j-1$. ■

The previous lemma presented the solution to a resource allocation problem that minimizes a rather general but convex objective function, subject to affine constraints. Our next theorem, Theorem 7.3, applies the results of this lemma to the special case of the objective function of problem (7.1). In doing so, we shall confirm the intuition

that there must exist a set of ordered pairs (ρ, σ) where the privacy risk vanishes and another set where it does not. We shall refer to the former set as the *critical-privacy region* and formally define it as

$$\mathscr{C} = \{(\rho, \sigma) \colon \mathcal{R}(\rho, \sigma) = 0\}.$$

The latter set will be the complementary set $\bar{\mathscr{C}}$ and we shall refer to it as the *noncritical-privacy region*.

Before proceeding with Theorem 7.3, first we shall introduce what we term *forgery* and *suppression thresholds*, two sequences of rates that will play a fundamental role in the characterization of the solution to the minimization problem defining the privacy-forgery-suppression function. Secondly, we shall investigate certain properties of these thresholds in Proposition 7.2. And thereafter, we shall introduce some definitions that will facilitate the exposition of the aforementioned theorem.

Let $Q_i = \sum_{k=1}^{i} q_k$ and $P_i = \sum_{k=1}^{i} p_k$ be the cumulative distribution functions corresponding to q and p. Denote by $\bar{Q}_i = \sum_{k=i}^{n} q_k$ and $\bar{P}_i = \sum_{k=i}^{n} p_k$ the complementary cumulative distribution functions of q and p. Define the *forgery thresholds* ρ_i as

$$\rho_i = \begin{cases} P_i \frac{q_i}{p_i} - Q_i & , \quad i = 1, \ldots, j-1 \\ \frac{P_{j-1}}{\bar{P}_j}(\bar{Q}_j - \sigma) - Q_{j-1}, & i = j \\ \infty & , \quad i = j+1 \end{cases},$$

for $j = 2, \ldots, n$. Additionally, define the *suppression thresholds* σ_j as

$$\sigma_j = \bar{Q}_j - \bar{P}_j \frac{q_j}{p_j}$$

for $j = 1, \ldots, n$, and $\sigma_0 = 1$. Observe that $\rho_1 = \sigma_n = 0$ and that the forgery threshold ρ_j is a linear function of σ. We shall refer to this latter threshold as the *critical forgery-suppression threshold* and denote it also by $\rho_{\text{crit}}(\sigma)$. The reason is that said threshold will determine the boundary of the critical-privacy region, as we shall see later. The following result, Proposition 7.2, characterizes the monotonicity of the forgery and the suppression thresholds.

Proposition 7.2 (Monotonicity of Thresholds)**.**

(i) *For $j = 3, \ldots, n$ and $i = 1, \ldots, j-2$, the forgery thresholds satisfy $\rho_i \leqslant \rho_{i+1}$, with equality if, and only if, $\frac{q_i}{p_i} = \frac{q_{i+1}}{p_{i+1}}$.*

(ii) *For $j = 2, \ldots, n$, the suppression thresholds satisfy $\sigma_j \leqslant \sigma_{j-1}$, with equality if, and only if, $\frac{q_j}{p_j} = \frac{q_{j-1}}{p_{j-1}}$.*

(iii) *Further, for any $j = 2, \ldots, n$ and any $\sigma \in (\sigma_j, \sigma_{j-1}]$, the critical forgery-suppression threshold satisfies $\rho_j(\sigma) \geqslant \rho_{j-1}$, with equality if, and only if, $\sigma = \sigma_{j-1}$.*

Proof: The first statement can be shown from the definition of the forgery thresholds by routine algebraic manipulation and under the labeling assumption (7.3). To this end, it is helpful to note that

$$P_i \frac{q_{i+1}}{p_{i+1}} - Q_i = P_{i+1} \frac{q_{i+1}}{p_{i+1}} - Q_{i+1}.$$

The second statement can be shown analogously, observing that

$$\bar{Q}_j - \bar{P}_j \frac{q_{j-1}}{p_{j-1}} = \bar{Q}_{j-1} - \bar{P}_{j-1} \frac{q_{j-1}}{p_{j-1}}.$$

For the last statement, use the definitions of the forgery and the suppression thresholds to note that the condition $\rho_j(\sigma) \geqslant \rho_{j-1}$ is equivalent to $\sigma \leqslant \sigma_{j-1}$. ■

Prior to investigate a closed-form solution to the problem (7.1), we introduce some definitions for ease of presentation. For $i = 1, \ldots, j-1$ and $j = 2, \ldots, n$, define

$$\begin{aligned}
\tilde{q} &= \big(\, Q_i \, , \, q_{i+1} \, , \ldots , \, q_{j-1} \, , \, \bar{Q}_j \, \big), \\
\tilde{r} &= \big(\, \rho \, , \, \, 0 \, \, , \ldots , \, \, 0 \, \, , \, 0 \, \big), \\
\tilde{s} &= \big(\, 0 \, , \, \, 0 \, \, , \ldots , \, \, 0 \, \, , \, \sigma \, \big), \\
\tilde{p} &= \big(\, P_i \, , p_{i+1} \, , \ldots , \, p_{j-1} \, , \, \bar{P}_j \, \big),
\end{aligned}$$

where $\tilde{q}$ and $\tilde{p}$ are distributions in the probability simplex of $j - i + 1$ dimensions, and $\tilde{r}$ and $\tilde{s}$ are tuples of the same dimension that represent a forgery strategy and a suppression strategy, respectively. Particularly, note that the indexes $i = 1$ and $j = n$ lead to $\tilde{q} = q$ and $\tilde{p} = p$.

Theorem 7.3. *Let $\partial\mathscr{C}$ be the boundary of $\mathscr{C}$, and $\mathrm{cl}\,\bar{\mathscr{C}}$ the closure of $\bar{\mathscr{C}}$.*

(i) *$\partial\mathscr{C} \subset \mathscr{C}$ and*

$$\partial\mathscr{C} = \{(\rho,\sigma)\colon \rho = \rho_j(\sigma), \sigma \in [\sigma_j, \sigma_{j-1}], \text{ for } j = 2,\dots,n\}.$$

(ii) *For any $(\rho,\sigma) \in \mathrm{cl}\,\bar{\mathscr{C}}$, either $\rho \in [\rho_i, \rho_{i+1}]$ for $i = 1$ or $\rho \in (\rho_i, \rho_{i+1}]$ for some $i = 2,\dots,j-1$, and either $\sigma \in [\sigma_j, \sigma_{j-1}]$ for $j = n$ or $\sigma \in (\sigma_j, \sigma_{j-1}]$ for some $j = 2,\dots,n-1$. Then, for the corresponding indexes i, j, the optimal forgery and suppression strategies are*

$$r_k^* = \begin{cases} \frac{p_k}{P_i}(Q_i + \rho) - q_k\,, & k = 1,\dots,i \\ 0 \qquad\qquad\quad , & k = i+1,\dots,n \end{cases},$$

$$s_k^* = \begin{cases} 0 \qquad\qquad\quad , & k = 1,\dots,j-1 \\ q_k - \frac{p_k}{P_j}(\bar{Q}_j - \sigma), & k = j,\dots,n \end{cases},$$

and the corresponding, minimum KL divergence yields the privacy-forgery-suppression function

$$\mathcal{R}(\rho,\sigma) = \mathrm{D}\left(\frac{\tilde{q} + \tilde{r} - \tilde{s}}{1 + \rho - \sigma}\middle\|\, \tilde{p}\right).$$

Proof: The proof is structured as follows. We begin by showing that the optimization problem (7.1) may be construed as a particular case of that stated in Lemma 7.1. Accordingly, we apply this lemma, namely the cases (ii) and (iii), to obtain the optimal forgery and suppression strategies. The application of the former case allows us to derive the solution for $(\rho,\sigma) \in \bar{\mathscr{C}}$. The latter case enables us, first, to confirm that this solution is also valid on $\partial\bar{\mathscr{C}}$, and secondly, to prove statement (i). Lastly, we complete the proof of (ii) by expressing function (7.1) in terms of the optimal apparent distribution.

Use the definition of KL divergence to write the objective function of the optimization problem as $\mathrm{D}(t\,\|\,p) = \sum_k t_k \log \frac{t_k}{p_k}$, with $t = \frac{q+r-s}{1+\rho-\sigma}$. Observe that the functions $f_k(r_k, s_k) = t_k \log \frac{t_k}{p_k}$ are twice differentiable on $\{(r_k, s_k)\colon q_k + r_k - s_k > 0\}$. Denote

by h_k the derivative of f_k with respect to r_k,

$$h_k(r_k, s_k) = \frac{1}{1+\rho-\sigma}\left(\log\frac{q_k + r_k - s_k}{(1+\rho-\sigma)p_k} + 1\right). \tag{7.10}$$

Then, note that the functions f_k and h_k satisfy the assumptions of Lemma 7.1, and that the inequality and equality constraints of function (7.1) coincide with those in the lemma. This exposes the structure of the optimization problem as a special case of the resource allocation lemma.

Before proceeding any further, notice from (7.10) that $h_k(r_k, 0)$ is a strictly increasing function of r_k and hence invertible. Note also that, according to the lemma, the solutions are completely determined by the inverse of this function, which is denoted by h_k^{-1} and yields

$$h_k^{-1}(\phi) = p_k(1+\rho-\sigma)2^{(1+\rho-\sigma)\phi-1} - q_k.$$

Finally, observe that the assumption $h_1(0,0) \leqslant \cdots \leqslant h_n(0,0)$ in the lemma is equivalent to the labeling assumption (7.3), as $h_k(0,0)$ is a strictly increasing function of $\frac{q_k}{p_k}$.

Next we apply Lemma 7.1 (ii), where it is assumed the condition $\psi < \omega$. We start with case (ii) (a). On account of part (i) of the lemma, the optimal forgery strategy must satisfy

$$\rho = \sum_{k=1}^{i} h_k^{-1}(\psi) = P_i(1+\rho-\sigma)2^{(1+\rho-\sigma)\psi-1} - Q_i,$$

or equivalently,

$$\psi = \frac{1}{1+\rho-\sigma}\left(\log\frac{Q_i + \rho}{(1+\rho-\sigma)P_i} + 1\right).$$

Analogously for the suppression strategy,

$$\sigma = -\sum_{k=j}^{n} h_k^{-1}(\omega) = \bar{Q}_j - \bar{P}_j(1+\rho-\sigma)2^{(1+\rho-\sigma)\omega-1},$$

and therefore

$$\omega = \frac{1}{1+\rho-\sigma}\left(\log\frac{\bar{Q}_j-\sigma}{(1+\rho-\sigma)\bar{P}_j}+1\right).$$

Then it suffices to substitute the expressions of ψ and ω into the function h_k^{-1}, to obtain the nonzero optimal solutions claimed in assertion (ii) of the theorem.

Now we proceed to confirm the interval of values of ρ and σ where these solutions are defined. In the case under study, ψ and ω satisfy $h_i(0,0) < \psi \leqslant h_{i+1}(0,0)$ for some $i = 1,\ldots,j-1$ and $h_{j-1}(0,0) \leqslant \omega < h_j(0,0)$ for some $j = 2,\ldots,n$. We split the discussion into two cases, namely $i < j-1$ and $i = j-1$.

Assume the former case. Observe that the condition $h_i(0,0) < \psi$ is equivalent to

$$\frac{1}{1+\rho-\sigma}\left(\log\frac{q_i}{(1+\rho-\sigma)p_i}+1\right) < \frac{1}{1+\rho-\sigma}\left(\log\frac{Q_i+\rho}{(1+\rho-\sigma)P_i}+1\right)$$

and finally, after routine algebraic manipulation, to

$$\rho > P_i\frac{q_i}{p_i} - Q_i.$$

Similarly, the upper-bound condition $\psi \leqslant h_{i+1}(0,0)$ leads to

$$\rho \leqslant P_i\frac{q_{i+1}}{p_{i+1}} - Q_i.$$

Hence, the intervals resulting from imposing $h_i(0,0) < \psi \leqslant h_{i+1}(0,0)$ are of the form $(\rho_i, \rho_{i+1}]$. The monotonicity of the thresholds ρ_i, demonstrated in Proposition 7.2, guarantees that these intervals are contiguous and nonoverlapping. In an analogous manner, it can be shown that the condition $h_{j-1}(0,0) \leqslant \omega < h_j(0,0)$ leads to intervals of the form $(\sigma_j, \sigma_{j-1}]$, also contiguous and nonoverlapping by virtue of Proposition 7.2.

Now assume the latter case, where $h_i(0,0) < \psi < \omega < h_j(0,0)$ with $i = j-1$. On the one hand, the assumption $h_{j-1}(0,0) < \psi$ is, as shown above, equivalent to the condition $\rho > \rho_{j-1}$. On the other hand, straightforward manipulation allows us to write the inequality $\psi < \omega$ as

$$\rho < \frac{P_{j-1}}{\bar{P}_j}(\bar{Q}_j - \sigma) - Q_{j-1}.$$

Combining these two bounds on ψ, we obtain the interval $(\rho_{j-1}, \rho_{\text{crit}}(\sigma))$. With this last interval, we complete the range of validity of the solution for the case (ii) (a) in the lemma. Ultimately, it is easy to verify that, in those intervals of ρ and σ, the optimal apparent profile $t = \frac{q+r-s}{1+\rho-\sigma}$ does not coincide with the population's profile p. In consequence, $\mathrm{D}(t \,\|\, p) > 0$.

Next, we turn to case (ii) (b) of the lemma. Here, the assumption $h_n(0,0) \leqslant \omega$ leads to $\sigma = 0$, or equivalently, to the solution $s = 0$. Note that, precisely, this is the solution given in the theorem for $\sigma = \sigma_j$ with $j = n$. On the other hand, the application of the condition $\sum_{k=1}^{i} r_k = \rho$ results in the same optimal forgery strategy obtained in case (ii) (a). Proceeding analogously as in this case, from the assumptions on ψ we derive the intervals of values of ρ where the solution is defined: $(\rho_i, \rho_{i+1}]$ for $i = 1, \ldots, n-1$ and (ρ_i, ρ_{i+1}) for $i = n$. Given these intervals, it is then straightforward to check that $\mathcal{R}(\rho, 0) = 0$ if, and only if, $\rho \geqslant \rho_n$. This provides us with the pairs $(\rho, 0)$ that belong to $\mathrm{cl}\,\bar{\mathscr{C}}$.

In case (ii) (c), the condition $\psi \leqslant h_1(0,0)$ means that $\rho = 0$, or equivalently, $r = 0$. Observe that this is the solution stated in the theorem for $\rho = \rho_i$ with $i = 1$. Then again, the condition $\sum_{k=j}^{n} s_k = \sigma$ leads to the same optimal suppression strategy found in case (ii) (a). From the assumptions in the lemma on ω, we obtain the intervals $(\sigma_j, \sigma_{j-1}]$ for $j = 2, \ldots, n$ and (σ_j, σ_{j-1}) for $j = 1$. Then, we verify that $\mathcal{R}(0, \sigma) = 0$ if, and only if, $\sigma \geqslant \sigma_1$, from which it follows the pairs $(0, \sigma)$ that belong to $\mathrm{cl}\,\bar{\mathscr{C}}$.

Finally, the case (ii) (d) in the lemma, in which $h_n(0,0) \leqslant \omega$ and $\psi \leqslant h_1(0,0)$, corresponds to the trivial case $\sigma = \sigma_j$ for $j = n$ and $\rho = \rho_i$ for $i = 1$, that is, the solution $r = s = 0$.

After having applied Lemma 7.1 (ii) to function (7.1), now we proceed with case (iii) (a). In applying it, we shall show that the solution claimed in the theorem is also valid for the extreme values of the intervals in case (ii) (a), specifically the set

$$\{(\rho, \sigma) \colon \rho = \rho_{\text{crit}}(\sigma), \sigma \in (\sigma_j, \sigma_{j-1}] \text{ for } j = 3, \ldots, n, \text{ and } \sigma \in (\sigma_j, \sigma_{j-1}) \text{ for } j = 2\}.$$

Assume the case (iii) (a) in which $h_i(0,0) < \psi = \omega < h_j(0,0)$ for some $j = 2, \ldots, n$ and $i = j - 1$. Under this assumption, the equality constraint $\sum_{k=1}^{i} r_k = \rho$ in the lemma is equivalent, after simple algebraic manipulation, to

$$\psi = \frac{1}{1+\rho-\sigma}\left(\log \frac{Q_{j-1}+\rho-\zeta}{(1+\rho-\sigma)P_{j-1}} + 1\right), \tag{7.11}$$

where we define $\zeta = \sum_{k=1}^{n} \alpha_k$. Similarly, the equality constraint $\sum_{k=j}^{n} s_k = \sigma$ becomes

$$\omega = \frac{1}{1+\rho-\sigma}\left(\log \frac{\bar{Q}_j-\sigma+\zeta}{(1+\rho-\sigma)\bar{P}_j} + 1\right).$$

But $\psi = \omega$, therefore

$$\frac{Q_{j-1}+\rho-\zeta}{P_{j-1}} = \frac{\bar{Q}_j-\sigma+\zeta}{\bar{P}_j},$$

or equivalently,

$$\rho = \rho_{\text{crit}}(\sigma) + \frac{\zeta}{\bar{P}_j}.$$

In short, the assumption $\psi = \omega$ imposes the condition $(\rho, \sigma) \succeq (\rho_{\text{crit}}(\sigma), \sigma)$ for some nonnegative sequence $\alpha_1, \ldots, \alpha_n$ satisfying the above equality. Next we examine, for a given σ, these two possibilities, $\rho = \rho_{\text{crit}}(\sigma)$ and $\rho > \rho_{\text{crit}}(\sigma)$.

Consider the former possibility and observe that $\rho = \rho_{\text{crit}}(\sigma)$ if, and only if, $\alpha_k = 0$ for $k = 1, \ldots, n$. According to the lemma, the nonzero optimal solutions yield

$$\begin{aligned} r_k = h_k^{-1}(\psi) &= p_k \frac{Q_{j-1}+\rho_{\text{crit}}(\sigma)}{P_{j-1}} - q_k \\ &= p_k(1+\rho_{\text{crit}}(\sigma)-\sigma) - q_k \end{aligned}$$

for $k = 1, \ldots, j-1$, and

$$s_k = -h_k^{-1}(\psi) = q_k - p_k(1+\rho_{\text{crit}}(\sigma)-\sigma)$$

for $k = j, \ldots, n$, that is, the solutions obtained after applying case (ii) (a), but evaluated at $\rho = \rho_{\text{crit}}(\sigma)$. From these expression for r and s, it is immediate to verify then that $t = p$ and thus $\mathcal{R}(\rho, \sigma) = 0$.

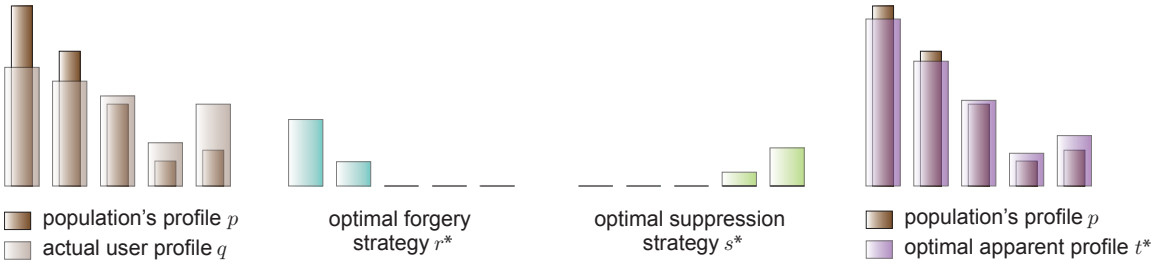


Figure 7.4: A user's item distribution is perturbed according to two optimal forgery and suppression strategies, in order for the resulting profile to minimize the KL divergence with respect to the population's distribution.

Now we assume the latter possibility, i.e., $(\rho, \sigma) \succ (\rho_{\text{crit}}(\sigma), \sigma)$, to show that the privacy-risk function also vanishes for these values of ρ and σ. On account of part (iii) (a) of the lemma and (7.11), we derive the optimal forgery and suppression strategies

$$r_k = p_k(1 + \rho_{\text{crit}}(\sigma) - \sigma) + \frac{p_k\,\zeta}{\bar{P}_j} - q_k + \alpha_k$$

and $s_k = \alpha_k$ for $k = 1, \ldots, j-1$, and

$$s_k = q_k - p_k(1 + \rho_{\text{crit}}(\sigma) - \sigma) - \frac{p_k\,\zeta}{\bar{P}_j} + \alpha_k$$

and $r_k = \alpha_k$ for $k = j, \ldots, n$. Then, we substitute r and s back into the apparent profile t and check that $\mathrm{D}(t \,\|\, p) = 0$. In doing so, we determine the pairs $(\rho, \sigma) \succ 0$ that belong to $\mathrm{cl}\,\bar{\mathscr{C}}$, and finally obtain the expression for the boundary of the critical-privacy region claimed in statement (i) of the theorem.

To conclude the proof, it remains only to write the privacy-risk function $\mathcal{R}(\rho, \sigma) = \sum_{k=1}^{n} t_k \log \frac{t_k}{p_k}$ in terms of the optimal apparent distribution. With this aim, we split the summation into three parts. The first part, corresponding to $t_k = \frac{p_k(Q_i+\rho)}{P_i(1+\rho-\sigma)}$, is

$$\sum_{k=1}^{i} t_k \log \frac{t_k}{p_k} = \frac{Q_i + \rho}{1 + \rho - \sigma} \log \frac{Q_i + \rho}{(1 + \rho - \sigma)P_i},$$

where we leverage on the fact that $\frac{t_k}{p_k}$ does not depend on k. The second part of the sum, corresponding to $t_k = \frac{q_k}{1+\rho-\sigma}$, yields

$$\sum_{k=i+1}^{j-1} t_k \log \frac{t_k}{p_k} = \sum_{k=i+1}^{j-1} \frac{q_k}{1+\rho-\sigma} \log \frac{q_k}{(1+\rho-\sigma)p_k}.$$

The last part, corresponding to $t_k = \frac{p_k(\bar{Q}_j-\sigma)}{\bar{P}_j(1+\rho-\sigma)}$, is

$$\sum_{k=j}^{n} t_k \log \frac{t_k}{p_k} = \frac{\bar{Q}_j-\sigma}{1+\rho-\sigma} \log \frac{\bar{Q}_j-\sigma}{(1+\rho-\sigma)\bar{P}_j},$$

where we also note that $\frac{t_k}{p_k}$ does not depend on k either. Now, it is straightforward to identify the terms of $\mathcal{R}(\rho,\sigma)$ as the KL divergence between the distributions

$$\left(\frac{Q_i+\rho}{1+\rho-\sigma}, \frac{q_{i+1}}{1+\rho-\sigma}, \ldots, \frac{q_{j-1}}{1+\rho-\sigma}, \frac{\bar{Q}_j-\sigma}{1+\rho-\sigma}\right)$$

and

$$\left(P_i, p_{i+1}, \ldots, p_{j-1}, \bar{P}_j\right),$$

precisely the distributions stated in the theorem. ■

In light of Theorem 7.3, we would like to remark the intuitive principle that both the optimal forgery and suppression strategies follow. On the one hand, the forgery strategy suggests adding ratings to those categories with a low ratio $\frac{q_k}{p_k}$, that is, to those in which the user's interest is considerably lower than the population's. On the other hand, the suppression strategy recommends eliminating ratings from those categories where the ratio $\frac{q_k}{p_k}$ is high, i.e., where the interest of the user exceeds that of the population. Further, we would like to highlight that the solution provided in the theorem is confined to the closure of the noncritical-privacy region. The reason is that the interior of the critical-privacy region is of no interest—the privacy risk attains its minimum value at the boundary of $\bar{\mathscr{C}}$ and therefore any $(\rho,\sigma) \succ (\rho_{\text{crit}}(\sigma),\sigma)$ cannot lower said privacy risk.

Another straightforward consequence of Theorem 7.3 is the role of the forgery and the suppression thresholds. In particular, we identify ρ_i as the forgery rate beyond which the components of r_k for $k = 1, \ldots, i$ become positive. A similar reasoning applies to σ_j, which indicates the suppression rate beyond which the components of s_k for $k = j, \ldots, n$ are positive. In a nutshell, these thresholds determine the number of nonzero components of the optimal strategies.

Also, from this theorem we deduce that the perturbation of the user profile does not *only* affect those categories where either $r_k > 0$ or $s_k > 0$. In fact, since we are dealing with relative frequencies, the components of the apparent distribution t_k belonging to the categories $k = i + 1, \ldots, j - 1$ are normalized by $\frac{1}{1+\rho-\sigma}$. Fig. 7.4 illustrates these three conclusions by means of a simple example with $n = 5$ categories of interest.

In this example we consider a user who is disposed to submit a percentage of false ratings $\rho \in (\rho_2, \rho_3]$, and to refrain from sending a fraction of genuine ratings $\sigma \in (\sigma_4, \sigma_3]$. Given these rates, the optimal forgery strategy recommends that the user forge ratings belonging to the categories 1 and 2, where clearly there is a lack of interest, compared to the reference distribution. On the contrary, the suppression strategy specifies that the user eliminate ratings from the categories 4 and 5, that is, from those categories where they show too much interest, again compared to the population's profile. In adopting these two strategies, the apparent user profile approaches the population's distribution, especially in those components where the ratio $\frac{q_k}{p_k}$ deviates significantly from 1. Finally, the component of the apparent profile t_3, which is not directly affected by the forgery and the suppression strategies, gets closer to p_3 as a result of the aforementioned normalization.

In the following subsections, we shall analyze a number of important consequences of Theorem 7.3.

7.6.2 Orthogonality, Continuity and Proportionality

In this subsection we study some interesting properties of the closed-form solution obtained in Sec. 7.6.1. Specifically, we investigate the orthogonality and continuity of the optimal forgery and suppression strategies, and then establish a proportionality

relationship between the optimal apparent user profile and the population's distribution.

Corollary 7.4 (Orthogonality and Continuity)**.**

(i) *For any* $(\rho, \sigma) \in \mathrm{cl}\,\bar{\mathscr{C}}$*, the optimal forgery and suppression strategies satisfy* $r_k^* \, s_k^* = 0$ *for* $k = 1, \ldots, n$*.*

(ii) *The components of* r^* *and* s^**, interpreted as functions of* ρ *and* σ *respectively, are continuous on* $\mathrm{cl}\,\bar{\mathscr{C}}$*.*

Proof: The proof of (i) is trivial from Theorem 7.3. To prove statement (ii) we also resort to this theorem. According to it, each component r_k^* may be regarded as a piecewise function of ρ defined on the contiguous, nonoverlapping intervals $[\rho_i, \rho_{i+1}]$ for $i = 1$ and $(\rho_i, \rho_{i+1}]$ for $i = 2, \ldots, j-1$. A direct verification shows that, for any $k = j, \ldots, n$, the component r_k^* is identically zero on the whole interval $[\rho_1, \rho_j]$ and hence continuous. For any $k = 1, \ldots, j-1$, we immediately check the continuity of r_k^* on the interior of each of the intervals parameterized by i. Now we examine the endpoints of such intervals. The continuity at the extreme points ρ_1 and ρ_j is verified straightforwardly as the intervals are closed at these points. Then, we check that the limit at the remaining endpoints ρ_i exists, since

$$\begin{aligned} \lim_{\rho \to \rho_i^-} r_k^*(\rho) &= \frac{p_k}{P_{i-1}}(Q_{i-1} + \rho_i) - q_k \\ &= \frac{p_k}{P_i}(Q_i + \rho_i) - q_k = \lim_{\rho \to \rho_i^+} r_k^*(\rho), \end{aligned}$$

for $i = 2, \ldots, j-1$. Because each limit coincides with the corresponding value $r_k^*(\rho_i)$, we prove the continuity of the components $r_1, \ldots, r_{j-1}$. The proof of the continuity of the components of s^* is analogous to that of r^*. ■

The orthogonality of the optimal forgery and suppression strategies, in the sense indicated by Corollary 7.4 (i), conforms to intuition—it would not make any sense to submit false ratings to items of a particular category and, at the same time, eliminate genuine ratings from this category. This intuitive result is illustrated in Fig. 7.4. The second part of Corollary 7.4 is applied to show our next result, Proposition 7.5.

Proposition 7.5 (Proportionality)**.** *Define the piecewise functions* $\phi(\rho,\sigma) = \frac{Q_i+\rho}{(1+\rho-\sigma)P_i}$ *and* $\chi(\rho,\sigma) = \frac{\bar{Q}_j-\sigma}{(1+\rho-\sigma)\bar{P}_j}$ *on the intervals* $[\sigma_j,\sigma_{j-1}]$ *for* $j=2,\dots,n$ *and* $[\rho_i,\rho_{i+1}]$ *for* $i=1,\dots,j-1$.

(i) *For any* $j=2,\dots,n$ *and* $i=1,\dots,j-1$, *and for any* $\sigma\in[\sigma_j,\sigma_{j-1}]$ *and* $\rho\in[\rho_i,\rho_{i+1}]$, *the optimal apparent profile* t^* *and the population's distribution* p *satisfy*

$$\frac{t_1^*}{p_1}=\cdots=\frac{t_i^*}{p_i}=\phi(\rho,\sigma),$$
$$\frac{t_j^*}{p_j}=\cdots=\frac{t_n^*}{p_n}=\chi(\rho,\sigma),$$

and

$$\phi(\rho,\sigma)\leqslant\frac{t_{i+1}^*}{p_{i+1}}\leqslant\cdots\leqslant\frac{t_{j-1}^*}{p_{j-1}}\leqslant\chi(\rho,\sigma).$$

(ii) *The function* ϕ *is continuous and strictly* increasing *in each of its arguments, and satisfies* $\phi(\rho,\sigma)\leqslant 1$, *with equality if, and only if,* $(\rho,\sigma)=(\rho_j(\sigma),\sigma)$.

(iii) *The function* χ *is continuous and strictly* decreasing *in each of its arguments, and satisfies* $\chi(\rho,\sigma)\geqslant 1$, *with equality if, and only if,* $(\rho,\sigma)=(\rho_j(\sigma),\sigma)$.

Proof: The continuity of the components of t^* on $\operatorname{cl}\bar{\mathscr{C}}$ follows from Corollary 7.4 (ii). This allows us to write the intervals in Theorem 7.3 as $[\rho_i,\rho_{i+1}]$ and $[\sigma_j,\sigma_{j-1}]$, in lieu of $(\rho_i,\rho_{i+1}]$ and $(\sigma_j,\sigma_{j-1}]$, respectively. From the expressions of r_k^* and s_k^* in the theorem, it is immediate to identify the ratios $\frac{t_k^*}{p_k}$ as either $\phi(\rho,\sigma)$ or $\chi(\rho,\sigma)$. The inner inequalities in statement (i) of this proposition also follow immediately from the labeling assumption (7.3). Direct manipulation shows that the outer inequalities $\frac{t_i^*}{p_i}\leqslant\frac{t_{i+1}^*}{p_{i+1}}$ and $\frac{t_{j-1}^*}{p_{j-1}}\leqslant\frac{t_j^*}{p_j}$ are equivalent to $\rho\leqslant\rho_{i+1}$ and $\sigma\leqslant\sigma_{j-1}$, respectively. This proves (i).

Next, we proceed to demonstrate the strict monotonicity of ϕ. A simple calculation shows that

$$\frac{\partial\phi}{\partial\rho}=\frac{\bar{Q}_{i+1}-\sigma}{(1+\rho-\sigma)^2P_i}.$$

To prove that $\frac{\partial\phi}{\partial\rho}>0$, it is sufficient to verify that $\bar{Q}_j>\sigma_{j-1}$, or equivalently, that $\bar{P}_j\frac{q_{j-1}}{p_{j-1}}>0$. Then, by the positivity assumption (7.2), we immediately see that this

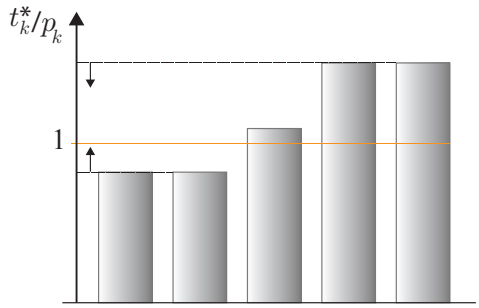


Figure 7.5: Proportionality relationship between the optimal user's apparent item distribution and the population's profile. In this figure we show the ratios $\frac{t_k^*}{p_k}$ of the example illustrated in Fig. 7.4, where the number of categories is $n = 5$, $\rho \in [\rho_2, \rho_3]$ and $\sigma \in [\sigma_4, \sigma_3]$.

latter inequality holds for any $j = 2, \ldots, n$. The strict monotonicity of ϕ in σ also follows from assumption (7.2).

To complete (ii), we write the condition $\phi(\rho, \sigma) \leqslant 1$ as

$$\rho \leqslant \frac{(1-\sigma)P_i - Q_i}{\bar{P}_{i+1}}.$$

A routine computation shows that the equality holds for $\rho_j(\sigma)$ and any $\sigma \in [\sigma_j, \sigma_{j-1}]$ with $j = 2, \ldots, n$. Therefore, for any fixed σ, the inequality holds strictly for any other ρ. The converse, that is, $\phi(\rho, \sigma) = 1$ implies $(\rho, \sigma) = (\rho_j(\sigma), \sigma)$, is immediate from the strict monotonicity of ϕ. The proof of statement (iii) proceeds along the same lines of that of (ii) and is omitted. ■

Our previous result tells us how perturbation operates. According to Proposition 7.5, the optimal strategies perturb the user profile in such a manner that, in those categories with the lowest and highest ratios $\frac{q_k}{p_k}$, the apparent profile becomes proportional to the population's distribution. More precisely, the common ratio $\frac{t_k^*}{p_k}$ increases with both ρ and σ in those categories affected by forgery, that is, $k = 1, \ldots, i$. Exactly the opposite happens in those categories affected by suppression, where the common ratio $\frac{t_j^*}{p_j}$ decreases with both rates. This tendency continues until $\rho = \rho_{\text{crit}}(\sigma)$, at which point $t^* = p$. Fig. 7.5 illustrates this proportionality property in the case of the example depicted in Fig. 7.4.

7.6.3 Critical-Privacy Region

One of the results of Theorem 7.3 is that the boundary of the critical-privacy region is determined by the critical forgery-suppression threshold $\rho_j(\sigma)$, which we also denote by $\rho_{\text{crit}}(\sigma)$ to highlight this fact. The following proposition leverages on this result and characterizes said region. In particular, Proposition 7.6 first examines some properties of this threshold and then investigates the convexity of the critical-privacy region.

Proposition 7.6 (Convexity of the Critical-Privacy Region)**.**

(i) *ρ_j is a convex, piecewise linear function of $\sigma \in [\sigma_j, \sigma_{j-1}]$ for $j = 2, \ldots, n$.*

(ii) *$\mathscr{C}$ is convex.*

Proof: From Theorem 7.3, it is routine to check the continuity of ρ_j on $[\sigma_n, \sigma_1]$. To show its convexity, we conveniently write this function as $\rho_j(\sigma) = m_j\,\sigma + b_j$, where $m_j = -\frac{P_{j-1}}{\bar{P}_j}$ and $b_j = \frac{P_{j-1} - Q_{j-1}}{\bar{P}_j}$. Next, we prove that the slopes satisfy $m_j < m_{j-1}$ for all $j = 3, \ldots, n$. We proceed by contradiction, assuming that $m_j \geqslant m_{j-1}$. Note that this inequality is equivalent to $P_{j-1}\bar{P}_{j-1} \leqslant \bar{P}_j - \bar{P}_j\bar{P}_{j-1}$ and, after algebraic simplification, to $p_{j-1} \leqslant 0$. This contradicts the positivity assumption (7.2), which, in turn, implies that $m_j < 0$ for all $j = 2, \ldots, n$. Therefore, since ρ_j is a piecewise linear function defined by the strictly increasing sequence of negative slopes $\{m_n, \ldots, m_2\}$, we can conclude that ρ_j is convex. This proves statement (i). The second statement follows from the first one. As ρ_j is convex, so is its epigraph, i.e., the critical-privacy region. ■

The conclusions drawn from Proposition 7.6 are illustrated in Fig. 7.6. In this figure we represent the critical and noncritical-privacy regions for $n = 5$ categories of interest; the distributions q and p assumed in this conceptual example are different from those considered in Figs. 7.4 and 7.5. That said, the figure in question shows a straightforward consequence of our previous proposition—the noncritical-privacy region is nonconvex.

In this illustrative example, the sequences of forgery thresholds $\{\rho_1 \ldots, \rho_5\}$ and suppression thresholds $\{\sigma_5, \ldots, \sigma_1\}$ are strictly increasing. By Proposition 7.2, we can conclude then that the inequalities of the labeling assumption (7.3) hold strictly. Related to these thresholds is also the number of nonzero components of the optimal

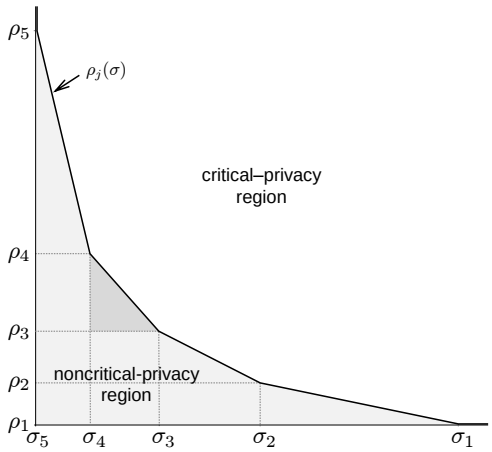


Figure 7.6: Conceptual plot of the critical and noncritical privacy regions for $n = 5$ categories.

strategies, as follows from Theorem 7.3. Fig. 7.6 shows the sets of pairs (ρ, σ) where the number of nonzero components of r^* and s^* is fixed. Thus, in the triangular area shown darker, corresponding to the Cartesian product of the intervals $[\rho_3, \rho_4]$ and $[\sigma_4, \sigma_3]$, the solutions r^* and s^* have $i = 3$ and $n - j + 1 = 2$ nonzero components, respectively.

7.6.4 Case of Low Forgery and Suppression

This subsection characterizes the privacy-forgery-suppression function in the special case when $\rho, \sigma \simeq 0$.

Proposition 7.7 (Low Rates of Forgery and Suppression)**.** *Assume the nontrivial case in which $q \neq p$. Then, there exist two indexes i, j such that $0 = \rho_1 = \cdots = \rho_i < \rho_{i+1}$ and $0 = \sigma_n = \cdots = \sigma_j < \sigma_{j-1}$. For any $\rho \in [0, \rho_{i+1}]$ and $\sigma \in [0, \sigma_{j-1}]$, the number of nonzero components of the optimal forgery and suppression strategies is i and $n - j + 1$, respectively. Further, the gradient of the privacy-forgery-suppression function at the origin is*

$$\nabla \mathcal{R}(0,0) = \begin{pmatrix} \frac{\partial \mathcal{R}(0,0)}{\partial \rho} \\ \frac{\partial \mathcal{R}(0,0)}{\partial \sigma} \end{pmatrix} = \begin{pmatrix} \log \frac{q_1}{p_1} - \mathrm{D}(q \,\|\, p) \\ \mathrm{D}(q \,\|\, p) - \log \frac{q_n}{p_n} \end{pmatrix}.$$

Proof: The existence of the indexes i and j is guaranteed by the assumption that $q \neq p$. The number of nonzero components of r^* and s^* is trivial from Theorem 7.3.

In view of this theorem, for any $\rho \in [0, \rho_{i+1}]$ and $\sigma \in [0, \sigma_{j-1}]$, we have

$$\mathcal{R}(\rho, \sigma) = \mathrm{D}\left(\frac{\tilde{q} + \rho(1, 0, \ldots, 0) - \sigma(0, \ldots, 0, 1)}{1 + \rho - \sigma} \middle\| \tilde{p} \right).$$

The continuity of the components of r^* and s^* proven in Corollary 7.4 (ii) ensures the continuity of the privacy-forgery-suppression function on $\bar{\mathscr{C}}$. It is routine to check its differentiability in this region and to obtain its derivative with respect to σ at the origin,

$$\frac{\partial \mathcal{R}(0,0)}{\partial \sigma} = Q_i \log \frac{Q_i \, \bar{P}_j}{P_i \, \bar{Q}_j} + \sum_{k=i+1}^{j-1} q_k \log \frac{\bar{P}_j \, q_k}{\bar{Q}_j \, p_k}.$$

On account of Proposition 7.2, the conditions $\rho_1 = \cdots = \rho_i$ and $\sigma_j = \cdots = \sigma_n$ imply

$$\frac{q_1}{p_1} = \cdots = \frac{q_i}{p_i} = \frac{Q_i}{P_i}$$

and

$$\frac{q_j}{p_j} = \cdots = \frac{q_n}{p_n} = \frac{\bar{Q}_j}{\bar{P}_j}.$$

Therefore,

$$\begin{aligned} \frac{\partial \mathcal{R}(0,0)}{\partial \sigma} &= \sum_{k=1}^{j-1} q_k \log \frac{q_k}{p_k} - Q_{j-1} \log \frac{q_n}{p_n} \\ &= \mathrm{D}(q \,\|\, p) - \log \frac{q_n}{p_n}. \end{aligned}$$

The derivative of $\mathcal{R}$ with respect to ρ at $\rho = \sigma = 0$ follows analogously. ■

Next, we shall derive an expression for the relative decrement of the privacy-risk function at $\rho, \sigma \simeq 0$. To this end, define the *forgery relative decrement factor*

$$\delta_\rho = -\frac{\frac{\partial \mathcal{R}(0,0)}{\partial \rho}}{\mathcal{R}(0,0)} = 1 - \frac{\log \frac{q_1}{p_1}}{\mathrm{D}(q \,\|\, p)},$$

and the *suppression relative decrement factor*

$$\delta_\sigma = -\frac{\frac{\partial \mathcal{R}(0,0)}{\partial \sigma}}{\mathcal{R}(0,0)} = \frac{\log \frac{q_n}{p_n}}{\mathrm{D}(q \,\|\, p)} - 1.$$

By dint of Proposition 7.7, the first-order Taylor approximation of function (7.1) around $\rho = \sigma = 0$ yields

$$\mathcal{R}(\rho, \sigma) \simeq \mathrm{D}(q \,\|\, p) + \rho \left(\log \frac{q_1}{p_1} - \mathrm{D}(q \,\|\, p) \right) + \sigma \left(\mathrm{D}(q \,\|\, p) - \log \frac{q_n}{p_n} \right),$$

or more compactly, in terms of the decrement factors,

$$\frac{\mathrm{D}(q \,\|\, p) - \mathcal{R}(\rho, \sigma)}{\mathrm{D}(q \,\|\, p)} \simeq \delta_\rho \, \rho + \delta_\sigma \, \sigma.$$

In words, the minimum and maximum ratios $\frac{q_k}{p_k}$ characterize the relative reduction in privacy risk. The following result, Proposition 7.8, establishes a bound on these relative decrement factors.

Proposition 7.8 (Relative Decrement Factors)**.** *In the nontrivial case when $q \neq p$, the relative decrement factors satisfy $\delta_\rho > 1$ and $\delta_\sigma > 0$.*

Proof: Observe that the statement $\delta_\rho > 1$ is equivalent to the condition $q_1 < p_1$. We prove this by contradiction. Suppose that $q_1 > p_1$. By the labeling assumption (7.3), it follows that $q_k > p_k$ for all k, which leads to the contradiction that $1 = \sum q_k > \sum p_k = 1$. Now assume that $q_1 = p_1$. Since $q \neq p$, there must exist an index i such that

$$\frac{q_1}{p_1} = \cdots = \frac{q_{i-1}}{p_{i-1}} < \frac{q_i}{p_i} \leqslant \cdots \leqslant \frac{q_n}{p_n}.$$

But this implies that

$$1 - \sum_{k=1}^{i-1} q_k = \sum_{k=i}^{n} q_k > \sum_{k=i}^{n} p_k = 1 - \sum_{k=1}^{i-1} q_k,$$

a contradiction. This proves the first part of the proposition.

For the second part, note that the statement $\delta_\sigma > 0$ is equivalent to

$$q_1 \log \frac{q_1}{p_1} + \cdots + q_n \log \frac{q_n}{p_n} < \log \frac{q_n}{p_n},$$

and, after algebraic manipulation, to

$$q_1 \log \frac{q_1}{p_1}\frac{p_n}{q_n} + \cdots + q_{n-1} \log \frac{q_{n-1}}{p_{n-1}}\frac{p_n}{q_n} < 0.$$

The positivity and labeling assumptions (7.2), (7.3) ensure that all terms in the sum are nonpositive. However, the additional assumption $q \neq p$ implies that $\frac{q_1}{p_1} < \frac{q_n}{p_n}$, which in turn implies that the first term is negative and so is, consequently, the entire summation. ■

Conceptually, the bound on δ_ρ tells us that the relative decrement in privacy risk is greater than the forgery rate introduced. This is under the assumption that $q \neq p$ and at low rates of forgery and suppression. The bound on δ_σ, however, is looser than the previous one and just ensures that an increase in the suppression rate always leads to a decrease in privacy risk, as one would expect.

7.6.5 Pure Strategies

In the previous subsections we investigated the forgery and the suppression of ratings as a *mixed* strategy that users may adopt to enhance their privacy. In this subsection we contemplate the case in which users may be reluctant to use these two mechanisms in conjunction; and as a consequence, they may opt for a *pure* strategy consisting in the application of either forgery or suppression. In this case, it would be useful to determine which is the most appropriate technique in terms of the privacy-utility trade-off posed. Our next result, Corollary 7.9, provides some insight on this, under the assumption that, from the user's perspective, the impact on utility due to forgery is equivalent to that caused by the effect of suppression.

Before showing this result, observe from Theorem 7.3 that $\rho_n = \frac{q_n}{p_n} - 1$ is the minimum forgery rate such that $\mathcal{R}(\rho, 0) = 0$. Analogously, $\sigma_1 = 1 - \frac{q_1}{p_1}$ is the minimum suppression rate satisfying $\mathcal{R}(0, \sigma) = 0$. In other words, ρ_n and σ_1 are the *critical rates* of the pure forgery and suppression strategies, respectively. Further, note that $\sigma_1 < \sigma_0 = 1$, on account of the positivity assumption (7.2). However, $\rho_n > 1$ if, and only if, $\frac{q_n}{p_n} > 2$.

Corollary 7.9 (Pure Strategies)**.** *Consider the nontrivial case when $q \neq p$.*

(i) *The critical rates of the pure forgery and suppression strategies satisfy* $\rho_n < \sigma_1$ *if, and only if,*

$$\frac{q_1/p_1 + q_n/p_n}{2} < 1.$$

(ii) *The forgery and the suppression relative decrement factors satisfy* $\delta_\rho > \delta_\sigma$ *if, and only if,*

$$\sqrt{\frac{q_1}{p_1}\,\frac{q_n}{p_n}} < 2^{\mathrm{D}(q\,\|\,p)}.$$

Proof: Both statements are immediate from the definitions of ρ_n and σ_1 on the one hand, and δ_ρ and δ_σ on the other. ■

In conceptual terms, the condition $\rho_n < \sigma_1$ means that the pure forgery strategy is the most appropriate mechanism in terms of causing the minimum distortion to attain the critical-privacy region. On the other hand, the condition $\delta_\rho > \delta_\sigma$ implies that, at low rates, the pure forgery strategy offers better privacy protection than the pure suppression strategy does. Therefore, the conclusion that follows from Corollary 7.9 is that, together with the quantity $\mathrm{D}(q\,\|\,p)$, the arithmetic and geometric mean of the ratios $\frac{q_1}{p_1}$ and $\frac{q_n}{p_n}$ determine which strategy to choose.

Another interesting remark is the duality of these two ratios $\frac{q_1}{p_1}$ and $\frac{q_n}{p_n}$. The former characterizes the minimum rate for the pure suppression strategy to reach the critical-privacy region and, at the same time, it establishes the privacy gain at low forgery rates. Conversely, the latter ratio defines the critical rate of the pure forgery strategy and determines the relative decrement in privacy risk at low suppression rates.

Lastly, we would like to establish a connection between our work and that of [95], where the *pure* forgery strategy is investigated in the context of information retrieval. In the cited work, the optimal trade-off between privacy risk and query redundancy is modeled by the function

$$\mathcal{R}(\rho') = \min_{r'} \mathrm{D}((1-\rho')q + \rho'\, r' \,\|\, p),$$

where ρ' is the ratio of forged queries to *total* number of queries, and r' is the distribution of the user's forged queries. Accordingly, it can be shown that $\rho' = \frac{\rho}{1+\rho}$ and that $\mathcal{R}(\rho, 0) = \frac{1}{\ln 2}\mathcal{R}(\rho')$. Similarly, we may formulate the problem of optimal

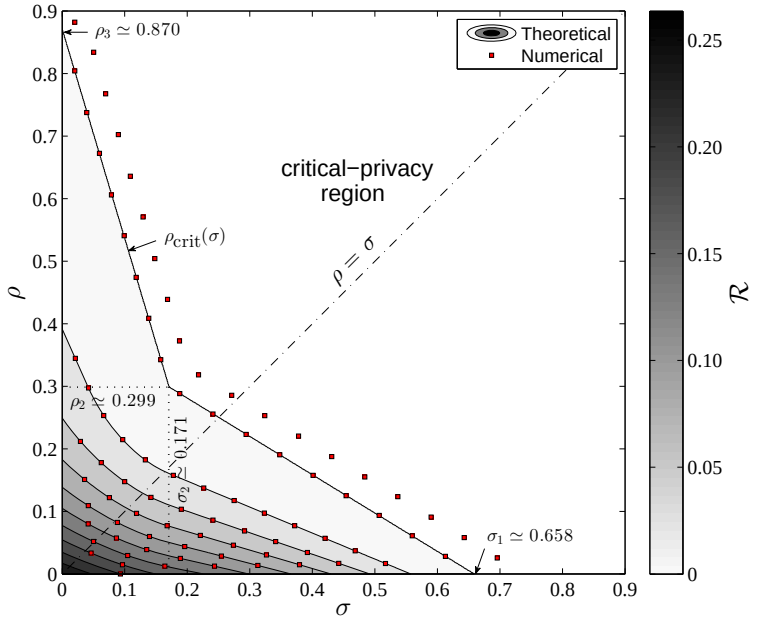


Figure 7.7: Contour lines of the privacy-forgery-suppression function, the corresponding forgery and suppression thresholds, and the critical and noncritical privacy regions.

tag suppression as a particular case of the optimization problem investigated in this chapter. Under the assumption that the population's profile is uniform, it can be proven that $\mathcal{R}(0,\sigma) = \log n - \frac{1}{\ln 2}\mathcal{P}(\sigma)$. In short, our formulation of the problem of optimal forgery and suppression of ratings encompasses, as particular cases, the pure forgery case of [95] and the pure suppression problem examined in Chapter 5.

7.6.6 Numerical Example

This subsection presents a numerical example that illustrates the theoretical analysis conducted in the previous subsections. Later in Sec. 7.7 we shall evaluate the effectiveness of our approach in a real scenario, namely in the movie recommendation system Movielens.

In this example we assume $n = 3$ categories of interests. Although the example shown here is synthetic, these three categories could very well represent interests across topics such as technology, sports and beauty. Accordingly, we suppose that

the user's item distribution is

$$q = (0.130, 0.440, 0.430),$$

and the population's,

$$p = (0.380, 0.390, 0.230).$$

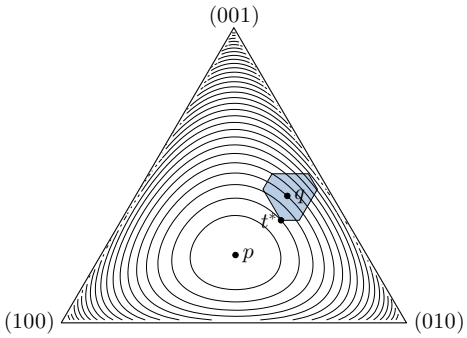
Note that these distributions satisfy the positivity and labeling assumptions (7.2), (7.3).

From Sec. 7.6.1, we easily obtain the forgery thresholds $\rho_1 = 0$, $\rho_2 \simeq 0.299$ and $\rho_3 \simeq 0.870$ on the one hand, and on the other the suppression thresholds $\sigma_3 = 0$, $\sigma_2 \simeq 0.171$ and $\sigma_1 \simeq 0.658$. The thresholds ρ_3 and σ_1 are the critical rates of the pure strategies. If we are to reach the critical-privacy region and do not have any preference for either forgery or suppression, the fact that $\rho_3 > \sigma_1$ leads us to opt for suppression as pure strategy. However, the geometric mean of $\frac{q_1}{p_1}$ and $\frac{q_3}{p_3}$ is approximately 0.799, which is lower than $2^{\mathrm{D}(q\,\|\,p)} \simeq 1.20$. On account of Corollary 7.9, this means that the pure forgery strategy contributes to a greater reduction in privacy risk at low rates than suppression does. In fact, the gradient of the privacy-forgery-suppression function at the origin is $\nabla\mathcal{R}(0,0)^{\mathrm{T}} \simeq (-1.81, -0.639)$, by virtue of Proposition 7.7.

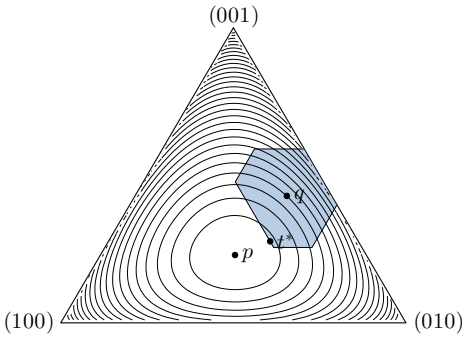
Fig. 7.7 shows the contour lines of this function, computed analytically from Theorem 7.3 and numerically [(e)]. The region plotted in gray shades corresponds to the noncritical-privacy region $\bar{\mathscr{C}}$. The initial privacy risk is $\mathcal{R}(0,0) \simeq 0.263$. The white area represents the critical-privacy region $\mathscr{C}$, where the apparent user profile coincides with the population's distribution and thus the privacy risk vanishes. In accordance with Proposition 7.6, we observe that the critical forgery-suppression threshold $\rho_{\mathrm{crit}}(\sigma)$ is convex and so is $\mathscr{C}$.

Another interesting observation arising from Fig. 7.7 is the synergistic effect of combining forgery and suppression. Just as an example, in the case when $\rho = \rho_2$ and $\sigma = \sigma_2$, we note that $\mathcal{R}(\rho, \sigma)$ is lower than $\mathcal{R}(\rho + \sigma, 0)$ and $\mathcal{R}(0, \rho + \sigma)$. Put differently, forgery and suppression provide better privacy for the same total rate than just forgery or suppression alone. This is true for this particular example, but it is

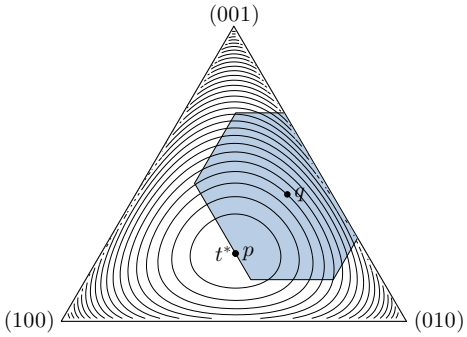
[(e)] The numerical method chosen is the interior-point algorithm [75, 194–196] implemented by the Matlab R2012b function `fmincon`.



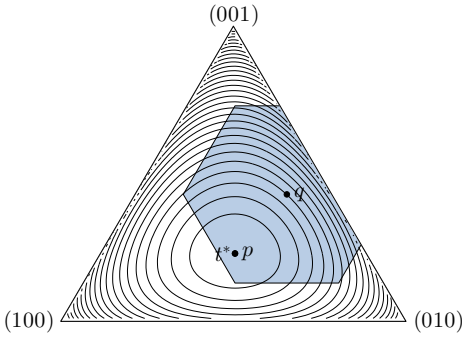
(a) $\rho = 0.050$, $\sigma = 0.100$, $\rho/\rho_{\text{crit}}(\sigma) \simeq 0.093$, $\mathcal{R}(\rho,\sigma) \simeq 0.131$, $\mathcal{R}(\rho,\sigma)/\mathcal{R}_0 \simeq 0.498$, $r^* = (0.050, 0, 0)$, $s^* = (0, 0, 0.100)$, $t^* \simeq (0.189, 0.463, 0.347)$.



(b) $\rho = 0.100$, $\sigma = 0.200$, $\rho/\rho_{\text{crit}}(\sigma) \simeq 0.356$, $\mathcal{R}(\rho,\sigma) \simeq 0.050$, $\mathcal{R}(\rho,\sigma)/\mathcal{R}_0 \simeq 0.190$, $r^* = (0.100, 0, 0)$, $s^* \simeq (0, 0.019, 0.181)$, $t^* \simeq (0.256, 0.468, 0.276)$.



(c) $\rho \simeq 0.219$, $\sigma = 0.300$, $\rho/\rho_{\text{crit}}(\sigma) = 1$, $\mathcal{R}(\rho,\sigma) = 0$, $\mathcal{R}(\rho,\sigma)/\mathcal{R}_0 = 0$, $r^* \simeq (0.219, 0, 0)$, $s^* \simeq (0, 0.081, 0.219)$, $t^* = p$.



(d) $\rho = 0.300$, $\sigma = 0.300$, $\rho/\rho_{\text{crit}}(\sigma) \simeq 1.368$, $\mathcal{R}(\rho,\sigma) = 0$, $\mathcal{R}(\rho,\sigma)/\mathcal{R}_0 = 0$, $r^* \simeq (0.260, 0.021, 0.019)$, $s^* = (0.010, 0.071, 0.219)$, $t^* = p$.

Figure 7.8: Probability simplices showing, for several interesting values of ρ and σ, the user's actual profile $q = (0.130, 0.440, 0.430)$, the population's distribution $p = (0.380, 0.390, 0.230)$, the optimal apparent distribution t^* and the set of feasible apparent distributions.

not a general rule. What is always true, however, is that the mixed strategy cannot be worse than the pure strategies. This is because the feasible set of the problem minimizing $\mathcal{R}(\rho,\sigma)$ subject to the constraint $\rho + \sigma = \tau$ includes the extreme values $\rho = \tau$ and $\sigma = \tau$, that is, the cases corresponding to the pure strategies.

Next, we examine the optimal apparent item distribution for different values of ρ and σ. For this purpose, the user's genuine distribution q, the population's distribution p and the optimal apparent distribution t^* are depicted in the probability

simplices shown in Fig. 7.8. In each simplex, we also represent the contour lines of the KL divergence $\mathrm{D}(\cdot \,\|\, p)$ between every distribution in the simplex and p. Further, we plot the set of feasible apparent user distributions, not necessarily optimal, for four different combinations of ρ and σ; in any of these cases, the set takes the form of a hexagon. Having said this, now we turn our attention to Fig. 7.8(a). In this case, the optimal forgery and suppression strategies have $i = n - j + 1 = 1$ nonzero component, since $\rho \in [0, \rho_2]$ and $\sigma \in [0, \sigma_2]$. This places the solution t^* at one vertex of the hexagon. A remarkable fact is that, for these rates, the privacy risk is approximately halved. In the end, consistently with Proposition 7.8, the forgery and the suppression relative decrement factors are $\delta_\rho \simeq 6.87 > 1$ and $\delta_\sigma \simeq 2.42 > 0$.

In the case shown in Fig. 7.8(b), r^* still has $i = 1$ nonzero components, while s^* contains $n - j + 1 = 2$ nonzero components. Geometrically, the optimal apparent distribution lies at one edge of the feasible region. This lowers privacy risk to a 19% of its initial value. The case in which $(\rho, \sigma) = (\rho_{\mathrm{crit}}(\sigma), \sigma)$ is depicted in Fig. 7.8(c). Here, the number of nonzero components of r^* and s^* remains the same as in the previous case, but the privacy risk becomes zero. The last case, illustrated in Fig. 7.8(d), does not have any practical application, as $\mathcal{R}(\rho, \sigma) = 0$ for any $(\rho, \sigma) \in \partial\mathscr{C}$. In this figure we can observe that the solution t^* is placed in the interior of the hexagon, and that the orthogonality principle of the strategies r^* and s^* stated in Corollary 7.4 is not satisfied.

7.7 Experimental Analysis

In this section we evaluate the extent to which the forgery and the suppression of ratings could enhance user privacy in a real-world personalized recommendation system. The system chosen to conduct this evaluation is Movielens, a popular movie recommender developed by the GroupLens Research Lab [197] at the University of Minnesota. As many other recommenders, Movielens allows users to both rate and tag movies according to their preferences. These preferences are then exploited by the recommender to suggest movies that users have not watched yet. The algorithms

used by Movielens to generate these recommendations are based on collaborative-filtering techniques.

7.7.1 Data set

The data set that we used to assess our data-perturbative mechanism is the *Movielens 10M* data set [198], which contains 10 000 054 ratings and 95 580 tags. The ratings and tags included in this data set were assigned to 10 681 movies by 71 567 users. In our series of experiments we only contemplate the ratings posted by users, i.e., tags are not taken into account in the characterization of users' interests. This is because it simplifies the modeling of user profiles, and because this chapter concentrates on the perturbation of ratings, rather than tags; a more sophisticated model of user profile would undoubtedly enrich this ratings-based profile with such semantic annotations.

The data set in question is organized in the form of quadruples (*username*, *movie*, *rating*, *time*), each one representing the action of a user rating a movie at a certain time. In fact, [197] replaced usernames were with numbers in an attempt to anonymize the data set. This is similar to the way user identifiers were processed in the data set used in Chapter 5. In that case, usernames were anonymized by applying a hash function. We would like to stress that such anonymization may not be sufficient to guarantee user privacy. As mentioned in Sec. 7.1, inferences about a user's movie rating history may be far more conclusive when cross-referencing several pieces of data from multiple sources [64].

For our purposes of experimentation, we just needed the data fields *username* and *movie*, together with the categories each movie belongs to. Movielens contemplates $n = 19$ categories or movies genres, listed in alphabetical order as follows: *action*, *adventure*, *animation*, *children's*, *comedy*, *crime*, *documentary*, *drama*, *fantasy*, *film-noir*, *horror*, *IMAX*, *musical*, *mystery*, *romance*, *sci-fi*, *thriller*, *war* and *western*. As we shall see later in Sec. 7.7.2, for each particular user, we shall have to rearrange those categories in such a way that the labeling assumption (7.3) is satisfied.

In our data set, all users rated, at least, 20 movies. This was the minimum number of ratings for the recommender to start working [(f)]. After the elimination of those users who exclusively tagged movies, the total number of users reduced to 69 878. Then, we used simple random sampling to select 10% of this group of users. After that sampling, we found that only 4 099 of these users satisfied the positivity assumption (7.2). Since the resulting group represents a relatively small fraction of the total number of users, we can assume that the application of our technique will have a negligible effect on the population's profile p, as supposed in Sec. 7.5.

7.7.2 Results

In this subsection we examine how the forgery and the suppression of ratings may help users of Movielens to enhance their privacy. With this aim, first, we analyze the effect of the perturbation of ratings on the privacy protection of a particular user from our data set. Secondly, we consider the entire set of 4 099 users and assess the relative reduction in privacy risk when these users apply the same forgery and suppression rates. Lastly, we investigate the forgery and the suppression strategies separately, and draw some conclusions about these two pure strategies.

To conduct our first experiments, we choose a particular user from our data set [(g)]. Before perturbing the movie rating history of this user, it is necessary that the components of the user's profile q and the population's distribution p be rearranged to satisfy the labeling assumption (7.3). Table 7.2 shows how the movie categories have been sorted and then indexed from 1 to n, to fulfill the assumption above. We would like to note that the index provided in this table does not have to coincide with the index of other users in our data set.

Fig. 7.9(a) depicts the user profile and the population profile, the latter being computed by averaging across the 69 878 users. From this figure we note that the user's interest far exceeds the population's in categories such as *musical*, *romance*,

[(f)] Nowadays, the algorithm implemented by Movielens requires only 15 ratings to start generating predictions.

[(g)] The user considered in this first series of experiments is identified by the number `3301` in [198].

Table 7.2: Category index of the particular user examined in our experiments. The categories of Movielens have been sorted and indexed in order to satisfy the labeling assumption (7.3).

Index	Category name	Index	Category name	Index	Category name
1	animation	7	sci-fi	13	war
2	action	8	comedy	14	mystery
3	film-noir	9	thriller	15	musical
4	children's	10	fantasy	16	romance
5	adventure	11	horror	17	IMAX
6	crime	12	western	18	drama
				19	documentary

IMAX, *drama* and *documentary*. More precisely, the ratios $\frac{q_k}{p_k}$ yield

$$\left(\frac{q_k}{p_k}\right)_{k=15,\ldots,19} \simeq (1.300, 1.306, 1.451, 1.728, 2.292).$$

In this figure, we also observe that the user's interest and the population's in the category 17 are nearly zero, namely $q_{17} \simeq 0.0005$ and $p_{17} \simeq 0.0003$.

On the other hand, Fig. 7.9(a) indicates that the user shows little interest, compared to the population's preferences, in categories such as *animation*, *action*, *film-noir* or *children's*, to name just a few. Specifically, the first six smallest ratios $\frac{q_k}{p_k}$ yield

$$\left(\frac{q_k}{p_k}\right)_{k=1,\ldots,6} \simeq (0.444, 0.599, 0.651, 0.691, 0.705, 0.714).$$

Figs. 7.9(b) and 7.9(c) show the optimal forgery and suppression strategies that this particular user should apply, in the case when $\sigma = 0.150$ and $\rho_{\text{crit}}(\sigma) \simeq 0.180$. The solutions plotted in these figures are consistent with our two previous observations: the optimal forgery strategy recommends that the user submit false ratings to movies falling into the categories where the ratio $\frac{q_k}{p_k}$ is low; and the optimal suppression strategy suggests that the user refrain from rating movies belonging to categories where the ratio $\frac{q_k}{p_k}$ is high. Just as an example, the fact that $s^*_{17} \simeq 0.0001$ means that the user at hand should eliminate one in five ratings to movies classified as *IMAX*.

The optimal trade-off surface among privacy, forgery rate and suppression rate is represented in Fig. 7.10. In this figure we plot the contour levels of the function $\mathcal{R}(\rho, \sigma)$, which we computed theoretically. The initial privacy risk is $\mathcal{R}(0, 0) \simeq 0.101$

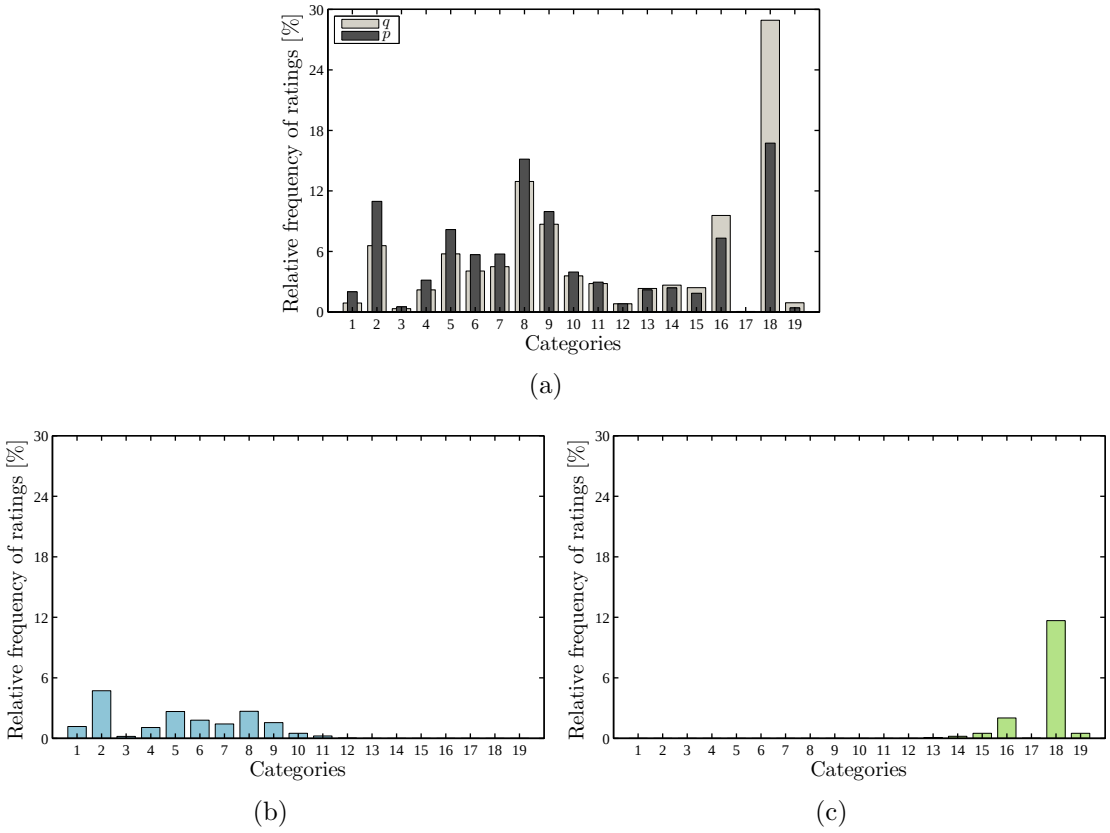


Figure 7.9: In this figure we represent (a) the item distribution q of a particular user as well as the population's item distribution p. In addition, we plot (b) the optimal forgery strategy r^* and (c) the optimal suppression strategy s^* that the user in question should adopt when they specify $\sigma = 0.150$ and $\rho = \rho_{\text{crit}}(\sigma) \simeq 0.180$.

and the arithmetic mean between the ratios $\frac{q_1}{p_1}$ and $\frac{q_{19}}{p_{19}}$ yields approximately 1.37. Since the mean is higher than 1, Corollary 7.9 tells us that the user should opt for suppression as pure strategy, in lieu of forgery. This is under the assumption that they wish to achieve the minimum privacy risk and do not have any preference for any of the pure strategies. Nevertheless, the fact that $\delta_\rho \simeq 12.6 > \delta_\sigma \simeq 10.9$ leads us to choose forgery as pure strategy for $\rho, \sigma \simeq 0$. When both strategies are combined, we note that a forgery and suppression rate of just 0.1% leads to a reduction in privacy risk of 2.35%, on account of the first-order Taylor approximation derived in Sec. 7.6.4.

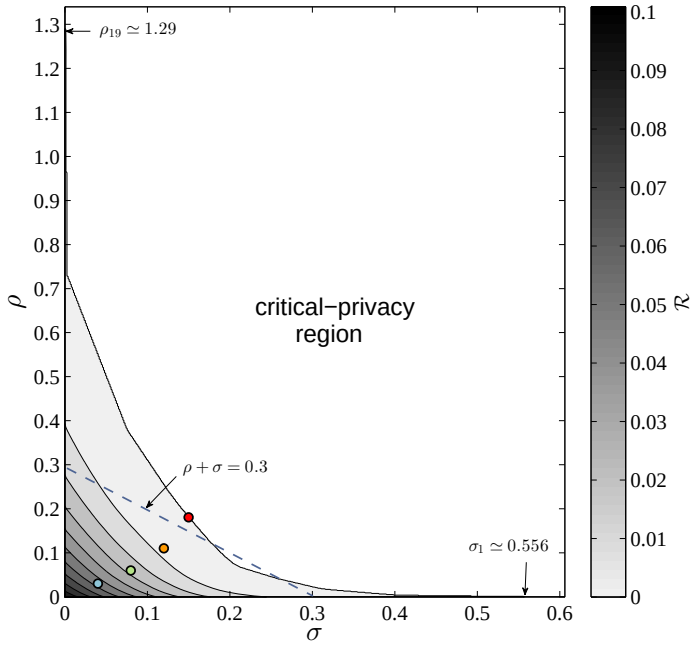


Figure 7.10: Optimal trade-off surface among privacy, forgery rate and suppression rate for one particular user of Movielens. The four points shown in this figure correspond to the pairs of values (ρ, σ) that we used to show the proportionality relationship between t^* and p in Fig. 7.11.

As in Sec. 7.6.6, we also observe the synergistic effect of forgery and suppression. For example, for a total rate $\tau = \rho + \sigma = 0.3$, the pure forgery and the pure suppression strategies reduce privacy risk by $\mathcal{R}(\tau, 0)/\mathcal{R}_0 \simeq 0.166$ and $\mathcal{R}(0, \tau)/\mathcal{R}_0 \simeq 0.048$, respectively, whereas the optimal strategy for simultaneous forgery and suppression removes any privacy risk. In Fig. 7.10, we depict the set of pairs (ρ, σ) such that $\rho + \sigma = 0.3$ and note that, for example, for $(\rho, \sigma) = (0.1, 0.2)$, the critical-privacy region is attained. Simply put, the combined use of these two perturbation techniques may result in a synergy that can help users protect their privacy more efficiently.

In Fig. 7.10 we have also plotted 4 points, which correspond to the following pairs of values (ρ, σ): $(0.03, 0.04)$, $(0.06, 0.08)$, $(0.11, 0.12)$ and $(0.18, 0.15)$. For each of these pairs, we have represented the quotient $\frac{t_k^*}{p_k}$ in Fig. 7.11. The aim is to show how the optimal apparent profile becomes proportional to the population's distribution, as the user approaches the critical-privacy region. Fig. 7.11(a) considers the first pair of values. Here, ρ and σ fall into the intervals $[\rho_6, \rho_7]$ and $[\sigma_{18}, \sigma_{17}]$, respectively.

Consistently with Proposition 7.5, we check that $\frac{t_1^*}{p_1} = \cdots = \frac{t_6^*}{p_6} \simeq 0.756 \leqslant 1$ and that $\frac{t_{18}^*}{p_{18}} = \frac{t_{19}^*}{p_{19}} \simeq 1.52 \geqslant 1$.

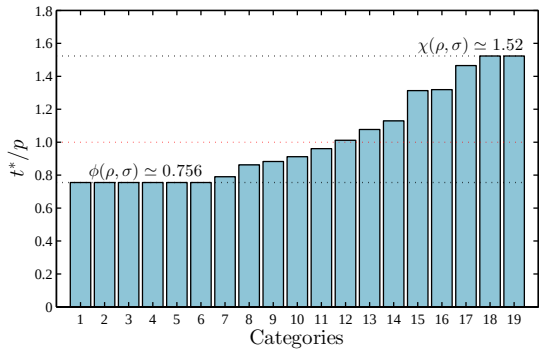
In Fig. 7.11(b) we double the rates of forgery and suppression. On the one hand, this leads to $\frac{t_1^*}{p_1} = \cdots = \frac{t_7^*}{p_7}$. On the other, the fact that $\sigma \in [\sigma_{15}, \sigma_{14}]$ implies that $\frac{t_{15}^*}{p_{15}} = \cdots = \frac{t_{19}^*}{p_{19}}$. It is also interesting to note that, for these relatively small values of ρ and σ, the final privacy risk is 26% of the initial value $\mathrm{D}(q \,\|\, p)$.

As ρ and σ increase, so does the function ϕ. The contrary happens with the function χ, which decreases with both rates. In Fig. 7.11(c), for example, the proportionality relationship between t^* and p holds for all except 4 categories. The last pair $(\rho, \sigma) \simeq (0.18, 0.15)$ lies at the boundary of $\mathscr{C}$, as shown in Fig. 7.10. This implies that $\frac{t^*}{p} = 1$ and therefore that $\mathcal{R}(\rho, \sigma) = 0$, as captured in Fig. 7.11(d).

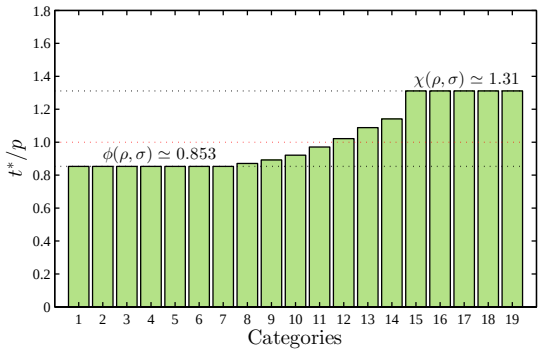
Having examined the case of a specific user, in our next series of experiments we evaluate the level of privacy protection that users can achieve if they are disposed to forge and eliminate a fraction of their ratings. For simplicity, we suppose that all users satisfying the positivity assumption (7.2) apply a common forgery rate and a common suppression rate. Fig. 7.12 depicts the contours of the 10^{th}, 50^{th} and 90^{th} percentile surfaces of relative reduction in privacy risk, for different values of ρ and σ. Two conclusions can be drawn from this figure.

- First, for relatively small values of ρ and σ (lower than 15%), a vast majority of users lowered privacy risk significantly. In quantitative terms, we observe in Fig. 7.12(a) that, for $\rho = \sigma = 0.05$, the 90% of users adhered to our technique obtained a reduction in privacy risk greater than 52.4%. For those same rates of forgery and suppression, the 50^{th} and 90^{th} percentiles are 73.9% and 94.8%. For higher rates, e.g., $\rho = \sigma = 0.13$, Fig. 7.12(b) shows that half of users experienced a reduction in privacy risk equal to 100%.

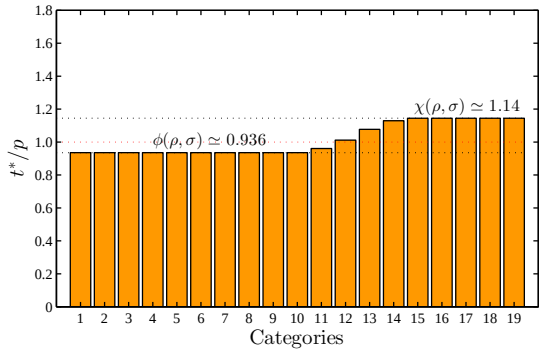
- Secondly, the three percentile surfaces exhibit a certain symmetry with respect to the line $\rho = \sigma$. If this symmetry were exact, the exchange of the rates of forgery and suppression would not have any impact on the resulting privacy protection achieved. However, we note that this is not the case. For example, Fig. 7.12(a) shows a lower reduction in privacy risk for $\rho < \sigma$, particularly



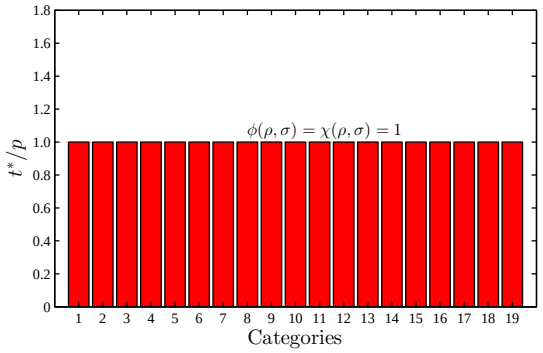
(a) $\rho = 0.03$, $\rho \in [\rho_6, \rho_7]$, $\sigma = 0.04$, $\sigma \in [\sigma_{18}, \sigma_{17}]$, $\rho/\rho_{\text{crit}}(\sigma) \simeq 0.055$, $\mathcal{R}(\rho,\sigma) \simeq 0.055$, $\mathcal{R}(\rho,\sigma)/\mathcal{R}_0 \simeq 0.549$.



(b) $\rho = 0.06$, $\rho \in [\rho_7, \rho_8]$, $\sigma = 0.08$, $\sigma \in [\sigma_{15}, \sigma_{14}]$, $\rho/\rho_{\text{crit}}(\sigma) \simeq 0.164$, $\mathcal{R}(\rho,\sigma) \simeq 0.026$, $\mathcal{R}(\rho,\sigma)/\mathcal{R}_0 \simeq 0.259$.



(c) $\rho = 0.11$, $\rho \in [\rho_{10}, \rho_{11}]$, $\sigma = 0.12$, $\sigma \in [\sigma_{15}, \sigma_{14}]$, $\rho/\rho_{\text{crit}}(\sigma) \simeq 0.434$, $\mathcal{R}(\rho,\sigma) \simeq 0.006$, $\mathcal{R}(\rho,\sigma)/\mathcal{R}_0 \simeq 0.061$.

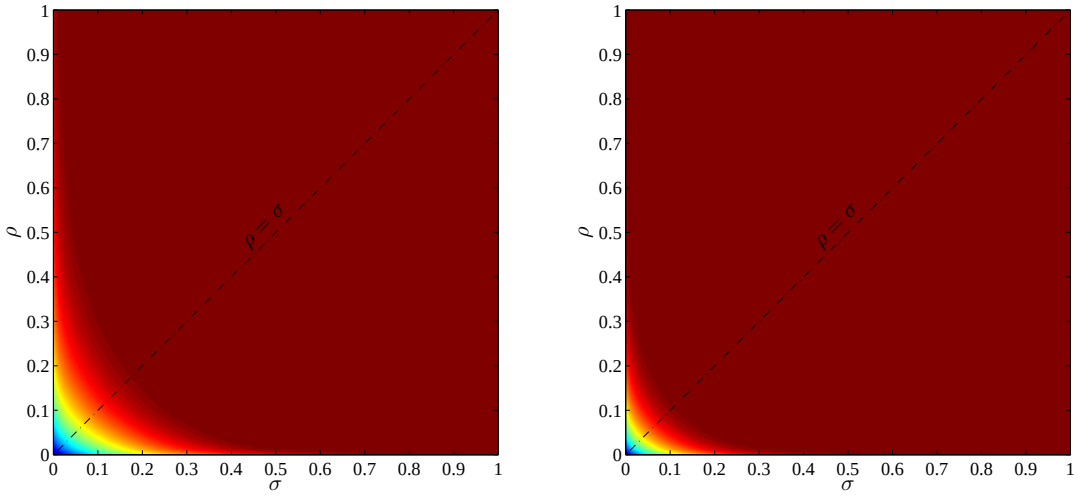


(d) $\rho \simeq 0.180$, $\rho \in [\rho_{12}, \rho_{13}]$, $\sigma = 0.15$, $\sigma \in [\sigma_{13}, \sigma_{12}]$, $\rho/\rho_{\text{crit}}(\sigma) = 1$, $\mathcal{R}(\rho,\sigma) = 0$, $\mathcal{R}(\rho,\sigma)/\mathcal{R}_0 = 0$.

Figure 7.11: Proportionality relationship between, on the one hand, the optimal apparent item distribution t^* of the user identified as `3301` in our data set, and on the other, the population's item distribution p.

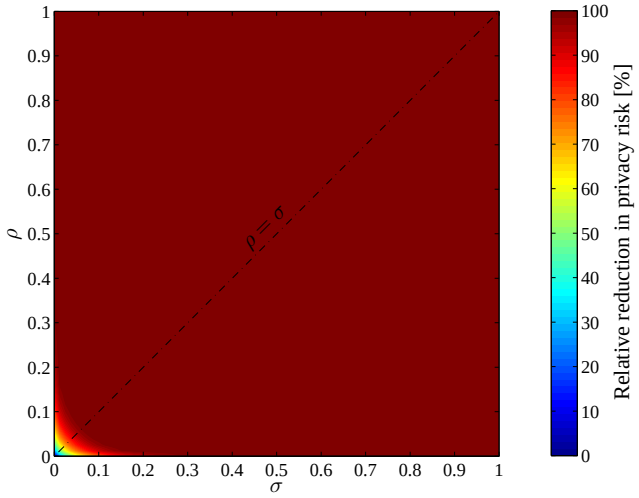
accentuated when $\sigma \simeq 0$. The reason for this may be found in the fact that, for most users, ρ_n is greater than σ_1. We shall elaborate more on this later when we consider forgery and suppression as pure strategies.

Next, we analyze the privacy protection provided by our technique for $\rho, \sigma \simeq 0$. In the theoretical analysis conducted in Sec. 7.6.4 we derived an expression for the relative reduction in privacy risk at low rates. Particularly, said expression was in terms of two factors, namely δ_ρ and δ_σ. In Fig. 7.13 we show the probability distribution of these factors. Consistently with Proposition 7.8, their minimum values are $\delta_\rho \simeq 3.12 > 1$ and $\delta_\sigma \simeq 2.30 > 0$. The maximum values attained by these forgery



(a) 10^{th} percentile.

(b) 50^{th} percentile.



(c) 90^{th} percentile.

Figure 7.12: We assume that the 4 099 users satisfying the positivity assumption (7.2) protect their privacy by using a common forgery rate and a common suppression rate. Under this assumption, we plot some percentiles surfaces of relative reduction in privacy risk, against these two common rates.

and suppression factors are approximately 324.98 and 266.13. On the other hand, in favor of suppression is the fact that the percentage of users with $\delta_\rho \geqslant 30$ is lower than the percentage of users with $\delta_\sigma \geqslant 30$. More precisely, these percentages yield 26.8% and 33.1%, respectively. In the end, an eye-opening finding is that $\delta_\rho > \delta_\sigma$ in 43.45%

of users, which suggests introducing a suppression rate higher than that of forgery, at least at low rates.

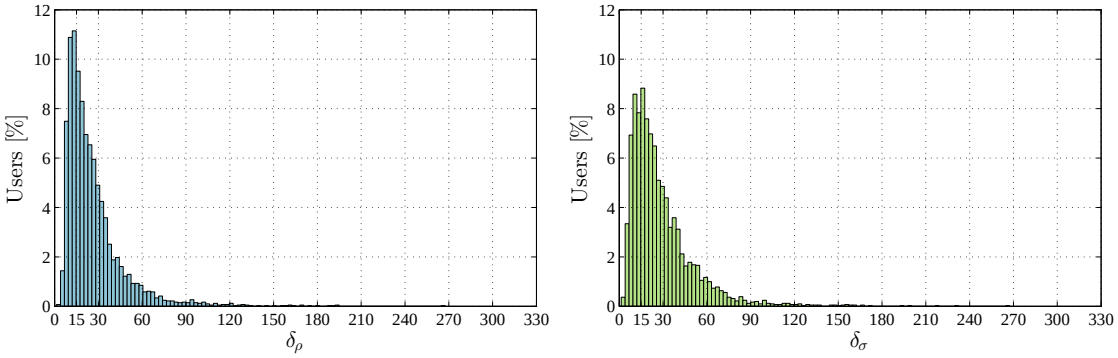


Figure 7.13: Probability distribution of the relative decrement factors of forgery and suppression.

After analyzing the forgery and the suppression of ratings as a mixed strategy, our last experimental results contemplate the application of forgery and suppression as pure strategies. In Fig. 7.14 we illustrate the probability distribution of the critical rates ρ_n and σ_1. The critical-forgery rate ranges approximately from 0.171 to 54.18, and its average is 3.45. The critical-suppression rate, on the other hand, goes from 0.153 to 0.963, and its average is 0.632. These figures indicate that, on average, a user will have either to refrain from rating an item six out of ten times, or submit nearly 3.45 false ratings per each original rating. This is, of course, when the user wishes to reach the critical-privacy region. Bearing these figures in mind, it is not surprising then that 95.3% of the users in our data set would opt for suppression as pure strategy, as it comes at the cost of a lower impact on utility.

7.8 Conclusions

In the literature of recommendation systems there exists a variety of approaches aimed at protecting user privacy. Among these approaches, the combined use of the forgery and the suppression of ratings emerges as a technique to hinder privacy attackers in their efforts to target peculiar users based on the items rated by these users. Our technique enhances users' privacy to a certain degree by blending their profiles into the crowd. Besides, it does not require users to trust neither the recommender nor the

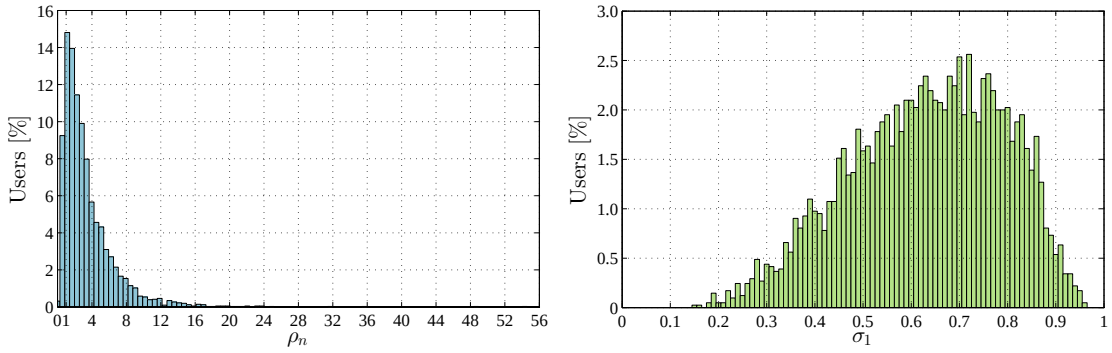


Figure 7.14: Probability distribution of the critical-forgery rate and the critical-suppression rate.

network operator, it is simple in terms of infrastructure requirements, and it can be used in combination with other approaches such as anonymous communications and user collaboration. However, as any data-perturbative mechanism, our PET comes at the expense of a loss in data utility, in particular a degradation of the quality of the recommender's predictions. The overall objective of this chapter is to engineer our mechanism to attain the optimal trade-off between privacy and utility, in the sense of maximizing privacy for an acceptable level of utility.

Our first contribution is an architecture that specifies, at a functional level, how our approach could be implemented as software. The purpose of this architecture is to help users determine which ratings should be made and which ones should be avoided. The core of our approach is a block that calculates the optimal forgery and suppression strategies, two tuples containing the percentage of items that should be forged and eliminated in each category. With these tuples, the proposed architecture warns the user when their profile deviates significantly from the population's item distribution.

The second contribution of this chapter is to investigate mathematically the aforementioned trade-off. With this aim, first we propose a quantitative measure of both privacy and utility. We quantify privacy risk as the KL divergence between the user's item distribution and the population's, and measure utility as the fraction of ratings the user is willing to forge and suppress. With these two quantities, we formulate a multiobjective optimization problem characterizing the trade-off between privacy risk on the one hand, and on the other forgery rate and suppression rate.

Our theoretical analysis provides a closed-form solution to this problem and characterizes the optimal trade-off surface between privacy and utility. The solution is confined to the closure of the noncritical-privacy region. The interior of the critical-privacy region is of no interest as the privacy risk attains its minimum value at the boundary of $\bar{\mathscr{C}}$. In the region of interest, our analysis finds that the optimal forgery and suppression strategies are orthogonal. In addition, these two strategies follow an intuitive principle. The forgery strategy recommends adding ratings to those categories where the user's interest is lower than the population's. The suppression strategy suggests eliminating those ratings belonging to the categories where the user shows too much interest compared to the reference distribution.

Our theoretical study also examines how these optimal strategies perturb user profiles. It is interesting to observe that the optimal apparent profile becomes proportional to the population's distribution in those categories with the lowest and highest ratios $\frac{q_k}{p_k}$. Our analysis also includes the study of the convexity of $\mathscr{C}$ and the characterization of $\mathcal{R}$ at low rates of forgery and suppression. More accurately, we provide a first-order Taylor approximation of the privacy-utility trade-off function, from which we conclude that the ratios $\frac{q_1}{p_1}$ and $\frac{q_n}{p_n}$ determine, together with the quantity $\mathrm{D}(q \,\|\, p)$, the privacy risk at low rates. An eye-opening fact is that, for low perturbation rates, the relative decrement in privacy risk is greater than the forgery rate introduced.

Further, we consider the special case when forgery and suppression are not used in combination. Under this consideration, we investigate which one is the most appropriate technique, first, in terms of causing the minimum distortion to reach the critical-privacy region, and secondly, in terms of offering better privacy protection at low rates. Our findings show that the arithmetic and geometric mean of the maximum and minimum ratios $\frac{q_k}{p_k}$ play a fundamental role in deciding the best technique to use. Afterwards, our formulation and theoretical analysis are illustrated with a numerical example.

In the end, the last section is devoted to the experimental evaluation of our data-perturbative mechanism in a real-world personalized recommendation system. In particular, we examine how the application of the forgery and the suppression of

ratings may preserve user privacy in Movielens. Among other results, we find that half of the users mitigate any privacy risk for forgery and suppression rates of just 13%. We also check that the mixed forgery and suppression strategy may provide better privacy protection for the same total rate than the pure forgery and suppression strategies. Further, the probability distributions of the relative decrement factors indicate that, at low rates, forgery provides a higher reduction in privacy risk than suppression does. By contrast, we observe that the suppression relative decrement factor is greater than that of forgery in 43.45% of users. Lastly, we consider the case when users opt for either forgery or suppression; and find that the latter is the best strategy to use in 95.3% of users who wish to vanish privacy risk while causing the minimum distortion.

Chapter 8

Conclusions and Future Work

In recent times we are witnessing the emergence of a new generation of information systems that adapt their functionalities to meet the unique needs of each individual. Personalization is revolutionizing the manner we access information but, at the same time, it is raising new privacy concerns with respect to user profiling.

The literature abounds with PETs aimed at safeguarding user privacy in a diverse range of applications, including among others the fields of SDC and anonymous communications. A wide variety of metrics have been proposed to assess the extent to which these general-purpose PETs may contribute to privacy enhancement. However, privacy researchers and users community lack a general framework that enables them to measure and compare the effectiveness of such technologies under a common perspective; frequently, the evaluation of a technology is done by using ad hoc metrics and adversary models specific to the application for which it has been conceived. In the particular context of personalized information systems, there are a few proposals for quantifying the privacy of user profiles, and those existing are not justified or fail to justify the choice.

The first part of this thesis tackles the issue of measuring user privacy. First, we propose a unifying view to compare and choose state-of-the-art privacy metrics in a systematic, rigorous manner. Secondly, we examine two information-theoretic quantities as privacy criteria in the context of personalized information systems.

The second part of this dissertation proposes data-perturbative privacy-protecting mechanisms for the applications of collaborative tagging and personalized recommendation systems. Equipped with quantitative measures of privacy and utility, we investigate the optimal privacy-utility trade-off posed by such mechanisms.

The remainder of this chapter summarizes the main results from our research and identifies some future research lines.

8.1 Conclusions

This section is organized according to the two parts this dissertation has been structured.

8.1.1 Privacy Metrics

We have presented a theoretical framework that permits comparing and interpreting privacy metrics from diverse fields of information privacy. Our framework provides a unifying view of privacy by measuring it as the estimation error of an adversary whose purpose is to unveil the private information that a system wishes to protect.

We have shown that a large number of privacy metrics from SDC, ACSs and LBSs are related to this estimation error. In particular, we have proven that there is a bijective relation between most of these metrics and our more general view of privacy, which implies that all these criteria are equivalent both in terms of comparison and optimization.

In our interpretations of such criteria as estimation errors, we have allowed for the geometry of the attacker's distortion function, the system's knowledge about this function and the nature of the private information to be protected. The arguments expounded in these interpretations build upon numerous concepts from information theory and BDT. For example, k-anonymity and l-diversity are connected to an estimation error through Rényi's entropy and MAP estimation. This is in the special case when the attacker's distortion function is the Hamming distance and users' private information are single-occurrence data. In the case of multiple-occurrence data,

fundamental results from AEP enable us to regard Shannon's conditional entropy as a measure of the cardinality of a high-confidence set.

When we consider non-Hamming distortion functions and assume that this function is known to the system, the total variation distance and Pinsker's inequality allow us to interpret t-closeness and mutual information as upper bounds on the reduction of the Bayes conditional risk. A greater upper bound on this risk is given by the δ-disclosure requirement, which may be deemed as a stricter privacy measure than mutual information and t-closeness. Another interesting result is the connection between our formulation of the privacy-utility trade-off and the rate-distortion problem, a well-known and extensively studied optimization problem appearing in information theory.

We have illustrated the applicability of our framework in the contexts of LBSs and ACSs by means of two numerical examples. In LBSs, we consider the squared error distance as the attacker's distortion function, and assume this information is available to the system. Under these assumptions, we show that the attacker's estimation error boils down to the MSE. An anonymous-communication protocol inspired by Crowds is then evaluated. Adopting the Hamming distance, we interpret Shannon's, Hartley's and min- entropy as particular cases of the MAP error.

A comprehensive guide is provided for those designers of SDC and ACSs who want to skip the mathematical details of our framework and wish to know which particular metric is the most suitable for their privacy requirements. The theoretical analysis, together with these guidelines, constitute a systematic approach to the problem of measuring user privacy and evaluating PETs.

We have also tackled the issue of quantifying privacy in those applications where user profiles are involved. Specifically, we have proposed and justified KL divergence and Shannon's entropy as measures of user privacy in personalized information systems.

Our justifications build on two adversary models defined according to the technical literature of profiling. In both models the attacker strives to profile users of those systems. The difference is in the ultimate objective of profiling. In the former model, the adversary is interested in finding users who deviate significantly from the average

profile of the population. In the latter model, the attacker aims at classifying users into certain groups of people or collectives.

Under the objective of individuation, Jaynes' argument behind entropy-maximization methods permits interpreting the KL divergence between the user's apparent profile and the population's profile as a measure of anonymity. Our criterion is a measure of anonymity not in the sense that the user's identity remains hidden, but in the sense that the lower the divergence between these two profiles, the higher the probability of the apparent profile, and therefore the larger the population of users in which the user's interests are blended. In a nutshell, the KL divergence is an (inverse) indicator of the commonness of the apparent profile in said population. If the population's distribution is not available, Shannon's entropy of the apparent profile is of special interest as it may be regarded as an anonymity criterion in a sense analogous to that of divergence. These two interpretations are based on the realistic assumption that a probabilistic model of profiles is not at the disposal of users.

Under the objective of classification, we propose measuring privacy as the divergence between the apparent profile and the profile of the group into which the user does not want to be classified. Our justification leverages on hypothesis testing and the Neyman-Pearson lemma. When a suitable representation of the group profile is not available or simply it is unknown, the maximization of the divergence between the perturbed, observed profile and the actual one describes the situation where the user wants the former profile to resemble as little as possible the genuine profile. This is in contrast to our previous interpretation of divergence as a measure of user-profile density: a profile already matching the population's distribution would not need any perturbation.

Further, we have presented a comparative analysis between our privacy metrics and other proposals measuring the privacy risks in personalized information systems. Our systematic classification of metrics shows that most of them may be described in terms of the classifier adversary model, and under the assumption that the group profile is unknown. Lastly, we interpret KL divergence and Shannon's entropy as an attacker's estimation error, which demonstrates the generality of the proposed theoretical framework.

8.1.2 Data-Perturbative Mechanisms and Privacy-Utility Trade-Off

We have proposed a data-perturbative method aimed at protecting user privacy in the semantic Web. Specifically, our tag-suppression strategy has the purpose of hindering a privacy attacker in its efforts to individuate users based on their tagging activity.

The proposed strategy may be used in combination with other PETs, can be implemented on the user's computer and does not require users to trust any external entity. We have presented a modular architecture describing how our approach could be implemented in practice.

Our model measures user privacy as Shannon's entropy of the apparent profile, and utility as the tag-suppression rate. With these quantitative measures of privacy and utility, we optimize the tag-suppression mechanism in terms of its privacy-utility trade-off. This trade-off is formulated as a multiobjective optimization problem, which turns to be a resource allocation problem.

We have found a closed-form solution and characterized the optimal trade-off. Among other results, we have shown that there exists a critical tag-suppression rate beyond which the apparent profile becomes the uniform distribution and privacy is therefore maximized. The optimal suppression strategy follows the intuitive principle of eliminating the tags from those categories where the user has shown too much interest in.

Further, we have analyzed the cases of low-suppression rate and high privacy. Our analysis demonstrates that the entropy of the user's actual profile, together with the maximum value of such profile, characterize the relative privacy gain at low rates. This is in contrast to the fact that the critical tag-suppression rate is determined by the minimum value of this profile. For suppression rates approaching this critical rate, we have found that the second-order Taylor approximation of the privacy-utility trade-off function is given by the Fisher information.

Experimental results in the collaborative tagging service BibSonomy show how our tag-suppression technique may contribute to privacy protection. These results indicate that users would need high suppression rates to attain the maximum privacy level. The distributions of suppression thresholds suggest, however, that smaller suppression rates would lead to significant gains in privacy.

In addition, we have proposed an architecture that extends the functionalities of the current collaborative tagging systems, by incorporating two additional layers placed on top of a social bookmarking application such as Delicious. The policy layer allows users to specify their preferences explicitly. The privacy layer implements our optimized tag-suppression mechanism.

On the one hand, we have examined how tag suppression enhances the privacy of the users of Delicious. On the other, we have investigated the influence of this privacy layer, first, on the semantic functionality of the underlying bookmarking application, and secondly, on two services enabled by the policy layer. These two services provide resource recommendations and content-filtering capabilities.

To conduct our experimental analysis, we have employed a more elaborate utility measure than the simplified but mathematically tractable tag-suppression rate, namely, the percentages of tags that each bookmark loses due to suppression. To assess the impact of suppression on the aforementioned content-filtering application, we have measured the number of false negatives and false positives, precision and recall. Our empirical evaluation indicates that the effect of tag suppression on the accuracy of a parental-control filter is relatively small. For example, recall exhibits a reduction by 0.11% when all users eliminate almost all their tags. This and further results are explained by the fact that the PMFs of a large number of resources concentrate a substantial amount of their masses in a reduced set of subcategories.

In the enthralling application of personalized recommendation systems, we have proposed a PET that simultaneously combines two data-perturbative strategies—the forgery and the suppression of ratings. Our mechanism builds on the same adversary model than that of tag suppression and thus aims at thwarting a privacy attacker in its efforts to individuate users based on the items rated.

Under this objective, and assuming that the population's item distribution is known, we measure privacy as the KL divergence between the user's perturbed profile and the population's profile. We also quantify the degradation in the quality of the recommendations due to perturbation, but use a simplified measure of utility —the forgery and the suppression rates— which allows us to formulate the privacy-utility trade-off by means of a mathematically tractable model.

Endowed with quantitative privacy and utility metrics, we design the proposed PET to achieve the optimal privacy-utility trade-off in the sense of maximizing privacy for a desired level of utility. The trade-off posed by our PET is modeled as a multiobjective optimization problem. We provide a closed-form solution to said problem.

Our mathematical analysis characterizes the optimal trade-off surface and expresses it in terms of a divergence between two distributions. We show that there exists a critical-privacy region where the privacy risk vanishes and another region where it does not. The optimal forgery and suppression strategies are confined to this latter region, which we prove to be nonconvex. Such strategies conform to intuition as they suggest forging ratings where the user has little interest and recommend eliminating ratings where the user shows too much interest, compared to the population's distribution.

We demonstrate that the solution is determined by a sequence of forgery and suppression thresholds, which specify the number of nonzero components of the optimal strategies. Further, we verify that the user's genuine profile is perturbed in such a manner that the apparent profile progressively becomes proportional to the population's distribution.

We explore the behavior of the function modeling the trade-off at low rates of forgery and suppression. To this end, we derive its first-order Taylor approximation at the origin and show that the relative decrement in privacy risk depends, on the one hand, on the initial privacy level, and on the other, on the minimum and the maximum ratio between the user's profile and the population's. Also, we prove that the reduction in privacy risk due to forgery is greater than the loss in utility.

In addition, we consider the case when users must opt for either forgery or suppression. We find that the arithmetic and geometric mean of the aforementioned minimum and maximum ratios determine the choice for low perturbation rates and when users want to attain the critical-privacy region. Our theoretical analysis for simultaneous forgery and suppression shows consistency with the results obtained for tag suppression.

Experimental results in the movie recommendation system Movielens show that, for relatively small rates of forgery and suppression, an important number of users obtain significant gains in privacy. For example, forgery and suppression rates of 13% lead half of the users to reach the critical-privacy region. Moreover, we observe that the mixed strategy may provide stronger privacy protection for the same total rate than the pure strategies. Another finding is that the minimum relative decrement factor of suppression is approximately 2.30, which means that, for low rates, the reduction in privacy risk is greater than the suppression rate introduced. Finally, our empirical analysis shows that, in 95.3% of cases, the pure suppression strategy reaches the critical-privacy region with a lower distortion than the pure forgery strategy does.

8.2 Future Work

In this section we explore possible improvements and open research directions based on ideas and results provided in this dissertation.

- **User-profile model**. In Sec. 4.3.2 we specified the model of user profile used both in Chapter 4 and in the second part of this dissertation. The model in question constitutes the profile of user interests that an attacker would create. Such model is defined based on (1) the information exploited by the attacker to profile users, and (2) the type of representation used to model the user interests.

  Our model of user profile as a PMF is a first-approximation, mathematically-tractable model. The proposed model captures the information provided explicitly by the user, and computes a histogram of relative frequencies of such information, classified according to a set of categories of interest. This model, however, does not consider a range of other factors that an attacker would possibly use to better characterize such interests.

  One of these factors is the user activity, that is, the total number of tags, ratings and any other data conveyed to the personalized information system in question. User activity would allow an attacker to estimate user interests in absolute terms, rather than relative, and therefore gain further knowledge on user preferences.

Moreover, our model considers just the information users communicate explicitly. In practice, a privacy attacker would enrich user profiles with implicit information such as the time it takes users to examine an item or the items purchased. Also, the adversary could leverage on the time of the day when the explicit information is submitted, or capitalize on the profiles of those users with whom social relationships are known to said attacker.

Our model of profile implicitly assumes that the user-generated data are statistically independent. Specifically, we assume that each user generates a sequence of actions (e.g., tags or ratings) and that these actions are modeled as i.i.d. r.v.'s distributed according to the user's profile. However, it is clear that there will exist some statistical dependence between those r.v.'s, and that a sophisticated attacker could exploit this fact. For example, the adversary could model such sequence of actions by assuming a stationary random process with memory.

In short, a promising future line of research would be studying user-profile models which take into consideration the aspects mentioned above. Tightly related to these models is undoubtedly the investigation of privacy metrics for these profiles as well as mechanisms designed for their protection.

- **Experimental validation of privacy metrics**. In Sec. 4.4 we gave a first, preliminary step towards the quantification of the privacy of user profiles. Particularly, we investigated Jaynes' rationale behind entropy-maximization methods to justify KL divergence and Shannon's entropy as metrics of profile privacy. Through Jaynes' argument we interpreted both information-theoretic quantities as measures of the relative frequency of a user profile. Concretely, our conjecture was that the probability $p_T(t)$ of a PMF t modeling the profile of a user was related to its divergence with respect to the average profile p.

  Although Jaynes' rationale provides a solid mathematical justification in favor of divergence, a line for further research could be to experimentally validate if our model is a good approximation, in the sense that minimizing $\mathrm{D}(t \,\|\, p)$ is a good criterion when we wish to maximize user anonymity, measured as $p_T(t)$.

- **Adversary model**. One of the suppositions of our adversary model is that the attacker assumes the observed, perturbed profile is the user's genuine profile. Said otherwise, our attacker is unable to discern whether a particular user is applying some data-perturbative mechanism to protect their privacy. As explained in Secs. 5.4 and 7.4, we based this assumption on the fact that such mechanisms may be implemented as software running on users' machines.

  Although we limited the scope of our work to this adversary model, we recognize that sophisticated attackers might exploit certain information such as user activity to ascertain whether a user is applying some perturbative strategy, and ultimately to guess their actual profile. For example, consider an adversary who attempts to estimate the tag-suppression rate of a user on the basis of observed differences in tagging activity; and according to this rate, the attacker strives to reverse the perturbation introduced by the privacy-utility optimized mechanism proposed in Chapter 5.

- **Practical implementation of our data-perturbative mechanisms**. In Secs. 5.4 and 7.4, we proposed two architectures describing how our data-perturbative mechanisms could be implemented as software. One was for tag suppression, and the other for the combination of the forgery and the suppression of ratings. In those sections, we provided a functional description of the internal components of such architectures. We did not explore, however, some crucial aspects that a successful implementation should take into account.

  Among other aspects, future research should delve further into the modules that estimate both the actual user profile and the population's distribution. In particular, it would be necessary to investigate computationally-efficient categorization algorithms that can be executed on the user's machine, without the need to access any external database such as the Open Directory Project. Another aspect that a practical implementation should consider is the initialization of the profile and the fact that this profile may vary substantially over time. The assumption of dynamic user profiles undoubtedly calls for new mathematical models.

- **Evaluation of the impact of perturbation on recommendation systems**. In Chapter 7 we proposed the rating-forgery rate and the rating-suppression rate as measures of utility loss in the context of recommender systems. These two measures enabled us to model the trade-off between privacy and utility as a mathematically-tractable optimization problem which we later solved.

  While these metrics are suitable for our mathematical modeling, it would be interesting to measure the impact that forgery and suppression actually have on the accuracy of the predictions generated by the recommender system. In other words, a possible line of future research would be the exploration of more sophisticated but computationally-feasible utility metrics.

- **Other data-perturbative strategies**. The second part of this dissertation investigates mechanisms that perturb users' information to enhance their privacy in the context of personalized information systems. Our analysis contemplates two mechanisms, namely, the suppression of tags in the semantic Web and the combination of the forgery and the suppression of ratings in personalized recommendation systems.

  As commented in Sec. 7.2, some perturbative mechanisms may be suitable for certain applications but not for others. For instance, the simultaneous use of forgery and suppression is a good strategy when the information to be perturbed are ratings, but it could not be the case in personalized video-streaming services such as YouTube, where user profiles are created from the history of watched videos. Playing videos the user is not actually interested in might use up their bandwidth and consequently degrade the quality of other Web services. By contrast, users may be reticent to suppression, in the sense of refusing to play those videos they wish to watch. Likewise, suppression is an appropriate strategy for tagging applications but not for Web search.

  The set of personalized information systems is very diverse and, for this reason, it would be desirable to investigate mechanisms with a broader scope of application. One of these mechanisms could be *generalization*, a data-perturbative

strategy whereby specific terms are transformed into more general terms. Conceptually, consider a user replacing the tag "depression" with "health". Generalization could be used not only in tagging applications but also in other contexts such as Web search. One direction for future work would be to study theoretically and experimentally the privacy-utility trade-off posed by such mechanism.

- **Message deferral against profiling based on time.** In Chapters 5, 6 and 7, we investigated data-perturbative mechanisms that protect user privacy against a class of adversaries who, first, analyzes the *content of the information* users sent to personalized information systems, then classifies this information into a given set of interest categories, and ultimately profiles them according to such interests.

  As a future research line, we could consider privacy attackers who, instead of profiling users based on their interests, exploit the time instants when users communicate with information providers. In other words, we could explore adversaries who profile users based on the time when they submit tags, ratings, messages, queries, etc.

  Although this kind of profiling clearly could be conducted on all sorts of information systems, we believe that online social networking services and microblogging services such as Twitter and Facebook are more prone to such attacks. In this kind of personalized information systems, it could be more burdensome for an attacker to analyze the content of users' messages, in which, in addition to text [(a)], users often include images and videos. Since processing all these data and mapping them into interest categories could require certain computational effort [(b)], the attacker could take advantage of timing information. Fig. 8.1 shows an example of user profile based on time.

[(a)] Twitter messages, also known as *tweets*, are limited to 140 characters. Facebook does not impose any limitation on message length.

[(b)] This is in contrast to other information systems where user data (e.g., tags, queries or ratings) are simpler to process

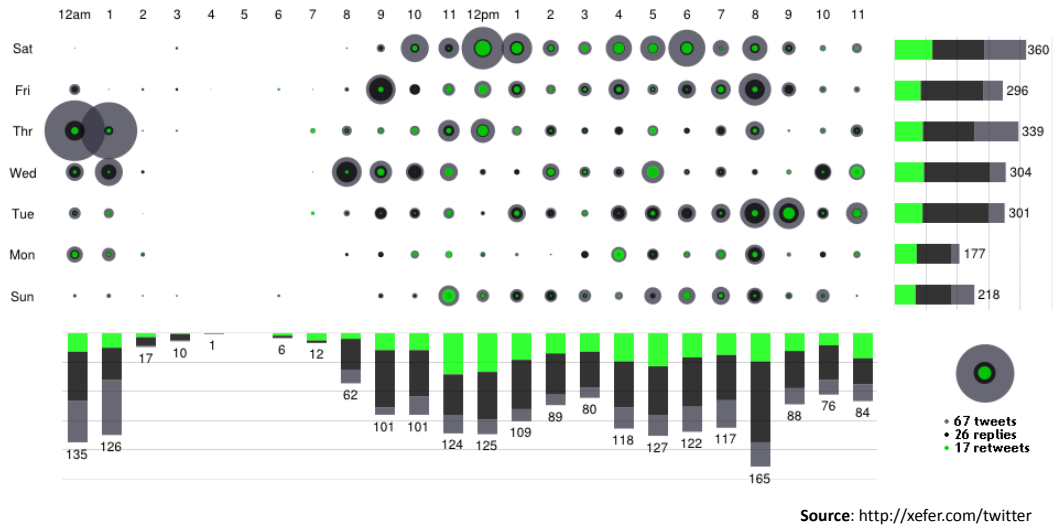


Figure 8.1: Profile of a Twitter user by hours and days of the week, as retrieved from `http://xefer.com/twitter`.

In this situation, the *deferral or delay of messages* appears a simple PET that could help users protect their privacy against such profiling attack. The proposed data-perturbative mechanism would allow users to delay the submission of certain messages by storing them locally and afterwards sending them to the information system in question. On the one hand, this would enable users to enhance their privacy to a certain extent, but on the other, the utility of the microblogging and social networking services mentioned above would be affected. For instance, consider a user posting a tweet to confirm a meeting this evening. If this tweet was delayed, the confirmation could not arrive on time and, if so, the information-exchange functionality would be useless. In short, delaying messages poses a trade-off between privacy and utility.

Our model of user profile would be similar to that presented in Sec. 7.5. We could represent the messages of a user as a sequence of i.i.d. r.v.'s taking on values in an alphabet of $n \geqslant 2$ time periods. The set of time periods could be, for example, the hours of a day or a week, or the days of a month. Consistently with the notation used in the second part of this thesis, q would denote the actual user profile and p the population's distribution, conceptually, histograms of relative frequencies of messages over those time periods.

A user adhered to this technique would first specify a *message-deferral rate* $\varphi \in [0, 1)$, that is, the ratio of messages to total number of messages that they would be willing to delay. When delaying messages, q would be seen from the outside as the apparent profile $t = q - s + r$, according to some *storing strategy* s and some *forwarding strategy* r. The former represents the percentage of messages that the user should temporarily store on a buffer at each time instant, and the latter the fraction of messages to total number of messages that should be output from the buffer at each time period. These strategies obviously must satisfy $r_i, s_i \geqslant 0$, $q_i - s_i + r_i \geqslant 0$ for $i = 1, \ldots, n$, and $\sum_i r_i = \sum_i s_i = \varphi$.

Assuming that p is available to users and that the attacker attempts to individuate them, we could measure privacy risk as the KL divergence between t and p. Our utility metric could be, on the other hand, the rate of messages delayed φ. Considering these two metrics, the formulation of the optimal privacy-utility trade-off turns to be a particular case of that for the forgery and the suppression of ratings (7.1), namely the case when $\rho = \sigma = \varphi$.

The trade-off between privacy and message-deferral rate is therefore characterized by the theoretical analysis presented in Sec. 7.6. However, it would be interesting and even necessary to investigate more elaborate utility measures such as the expected delay or the capacity of the buffer, and study how these metrics are related to each other. During my research stay at NEC Laboratories Europe, we have already started exploring the relationship among these metrics and have obtained some preliminary research results, to be published shortly.

Acronyms

ACS	anonymous-communication system
AEP	asymptotic equipartition property
BDT	Bayes decision theory
HTTP	hypertext transfer protocol
IP address	Internet protocol address
ISP	Internet service provider
KKT conditions	Karush-Kuhn-Tucker conditions
KL divergence	Kullback-Leibler divergence
LBS	location-based service
MSE	mean squared error
P2P	peer to peer
PET	privacy-enhancing technology
PIR	private information retrieval
PMF	probability mass function
SDC	statistical disclosure control
SME	small and medium enterprise
TF-IDF	term frequency-inverse document frequency
TTP	trusted third party

Bibliography

[1] P. C. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, and G. Palis, *Understanding big data.* McGraw-Hill, 2012. [Online]. Available: http://www-01.ibm.com/software/data/bigdata

[2] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Comput. Mag.*, vol. 7, no. 1, pp. 76–80, Jan. 2003.

[3] "Montly unique visitors to u. s. retail web sites in 4th quarter 2012." [Online]. Available: http://www.statista.com/statistics/172685/monthly-unique-visitors-of-us-retail-websites

[4] T. Geron, "Facebook q1 earnings: $1.46 billion revenue, up 38 percent," May 2013. [Online]. Available: http://www.forbes.com/sites/tomiogeron/2013/05/01/facebook-q1-earnings-1-46-billion-revenue-up-38

[5] "Global online advertising industry 2013-2018: Trend, profit, and forecast analysis," Jan. 2013. [Online]. Available: http://www.reportlinker.com/p01162298/Global-Online-Advertising-Industry-Trend-Profit-and-Forecast-Analysis.html

[6] L. Cottrell, "Mixmaster and remailer attacks," 1994. [Online]. Available: http://obscura.com/~loki/remailer/remailer-essay.html

[7] A. Serjantov and R. E. Newman, "On the anonymity of timed pool mixes," in *Proc. Workshop Priv., Anon. Issues Netw., Distrib. Syst.* Kluwer, 2003, pp. 427–434.

[8] U. Möller, L. Cottrell, P. Palfrader, and L. Sassaman, "Mixmaster protocol – Version 2," Internet Eng. Task Force, Internet Draft, Jul. 2003. [Online]. Available: http://www.freehaven.net/anonbib/cache/mixmaster-spec.txt

[9] D. Kesdogan, J. Egner, and R. Büschkes, "Stop-and-go mixes: Providing probabilistic anonymity in an open system," in *Proc. Inform. Hiding Workshop (IH)*. Springer-Verlag, 1998, pp. 83–98.

[10] D. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Commun. ACM*, vol. 24, no. 2, pp. 84–88, 1981.

[11] M. Rennhard and B. Plattner, "Practical anonymity for the masses with mix-networks," in *Proc. Int. Workshop Enabling Technol.: Infra. Col. Enterprises (WETICE)*. IEEE Comput. Soc., 2003, pp. 255–260.

[12] G. Danezis, "Mix-networks with restricted routes," in *Proc. Int. Symp. Priv. Enhanc. Technol. (PETS)*. Lecture Notes Comput. Sci. (LNCS), 2003, pp. 1–17.

[13] D. Goldschlag, M. Reed, and P. Syverson, "Hiding routing information," in *Proc. Inform. Hiding Workshop (IH)*, 1996, pp. 137–150.

[14] M. G. Reed, P. F. Syverson, and D. M. Goldschlag, "Proxies for anonymous routing," in *Proc. Comput. Secur. Appl. Conf. (CSAC)*, San Diego, CA, Dec. 1996, pp. 9–13.

[15] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The second-generation onion router," in *Proc. Conf. USENIX Secur. Symp.*, Berkeley, CA, 2004, pp. 21–21.

[16] D. Chaum, "Security without identification: Transaction systems to make big brother obsolete," *Commun. ACM*, vol. 28, no. 10, pp. 1030–1044, Oct. 1985.

[17] V. Benjumea, J. López, and J. M. T. Linero, "Specification of a framework for the anonymous use of privileges," *Telemat., Informat.*, vol. 23, no. 3, pp. 179–195, Aug. 2006.

[18] G. Bianchi, M. Bonola, V. Falletta, F. S. Proto, and S. Teofili, "The SPARTA pseudonym and authorization system," *Sci. Comput. Program.*, vol. 74, no. 1–2, pp. 23–33, 2008.

[19] G. Fuchsbauer, D. Pointcheval, and D. Vergnaud, "Transferable constant-size fair e-cash," in *Proc. Cryptology, Netw. Secur. (CNS)*. Springer-Verlag, 2009, pp. 226–247.

[20] P. Bogetoft, D. L. Christensen, I. Damgård, M. Geisler, T. Jakobsen, M. Krøigaard, J. D. Nielsen, J. B. Nielsen, K. Nielsen, J. Pagter, M. Schwartzbach, and T. Toft, *Financial Cryptography and Data Security*. Springer-Verlag, 2009, ch. Secure Multiparty Computation Goes Live, pp. 325–343.

[21] A. Rial and B. Preneel, "Optimistic fair priced oblivious transfer," in *Proc. Int. Conf. Cryptology Africa (AFRICACRYPT)*. Springer-Verlag, 2010, pp. 131–147.

[22] J. Borking, "Why adopting privacy enhancing technologies (PETs) takes so much time," in *Proc. Comput. Priv., Data Prot. (CPD)*, S. Gutwirth, Y. Poullet, P. Hert, and R. Leenes, Eds. Springer-Verlag, 2011, pp. 309–341.

[23] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 6, pp. 1010–1027, 2001.

[24] L. Sweeney, "k-Anonymity: A model for protecting privacy," *Int. J. Uncertain., Fuzz., Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[25] T. M. Truta and B. Vinay, "Privacy protection: p-sensitive k-anonymity property," in *Proc. Int. Workshop Priv. Data Manage. (PDM)*, Atlanta, GA, 2006, p. 94.

[26] X. Sun, H. Wang, J. Li, and T. M. Truta, "Enhanced p-sensitive k-anonymity models for privacy preserving data publishing," *Trans. Data Priv.*, vol. 1, no. 2, pp. 53–66, 2008.

[27] A. Machanavajjhala, J. Gehrke, D. Kiefer, and M. Venkitasubramanian, "*l*-Diversity: Privacy beyond *k*-anonymity," in *Proc. IEEE Int. Conf. Data Eng. (ICDE)*, Atlanta, GA, Apr. 2006, p. 24.

[28] H. Jian-min, C. Ting-ting, and Y. Hui-qun, "An improved V-MDAV algorithm for *l*-diversity," in *Proc. IEEE Int. Symp. Inform. Process. (ISIP)*, Moscow, Russia, May 2008, pp. 733–739.

[29] N. Li, T. Li, and S. Venkatasubramanian, "*t*-Closeness: Privacy beyond *k*-anonymity and *l*-diversity," in *Proc. IEEE Int. Conf. Data Eng. (ICDE)*, Istanbul, Turkey, Apr. 2007, pp. 106–115.

[30] J. Brickell and V. Shmatikov, "The cost of privacy: Destruction of data-mining utility in anonymized data publishing," in *Proc. ACM SIGKDD Int. Conf. Knowl. Disc., Data Min. (KDD)*, Las Vegas, NV, Aug. 2008, pp. 70–78.

[31] C. Dwork, "Differential privacy," in *Proc. Int. Colloq. Automata, Lang., Program.* Springer-Verlag, 2006, pp. 1–12.

[32] D. Rebollo-Monedero, J. Forné, and J. Domingo-Ferrer, "From *t*-closeness-like privacy to postrandomization via information theory," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 11, pp. 1623–1636, Nov. 2010. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TKDE.2009.190

[33] O. Berthold, A. Pfitzmann, and R. Standtke, "The disadvantages of free MIX routes and how to overcome them," in *Proc. Design. Priv. Enhanc. Technol.: Workshop Design Issues Anon., Unobser.*, ser. Lecture Notes Comput. Sci. (LNCS). Berkeley, CA: Springer-Verlag, Jul. 2000, pp. 30–45.

[34] C. Díaz, S. Seys, J. Claessens, and B. Preneel, "Towards measuring anonymity," in *Proc. Int. Symp. Priv. Enhanc. Technol. (PETS)*, ser. Lecture Notes Comput. Sci. (LNCS), vol. 2482. Springer-Verlag, Apr. 2002, pp. 54–68.

[35] A. Serjantov and G. Danezis, "Towards an information theoretic metric for anonymity," in *Proc. Int. Symp. Priv. Enhanc. Technol. (PETS)*, vol. 2482. Springer-Verlag, 2002, pp. 41–53.

[36] S. Steinbrecher and S. Kopsell, "Modelling unlinkability," in *Proc. Int. Symp. Priv. Enhanc. Technol. (PETS)*. Springer-Verlag, 2003, pp. 32–47.

[37] G. Tóth, Z. Hornák, and F. Vajda, "Measuring anonymity revisited," in *Proc. Nordic Workshop Secure IT Syst.*, Nov. 2004, pp. 85–90.

[38] G. Tóth and Z. Hornák, "Measuring anonymity in a non-adaptive, real-time system," in *Proc. Int. Symp. Priv. Enhanc. Technol. (PETS)*, ser. Lecture Notes Comput. Sci. (LNCS), vol. 3424. Toronto, Canada: Springer-Verlag, May 2004, pp. 226–241.

[39] V. Shmatikov and M. H. Wang, "Measuring relationship anonymity in mix networks," in *Proc. Workshop Priv. Electron. Soc.* ACM, 2006, pp. 59–62.

[40] S. Clauß and S. Schiffner, "Structuring anonymity metrics," in *Proc. ACM Workshop on Digit. Identity Manage.* Fairfax, VA: ACM, Nov. 2006, pp. 55–62.

[41] P. Syverson and S. Stubblebine, "Group principals and the formalization of anonymity," in *Proc. World Congr. Formal Methods*, 1999, pp. 814–833.

[42] S. Mauw, J. Verschuren, and E. P. de Vink, "A formalization of anonymity and onion routing," in *Proc. European Symp. Res. Comput. Secur. (ESORICS)*, vol. 3193. Lecture Notes Comput. Sci. (LNCS), 2004, pp. 109–124.

[43] J. Feigenbaum, A. Johnson, and P. Syverson, "A model of onion routing with provable anonymity," in *Proc. Financ. Cryptogr., Data Secur. (FI)*. Springer-Verlag, 2007.

[44] M. Edman, F. Sivrikaya, and B. Yener, "A combinatorial approach to measuring anonymity," *IEEE J. Intell., Secur. Inform.*, pp. 356–363, 2007.

[45] J. Parra-Arnau, A. Perego, E. Ferrari, J. Forné, and D. Rebollo-Monedero, "Privacy-preserving enhanced collaborative tagging," *IEEE Trans. Knowl. Data Eng.*, 2012, to appear. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2012.248

[46] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "Measuring the privacy of user profiles in personalized information systems," *Future Gen. Comput. Syst. (FGCS), Special Issue Data, Knowl. Eng.*, 2013, to appear. [Online]. Available: http://dx.doi.org/10.1016/j.future.2013.01.001

[47] J. Parra-Arnau, D. Rebollo-Monedero, J. Forné, J. L. Muñoz, and O. Esparza, "Optimal tag suppression for privacy protection in the semantic Web," *Data, Knowl. Eng.*, vol. 81–82, pp. 46–66, Nov. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.datak.2012.07.004

[48] D. Rebollo-Monedero, J. Parra-Arnau, C. Diaz, and J. Forné, "On the measurement of privacy as an attacker's estimation error," *Int. J. Inform. Secur.*, vol. 12, no. 2, pp. 129–149, Apr. 2012. [Online]. Available: http://link.springer.com/article/10.1007/s10207-012-0182-5

[49] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "Optimal forgery and suppression of ratings for privacy enhancement in recommendation systems," 2013, submitted. [Online]. Available: http://arxiv.org/abs/1302.2501

[50] D. Rebollo-Monedero, J. Parra-Arnau, and J. Forné, "An information-theoretic privacy criterion for query forgery in information retrieval," in *Proc. Int. Conf. Secur. Technol.(SecTech)*, ser. Commun. Comput., Inform. Sci. (CCIS), vol. 259. Jeju Island, South Korea: Springer-Verlag, Dec. 2011, pp. 146–154.

[51] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "A privacy-preserving architecture for the semantic Web based on tag suppression," in *Proc. Int. Conf. Trust, Priv., Secur., Digit. Bus. (TrustBus)*, ser. Lecture Notes Comput. Sci. (LNCS), vol. 6264, Bilbao, Spain, Aug. 2010, pp. 58–68.

[52] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "A privacy-protecting architecture for collaborative filtering via forgery and suppression of ratings," in *Proc. Int. Workshop Data Priv. Manage. (DPM)*, ser. Lecture Notes Comput. Sci. (LNCS), vol. 7122, Leuven, Belgium, Sep. 2011, pp. 42–57.

[53] D. Rebollo-Monedero, J. Forné, E. Pallarès, and J. Parra-Arnau, "A modification of the lloyd algorithm for k-anonymous quantization," *Inform. Sci.*, vol. 222, pp. 185–202, Feb. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.ins.2012.08.022

[54] C. Tripp-Barba, L. Urquiza, M. Aguilar, J. Parra-Arnau, D. Rebollo-Monedero, and E. P. J. Forné, "A collaborative protocol for anonymous reporting in vehicular ad hoc networks," *Comput. Stand. & Interf.*, 2013, to appear. [Online]. Available: http://dx.doi.org/10.1016/j.csi.2013.06.001

[55] D. Rebollo-Monedero, J. Forné, E. Pallarès, J. Parra-Arnau, C. Tripp, L. Urquiza, and M. Aguilar, "On collaborative anonymous communications in lossy networks," *Security, Commun. Netw. (SCN), Special Issue Security Completely Interconnect. World*, 2013, to appear.

[56] "Things that happen on internet every sixty seconds," Jun. 2011. [Online]. Available: http://www.go-gulf.com/blog/60-seconds

[57] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scient. Amer.*, pp. 35–43, May 2001.

[58] J. Constine, "Facebooks growth since IPO in 12 big numbers," May 2013. [Online]. Available: http://techcrunch.com/2013/05/17/facebook-growth

[59] "Internet world stats," May 2013. [Online]. Available: http://www.internetworldstats.com/stats.htm

[60] J. Manyika and C. Roxburgh, "The great transformer: the impact of the internet on economic growth and prosperity," McKinsey Global Inst, Tech. Rep., Oct. 2011. [Online]. Available: http://www.mckinsey.com/insights/high_tech_telecoms_internet/the_great_transformer

[61] "Internet world stats," Jul. 2012. [Online]. Available: http://www.idc.com/getdoc.jsp?containerId=prUS23624612#.UPBBOG_hKSo

[62] C. Edwards and A. Fixmer, "Pandora quarterly revenue advances helped by mobile users," May 2012. [Online]. Available: http://www.bloomberg.com/news/2013-05-23/pandora-quarterly-loss-widens-as-costs-outpace-ad-sales.html

[63] M. P. du Rausas, J. Manyika, E. Hazan, J. Bughin, M. Chui, and R. Said, "Internet matters: The net's sweeping impact on growth, jobs, and prosperity," McKinsey Global Inst, Tech. Rep., May 2011. [Online]. Available: http://www.mckinsey.com/insights/high_tech_telecoms_internet/internet_matters

[64] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Secur., Priv. (SP)*. Washington, DC: IEEE Comput. Soc., 2008, pp. 111–125. [Online]. Available: http://dx.doi.org/10.1109/SP.2008.33

[65] "Netflix prize." [Online]. Available: http://en.wikipedia.org/wiki/Netflix_Prize

[66] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private information retrieval," in *Proc. IEEE Annual Symp. Found. Comput. Sci. (FOCS)*, Milwaukee, WI, 1995, pp. 41–50.

[67] E. Kushilevitz and R. Ostrovsky, "Replication is not needed: single database, computationally-private information retrieval," in *Proc. IEEE Annual Symp. Found. Comput. Sci. (FOCS)*. IEEE Comput. Soc., 1997, pp. 364–373. [Online]. Available: http://dl.acm.org/citation.cfm?id=795663.796363

[68] M. K. Reiter and A. D. Rubin, "Crowds: Anonymity for Web transactions," *ACM Trans. Inform. Syst. Secur.*, vol. 1, no. 1, pp. 66–92, 1998.

[69] C. Chow, M. F. Mokbel, and X. Liu, "A peer-to-peer spatial cloaking algorithm for anonymous location-based services," in *Proc. ACM Int. Symp. Adv. Geogr. Inform. Syst. (GIS)*, Arlington, VA, Nov. 2006, pp. 171–178.

[70] D. Rebollo-Monedero, J. Forné, A. Solanas, and T. Martnez-Ballesté, "Private location-based information retrieval through user collaboration,"

Comput. Commun., vol. 33, no. 6, pp. 762–774, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.comcom.2009.11.024

[71] C. A. W. Citteur and L. C. R. J. Willenborg, "Public use microdata files: Current practices at national statistical bureaus," *J. Official Stat.*, vol. 9, no. 4, pp. 783–794, 1993.

[72] J. Domingo-Ferrer and J. M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 1, pp. 189–201, 2002.

[73] J. Domingo-Ferrer and V. Torra, "Ordinal, continuous and heterogenerous k-anonymity through microaggregation," *Data Min., Knowl. Disc.*, vol. 11, no. 2, pp. 195–212, 2005.

[74] A. Solanas, A. Martínez-Ballesté, and J. Domingo-Ferrer, "VMDAV: A multivariate microaggregation with variable group size," in *Proc. Comput. Stat. (COMPSTAT).* Rome, Italy: Springer-Verlag, 2006.

[75] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge, UK: Cambridge University Press, 2004.

[76] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.

[77] W. M. Grossman, "alt.scientology.war," 1996. [Online]. Available: www.wired.com/wired/archive/3.12/alt.scientology.war_pr.html

[78] "AOL search data scandal," Aug. 2006. [Online]. Available: http://en.wikipedia.org/wiki/AOL_search_data_scandal

[79] X. Shen, B. Tan, and C. Zhai, "Privacy protection in personalized search," *ACM Spec. Interest Group Inform. Retrieval (SIGIR) Forum*, vol. 41, no. 1, pp. 4–17, Jun. 2007. [Online]. Available: http://doi.acm.org/10.1145/1273221.1273222

[80] P. Srisuresh and M. Holdrege, "IP Network Address Translator (NAT) Terminology and Considerations," RFC 2663 (Informational), Internet Engineering Task Force, Aug. 1999. [Online]. Available: http://www.ietf.org/rfc/rfc2663.txt

[81] R. Droms, "Dynamic Host Configuration Protocol," RFC 2131 (Draft Standard), Internet Engineering Task Force, Mar. 1997, updated by RFCs 3396, 4361, 5494, 6842. [Online]. Available: http://www.ietf.org/rfc/rfc2131.txt

[82] R. Ostrovsky and W. E. Skeith III, "A survey of single-database PIR: Techniques and applications," in *Proc. Int. Conf. Practice, Theory Public-Key Cryptogr. (PKC)*, ser. Lecture Notes Comput. Sci. (LNCS), vol. 4450. Beijing, China: Springer-Verlag, Sep. 2007, pp. 393–411.

[83] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.-L. Tan, "Private queries in location based services: Anonymizers are not necessary," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Vancouver, Canada, Jun. 2008, pp. 121–132.

[84] S. Yekhanin, "Private information retrieval," *Commun. ACM*, vol. 53, no. 4, pp. 68–73, Apr. 2010. [Online]. Available: http://doi.acm.org/10.1145/1721654.1721674

[85] J. Canny, "Collaborative filtering with privacy via factor analysis," in *Proc. ACM SIGIR Conf. Res., Develop. Inform. Retrieval.* Tampere, Finland: ACM, 2002, pp. 238–245.

[86] J. F. Canny, "Collaborative filtering with privacy," in *Proc. IEEE Symp. Secur., Priv. (SP)*, 2002, pp. 45–57.

[87] W. Ahmad and A. Khokhar, "An architecture for privacy preserving collaborative filtering on Web portals," in *Proc. IEEE Int. Symp. Inform. Assurance, Secur. (IAS).* Washington, DC: IEEE Comput. Soc., 2007, pp. 273–278.

[88] J. Zhan, C. L. Hsieh, I. C. Wang, T. S. Hsu, C. J. Liau, and D. W. Wang, "Privacy-preserving collaborative recommender systems," *IEEE Trans. Syst. Man, Cybern.*, vol. 40, no. 4, pp. 472–476, Jul. 2010.

[89] A. Erola, J. Castellà-Roca, A. Viejo, and J. M. Mateo-Sanz, "Exploiting social networks to provide privacy in personalized Web search," *J. Syst., Softw.*, vol. 84, no. 10, pp. 1734–745, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0164121211001117

[90] D. Rebollo-Monedero, J. Forné, and J. Domingo-Ferrer, "Coprivate query profile obfuscation by means of optimal query exchange between users," *IEEE Trans. Depend., Secure Comput.*, 2012. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TDSC.2012.16

[91] J. Domingo-Ferrer and Ú. González-Nicolás, "Rational behavior in peer-to-peer profile obfuscation for anonymous keyword search," *Inform. Sci.*, vol. 185, no. 1, pp. 191–204, 2012.

[92] B. Miller, N. Bradley, and J. A. K. J. Riedl, "Pocketlens: Toward a personal recommender system," *ACM Trans. Inform. Syst.*, vol. 22, no. 3, pp. 437–476, Jul. 2004.

[93] S. Berkovsky, Y. Eytani, T. Kuflik, and F. Ricci, "Enhancing privacy and preserving accuracy of a distributed collaborative filtering," in *Proc. ACM Conf. Recommender Syst. (RecSys)*. ACM, 2007, pp. 9–16.

[94] H. Polat and W. Du, "Privacy-preserving collaborative filtering using randomized perturbation techniques," in *Proc. SIAM Int. Conf. Data Min. (SDM)*. IEEE Comput. Soc., 2003.

[95] D. Rebollo-Monedero and J. Forné, "Optimal query forgery for private information retrieval," *IEEE Trans. Inform. Theory*, vol. 56, no. 9, pp. 4631–4642, 2010.

[96] Y. Elovici, B. Shapira, and A. Maschiach, "A new privacy model for hiding group interests while accessing the Web," in *Proc. Workshop Priv. Electron. Soc.* Washington, DC: ACM, 2002, pp. 63–70.

[97] Y. Elovici, B. Shapira, and A. Maschiach, "A new privacy model for Web surfing," in *Proc. Int. Workshop Next-Gen. Inform. Technol., Syst. (NGITS).* Springer-Verlag, 2002, pp. 45–57.

[98] T. Kuflik, B. Shapira, Y. Elovici, and A. Maschiach, "Privacy preservation improvement by learning optimal profile generation rate," in *User Modeling*, ser. Lecture Notes Comput. Sci. (LNCS), vol. 2702. Springer-Verlag, 2003, pp. 168–177.

[99] B. Shapira, Y. Elovici, A. Meshiach, and T. Kuflik, "PRAW – The model for PRivAte Web," *J. Amer. Soc. Inform. Sci., Technol.*, vol. 56, no. 2, pp. 159–172, 2005.

[100] Y. Elovici, C. Glezer, and B. Shapira, "Enhancing customer privacy while searching for products and services on the World Wide Web," *Internet Res.*, vol. 15, no. 4, pp. 378–399, 2005.

[101] Y. Elovici, B. Shapira, and A. Meshiach, "Cluster-analysis attack against a private Web solution (PRAW)," *Online Inform. Rev.*, vol. 30, pp. 624–643, 2006.

[102] S. Ye, F. Wu, R. Pandey, and H. Chen, "Noise injection for search privacy protection," in *Proc. Int. Conf. Comput. Sci., Eng.* IEEE Comput. Soc., 2009, pp. 1–8.

[103] D. C. Howe and H. Nissenbaum, *Lessons from the Identity Trail: Privacy, Anonymity and Identity in a Networked Society.* NY: Oxford Univ. Press, 2009, ch. TrackMeNot: Resisting surveillance in Web search, pp. 417–436. [Online]. Available: http://mrl.nyu.edu/~dhowe/trackmenot

[104] R. Chow and P. Golle, "Faking contextual data for fun, profit, and privacy," in *Proc. Workshop Priv. Electron. Soc.* ACM, 2009, pp. 105–108. [Online]. Available: http://doi.acm.org/10.1145/1655188.1655204

[105] J. Domingo-Ferrer, A. Solanas, and J. Castellà-Roca, "$h(k)$-private information retrieval from privacy-uncooperative queryable databases," *Online Inform. Rev.*, vol. 33, no. 4, pp. 720–744, 2009.

[106] E. Balsa, C. Troncoso, and C. Daz, "OB-PWS: Obfuscation-based private Web search," in *Proc. IEEE Symp. Secur., Priv. (SP).* IEEE Comput. Soc., 2012, pp. 491–505.

[107] Y. Xu, K. Wang, B. Zhang, and Z. Chen, "Privacy-enhancing personalized Web search," in *Proc. Int. WWW Conf.* ACM, 2007, pp. 591–600.

[108] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in *Proc. IEEE Int. Conf. Data Min. (ICDM).* Washington, DC: IEEE Comput. Soc., 2003, pp. 99–106.

[109] Z. Huang, W. Du, and B. Chen, "Deriving private information from randomized data," in *Proc. ACM SIGMOD Int. Conf. Manage. Data.* ACM, 2005, pp. 37–48.

[110] H. Polat and W. Du, "SVD-based collaborative filtering with privacy," in *Proc. ACM Int. Symp. Appl. Comput. (SASC).* ACM, 2005, pp. 791–795.

[111] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Santa Barbara, CA, 2001, pp. 247–255.

[112] T. B. Jabine, "Statistical disclosure limitation practices at united states statistical agencies," *J. Official Stat.*, vol. 9, no. 2, pp. 427–454, 1993.

[113] L. Willenborg and T. DeWaal, *Elements of Statistical Disclosure Control.* New York: Springer-Verlag, 2001.

[114] J. Domingo-Ferrer and V. Torra, "A critique of k-anonymity and some of its enhancements," in *Proc. Workshop Priv., Secur., Artif. Intell. (PSAI)*, Barcelona, Spain, 2008, pp. 990–993.

[115] J. F. Raymond, "Traffic analysis: Protocols, attacks, design issues and open problems," in *Proc. Design. Priv. Enhanc. Technol.: Workshop Design Issues Anon., Unobser.* Springer-Verlag, 2001, pp. 10–29.

[116] C. E. Shannon, "Communication theory of secrecy systems," Bell Syst., Tech. J., 1949.

[117] A. Wyner, "The wiretap channel," Bell Syst., Tech. J. 54, 1975.

[118] I. Csiszár and J. Körner, "Broadcast channels with confidential messages," *IEEE Trans. Inform. Theory*, vol. 24, pp. 339–348, May 1978.

[119] A. Pfitzmann and M. Hansen, "A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management," Aug. 2010, v0.34. [Online]. Available: http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf

[120] B. Gierlichs, C. Troncoso, C. Díaz, B. Preneel, and I. Verbauwhede, "Revisiting a combinatorial approach toward measuring anonymity," in *Proc. Workshop Priv. Electron. Soc.* ACM, 2008, pp. 111–116.

[121] M. Halkidi and I. Koutsopoulos, "A game theoretic framework for data privacy preservation in recommender systems," in *Proc. European Mach. Learn., Prin., Pract. Knowl. Disc. Databases (ECML PKDD).* Springer-Verlag, 2011, pp. 629–644.

[122] M. Duckham, K. Mason, J. Stell, and M. Worboys, "A formal approach to imperfection in geographic information," *Comput., Environ., Urban Syst.*, vol. 25, no. 1, pp. 89–103, 2001.

[123] E. L. Lehmann, *Theory of Point Estimation.* New York: Springer-Verlag, 1983.

[124] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag, 1985.

[125] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[126] P. H. Algoet and T. M. Cover, "A sandwich proof of the Shannon-McMillan-Breiman theorem," *Annals Prob.*, vol. 16, no. 2, pp. 899–909, 1988.

[127] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *IRE Nat. Conv. Rec.*, vol. 7 Part 4, 1959, pp. 142–163.

[128] R. Shokri, G. Theodorakopoulos, J. Y. L. Boudec, and J. P. Hubaux, "Quantifying location privacy," in *Proc. IEEE Symp. Secur., Priv. (SP)*. Washington, DC, USA: IEEE Comput. Soc., 2011, pp. 247–262.

[129] G. Danezis, "Statistical disclosure attacks: Traffic confirmation in open environments," in *Proc. Secur., Priv., Age Uncertainty, (SEC)*, Athens, Greece, May 2003, pp. 421–426.

[130] D. Rebollo-Monedero, J. Parra-Arnau, and J. Forné, "Un criterio de privacidad basado en teoría de la información para la generación de consultas falsas," in *Proc. XI Spanish Meeting Cryptology, Inform. Security (RECSI)*, Tarragona, Spain, Sep. 2010, pp. 129–134.

[131] M. Hildebrandt and S. Gutwirth, Eds., *Profiling the European Citizen: Cross-Disciplinary Perspectives*. Springer-Verlag, 2008.

[132] M. Hildebrandt, J. Backhouse, V. Andronikou, E. Benoist, A. Canhoto, C. Diaz, M. Gasson, Z. Geradts, M. Meints, T. Nabeth, J. P. V. Bendegem, S. V. der Hof, A. Vedder, and A. Yannopoulos, "Descriptive analysis and inventory of profiling practices – deliverable 7.2," Future Identity Inform. Soc. (FIDIS), Tech. Rep., 2005.

[133] G. Elmer, *Profiling Machines: Mapping the Personal Information Economy*. MIT Press, Jan. 2004.

[134] A. Domingos and J. Backhouse, "Constructing categories, construing signs—analysing differences in suspicious transaction reporting practice," in *Inform. Syst. Cogn. Res. Exchange (IS-CORE)*, Jan. 2004.

[135] M. Volkamer, *Data Protection in a Profiled World*, S. Gutwirth, Y. Poullet, and P. Hert, Eds. Springer-Verlag, 2010.

[136] N. Andrade, *Privacy and Identity Management for Life.* Springer-Verlag, 2011, ch. Data Protection, Privacy and Identity: Distinguishing Concepts and Articulating Rights, pp. 90–107. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-20769-3_8

[137] D. Lyon, Ed., *Surveillance as Social Sorting: Privacy, Risk and Automated Discrimination.* Routledge, Dec. 2002.

[138] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," in *The adaptive web*, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Springer-Verlag, 2007, pp. 325–341.

[139] J. Liu, P. Dolan, and E. R. Pedersen, "Personalized news recommendation based on click behavior," in *Proc. Int. Conf. Intell. User Interf. (IUI).* ACM, 2010, pp. 31–40.

[140] V. Toubiana, A. Narayanan, D. Boneh, H. Nissenbaum, and S. Barocas, "Adnostic: Privacy preserving targeted advertising," in *Proc. IEEE Symp. Netw. Distrib. Syst. Secur. (SNDSS)*, 2010, pp. 1–21.

[141] M. Fredrikson and B. Livshits, "RePriv: Re-envisioning in-browser privacy," in *Proc. IEEE Symp. Secur., Priv. (SP)*, May 2011, pp. 131–146.

[142] A. Viejo, D. Sánchez, and J. Castellà-Roca, "Using profiling techniques to protect the user's privacy in twitter," in *Proc. Int. Conf. Model. Decisions Artif. Intell.* Springer-Verlag, 2012, pp. 161–172.

[143] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli, *The adaptive Web*. Springer-Verlag, 2007, ch. User profiles for personalized information access, pp. 54–89.

[144] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proc. IEEE*, vol. 70, no. 9, pp. 939–952, Sep. 1982.

[145] L. Brillouin, *Science and Information Theory*. New York: Academic-Press, 1962.

[146] E. T. Jaynes, *Papers on Probability, Statistics and Statistical Physics*. Dordrecht: Reidel, 1982.

[147] J. P. Burg, "Maximum entropy spectral analysis," Ph.D. dissertation, Stanford Univ., 1975.

[148] A. L. Berger, J. della Pietra, and A. della Pietra, "A maximum entropy approach to natural language processing," *MIT Comput. Ling.*, vol. 22, no. 1, pp. 39–71, Mar. 1996.

[149] C. Díaz, "Anonymity and privacy in electronic services," Ph.D. dissertation, Katholieke Univ. Leuven, Dec. 2005.

[150] E. T. Jaynes, "Information theory and statistical mechanics II," *Phys. Review Ser. II*, vol. 108, no. 2, pp. 171–190, 1957.

[151] P. A. Olsen and S. Dharanipragada, "An efficient integrated gender detection scheme and time mediated averaging of gender dependent acoustic models," in *Proc. European Conf. Speech Commun., Technol. (EUROSPEECH)*, 2003, pp. 2509–2512.

[152] H. Printz and P. Olsen, "Theory and practice of acoustic confusability," *Comput. Speech, Lang.*, vol. 16, no. 1, pp. 131–164, 2002.

[153] J. Silva and S. Narayanan, "Average divergence distance as a statistical discrimination measure for hidden markov models," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 890–906, 2006.

[154] Q. Huo and W. Li, "A dtw-based dissimilarity measure for left-to-right hidden markov models and its application to word confusability analysis," in *Proc. Interspeech*, 2006, pp. 2338–2341.

[155] J. Goldberger, S. Gordon, and H. Greenspan, "An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures," in *Proc. Int. Conf. Comput. Vision(ICCV)*, 2003, pp. 487–493.

[156] D. Olszewski, "Fraud detection in telecommunications using kullback-leibler divergence and latent dirichlet allocation," in *Proc. Adap., Nat. Comput. Alg.* Springer-Verlag, 2011, pp. 71–80.

[157] Q. Mei and C. Zhai, "Discovering evolutionary theme patterns from text: an exploration of temporal text mining," in *Proc. ACM SIGKDD Int. Conf. Knowl. Disc., Data Min. (KDD)*. ACM, 2005, pp. 198–207.

[158] K. Waikit and M. Lik, "An information theoretic approach for ontology-based interest matching," in *Proc. Workshop Ontology Learn.*, A. Maedche, S. Staab, C. Nedellec, and E. H. Hovy, Eds., vol. 38, 2001. [Online]. Available: http://SunSITE.Informatik.RWTH-Aachen.DE/Publications/CEUR-WS/Vol-38/koh.pdf

[159] T. C. Zhou, H. Ma, M. R. Lyu, and I. King, "UserRec: A user recommendation framework in social tagging systems," in *Proc. AAAI Conf. Artif. Intell.*, M. Fox and D. Poole, Eds. Assoc. Adv. Artif. Intell., 2010.

[160] A. Livne, M. P. Simmons, E. Adar, and L. A. Adamic, "The party is over here: Structure and content in the 2010 election," in *Proc. Int. Conf. Weblogs, Social Media*, L. A. Adamic, R. A. Baeza-Yates, and S. Counts, Eds. Assoc. Adv. Artif. Intell., 2011.

[161] L. Feigenbaum, I. Herman, T. Hongsermeier, E. Neumann, and S. Stephens, "The semantic web in action," *Scient. Amer.*, vol. 297, pp. 90–97, Dec. 2007.

[162] M. Baldoni, C. Baroglio, and N. Henze, "Personalization for the semantic web," in *Proc. Reasoning Web Summer School 2005*, 2005, pp. 173–212.

[163] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds., *Recommender Systems Handbook*. Springer-Verlag, 2011.

[164] E. Michlmayr and S. Cazer, "Learning user profiles from tagging data and leveraging them for personal(ized) information access," in *Proc. Workshop Tagging and Metadata for Social Inform. Org. Workshop in Int. WWW Conf.*, 2007.

[165] A. John and D. Seligmann, "Collaborative tagging and expertise in the enterprise," in *Proc. Col. Web Tagging Workshop WWW*, 2006.

[166] "Fact sheet 18: Online privacy: Using the internet safely," Privacy Rights Clearinghouse, Tech. Rep., Jul. 2013.

[167] N. Bilton and B. Stelter, "Sony says playstation hacker got personal data," Apr. 2011. [Online]. Available: http://www.nytimes.com/2011/04/27/technology/27playstation.html

[168] S. Ovide, "Evernote discloses security breach," Mar. 2013. [Online]. Available: http://online.wsj.com/article/SB10001424127887323478304578336373531236296.html

[169] S. T. Peddinti and N. Saxena, "On the privacy of web search based on query obfuscation: a case study of trackmenot," in *Proc. Int. Symp. Priv. Enhanc. Technol. (PETS)*. Springer-Verlag, 2010, pp. 19–37.

[170] R. Al-Rfou', W. Jannen, and N. Patwardhan, "Trackmenot-so-good-after-all," Stony Brook Univ., Tech. Rep., 2012.

[171] "European data protection supervisor," May 2013. [Online]. Available: http://www.edps.europa.eu

[172] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[173] "Knowledge and Data Engineering Group, University of Kassel: Benchmark folksonomy data from BibSonomy," Dec. 2007. [Online]. Available: http://www.kde.cs.uni-kassel.de/bibsonomy/dumps

[174] M. Grahl, A. Hotho, and G. Stumme, "Conceptual clustering of social bookmarking sites," in *Proc. Int. Conf. Knowl. Manage. (I-KNOW)*, Graz, Austria, Sep. 2007, pp. 356–364.

[175] L. Specia and E. Motta, "Integrating folksonomies with the semantic web," in *Proc. Int. Semantic Web Conf.*, 2007, pp. 624–639.

[176] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 129–137, Mar. 1982.

[177] T. Berners-Lee, R. Fielding, and L. Masinter, "Uniform Resource Identifier (URI): Generic Syntax," RFC 3986 (INTERNET STANDARD), Internet Engineering Task Force, Jan. 2005, updated by RFC 6874. [Online]. Available: http://www.ietf.org/rfc/rfc3986.txt

[178] D. Maltz and K. Ehrlich, "Pointing the way: active collaborative filtering," in *Procc SIGCHI Conf. Hum. Fact. Comput. Syst.* ACM, 1995, pp. 202–209.

[179] B. Carminati, E. Ferrari, and A. Perego, "Combining social networks and Semantic Web technologies for personalizing Web access," in *Collaborative Computing: Networking, Applications and Worksharing*, ser. LNICST. Springer, 2009, vol. 10, pp. 126–144.

[180] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," in *WPES 2005.* ACM Press, 2005, pp. 71–80.

[181] S. B. Barnes, "A privacy paradox: Social networking in the United States," *First Monday*, vol. 11, no. 9, Sep. 2006.

[182] P. Lops, M. Gemmis, and G. Semeraro, *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Springer-Verlag, 2011.

[183] E. Ferrari and B. Thuraisingham, "Secure database systems," in *Advanced Database Technology and Design*, M. Piattini and O. Diaz, Eds. Norwood, MA: Artech House, Inc., 2000, ch. 11, pp. 353–403.

[184] [Online]. Available: http://www.dai-labor.de/en/competence_centers/irml/datasets/

[185] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stum, "Evaluating similarity measures for emergent semantics of social tagging," in *Proc. Int. WWW Conf.* ACM, 2009, pp. 641–650.

[186] D. S. Hochbaum and D. B. Shmoys, "A best possible heuristic for the k-center problem," *Math. Oper. Res.*, vol. 10, no. 2, pp. 180–184, 1985.

[187] G. Hamerly and C. Elkan, "Alternatives to the k-means algorithm that find better clusterings," in *Proc. Int. Conf. Inform., Knowl. Manage. (CIKM)*. ACM, 2002, pp. 600–607.

[188] W. E. Mackay, "Triggers and barriers to customizing software," in *Procc SIGCHI Conf. Hum. Fact. Comput. Syst.* ACM, 1991, pp. 153–160.

[189] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, 2005.

[190] S. Fox, "Trust and privacy online: Why americans want to rewrite the rules," Pew Internet, Amer. Life Project, Res. Rep., Aug. 2000.

[191] D. L. Hoffman, T. P. Novak, and M. Peralta, "Building consumer trust online," *Commun. ACM*, vol. 42, no. 4, pp. 80–85, Apr. 1999.

[192] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "A privacy-protecting architecture for recommendation systems via the suppression of ratings," *Int. J. Security, Appl. (IJSIA)*, vol. 6, no. 2, pp. 61–80, Apr. 2012.

[193] T. Ibaraki and N. Katoh, *Resource allocation problems: algorithmic approaches.* MIT Press, 1988.

[194] R. H. Byrd, M. E. Hribar, and J. Nocedal, "An interior point algorithm for large-scale nonlinear programming," *(SIAM) J. Optim.*, vol. 9, no. 4, pp. 877–900, 1999.

[195] R. H. Byrd, J. C. Gilbert, and J. Nocedal, "A trust region method based on interior point techniques for nonlinear programming," *Math. Program.*, vol. 89, no. 1, pp. 149–185, 2000.

[196] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, "An interior algorithm for nonlinear optimization that combines line search and trust region steps," *Math. Program.*, vol. 107, no. 3, pp. 391–408, 2006.

[197] "GroupLens research." [Online]. Available: http://www.grouplens.org

[198] "MovieLens 10M data set," Aug. 2011. [Online]. Available: http://www.grouplens.org/system/files/ml-10m-README.html