Future Generation Computer Systems 33 (2014) 53–63

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Measuring the privacy of user profiles in personalized information systems☆

Javier Parra-Arnau *, David Rebollo-Monedero, Jordi Forné

Department of Telematics Engineering, Universitat Politècnica de Catalunya (UPC), C. Jordi Girona 1-3, 08034 Barcelona, Spain

ARTICLE INFO

Article history:
Received 15 June 2012
Received in revised form
19 November 2012
Accepted 4 January 2013
Available online 17 January 2013

Keywords:
Personalized information systems
User profiling
Privacy-enhancing technologies
Privacy criterion
Shannon's entropy
Kullback–Leibler divergence

ABSTRACT

Personalized information systems are information-filtering systems that endeavor to tailor information-exchange functionality to the specific interests of their users. The ability of these systems to profile users is, on the one hand, what enables such intelligent functionality, but on the other, the source of innumerable privacy risks. In this paper, we justify and interpret KL divergence as a criterion for quantifying the privacy of user profiles. Our criterion, which emerged from previous work in the domain of information retrieval, is here thoroughly examined by adopting the beautiful perspective of the method of types and large deviation theory, and under the assumption of two distinct adversary models. In particular, we first elaborate on the intimate connection between Jaynes' celebrated method of entropy maximization and the use of entropies and divergences as measures of privacy; and secondly, we interpret our privacy metric as false positives and negatives in a binary hypothesis testing.

1. Introduction

Recent years have witnessed the accelerated growth of a rich variety of personalized information systems of unprecedented sophistication, which have been integrating seamlessly into our daily lives. Examples of these systems comprise personalized Web search and news, resource tagging in the semantic Web and multimedia recommendation systems. The key enabling technology of such systems is personalization, a research area that has received great attention lately and whose aim is to tailor information-exchange functionality to the specific interests of their users. To accomplish this functionality, most personalized information 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.

Personalized services therefore allow 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. Besides, the enrichment of these services with data from social networks creates additional opportunities with respect to information sharing but, at the same time, increases the user privacy risks.

But the advent of these information systems is not only changing people's habits and stressing our concerns about privacy—it is also leading to a profound transformation of the traditional business model. As a matter of fact, the technologies enabling personalization as a solution of the one-size-fits- all are contributing to unprecedented performance improvements in large business and small and medium enterprises. These technologies are having an impact not only on how products are sold but also, and more importantly, on how companies approach users in a personalized manner, attending their specific and particular needs more effectively. Amazon, for example, who invented item-to-item collaborative-filtering algorithms [1], one of the most widely used personalization techniques, is visited by more than 93 million users per day. Another example that illustrates this transformation is Facebook, which will surpass 4.27 billion dollars in revenue this year, 89% of its income will come from selling access to their data so that advertisers can personalize their digital content [2]. The information used to provide such personalization ranges from location, education, likes and interests, to friends and relationship status [3]. Pushed by these personalization techniques, online advertising is expected to grow by 10.6% each year through 2016, with $70.9 billion in global advertising during 2011.

The impact of personalized information systems on society and economy is therefore undeniable. Nowadays, personalization is present in a myriad of applications we frequently use on the Internet, when submitting queries to a Web search engine, rating products at an online store or posting tags in a collaborative

☆ The material in this paper has been published in part in the proceedings of the International Conference on Security Technology (SecTech), Jeju, South Korea, Dec. 2011.

* Corresponding author. Tel.: +34 93 401 7041.
E-mail addresses: javier.parra@entel.upc.edu (J. Parra-Arnau), david.rebollo@entel.upc.edu (D. Rebollo-Monedero), jforne@entel.upc.edu (J. Forné).

0167-739X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2013.01.001

tagging system. But this is only the tip of the iceberg – in the near future a much wider spectrum of services such as personalized medicine will become a reality. However, we must not forget that the cornerstone of these current and future systems is the ability to profile users, which poses serious threats to one of our fundamental rights – the right to privacy.

1.1. The need for measuring the privacy of user profiles

A variety of privacy-enhancing technologies (PETs) have been proposed to enable the provision of new services and functionalities aimed at mitigating those privacy threats. Anonymous-communication networks [4,5], anonymous credentials [6], anonymous electronic cash [7], multiparty computation [8] and oblivious transfer protocols [9] are some examples of general-purpose PETs whose development roughly originates from the fields of security and cryptography. Unfortunately, these technologies have not yet gained wide adoption. This is because it remains unclear whether their overall benefits outweigh their typically costly deployment and/or integration, as well as the operational cost that arises due to the fact that PETs typically come with penalties in terms of utility and performance, when compared to more privacy-invasive alternatives [10].

Assessing the privacy provided by a PET is, therefore, crucial to both determine its overall benefit and compare its effectiveness with other technologies. In other words, privacy metrics, accompanied with utility metrics, provide a quantitative means of contrasting 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.

A great effort has been devoted to the investigation of privacy metrics, especially in the scenario of statistical disclosure control (SDC) [11–18]. Although some of those metrics might be applied to our context of personalized information systems, the fact is that there are few proposals specifically conceived for measuring the privacy of user profiles; and not only that, but also they are often not appropriately justified and are defined in an ad hoc manner [19–29].

1.2. Contribution and organization

This paper approaches the fundamental problem of proposing quantitative measures of the privacy of user profiles. We have established the critical importance of quantifying privacy in order to assess, compare, improve and optimize privacy-enhancing technologies. In application scenarios involving user profiles, there exists no general framework systematically leading to a formal metric, but merely ad hoc proposals for a few specific applications. The main contribution of this work identifies the need for such quantitative measures of privacy for user profiles in personalized information systems.

Bearing this need in mind, we explore the privacy risks inherent in such systems, and then provide a thorough justification of a common, generalized framework to measure those risks. Our justification relies on fundamental principles from information theory and statistics, thereby drawing intriguing links between said fields and information privacy. In practice, the impact of a privacy mechanism on information-exchange functionality, traffic and processing overhead, and general usability cannot simply be overlooked. We would like to stress that quantitative measures of privacy on the one hand, and utility on the other, allow researchers to optimize their technologies in terms of the trade-off posed by these contrasting aspects.

Specifically, we tackle two adversary models. The first model considers 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 Kullback–Leibler (KL) divergence as a measure of privacy is justified by elaborating on Jaynes' rationale behind entropy-maximization methods and the method of types, a justification that we introduced in [30]. Under the latter adversary model, a riveting argument in favor of divergence stems from hypothesis testing and large deviation theory.

Section 2 illustrates the privacy concerns that arise in the motivating scenario of this work. Section 3 examines several approaches to model user profiles and specifies the adversary capabilities assumed in our interpretation of divergence as a measure of privacy. The use of divergence is justified on the one hand in Section 4 when the attacker strives to identify users, and on the other in Section 5 when the adversary endeavors to classify users. Section 6 then overviews some of the most relevant privacy criteria in the literature. Finally, conclusions are drawn in Section 7.

2. Illustration of privacy risks in personalized information systems

In this section, we carefully examine the privacy risks posed by the personalized information systems that proliferate these days on the Internet. The following example illustrates those privacy threats.

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 the 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 as a reflex response. Of course, her action will be diligently reflected back in her profile. In a way,