Privacy in Big Data

Benjamin Habegger[1], Omar Hasan[1], Thomas Cerqueus[1], Lionel Brunie[1],

Nadia Bennani[1], Harald Kosch[2], Ernesto Damiani[3]

[1] University of Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France

{benjamin.habegger, omar.hasan, thomas.cerqueus, lionel.brunie, nadia.bennani}@insa-lyon.fr

[2] Faculty of Informatics and Mathematics, University of Passau, Germany

harald.kosch@uni-passau.de

[3] Department of Computer Technology, University of Milan, Italy

ernesto.damiani@unimi.it

July 7, 2015

Contents

Chapter 1

Privacy in Big Data

1.1 Introduction

Personalization consists of adapting outputs to a particular context and user. It may rely on user profile attributes such as the geographical location, academic and professional background, membership in groups, interests, preferences, opinions, etc. Personalization is used by a variety of web based services for different purposes. A common form of personalization is the recommendation of items, elements or general information that a user has not yet considered but may find useful.

General purpose social networks such as Facebook.com use personalization techniques to find potential friends based on the existing relationships and group memberships of the user. Professional social networks such as LinkedIn.com exploit the skills and professional background information available in a user profile to recommend potential employees. Search engines such as Google.com use the history of user searches to personalize the current searches of the user.

Big data analysis techniques are a collection of various techniques that can be used to discover knowledge in high volume, highly dynamic, and highly varied data. Big data techniques

offer opportunities for personalization that can result in the collection of very comprehensive user profiles. Big data analysis techniques have two strengths in particular that enable collecting accurate and rich information for personalization: (1) Big data analysis techniques process unstructured data as well as structured data. Unstructured data of different varieties generated by users is growing in volume with high velocity and contains lots of useful information about the users. (2) Big data analysis techniques can process high volume data from multiple sources. This enables linking user attribute data from different sources and aggregating them into a single user profile. Moreover, user information from different sources can be correlated to validate or invalidate the information discovered from one source.

On one hand, user profiling with big data techniques is advantageous for providing better services as we have discussed above. On the other hand, user profiling poses a significant threat to user privacy. One can assume that an ethical and trustworthy service would use the information collected for personalization purposes with the user's explicit consent and only for the benefit of the user. However, services that are less inclined toward protecting user privacy, may use personalization data for a number of purposes which may not be approved by the user and which may result in loss of private information. One example is the utilization of personalization data for targeted advertising [15]. Another example is the selling of private information in user profiles to third parties for a profit. The third parties may then use the private information for commercial or even malicious purposes [35]. Privacy breaches may occur even when a service is willing to protect a user's privacy [24].

In this chapter, which is in continuation to our article [12] previously presented on the topic, we highlight some of the privacy issues related to personalization using big data techniques. We also present the use case of the EEXCESS research project[1], which aims to personalize user recommendations by making intensive use of user profiling and therefore collecting detailed information about users. In this chapter, we supplement the description of the EEXCESS project with an up to date account of real experiences gained from the project on implementing personalization with privacy.

The rest of the chapter is organized as follows. Section 1.2 discusses the objectives behind

[1]The presented work was developed within the EEXCESS project funded by the EU Seventh Framework Program, grant agreement number 600601.

personalization and how it can be achieved through different big data techniques. Section 1.3 presents an introduction to privacy and discusses a user scenario where privacy becomes an issue in the context of personalization. Section 1.4 discusses privacy challenges that may appear in systems relying on information about users and in particular, personalization systems. Section 1.5 recalls the role of the user in relationship to personalization and privacy. Section 1.6 gives an overview of the current state of privacy solutions. Section 1.7 describes the EEXCESS project, its goal of providing personalized content recommendation and the impacts considering privacy. We conclude in Section 1.8.

1.2 Personalization with big data techniques

1.2.1 What is personalization?

The goal of personalization is to provide the most adapted response to a user's current need with the least amount of explicit information provided by him/her. Many existing systems provide some form of personalization. Google search personalizes search results using information such as the user's geo-location, IP address, search history and result click-thru. Facebook provides "friend" recommendations based on a user's social network already known by the service. Many location-based services, at a very minimum, use a user's geo-location to provide results near the user's current position. Personalized advertisements and marketing solutions attempt to better understand buying habits in order to propose advertisements to users for products they could likely be interested in.

Personalization is not limited to online services. For example, medical analysis systems try to build patient profiles which are as fine-grained as possible (e.g. taking into account genetic information) in order to propose the most adapted treatment to the patient. Personalization even reaches industrial processes, e.g., the industrial process of printing. Many printing firms offer the possibility to personalize statement documents such as bank statements, with adapted advertisements and offers. With the arrival of technologies such as 3D printers, it is very likely that the near future makes even more room for personalization.

There are clear advantages to personalization. A typical example is the utilization of user profile data for targeted advertising [15]. This way, users only receive advertisements that they have the most interest in and are not overwhelmed with advertisements for products they wouldn't even consider buying. Another example is filtering out spam emails. Personalization also improves the impact of a given service. In search systems, it allows users to more quickly find the information they are looking for. More generally it relieves users from the information overload they face every day by letting the systems dig through the massive amounts of data on their behalf and letting them find the relevant data for the users.

1.2.2 How big data techniques allow for personalization?

Big data techniques are a collection of various techniques that can be used to discover knowledge in high volume, highly dynamic, and highly varied data. Big data techniques offer opportunities for personalization that can result in very comprehensive user profiles. Big data techniques have two strengths in particular that enable collecting accurate and rich information for user profiles: (1) Big data techniques process unstructured data as well as structured data. Unstructured data of different varieties generated by users is growing in volume with high velocity and contains lots of useful information about the users. (2) Big data techniques can process high volume data from multiple sources. This enables linking user attribute data from different sources and aggregating them into a single user profile. Moreover, user information from different sources can be correlated to validate or invalidate the information discovered from one source.

We list some of the big data analyses techniques below that can be used for collecting information about a user and building a user profile. An extended list of big data techniques that can be used for personalization can be found in [21].

Network analysis. Network analysis algorithms are used to discover relationships between the nodes in a graph or a network. Network analysis is particularly useful in the context of social networks where important information about the user such as his friends, co-workers, relatives, etc. can be discovered. Social network analysis can also reveal

central users in the network, i.e., users who exert the most influence over other users. This information can be used to populate the attributes of social and environmental contexts, individual characteristics, etc. in a user profile.

Sentiment analysis. Sentiment analysis is a natural language processing technique that aims to determine the opinion and subjectivity of reviewers. The Internet is replete with reviews, comments and ratings due to the growing popularity of web sites such as `Amazon.com`, `Ebay.com`, and `Epinion.com` where users provide their opinion on others users and items. Moreover, micro-blogging sites such as `Twitter.com` and social network sites such as `Facebook.com` also hold a large amount of user opinions. The goal of sentiment analysis is to classify user opinions. The classification may be a simple polarity classification, i.e., negative or positive, or a more complex one, e.g., multiple ratings. Sentiment analysis can be used to process unstructured text written by a user to discover their interests, opinions, preferences, etc. to be included into their profile.

Trust and reputation management. Trust and reputation management is a set of algorithms and protocols for determining the trustworthiness of a previously unknown user in the context of his reliability in performing some action. For example, a reputation management system could be used for computing the trustworthiness of an online vendor who may or may not deliver the promised product once he receives payment. The reputation of a user is computed as an aggregate of the feedback provided by other users in the system. Trust and reputation information can be an important part of a user profile. It can convey the user's trust in other users as well as his own reputation in various contexts. This information can be subsequently used as a basis for recommending trustworthy users and avoiding those who are untrustworthy. Trust and reputation management systems can function in conjunction with sentiment analysis for obtaining user opinions and then computing trustworthiness and reputation.

Machine learning. Machine learning is a sub-field of artificial intelligence that aims to build algorithms that can make decisions not based on explicit programming but instead based on historical empirical data. An example often cited is the algorithmic classification of email into spam and non-spam messages without user intervention. In the context of personalization, machine learning can be used for learning user behavior

by identifying patterns. Topics in machine learning include: supervised learning approaches, e.g., neural networks, parametric/non-parametric algorithms, support vector machines, etc.; and unsupervised learning approaches, e.g., cluster analysis, reduction of dimensionality, etc.

Cluster analysis. Cluster analysis is the process of classifying users (or any other objects) into smaller subgroups called clusters given a large single set of users. The clusters are formed based on the similarity of the users in that cluster in some aspect. Cluster analysis can be applied for discovering communities, learning membership of users in groups, etc. Cluster analysis can be considered as a sub-topic of machine learning.

1.3 Privacy

On one hand, personalization with big data techniques is advantageous for providing better services as we have discussed above. On the other hand, big data poses a significant threat to user privacy. One can assume that an ethical and trustworthy service providing personalization would use the information collected about users with their explicit consent. However, services that are less inclined towards protecting user privacy, may use such data for a number of purposes which may not be approved by the user and which may result in loss of private information. An example is the selling of private information to third parties for a profit. The third parties may then use the private information of the users for commercial or even malicious purposes [35].

1.3.1 What is privacy?

Depending on the application and the targeted privacy requirement we can have different levels of information disclosure. Let's take privacy preserving reputation systems (e.g. [14]) as an example. We can have five different levels for privacy depending on whether identities, votes and aggregated reputation score are disclosed and linked or not. For example, in the context of calculating the reputation of a user Alice by three other users Bob, Carol and

David, which respectively have the votes +1, +1 and -1, the reputation system may disclose the following information to Alice.

Full disclosure. All tuples (Bob,+1), (Carol,+1), (David,-1) as well as the aggregated score (+1 if sum is used) are known by Alice.

Permuted disclosure. All voters Bob, Carol, David are known by Alice as well as the scores but permuted so Alice cannot determine who voted what.

Identity disclosure. All voters Bob, Carol, David are known by Alice, however individual votes are hidden and only the aggregated score is known by Alice.

Votes disclosure. All votes are known by Alice but the voters are hidden.

Result disclosure. No details are disclosed except the aggregated score.

No disclosure. An aggregated score for Alice is calculated but she does not have access to it.

More generally, we can subdivide privacy objectives in two:

User anonymity. The first objective is preserving user anonymity. In this setting, untrusted peers should not be able to link the identity of the user to the requests that they receive. For example, if Bob is navigating the web, any request that a content provider receives should not be linkable to the real user Bob. Information such as his IP address, email identifiers, or any other such information which may help identify Bob should not be made available.

Disclosure of private information about known users. The second objective is preventing the disclosure of private information. Let's take the same example of Bob searching on the web but desiring his age to be kept private. However, let's suppose that he does not mind the origin of his query being revealed. In this case, privacy preservation does not necessarily require anonymity but rather providing guarantees that Bob's age will not be disclosed.

Ideally, the user would like to obtain quality and personalized recommendations without revealing anything about himself. Attaining such an objective means ensuring that a user remains anonymous with respect to the peers he considers non-trustworthy. Works have shown that in some restricted cases anonymization is possible [6, 10, 11]. This, however, often comes at the cost of quality or utility of the disclosed information [4, 19, 24, 31]. It may also be the case that users do not necessarily require anonymity (for example, in social networks), but rather have control over what is disclosed or not disclosed.

1.3.2 How can privacy be breached?

Depending on the definition of privacy, different techniques can be used to breach privacy even within systems which intend to protect it. We identify two types of privacy attacks: (1) "protocol" attacks are those relying on protocol exchanges between peers, in particular using connection information (IP address, cookies, etc.), to identify users or information about them; (2) "statistical" attacks are those relying on statistical techniques (in particular statistical machine learning) to analyze flows of information reaching a peer and using automated reasoning techniques to deduce user identity or private characteristics.

Protocol attacks. Protocol attacks are those relying on the fact that since a user wants to obtain an information from a peer, then the peer will have to be contacted by some means. For example, a user wanting to access a web page on "looms" will have his browser making a request to the hosting server. Having been contacted, the server has a trace of the user's IP and knows that this IP has requested the page on looms. Protection from such attacks can be obtained by using proxies but this just moves the problem of trust from the content provider to the proxy provider. It is then the proxy which must be trusted. This very basic example gives an initial intuition on the fact that protecting from protocol attacks can get complex.

Statistical attacks. Statistical attacks are those relying on the information which legitimately flows to a given peer. Even if users are protected by a privacy preserving protocol, the data which ends in the hands of a potentially malicious or curious peer

may be used to break this anonymity. For example, to be able to find interesting documents for a user, a search engine must be provided with a search query. This query in itself provides information about the user from which it originates (be it only that he is interested in the topic of the query). By correlating together the information that an untrusted peer has collected and linked together about a user, it can become possible to de-anonymize the user [27].

1.3.3 User scenario

Let us consider a user scenario in which a recommender system pushes information to a user in order to help her with her work. Alice is an economist employed by a consulting firm. She is currently working on a business plan for one of her customers on a market which is new to her. She uses a search engine that integrates a recommender system to investigate on the different actors of the market and in particular the potential competitors for her client. The recommender system component of the search engine pushes relevant content from an economic database to Alice. This content includes detailed descriptions of companies found in the target market of her client and strategic economic data.

In this scenario, the recommender system will have collected significant information about Alice: her interests (economic information), some comprehension of her goal (writing a business plan), her knowledge (expert in economics), her context of work (information about her customer, the target market, the information she has already collected, etc.). Knowing as much as possible about Alice and her customer will allow the recommender system to provide her with adapted recommendations. For example, instead of presenting general-purpose information about the market, the system will propose more detailed technical data which Alice needs and understands.

However, Alice requires that a high level of privacy is ensured by the system. In fact, she is legally-tied by a non-disclosure policy with her customer. In particular, it should not be learned that Alice's customer is taking a move toward the new market.

It would be unacceptable to Alice that any information about herself or her customer leak

out of the system. Alice's project may even be so sensitive that simply the fact that *someone* (without particularly knowing who) is setting up a business plan on the target market may be an unacceptable leak. Such a disclosure could lead to competitors taking strategic moves. This emphasizes the fact that preserving only anonymity may not be sufficient in some cases.

1.4 Privacy challenges

Providing users with quality recommendations using big data techniques is a seemingly conflicting objective with the equally important goal of privacy preservation. Even a small amount of personal information may lead to identifying a user with high probability in the presence of side channel external data [24].

Currently, systems which provide personalization, function as a black box from the user's perspective. Users do not know what is really collected about them, what is inferred about them by the system, with which other data sources their private data may be combined, what are their benefits of disclosure. Furthermore, faced with the multitude and growing number of external data sources, even limited disclosure of information to a given system may reveal enough about them for the same system to be able to infer knowledge they would have otherwise preferred to remain private. We list below some of the main categories of challenges that users face concerning their privacy in the existing systems using big data techniques for personalization.

1.4.1 Transparency

Users are often unable to monitor and follow precisely what information about them the system has collected. For example, it is common knowledge that different services, such as Google, Facebook, Amazon, etc., use big data analytics to provide personalization in many of their services. However, it is not always transparent to users what information has been collected, inferred and how it is used by whom. Even if these services wish to provide more transparency it is often technically challenging to provide tools to visualize complex

processing and manipulation (and in particular aggregation) of user information.

1.4.2 Control

Users are often unable to express their private information disclosure preferences. This can either be due to the unavailability of such options, the complexity of the provided tools or even their unawareness of privacy issues. They should be able to specify what is disclosed and how detailed the disclosure should be as well as to whom it is disclosed. A big challenge for control is that the more fine-grained privacy settings are the more complex and time consuming it becomes for users to set them. Furthermore, not all users have the same level of requirements, some desire such fine-grained control, whereas others would be satisfied with simpler high level control.

Another related issue is that users have different views on what should be private and give privacy varying importance. Some may prefer having very good personalization whereas others favor privacy. Privacy preservation is already a challenge in itself. Taking user privacy preferences into account requires that privacy algorithms should be able to dynamically adapt to the user's preferences.

1.4.3 Feedback

Users often have difficulties understanding the impacts of disclosing or not disclosing certain pieces of information on personalization. Personalization is impacted by the type, quantity and quality of information users provide. It is difficult for users to clearly perceive how their inputs impact personalization. This is amplified by the fact that often, these impacts are differed in time and their effects come only later. Also, in many cases, when they do perceive the advantages or lack of value of providing some piece of information, it is long after they have provided it. To make things worse, once the information is released, it is hard for it to be completely retracted.

1.4.4 Re-identification

Because of big data techniques, such as machine learning, very few discriminant data allow to (re)identify the user at the origin of a request. For example, it is possible for a search engine to re-identify some queries sent by a single user among all the queries. This is true even if the user is connected to the search engine via an anonymous network such as TOR [9]. This is done by using the content of the messages (rather than who they are coming from) and using classification techniques to re-identify their likely origin. This suggests that anonymous networks or query shuffling to guarantee unlinkability between users and their requests may not be enough. Therefore, within the context of personalization we are faced with a paradox: on one hand we want to adapt results to specific users, which requires discriminating the user from the others, and on the other hand, to preserve user privacy we should rather not discriminate them.

1.4.5 Discovery

Big data techniques can be utilized for discovering previously unknown information about a given individual. For example, through statistical reasoning, having access to the list of visited web sites may reveal the gender of the users even if they have not given them explicitly.

1.4.6 Privacy and utility balance

On one hand, personalization pushes towards providing discriminant data (the more the better) about users whereas privacy pushes to have non-discriminant data (the less the better). However, many personalization techniques rely on using data from similar users. If groups of similar users are sufficiently wide, it becomes difficult to distinguish users among these groups.

Ideally, privacy-preservation mechanisms should not impact the quality of personalization obtained from the user. However, this is likely not easily achievable. A less restrictive

requirement is that the privacy-preservation mechanisms should minimize the impacts of privacy-preservation on the quality of personalization. This implies, of course, being capable of measuring such quality which in itself could be a challenge.

1.4.7 Collusion

Collusion between peers is another risk for privacy. Indeed, the information which may not be individually discoverable through two uncombined sources of information, when combined through collusion, could lead to new discoveries and therefore privacy breaches.

1.4.8 Providing privacy guarantees

At all levels within the system, user privacy guarantees should be provided. This is most likely one of the hardest tasks. Indeed, as soon as information flows out of a system, sensitive information leaks become a risk. Solutions which may seem trivial, such as anonymization have been shown to be ineffective or inefficient. A well known example showing that simple anonymization is insufficient to protect privacy is the de-anonymization of the data of the Netflix contest [24]. Furthermore, Dwork [10] has shown that the published results of a statistical database may lead to privacy breaches even for users who are not originally part of the database. These examples show the difficulties which will have to be overcome in order to provide a privacy-safe system. Furthermore, these works show that research on privacy has shifted from totally preventing privacy breaches to minimizing privacy risks. One of the difficulties to overcome is to ensure that the collection of information flowing out of the system to potentially malicious peers, limits the risks in breaching any of the users' policies. It goes without saying that the attackers themselves very likely have access to big data techniques and that this aspect should be taken into account.

1.4.9 Flexible privacy policies

Users are different, in particular with respect to privacy. Some may not have any privacy concerns at all where as others may not want to disclose a single piece of information about themselves. For example, in one hypothesis, our user Alice may simply wish to remain anonymous. In another hypothesis, Alice may not be concerned by her identity being revealed, but wish that some information about her be kept private (e.g. she may wish to keep private that she is affected by a particular disease). One big challenge will be to define a policy model which allows for such flexibility and at the same time allows to ensure the policy is respected. Preventing direct disclosure of information marked private is quite straight forward. However, a real challenge is preventing the disclosure of the same information *indirectly*. Indeed, leaking other non-private information of a user's profile can lead, through inference, to unwanted disclosures.

1.4.10 Evaluating trust and reputation

What user profile information is disclosed, or at which granularity it is disclosed, may depend on the trust (with respect to privacy concerns) that the user has in a recommender system or a content provider. Calculating an entity's reputation and trustworthiness in a privacy preserving manner is thus an issue.

1.4.11 Performance

Supplementing big data techniques for personalization with privacy preserving functionalities often entails added complexity and consequently an increase in the computational resources required. It is imaginable that the raise in computational requirements may be prohibitive enough for a provider that they are forced to sacrifice privacy of users for the sake of practicality. This could be the case even when the provider has the will to implement privacy preserving features. Thus, a principal challenge in personalization with privacy is to limit the demand on computing resources while maintaining an acceptable level of performance.

1.5 User specific considerations for personalization with privacy

Personalization with privacy aims to provide users a service that better fits their needs. Users therefore need to be implied in the process. In particular, users play an important role in information disclosure, which, of course, has an impact on personalization with privacy.

1.5.1 Information disclosure

Even though privacy is often considered as a technical issue, it is also important to understand the user's perspective. In a study on user behavior, [17] have shown that user's globally tend to disclose less information when users are faced with a system explicitly talking about privacy. The interpretation given is that when privacy issues are put in the focus, users tend to become more suspicious and therefore leak less information. This is quite a paradox as a system willing to be transparent about privacy finds itself disadvantaged with respect to one not mentioning privacy at all. However, the same work studies how to improve disclosure (compared to a system not mentioning privacy issues). Giving the same explanations to everyone will lead to the tendency of users disclosing less because of the invocation of privacy. However, adapting explanations to the users can allow to improve disclosure. For example, within the test groups of [17], giving an explanation to men about what the data will be used for, and giving information to women about the percentage of users the data will be disclosed to, tended to globally improve disclosure.

1.5.2 Impact of including users in the process

A system aiming to have its users disclose information willingly and at the same time respect their privacy must have solutions which adapt to their needs. Furthermore, giving high and precise control to users can on one hand show a will for transparency from the service provider. However, on the other hand, this may make the system seem too complex. Therefore, users should be provided with a system allowing them to set their privacy settings *simply* but without losing *flexibility*. To this effect, users should be able to specify

their privacy concerns at a high level, but also be allowed more fine grained settings.

Another important aspect to consider is providing users with elements to understand the effects of disclosing information. As discussed previously, this involves providing the appropriate explanations to the appropriate users. In our user scenario, the objective of user information disclosure is mainly to improve the quality of the recommendations for the user. This can, for example, be obtained through a tool allowing to compare results using different privacy settings.

Given a user's preferences it is then necessary to have a system capable of optimizing the use of the disclosed information. In the user scenario, this means that the quality of the recommendations should be maintained as close as possible to those that the user could have expected with a more detailed profile. Furthermore, providing recommendation quality will also rely on user profiling. Such deep user profiling entails many privacy concerns. Indeed, while users are likely to be interested in having very precise recommendations, they may not at the same time be willing that a third-party collects private information about them.

1.6 Privacy solutions

There is a significant amount of research currently in progress to achieve the goal of preserving user privacy while collecting personal information. Big data techniques offer excellent opportunities for more accurate personalization. However, privacy is an issue that can hinder acceptance by users of personalization with big data techniques. Therefore, there is a need to develop big data techniques that can collect information for user profiles while respecting the privacy of the users. Privacy preserving big data techniques for personalization would raise the confidence of users toward allowing services to collect data for personalization purposes. Below, we list some of the works on privacy preservation in domains related to big data.

1.6.1 Differential privacy

In the domain of statistical databases, a major shift occurred with the work of Dwork and the introduction of differential privacy [10, 11, 25]. Through a theoretical framework, the authors demonstrate that, as soon as we consider external knowledge, privacy breaches can occur even for people who do not participate in a statistical database. This has introduced a shift in the way privacy is perceived. The objective is no longer to preserve privacy in an absolute manner, but rather limit the risk of increasing the privacy breach for a participant of a statistical database. To this effect, differentially private mechanisms are those that ensure that the statistical outputs of two databases which are only different by a single participant return similar statistical results. This most often consists in adding sufficient noise to the outputs. Even though there are situations in which differential privacy is attainable, in particular count queries, there are many constraints imposed by differential privacy [4, 31]. In particular, in situations which should allow multiple queries, noise must be augmented proportionally to the number of queries to prevent noise reduction techniques to be applied. However, adding too much noise can deprive the outputs of the system of any utility. Therefore much research is ongoing to evaluate the trade-offs between privacy and utility [30]. However, in practice, differential privacy can render some subsets of the randomized data less useful while poorly preserving the privacy of specific individuals. This has been demonstrated for instance in [31]. Thus, differential privacy preserving techniques still have much to achieve in order to render personal information of users truly private.

1.6.2 k-anonymity

Recommender systems need to gather massive instances of past user interactions and their ratings about objects that they were concerned with. This allows them to propose a selection of predicted objects to a current user, based on profile similarity analysis with the current user, using techniques like collaborative filtering. While this allows having good recommendation quality, it also creates user privacy concerns. k-anonymity is one of the well-known techniques to preserve user privacy. The recommender in this case should ensure that each selected object has been selected by at least k users and that each object has been rated sim-

ilarly by at least k users. This allows avoiding structure-based and label-based attacks [7]. Several methods have been proposed to ensure k-anonymity. We can cite [1, 7, 20, 29, 33]. Many solutions are aimed at resolving k-anonymity problems in databases [1, 29, 33]. [7, 20] both proposed using k-anonymity for privacy preserving recommenders. In both, past user ratings are represented using a bi-partite graph, where nodes are subdivided into user nodes and object nodes. A graph edge represents the rated selection of an object by a user. Projecting the graph on a single user gives the knowledge that the system has about that user's ratings and selections. The k-anonymity is obtained then by padding the graph such that user clustering with less recommendation accuracy can be obtained. Whereas, most solutions proposed for recommenders are based on a centralized gathering of user rating, [20] propose a user-centric distributed and anonymous solution to gather useful information to make recommendations. Interestingly, recent work has shown that it can be linked with differential privacy under certain circumstances [19].

1.6.3 Anonymization protocols

[8] introduced a routing protocol allowing the anonymization of communications between two peers by shuffling messages and therefore disabling a server from knowing where a given message comes from. The onion router [9] improves anonymity by using cryptography. A client message is encrypted multiple times with the keys of the peers of the routing path. This protocol preserves the target server from knowing the address of the client as long the intermediate peers do not collude. However, it is often possible to still identify the original user through information provided within the message itself. This is typically the case of web cookies and/or protocol headers. Solutions exist through browser extensions such as FoxTor or TorButton cookies and headers. However, the body of the message itself (e.g. a search query) which is required for the target server to provide a response (e.g. search results) itself reveals information about the user which in some cases may lead to user re-identification [27].

1.6.4 Data obfuscation

To tackle attacks based on the content of the message, works in the literature have proposed to rely on data obfuscation. Different works have suggested such an approach in the case of web search [22, 26]. In the case of search queries, seen as a collection of terms, the idea is to build an obfuscated query by adding extra decoy terms to the query. The obfuscated query is sent to the search engine which can therefore not know what the original query was. Search results are then filtered by the client in order to restore the accuracy of the original request.

1.6.5 Privacy preserving reputation management

A privacy preserving reputation management system operates such that the opinions used to compute a reputation score remain private and only the reputation score is made public. This approach allows users to give frank opinions about other users without the fear of rendering their opinions public or the fear of retaliation from the target user. Privacy preserving reputation management systems for centralized environments include those by Kerschbaum [16] and by Bethencourt et al. [2]. The system by Kerschbaum introduces the requirement of authorizability, which implies that only the users who have had a transaction with a ratee are allowed to rate him even though rating is done anonymously. Bethencourt's system lets a user verify that the reputation of a target user is composed of feedback provided by distinct feedback providers (implying no collusion) even when users are anonymous. Hasan et al. [14, 13] propose privacy preserving reputation management systems for environments where the existence of centralized entities and trusted third parties cannot be assumed. Current privacy preserving reputation management systems still face a number of open issues. These include attacks such as self-promotion and slandering, in which a user either submits unjustified good opinions about himself or unwarranted bad opinions about a competitor.

1.6.6 User groups

Based on the ideas of [6], their exist many works relying on providing personalization for groups of similar users rather than individual users themselves. For example, [23, 32], propose aggregating data of multiple users belonging to similar interest groups. A group profile is built anonymously using distributed and cryptographic techniques.

1.7 The EEXCESS use-case

1.7.1 What is EEXCESS?

EEXCESS (Enhancing Europe's eXchange in Cultural Educational and Scientific resources) (eexcess.eu) is a European Union FP7 research project that commenced in February 2013. The project consortium comprises of INSA Lyon (insa-lyon.fr), Joanneum Research (joanneum.at), University of Passau (uni-passau.de), Know-Center (know-center.tugraz.at), ZBW (zbw.eu), Bit media (bit.at), Archäologie und Museum Baselland (archaeologie.bl.ch), Collections Trust (collectionstrust.org.uk), Mendeley (mendeley.com), and Wissenmedia (wissenmedia.de). In this section we present the EEXCESS project to illustrate how user profiling can benefit recommender systems particularly with the use of big data techniques. We also discuss the associated privacy issues and the approaches currently being considered in the project for tackling them.

The main objective of EEXCESS is to promote the content of existing rich data sources available throughout Europe. While user context is more and more present, the current response of web search engines and recommendation engines to the massive amount of data found on the web has been to order query results based on popularity. Obviously, the introduction of the PageRank algorithm [5] in search engines has changed the landscape of online searching. However, this has made more difficult to access some valuable content, as it gets buried in the mass of data. This unseen data is sometimes referred to as "long-tail content" in reference to the long-tail of a power-law distribution, which in many cases

characterizes the distribution of user interest. This type of long-tail content is provided by the EEXCESS partners. This content includes precise and rich content such as museum object descriptions, scientific articles, and business articles. Currently, this high-quality content is not very visible, even though it would be invaluable in the appropriate contexts where fine-grained and precise information is sought for.

The aim of EEXCESS is to push such content made available by the partners to users when it is relevant for them. However, this relies on having a precise understanding of a given user's interests and her current context. Different levels of user profiling can help to characterize a user's interests. In EEXCESS, basic, yet informative, user profiles are collected to improve the recommendation results.

1.7.2 Architecture

Figure 1.1 presents a sketch of the current architecture of the EEXCESS system. From this perspective, EEXCESS is made of four components: (1) A plugin added to the user's client whose role is to collect and transfer the user's context, trigger recommendation requests and render them through rich visualizations, (2) a privacy proxy which collects the user's privacy policy and ensures that it is respected, (3) a usage mining component allowing to identify common usage patterns and enrich user profiles accordingly, and (4) a federated recommender service composed of individual data-sources hosting a specific data collection. The circled numbers on the figure give the information flow when content is being recommended.

A privacy proxy is part of the system, as preserving users' privacy is a crucial concern in the project. More specifically, no information about a user profile data should leak out of the system without the user's consent. As it is generally the case when considering privacy in data management systems, the mechanisms implemented to guarantee a certain level of privacy jeopardize some other features of the system (e.g., accuracy and performance).

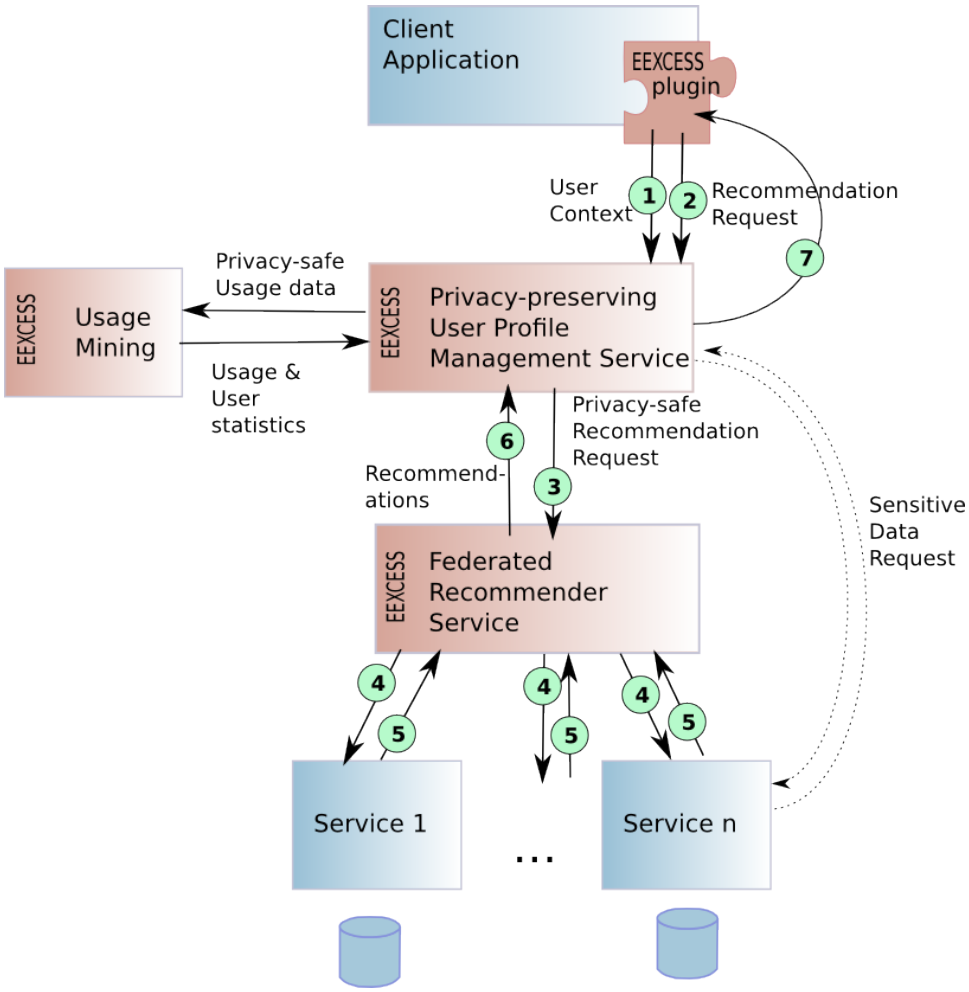


Figure 1.1: Architecture of the EEXCESS system

1.7.3 Personalization

One of the major objectives of EEXCESS is to provide users with relevant and personalized content from the EEXCESS partner data sources. To achieve this goal, accurate user-profiling is an important part of the project and consists in collecting sensitive data about users. An important usage-mining component is responsible for collection or enrichment of user profiles using big data techniques as those described in Section 1.2. Some basic information (e.g., age range, location and interests) is explicitly given by users. These *individual characteristics* form her profile. User interactions with the system are monitored to keep track of a user's *behavior*. Among the partners of EEXCESS, BitMedia is an e-learning platform. In this case, it is clear that the user's learning *goals* and current *knowledge*

(e.g., in the form of courses already taken) will be part of the user's profile. In EEXCESS, the user's *context* consists of information such as her current geo-location, the document or web page (both URL and content) she is working on and the navigation page which lead to the current page.

To capture a more comprehensive understanding of the user, different big data techniques can be applied to further enrich his profile. For example, usage mining aims at identifying usage trends, as well as information about the user's implicit goals and knowledge. On-the-fly analysis of user interests, context, and expectations is also planned. Clustering techniques may be used to identify communities within EEXCESS users. This profiling and better understanding of users only aims at providing them with personalized services, and in particular *personalized recommendations*. Indeed, the content of the EEXCESS partners being very specific (i.e., being in the long-tail of documents when ordered by popularity), having a fine-grained understanding of EEXCESS users is essential to link the correct users to the correct content.

1.7.4 EEXCESS specific challenges

The EEXCESS context brings in different specific privacy related constraints and requirements. This section presents these constraints and their impacts on privacy, as similar "real world" constraints are not restricted to the EEXCESS case and should be considered when trying to reconcile personalization and privacy.

Providers as black box recommenders. Currently, all the content providers give an access to their content in the form of a standard search. Only one of them, namely Mendeley, provides collaborative filtering. Therefore, in a first step, the focus has been put on recommendation through search. In the future, other forms of recommendation (collaborative filtering and hybrid recommendation) will be envisaged. However, in any case, the content-providers are considered as black boxes; the internal recommendation mechanism is hidden. Therefore, privacy-preserving mechanisms have to be modular to work with different recommendation solutions, and cannot be limited to one form.

Providers with existing profiles. Some of the content providers already have user-bases for which they may already have pre-calculated recommendations available. If privacy is limited to anonymization, then through EEXCESS, users will loose access to those recommendations as the recommenders of these providers will not be aware of the users they are sending recommendations to. Therefore, the privacy solutions proposed go beyond simple anonymity and allow users to issue queries anonymously.

Provider recommenders needing feedback to quality. An important objective of recommender systems is to continuously improve themselves through user feedback. To this effect, it is necessary for them to have access to such user feedback. However, this feedback should not lead to privacy leaks. This is a challenge as many attempts towards anonymizing recommendation data have failed in that the data could be de-anonymized [24].

Let us consider that a user wishes to remain anonymous to all the recommenders. In this case, the attacker could be one of the content-providers trying to collect information about the user that it receives queries from. The EEXCESS privacy requirements for such a user would include:

Content anonymity. To guarantee privacy, the attacker should not be able to identify the user from the provided data. Therefore, the system should ensure that an attacker cannot infer from the content of a request who issued it.

Request unlinkability. If multiple queries can be linked together, even while having content-anonymity for each individual query, the combination of the two could reveal information about the user. Therefore, it is required that the protocols guarantee that two independent requests coming from the same user are unlinkable.

Origin unlinkability. This should be feasible by anonymizing the origin of the request, but under the condition that the origin is not revealed by the application level protocols. Therefore, we also need to guarantee that the application level protocols are privacy-preserving (i.e., an attacker cannot link a given request to the user who issued it).

Respecting these three constraints is an ideal goal which requires to limit the information transmitted in each request. Such limitations have a high impact on the utility of the profile information disclosed. Thus the challenge is rather to find a balance between privacy and utility than to ensure complete privacy.

In information systems (such as recommender systems and statistical databases), the main goal of privacy preservation is not to reveal sensitive information about a single entity within the underlying data. This has been shown to be a difficult goal [10, 24]. In a survey on privacy in social networks, Zheleva and Getoor [36] describe some of the common approaches for preserving privacy: *differential privacy* and *k-anonymity*. In the context of recommender systems using collaborative filtering, an approach is to use big data techniques such as clustering to group users together in order to provide privacy [6, 18, 3] with the theory of k-anonymity.

In our particular setting, we are faced with a federated recommender system in which trusted and untrusted peers may exchange information. This requires that both the protocols for exchanging information and the content disclosed are privacy-safe. Furthermore, recommendations may not always be limited to a single recommendation technique among the peers. Each content source may use its own approach. In the context of EEXCESS, few hypotheses can be made on the computational capacities or the background knowledge that an untrusted peer may have access to.

Our work in the EEXCESS project includes the development of mechanisms to allow the definition of flexible user privacy policies, guarantees based on the user privacy policies for non-disclosure of private information, quantification of the risk of disclosing private information, mechanisms for exchange of information based on the reputation and trustworthiness of partners, as well as the definition of the relationship between the amount of information revealed and the quality of recommendations.

1.7.5 Impacts within the EEXCESS use-case

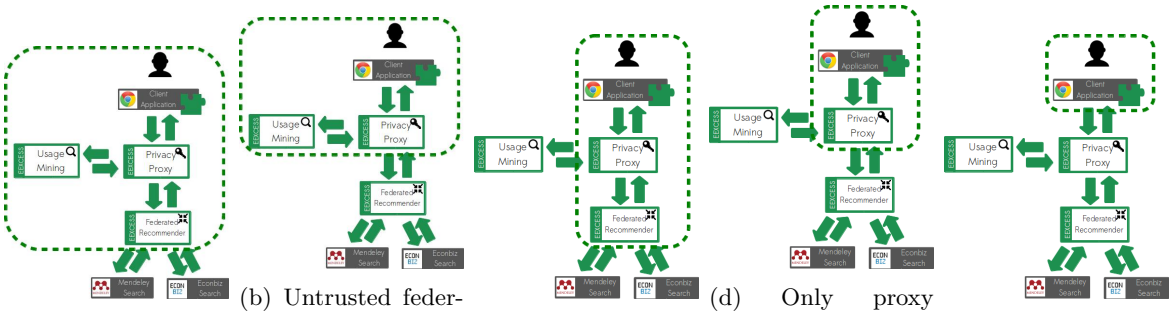
Privacy has multiple impacts on the design of systems heavily relying on personalization. Much depends on the trustworthiness of the peers, but most of all, the legal entities running these peers. In the case of EEXCESS, the architecture of the system and the recommendation algorithms are highly dependent on the trust put in the legal entity which will host EEXCESS software. If the federated recommender component could be hosted by possibly untrustworthy peers, then it could be required that the component be distributed and/or make use of cryptographic solutions.

The impacts of privacy mechanisms on personalization within the EEXCESS system can be summarized as follows:

Adapting personalization algorithms. Providing privacy-preserving personalization implies adapting existing personalization algorithms. Many approaches include cryptographic mechanisms, distribution over a network of peers, working with partial and incomplete data, working with groups of users or pseudonyms.

Architectural impacts. In general, privacy-preservation is not limited to inventing a new version of an algorithm. It has impacts on the global architecture of the privacy preserving system. Indeed, many privacy-preservation mechanisms rely on the fact that all the data does not reside on a single peer. This is particularly true to allow relaxing trustworthiness assumptions on some or part of the peers. Figure 1.2 gives different options of the trustworthiness assumptions within the EEXCESS architecture. Depending on the chosen trust scenarios, the privacy preserving strategies need to be adapted. For example, if the privacy proxy cannot be trusted (scenario (e) of figure 1.2), then it requires to distribute the proxy component over multiple non-colluding authorities to ensure that a single authority does not have all the information.

Making privacy-preservation dynamic. Finally, taking user preference into the privacy preservation mechanisms requires that the personalization algorithms dynamically adapt to each user. In particular, the information provided for two similar users but with different privacy preferences implies that the data available for each of those users



(a) All trusted (b) Untrusted federator (c) Untrusted mining (d) Only proxy trusted (e) Untrusted proxy

Figure 1.2: EEXCESS trustworthiness scenarios

be not as detailed. For example, users may provide some information at different levels of granularity. One user may allow providing a complete birth date whereas another may only allow revealing her age range.

Performance. The introduction of mechanisms to ensure privacy comes at a price. For instance, encryption mechanisms aiming at hiding users' identity when using a system add computation complexity. Similarly, query obfuscation techniques which create fake queries to prevent attackers to construct accurate user profiles generate network traffic and tend to overload the systems they target. TrackMeNot in an example of such a system [34]. The challenge system designers have to face is then to offer the best trade-off between privacy preservation and performance (i.e., computation cost, resource consumption, response time). The approach envisioned in the context of EEXCESS is to raise users' awareness regarding their privacy. In other words, it consists in informing users of the private data they divulge by using the system. Then, users are equipped to tune the privacy settings according to their expectations, and therefore achieve a satisfactory trade-off between privacy preservation and performance.

1.7.6 Implementation

In this section, we focus on the mechanisms used to preserve users' privacy in the EEXCESS recommender system. These mechanisms are based on the PEAS protocol [28]. PEAS aims at protecting users privacy in the context of Web search. It is composed of two protocols to

ensure users privacy: an unlinkability protocol and an indistinguishability protocol. They are respectively implemented on a privacy proxy and the client application.

Privacy proxy. The privacy proxy is located between the client application (where the queries are issued) and the federated recommender. The privacy proxy allows hiding the identity of the user who sent the query. The privacy proxy itself is composed of two components: the receiver and the issuer. By using an encryption protocol, PEAS assures that the receiver knows the user who sent the query, but not the content of the query; and the issuer knows the content of the query, but not the identity of the user. This protocol is made available through Web services.

Client application. The main application developed in the EEXCESS project is a Google Chrome extension. When a user is browsing the Web, queries are sent automatically (according to the page content) to the federated recommender and resources are suggested. In order to prevent the federated recommender to retrieve the identity of users by analyzing the content of the queries, PEAS offers a technique to obfuscate queries and filter the results. This technique uses a group profile, which is built by the issuer, and generates fake queries that are added to the original query. To remove irrelevant results introduced by the obfuscation technique, a filtering technique is implemented. The protocol is provided through a Javascript component.

1.8 Conclusion

In this chapter, we discussed the challenges raised when building systems, which simultaneously require a deep level of personalization as well as a high level of user privacy. Big data analysis techniques play an important role in making such personalization possible. On the other hand, this raises the issue of respecting a given user's privacy. Big data may even increase this risk by providing attackers the means of circumventing privacy-protective actions. We illustrated these issues by introducing the challenges raised by EEXCESS, a concrete project aiming both to provide high quality recommendations and to respect user privacy.

References

[1] Claudio A. Ardagna, Giovanni Livraga, and Pierangela Samarati. Protecting Privacy of User Information in Continuous Location-Based Services. In *2012 IEEE 15th International Conference on Computational Science and Engineering*, pages 162–169. IEEE, December 2012.

[2] John Bethencourt, Elaine Shi, and Dawn Song. Signatures of reputation: Towards trust without identity. In *14th International Conference on Financial Cryptography and Data Security*, pages 400–407, 2010.

[3] Antoine Boutet, Davide Frey, Arnaud Jegou, and Anne-marie Kermarrec. Privacy-Preserving Distributed Collaborative Filtering. Technical Report February, INRIA, Rennes, 2013.

[4] Hai Brenner and Kobbi Nissim. Impossibility of differentially private universally optimal mechanisms. In *51th Annual IEEE Symposium on Foundations of Computer Science*, pages 71–80, 2010.

[5] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.

[6] John F. Canny. Collaborative filtering with privacy. In *2002 IEEE Symposium on Security and Privacy*, pages 45–57, 2002.

[7] Chih-Cheng Chang, Brian Thompson, Hui (Wendy) Wang, and Danfeng Yao. Towards publishing recommendation data with predictive anonymization. In *5th ACM Symposium on Information, Computer and Communications Security*, page 24. ACM Press, April 2010.

[8] David Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM*, 24(2):84–90, February 1981.

[9] Roger Dingledine, Nick Mathewson, and Paul F. Syverson. Tor: The second-generation onion router. In *13th USENIX Security Symposium*, pages 303–320, 2004.

[10] Cynthia Dwork. Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming (ICALP)*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer Verlag, July 2006.

[11] Cynthia Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation*, volume 4978 of *Lecture Notes in Computer Science*, pages 1–19. Springer Verlag, April 2008.

[12] Benjamin Habegger, Omar Hasan, Lionel Brunie, Nadia Bennani, Harald Kosch, and Ernesto Damiani. Personalization vs. privacy in big data analysis. *International Journal of Big Data*, 1:25–35, 2014.

[13] Omar Hasan, Lionel Brunie, and Elisa Bertino. Preserving privacy of feedback providers in decentralized reputation systems. *Computers & Security*, 31(7):816–826, October 2012.

[14] Omar Hasan, Lionel Brunie, Elisa Bertino, and Ning Shang. A decentralized privacy preserving reputation protocol for the malicious adversarial model. *IEEE Transactions On Information Forensics And Security*, 8(6), 2013.

[15] Paul Jessup. Big data and targeted advertising. http://www.unleashed-technologies.com/blog/2012/06/28/big-data-and-targeted-advertising, June 2012.

[16] Florian Kerschbaum. A verifiable, centralized, coercion-free reputation system. In *8th ACM Workshop on Privacy in the E-Society*, pages 61–70, 2009.

[17] Alfred Kobsa. Privacy-enhanced personalization. *Communications of the ACM*, 50(8):24–33, 2007.

[18] Dongsheng Li, Qin Lv, Huanhuan Xia, Li Shang, Tun Lu, and Ning Gu. Pistis: A Privacy-Preserving Content Recommender System for Online Social Communities. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, pages 79–86, 2011.

[19] Ninghui Li, Wahbeh Qardaji, and Dong Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *7th ACM Symposium on Information, Computer and Communications Security*, pages 32–33, 2012.

[20] Zhifeng Luo, Shuhong Chen, and Yutian Li. A distributed anonymization scheme for privacy-preserving recommendation systems. In *2013 IEEE 4th International Conference on Software Engineering and Service Science*, pages 491–494. IEEE, May 2013.

[21] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. Big data: The next frontier for innovation, competition, and productivity. Technical report, The McKinsey Global Institute, May 2011.

[22] Mummoorthy Murugesan. *Privacy through deniable search*. PhD thesis, Purdue University, January 2010.

[23] Animesh Nandi, Armen Aghasaryan, and Makram Bouzid. P3: A privacy preserving personalization middleware for recommendation-based services. *Hot Topics in Privacy Enhancing Technologies*, pages 1–12, 2011.

[24] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy*, pages 111–125, 2008.

[25] Kobbi Nissim. Private data analysis via output perturbation - *A Rigorous Approach to Constructing Sanitizers and Privacy Preserving Algorithms*. In *Privacy-Preserving Data Mining - Models and Algorithms*, pages 383–414. Springer, 2008.

[26] HweeHwa Pang, Xuhua Ding, and Xiaokui Xiao. Embellishing text search queries to protect user privacy. *Proceedings of the VLDB Endowment*, 3(1):598–607, 2010.

[27] Sai Teja Peddinti and Nitesh Saxena. On the limitations of query obfuscation techniques for location privacy. In *UbiComp'11*, page 187, New York, New York, USA, 2011. ACM Press.

[28] Albin Petit, Thomas Cerqueus, Sonia Ben Mokhtar, Lionel Brunie, and Harald Kosch. PEAS: Private, Efficient and Accurate Web Search. In *14th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, 2015.

[29] P. Samarati. Protecting respondents identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

[30] Lalitha Sankar, S. Raj Rajagopalan, and H. Vincent Poor. Utility-Privacy Tradeoffs in Databases: An Information-Theoretic Approach. *IEEE Transactions on Information Forensics and Security*, 8(6):838–852, June 2013.

[31] Rathindra Sarathy and Krishnamurty Muralidhar. Evaluating Laplace Noise Addition to Satisfy Differential Privacy for Numeric Data. *Transactions on Data Privacy*, 4(1):1–17, April 2011.

[32] Shang Shang, Y Hui, Pan Hui, Paul Cuff, and Sanjeev Kulkarni. Privacy Preserving Recommendation System Based on Groups. *arXiv preprint arXiv:1305.0540*, pages 1–28, 2013.

[33] Latanya Sweeney. k-Anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, October 2002.

[34] Vincent Toubiana, Lakshminarayanan Subramanian, and Helen Nissenbaum. Trackmenot: Enhancing the privacy of web search. *arXiv preprint arXiv:1109.4677*, 2011.

[35] Jamie Yap. User profiling fears real but paranoia unnecessary. http://www.zdnet.com/user-profiling-fears-real-but-paranoia-unnecessary-2062302030/, September 2011.

[36] Elena Zheleva and Lise Getoor. Privacy in social networks: A survey. In *Social Network Data Analytics*, pages 277–306. Springer, 2011.