# Privacy Analysis of a Networked Collaborative Recommendation System

Ville Ollikainen and Valtteri Niemi

***Abstract*—**The rapid expansion of available online services has raised concerns about user privacy. As a response to this concern, EU Parliament has recently approved General Data Protection Regulation, which aims to give citizens back control of their personal data. Built upon a recently developed token-based recommendation method (UPCV), we introduce in this paper a novel approach of networking collaborative recommendation engines and present the first results of a series of studies regarding its capability to protect user privacy.

***Keywords*—**data protection, general data protection regulation, privacy, recommendations, targeted advertising, upcv.

## I. Introduction

While recommendations have become an integral part of successful web services, the rapid expansion of available online services has raised concerns about user privacy. Ironically, at the same time when users get added value by successful recommendations, helping them to discover and organize vast amounts of content, services and offerings, the same information also threatens their privacy. For example, recommendations play a fundamental role in targeted advertising, while targeted adverts may feel too intrusive from privacy point of view.

On a global scale, personal data is stored in and transferred between numerous services, companies and countries, with inconsistent legislation. Eventually, users do not know where their data is stored, and which pieces of it, and the ethics, how it will be eventually used.

After four years of work, EU Parliament approved on April 14th 2016 new EU data protection rules, which aim to give citizens control of their personal data [1]. These rules are known as General Data Protection Regulation, or "GDPR". GDPR defines globally how to deal with any data regarding EU citizens, no matter where it is processed and stored. Unlike EU directives, the provisions in the regulation will be directly applicable in all member states after two years, without either requiring any national legislation or allowing local modification. The same exact regulation will apply everywhere simultaneously.

This paper presents a study on a collaborative token-based recommendation method (UPCV) [2] that creates an interoperable abstraction layer for both user preferences and item properties. This layer separates user data from item data, enabling distributed ownership and storage, including true ownership of personal data. At the same time all actions become bilateral between a user and a service. As such the approach relates to multi-domain collaborative filtering proposed by [3] and cross-domain recommendations proposed by [4] by addressing sparsity problems that are often experienced in single domain collaborative recommenders. The general goal is to improve the performance of recommendations in all domains simultaneously.

The token-based abstraction layer also provides independence from domain knowledge: When it comes to collecting and using personal data, some approaches, such as [5], suggest storing personal profiles in a database in order to enable authorized parties to provide personalized services and user control upon whom to trust. However, after this point these services provide only little privacy.

Furthermore, privacy concerns have been raised because of several recommendation/targeting systems integrate personal usage data from multiple services. In legacy approaches, such as tracking cookies, it is often possible to identify people visiting service A, if they have been visiting service B before.

Despite recent development in distributed and cloud computing, single repositories pose an inherent problem in terms of scalability, while they also are single points of failure from technical and privacy perspectives. The presented token-based method is capable of operating bilaterally between a single user and a single service, each utilizing their own computing resources and making the approach inherently scalable.

The rest of this paper is structured as follows. Section II provides preliminaries about our use case, i.e. about ISBN semantics, Book-Crossing dataset and a token-based recommender. It also includes a brief introduction to relevant privacy concepts. In Section III we introduce a networked exchange mechanism of randomly generated tokens, and explain how the mechanism is used for recommendation purposes. Section IV describes our study with the dataset, while results are presented in Section V. Conclusions and discussions about future directions are included in Section VI.

## II. Preliminaries

### A. Recommendations

The recommendation problem can be defined as estimating the user's response to new items based on historical information stored in the system, and suggesting novel and original items for which the predicted response for that user is high [6]. Prediction of user interests is traditionally based on

Ville Ollikainen is with the VTT Technical Research Centre of Finland, 02150 Espoo, Finland

Valtteri Niemi is with the Department of Computer Science, University of Helsinki, 00014 Helsinki, Finland

.

demographic data, such as age, sex, income level and matrimonial status. The availability of more data has led to more sophisticated recommending algorithms being proposed in the literature, most commonly classified into two basic categories: content-based and collaborative recommendations.

Content-based recommenders are based on representing the items with a set of attributes and using these attributes to find the most relevant content for a particular user. When it comes to networking different recommendation engines, content-based recommenders need content analysis and domain specific vocabulary, making them less usable for general use as such. For instance, music requires quite a different vocabulary than text; how to make them interoperable?

Collaborative recommendations, on the other hand, learn from the behavior of users without any need to analyze items. Instead, they recommend items that have been preferred by users who have had similar behavior in the past. Collaborative recommendations are based on the intuition that people tend to like similar items, e.g. people who have liked a particular book or movie are likely to have the same taste also for other items, compared to taste of all people in average. As a challenge in collaborative methods, they can hardly recommend any new items that no-one else has accessed before, making them less useful as the sole method for presenting items to users. Also, most common collaborative methods are based on sparse matrices in which users and items intersect; from these legacy matrices it is difficult to separate item-only and user-only data.

*B. Token-based recommender*

A recently developed token-based recommendation method [2] associates both users and items with collections of tokens, each token carrying a random value. The method addresses the problem of sharing personal data with the help of storing the personal data as a collection of random values that we call "tokens", instead of explicit profiles. Thus, token collections provide an abstraction which is privacy-preserving by design.

Unlike cookies and other tracking means, tokens have no association with the real world. In particular, they do not have any association with persons. Tokens are mere random values that will be copied to and eventually deleted from collections.

Interaction between a user and an item triggers randomly selected tokens that are copied from the token collection of the user to the token collection of the item, and vice versa. When the same user interacts with several items, or the same item is involved in interactions with several users, tokens spread around, resulting in similarities among different token collections in the system. Since tokens spread in user-item interactions only, it is likely that similarities between two token sets originate from similar user behavior. The method is collaborative by nature and requires no content analysis.

Since tokens are exchanged in each transaction, collections are dynamic by nature, yet they can be universally compared, creating a foundation for collaborative recommendations that can be used for multiple purposes. This enables behavioral recommendations, personalized services and targeted advertising, which are the most important application areas for personal data.

Tokenized profiles provide an abstraction layer that is at the core of our approach. We will use an example to explain how the tokens work, and how they are able to provide personalized services, which are recommendations in the example presented in Fig. 1.

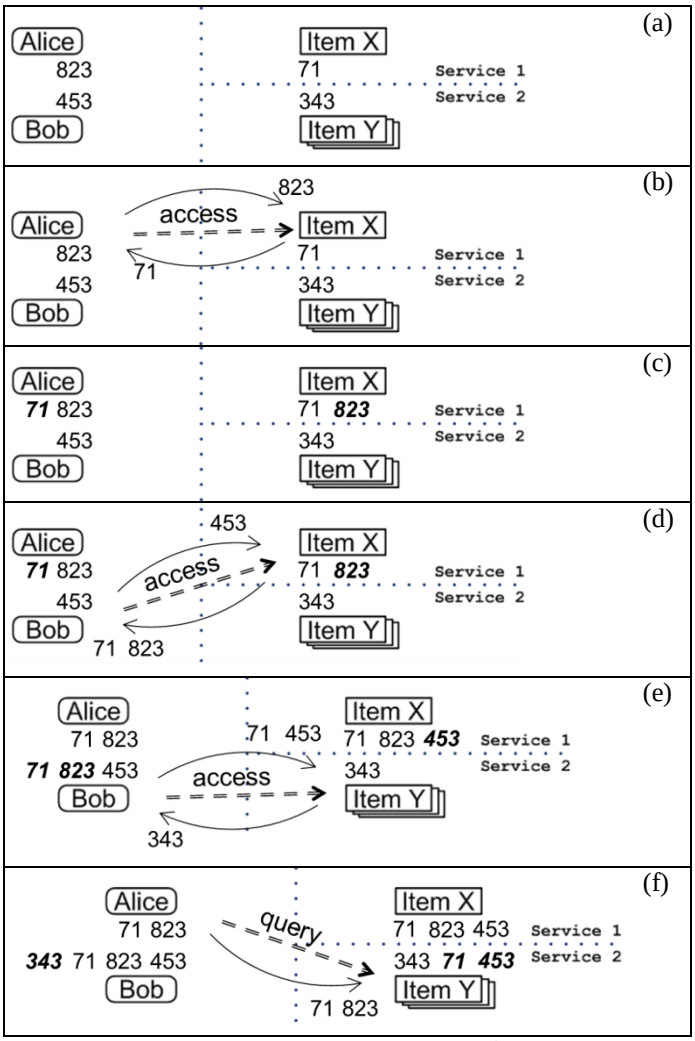


Fig. 1 Token exchange in an exemplary case of two users and two services.

Step (a) in Fig. 1 illustrates a simple case where there are two independent services on the top, the upper service ("service 1") having just a single item ("item X") and the lower service ("service 2") having multiple items, including "item Y". Only two users, "Alice" and "Bob", are presented for simplicity; in real life the method requires a number of users, since it is based on statistical phenomena caused by similarly behaving users. In the beginning each user and item has a random number, or "token", in their collections.

The different steps illustrate a sequence of actions by Alice and Bob. In step (b), Alice is accessing item X. If item X would be an online news article, "access" would mean clicking it using a web browser. This will trigger a token exchange procedure in which a couple of tokens are copied from the user to the item and vice versa. Since both Alice and item X have only one token each in the beginning, these are the tokens to be exchanged.

After this action, in (c), both Alice and item X have common tokens. (In the table, bold typography highlights the most recently acquired tokens.) It should be noted that these

tokens are random values only, thus they do not carry any history with them; we claim that after a while it is impossible to say from where item X and Alice got their tokens.

Next, Bob is accessing the same item X in the first service, and once again a couple of tokens are copied over (step d); this time item X is able to provide more than one token. After Bob's action it should be noted that Bob and Alice have similar tokens, as can be seen in step (e). Again, similarity exists in the data without any clue for its reason.

Still in step (e), Bob is accessing item Y in the second service. A couple of tokens are requested for exchange; since Bob has more than one token to give, some tokens are picked randomly from his collection.

As the last action in this example, Alice requests a recommendation for herself from the second service, which she is now visiting the very first time (step f) The query contains at least part of her tokens, in this example all of them. Finally, service 2 goes through all available items and compares their tokens with the provided tokens to create recommendations: The recommended item is the one that corresponds to the token collection with maximum Jaccard similarity. Jaccard similarity is a well know measure for comparing similarities of two sets. It is defined as the ratio of two numbers: size of the intersection of the two sets and size of their union.

There is a substantial likelihood that item Y will be in the recommendation list, since there are tokens in common.

In real cases collections may carry hundreds of tokens for each user and item, and the amount of exchanged tokens may vary e.g. between 0.5% and 15%, depending on the application; the percentage refers to the amount of tokens on sending side. Also, it is possible to create only one initial token for each new user and item, yet there can also be more than one initial token, say 32 of them.

There may be different strategies in processing tokens. At its simplest, when a collection reaches its maximum size, tokens can be deleted on a random basis in order to make space, eventually deleting also the initial tokens. Furthermore, tokens may expire at a certain time, but in a limited experiment this feature can be omitted.

### *C. International Standard Book Number*

ISBN is a hierarchical book identifier which contains agency, publisher and publication codes. Agencies are typically country-based, for instance Japanese agency allocates publisher codes in Japan. As exceptions to this rule, English, French and German languages are under their own specific agency codes. In addition, Germany as a country has its own agency. Each code is associated with exactly one agency but several codes may be associated with the same agency.

### *D.Book-Crossing dataset*

The Book-Crossing (BX) dataset is collected from a "Book-Crossing" literature exchange service, based on leaving books in public places to be found by other potential readers. Each book has a sticker with instructions about, what to do when found.

The BX [7] dataset is commonly used for developing and validating collaborative filtering methods. BX contains three different tables: "BX-Users" containing 278,859 users, "BX-Books" with 271,379 valid books and "BX-Ratings" of 1,149,780 user "ratings", sorted by user id. It should be noted that users have entered ISBN's without validation, so a percentage of ISBN's in "BX-Ratings" are invalid.

In the dataset there are two types of "ratings", "implicit" (value 0) and "explicit" (values 1-10, 10 being the best). "Implicit ratings" count for 62% of all "ratings" (716,109 instances). The minority of ratings, 433,671, are explicit.

BX is referring to books by their ISBN, thus indicating also the agency under which they have been published.

### *E. Privacy and deniability*

One of the fundamental principles of privacy design is that personally identifiable information is not collected without user consent. Furthermore, when giving the consent, the user should understand why the information is collected. Naturally, the collected information can only be used for purposes which the user has given consent.

In an ideal privacy-preserving setting, collected data should not provide any further information about individuals even when combined with extra information. This principle is also behind GDPR: if any information *can* be associated with a user even by someone else, it is no longer anonymous data; it becomes personal data.

A weaker form is deniability which means that the collected data may provide some probability information about individuals but at minimum a reasonable doubt would remain about whether the information is true. This would enable the individual to successfully deny correctness of the information.

## III. A TOKEN-BASED RECOMMENDER SYSTEM

Since token exchange leaves no traces of the origin of tokens, we assume in general that it is fairly safe to disclose tokens to third parties without disclosing any history.

In the example presented in Fig. 1 the system consisted of two independent services that had nothing in common, except two users. All token exchange took place bilaterally, between a user and a service, without common repositories or computing facilities.

Instead of one recommender, let's now illustrate two recommenders (Fig. 2). Since all token exchange takes place between a user and an item in a service, we can now divide the item set into two separate sets without affecting token exchange operations: when a user is accessing any item in set 1, token exchange is carried out by recommendation engine 1, and the same goes for set 2 and recommendation engine 2. Of course, when making a query for recommendations, the recommendations may cover only the items within the particular service.

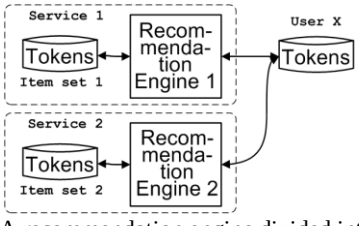


Fig. 2 A recommendation engine divided into two.

Otherwise, a recommendation engine can be divided. From a user's point of view, this means that independent recommendation engines act as a single recommendation engine as far as token exchange is considered.

Fig. 3 expands this concept a bit further illustrating an exemplary setup in which users X and Z aggregate their tokens in a public library service, while user Y gets his tokens from an online video service. Since service 1 and service 2 have at least one user in common, tokens are compatible in these services. In this illustration there is also a dating service into which users can upload their tokens, requesting the recommendation engine X to find other users with similar tokens (something in common, that is).

We could easily continue this expansion into a scenario in which users can aggregate their token collections in numerous services, and numerous services can benefit from user tokens provided to them in order to offer better user experience.

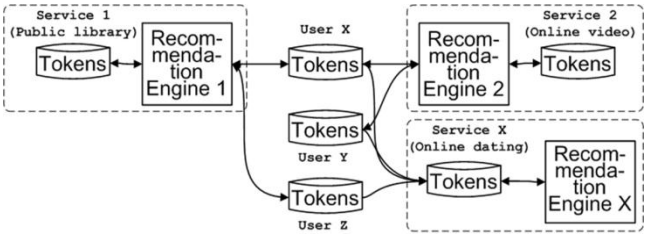


Fig. 3 A token based recommendation engine system

## IV. METHODOLOGY

### *A. Preparing BX data*

We used ISBN agencies in the "BX-ratings" table to create our training and validation sets. First, we ignored ratings and filtered out invalid ISBN's, resulting in 1,135,377 ratings. Second, we converted agency codes to agencies, with the exception that the agency code for "German language" was mapped to "Germany", and using the conversion we replaced ISBN's by agencies. Third, we removed one-time users, resulting in 1,077,310 ratings. Fourth, we shuffled the ratings to random order and got our transaction data set.

In our evaluation we imported the transaction data set into the token-based recommendation engine as two independent test cases with the following parameters. The maximum size of each collection was 1024 tokens; in each transaction 0.58%+1 (i.e. at most 6) randomly selected tokens were copied in both directions. In the first test case each user and item got a *single initial token*; in the second test case we created *32 initial tokens*.

### *B. Privacy aspects*

Privacy features should be defined with respect to adversary models. In such a model we first describe what the capabilities of an attacker are and what he may try to do. For instance, we could allow the adversary to observe all tokens that are exchanged between different parties while the target is to determine which token exchanges are carried out by the same party. In other words, the adversary tries to break unlinkability between different transactions.

In our system there is no need to allow outside observers to gain access to the tokens that are exchanged. This could be guaranteed by transferring all tokens inside encrypted communications. For example, we could define the mandatory use of a secure protocol, for instance HTTPS, in all communications related to tokens.

Another adversary model could allow disclosing all tokens of all parties to the attacker. In this scenario the target of the attacker would be to determine which parties have been in contact with each other. If a recommender system would be built upon an assumption that at least some parties make their tokens public, this adversary model would not be relevant.

In reality, there will be services that are trusted by users to an extent that a portion of users have disclosed their identities to the service under explicit consent. For instance, access to online archives of a weekly magazine may be associated to the subscription of the paper format, with specific name, address and other contact information. Since there is a trust relation between the service and these users, privacy is a lesser issue.

However, there will be a myriad of services where users just pop in, check something and leave. In these services it will be safer for users to disclose as few tokens as possible, especially if no recommendations are requested.

In our methodology we concentrate on studying, whether a non-trusted service can detect returning users by their tokens.

### *C. Evaluating privacy in terms of successfully detecting a returning customer*

In our study we present a scenario in which the BX agencies mimic real world services. In this scenario we create agency-to-agency recommendations by comparing token collections of different agencies; therefore agencies would need to reveal their token collections to each other.

We could alternatively use a trusted third party who would receive a view of every agency's tokens and who is expected to provide recommendations in return. This would put this trusted third party into an excellent position as an adversary but we would still trust that it would not use its position against any party in the system. Even in this case there would be no need to see readers' tokens, not even by the trusted party.

In our example scenario it is assumed that the agencies do not want to retain their privacy against the readers. On the contrary, each agency would prefer engaging with as many readers as possible, and it would be beneficial for the agency to use its own identity e.g. for reputation. However, we expect that readers want to remain anonymous towards agencies.

Now we may describe an attack against readers' privacy that could not be eliminated either by usage of HTTPS connections for the purpose of hiding token exchanges or by usage of a trusted third party for the purpose of computing recommendations. The attacker is a single agency (in our case the one responsible for French language books) *who keeps a record of tokens received from various readers and tries to determine which of the transactions are initiated by the same reader*. In a certain sense this kind of attacker can be classified as an "honest-but-curious" adversary.

Please note that if the readers would not try to remain anonymous, and would use their permanent identities instead of tokens when communicating with the agency, it would be trivial for the agency to determine which transactions would be with the same reader. Thus we consider, what could be disclosed from tokens alone.

We run a simulation of token exchanges with the BX dataset and chose one relatively popular agency as the adversary. Then we took the point of view of the adversary and, for each pair of transactions, compared tokens exchanged.

## V.RESULTS

The vast majority of the transaction pairs did not have any tokens in common. On the other hand, the vast majority of transaction pairs were such that the two transactions were done by two different readers.

These two observations were by no means surprising. We put our focus on those pairs of transactions where common tokens were found. The portion of pairs carried out by the same reader was considerably bigger among this set of pairs than among all pairs of transactions. But still the portion of pairs carried out by the same reader was smaller than the portion of pairs carried out by two different users. More specifically, in the case of *32 initial tokens*, around 41 % of pairs with common tokens were carried out by the same reader while around 59 % were carried out by two different readers. In the case of *one initial token* the corresponding figures are 17 % (same reader) and 83 % (different reader).

This finding serves as strong evidence for the claim that our token exchange mechanism is indeed privacy-preserving.

We separately studied those pairs that had at least two tokens in common, which led to different results. In the case of 32 initial tokens more than 99 % of the pairs were such that both transactions were done by the same reader while less than 1% was such that the two transactions were done by different readers. In the case of one initial token the corresponding figures were 95 % (same reader) and 5 % (different readers). Moreover, all pairs that had at least three common tokens were such that both transactions were done by the same reader. (This holds for both cases of 32 initial tokens and one initial token.) Actually the main reason for this phenomena was the fact that the amount of transactions where the reader provided more than one token to an agency were in a small minority, carried out by a handful of the most active readers.

One conclusion from the simulation is that the number of tokens exchanged should not depend on how many tokens the parties possess but should rather be kept constant. Otherwise, the number of tokens sent could be an even better indicator (than the tokens themselves) of the fact that two transactions are done by the same reader.

## VI. CONCLUSIONS

We have studied privacy aspects of a method of creating recommendations by exchanging random-looking tokens. Use of a secure channel for protecting token exchanges (e.g. by HTTPS protocol) provides privacy protection against outside observers. Use of a trusted party for computation of recommendations would remove the need of making tokens public between different services. Finally, we studied to what extent a single service is able to break the privacy of its users.

One can ask, in respect to GDPR, are tokens truly anonymous data? From the point of view of this study we can say that if a service (agency) does not memorize token transactions, but merely stores the tokens in the form of modified token collections, a returning individual user can't be detected. Memorizing tokens would require user consent in the spirit of GDPR, since there is a *possibility* to detect returning users from the crowd. However, even in that adverse case, users have a degree of privacy, since detection is not definite.

In the future we are going to study privacy properties more thoroughly in different adversary models. We are also planning to find various trade-offs between recommendation performance and privacy preservation by changing the parameters and functionality in the token exchange mechanism, specifically how to select outgoing tokens. Measuring privacy from simulations in various data sets is another important future direction.

## REFERENCES

[1] European Parliament News, "Data protection reform - Parliament approves new rules fit for the digital era", available at http://www.europarl.europa.eu, REF: 20160407IPR21776

[2] V. Ollikainen, A. Mensonen, M. Tavakolifard, "UPCV Distributed recommendation system based on token exchange," Journal of Print and Media Technology Research, Vol. 2, No. 3, pp. 195 – 201, 2013. Available at http://www.vtt.fi/inf/julkaisut/muut/2013/OA_JPMTR_1314_Ollikainen.pdf

[3] Y. Zhang., B. Cao, D.-Y. Yeung, "Multi-domain collaborative filtering". arXiv Preprint arXiv:1203.3535. 2012

[4] S. Gao, H. Luo, D. Chen, S. Li, P. Gallinari, J. Guo, "Cross-domain recommendation via cluster-level latent factor model," Lecture Notes in Computer Science, 8189 LNAI(PART 2), 161-176. doi:10.1007/978-3-642-40991-2_11. 2013

[5] T. Kirkham, S. Winfield, S. Ravet, S. Kellomaki, "The personal data store approach to personal data security," IEEE Security and Privacy, 11(5), pp. 12-19. 2013.

[6] C. Desrosiers, G. Karypis, "A Comprehensive Survey of Neighborhood-based Recommendation Methods," Boston, MA: Springer US, pp. 107-144. 2011.

[7] C.-N. Ziegler, S.M. McNee, J.A. Konstan, G. Lausen, "Improving Recommendation Lists Through Topic Diversification," Proceedings of the 14th International World Wide Web Conference (WWW '05), May 10-14, 2005, Chiba, Japan