Measuring privacy in high dimensional microdata collections

Spyros Boukoros and Stefan Katzenbeisser
Department of computer science
TU Darmstadt
Darmstadt, Germany
{boukoros,katzenbeisser}@seceng.informatik.tu-darmstadt.de

ABSTRACT

Microdata is collected by companies in order to enhance their quality of service as well as the accuracy of their recommendation systems. These data often become publicly available after they have been sanitized. Recent reidentification attacks on publicly available, sanitized datasets illustrate the privacy risks involved in microdata collections. Currently, users have to trust the provider that their data will be safe in case data is published or if a privacy breach occurs. In this work, we empower users by developing a novel, user-centric tool for privacy measurement and a new lightweight privacy metric. The goal of our tool is to estimate users' privacy level *prior to* sharing their data with a provider. Hence, users can consciously decide whether to contribute their data. Our tool estimates an individuals' privacy level based on published popularity statistics regarding the items in the provider's database, and the users' microdata. In this work, we describe the architecture of our tool as well as a novel privacy metric, which is necessary for our setting where we do not have access to the provider's database. Our tool is user friendly, relying on smart visual results that raise privacy awareness. We evaluate our tool using three real world datasets, collected from major providers. We demonstrate strong correlations between the average anonymity set per user and the privacy score obtained by our metric. Our results illustrate that our tool which uses minimal information from the provider, estimates users' privacy levels comparably well, as if it had access to the actual database.

KEYWORDS

privacy, privacy metrics, microdata, user empowerment

ACM Reference format:
Spyros Boukoros and Stefan Katzenbeisser. 2017. Measuring privacy in high dimensional microdata collections. In *Proceedings of ARES '17, Reggio Calabria, Italy, August 29-September 01, 2017,* 8 pages.
DOI: 10.1145/3098954.3098977

1 INTRODUCTION

In the era of information explosion, big parts of human life take place on the web and online personalized services have become an integral part of everyday life. These services include social interaction and matching, health care, entertainment, shopping, etc., and are accessible even on the go with smartphones. In many services, recommender systems are available, assisting users by limiting the space of available choices to only those relevant to their preferences, surprise them with serendipitous choices and significantly speed up the decision process. In order to provide accurate recommendation results and more targeted advertisements, providers need to store data about personal preferences and past interactions with the service, for example, the movies that someone has watched or the pages she likes on Facebook. This kind of data regarding users is called microdata and is a lucrative asset for a variety of companies and organizations, because a plethora of information can be extracted, such as shopping trends, health indicators, etc.

Microdata poses a privacy risk in case parts of databases are being compromised, or even when collections are published sanitized (i.e., anonymized). Inadequate procedures are typically used in the anonymization process, such as the simple removal of personal identifiable information, i.e., names, tax IDs, etc. In a demonstration to show that these procedures are not safe, a US politician was reidentified through published anonymized health records [16], while in [11] the authors demonstrated that half a million users of a popular movie database could be reidentified with high accuracy.

Whenever such personal data/preferences become publicly available, individuals can be affected in a variety of ways. It is known that political orientation, sexual preferences, age, intelligence and other personal traits can be statistically inferred with high accuracy just by accessing someones online profile [7]. In addition, unexpected connections in social graphs, so called weak ties, can uncover identities and reveal sensitive data [14]. For example, if the online profiles of two users share some common interests, we can predict traits about one profile by having access only to the other.

Currently, users can only *hope* that their data will be safe in case of a data breach. In addition, a privacy breach has implications for the future privacy [11]. In a hypothetical scenario where Alice's identity has been revealed, she can no longer hide behind another online account with a different nickname. Even if she creates a brand new account, her microdata will *always* identify her.

Motivated by recent linkage attacks and focusing primarily on user's empowerment, we develop a novel lightweight tool for individuals' privacy assessment based on their microdata. Our tool measures the risk of being reidentified based on their microdata, *prior to* sharing their data, based on two factors: The data users want to *disclose to the provider* and *published statistics about those data.* We consider the following scenario: Alice wants to use a service for music recommendation. However, Alice heard on the news about recent reidentification attacks on similar services and

ARES '17, Reggio Calabria, Italy

DOI: 10.1145/3098954.3098977

is worried about her online privacy. By using our tool, she can obtain an estimation and a visual representation of the risks being reidentified based on her microdata (privacy level), without disclosing anything to the provider. With this extra information, she no longer has to trust the provider but can *consciously decide* whether she wants to contribute her data.

Our tool is composed of two instances, one for the service provider and one for clients, and a privacy metric. On the provider's side, our tool clusters clients based on their microdata into privacy groups and publishes information regarding each group. In addition, statistics regarding the items' popularities are published. These kind of statistics already exist in many services in the form of "another x users liked this item". The users' instance measures the privacy level of the microdata of a user, based on the published information by the provider. In contrast to existing metrics, the privacy metric developed in this work is lightweight, hence is easily computable by users' devices, and does not rely on the entire database of the service provider (just on published statistics). As a last step, the tool provides a visual and easily comprehensible result to the user. Our system requires minimal configuration effort and computing resources, while it provides immediate value to users. The main feature of our design is that users are able to determine their privacy score locally, *without* full access to the data of other clients.

Contributions. Our main contribution is a novel lightweight tool that measures users' privacy, prior to sharing their data with a provider. In addition, we develop a lightweight, user-centric privacy metric, because traditional database privacy metrics fall short in our setting.

Furthermore, we perform real world evaluation of our tool. We use three datasets from online services, which have in total over 1.5 million ratings and 30 thousand users. By doing extensive simulations of reidentification attacks for every user, we measure their average anonymity set. The anonymity set (in our scenario) is the number of users that share the same amount of information (for a specific query) with the user we want to deanonymize. We demonstrate a strong correlation between the anonymity of a user, measured by the size of his anonymity set, and his privacy score obtained by our tool.

Finally, we present a smart way to represent the privacy levels to the user and raise privacy awareness. Focusing on the adoption and usability of our tool, we propose a visual method. Hence, we do not require users to understand the details of our tool.

Outline. We proceed by reviewing related work in Section 2 and we describe the tool and the privacy metric in Section 3. In Section 4, we introduce the datasets used for evaluation, and present the experiments demonstrating that our tool can capture the reidentification risk of someone adequately, even without full access to the database. In Section 5 we describe the proposed visual way in which the privacy levels are presented to users. We conclude in Section 6.

2 RELATED WORK

2.1 Reidentification Attacks

Various reidentification attacks, mostly on publicly available sanitized datasets, demonstrate the risks involved in data publishing. Sweeney [16] was able to identify the profile of a US governor in an anonymized health database, by combining it with a voters registration list. In this work, the concept of k-anonymity was firstly introduced. Frankowski et al. [6], using the Movie-lens database, correlated public posts regarding movies to private profiles in the website's database. Narayanan and Shmatikov [11] were able to deanonymize over 80% of the published Netflix dataset, a big movie database measuring more than half a million accounts. The attack succeeded even with imprecise auxiliary information. Merener [10] elaborates on the algorithms used in [11]. In addition, he demonstrates the importance of sparsity and the long tail phenomenon, in the efficiency of the attacks. Al-Azizy et al. [2] survey data deanonymization techniques and cluster them according to they type of auxiliary information, and the structure of the datasets. In this work we are motivated by these attacks. However, we do not try to develop a privacy mechanism but rather give users a tool that they can use to estimate their privacy.

2.2 Privacy metrics

Many privacy metrics have been proposed for the field of statistical databases such as k-anonymity [16], l-diversity [9], or t-closeness [8]. However, due to multidimensionality and sparsity in microdata publication, traditional anonymization approaches and privacy metrics are not effective [1]. Parra-Arnau et al. [12] propose the use of the Kullback-Leibler divergence as a privacy metric. The rationale behind their metric is that profiles that deviate less from an "average" profile, are less likely to be selected for reidentification. In our setting however, we do not have access to the database in order to create an average profile. A detailed analysis of the most used privacy metrics was done by Wagner and Eckhoff [17], while a survey of privacy preserving data publishing techniques was presented by Chen et al. [3]. To the best of our knowledge, this is the first work that addresses the problem of estimating privacy levels based mainly on published statistics, and the first tool proposed that is user-centric and requires minimal information from the provider's side.

3 PRIVACY ASSESSMENT MECHANISM

3.1 Overview of the system

An overview of our system in presented in Figure 1. The purpose of the tool is to give customers the ability to estimate their privacy level (i.e., their reidentification risk) before using a recommendation service. For this, we require a provider that has a microdata database. The provider creates privacy groups, based on his existing clients, using our proposed metric (Figure 1: A). These groups consist of users whose risk of reidentification is on roughly the same level. In addition, the provider publishes statistics regarding the popularity of all items (Figure 1: B). On the client's side, the tool estimates the privacy score using those public statistics and his microdata (Figure 1: C and D). Finally, the user gets a visual representation of his privacy level, illustrating the privacy group in which he was categorized.

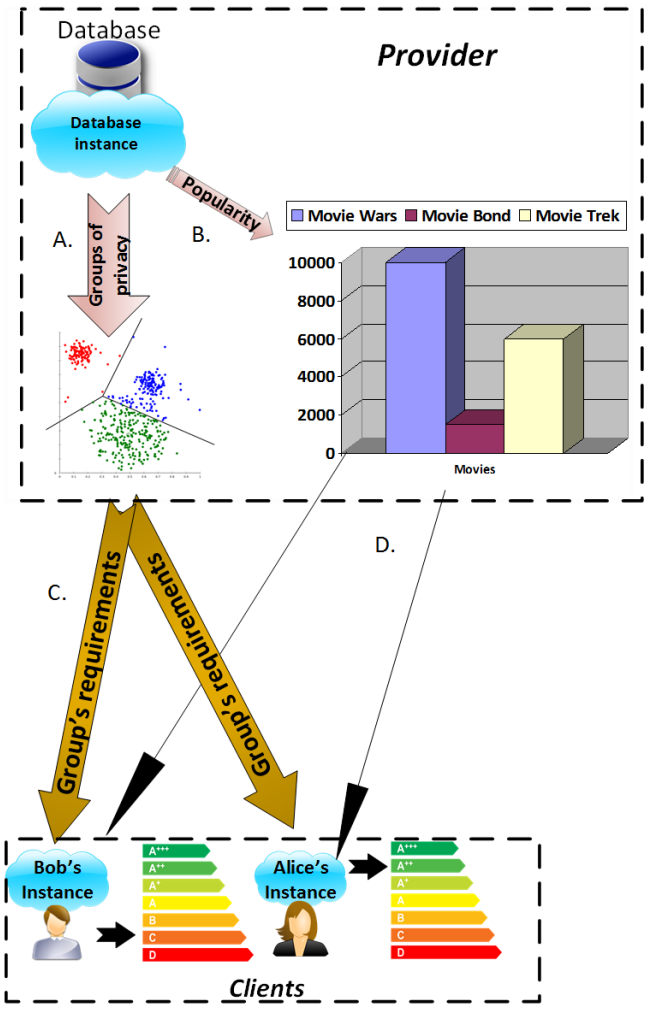

Figure 1: The two parts of our system. The database instance, publishes popularity statistics for all items. In addition, using our metric it classifies existing users to groups of privacy. The client's part of our system receives those statistics and calculates the user's privacy score.

3.2 Algorithm and privacy metric

3.2.1 Privacy metric. Our tool requires a metric that can estimate individuals' privacy level locally, using minimal information from the provider. Hence, the need for a new privacy metric arises, because existing privacy metrics used in databases are not applicable due to two reasons. First, we work with providers that store microdata, characterized by sparsity and high dimensionality. Hence, exclusive combinations of even popular choices might identify clients. Second, all of the proposed metrics require access to other user's profiles in the database, in order to compute a privacy score, which is not possible in a setting where the score is computed at the client's side. We identify the following requirements for a new privacy metric necessary for our tool:

- It should use mainly the user's preferences for computations,
- be lightweight, in order to be computed on user's devices, and
- require minimum information from the provider's side.

Intuitively, we expect that rare items contribute a lot to the distinguishability of users, as some combinations of items can directly identify them (e.g., items preferred by just one user). The total amount of items, however, also plays a significant role on the distinguishability of users, as users with many items usually have more identifying combinations of their microdata. We combine those two factors in a privacy metric. In our experiments an item is considered as rare if it had less than 100 users interacting with it, otherwise we consider it as popular. After extensive testing, we identify the following formula as the most promising in capturing the reidentification risk using only information from the user's side and popularity statistics about the items:

$$privacy\_score = \frac{all - popular}{all} + log(all). \quad (1)$$

Here, *all* refers to the total amount of items the user has, while *popular* refers to the number of his popular items. Since we have only two categories, $all - popular$ equals the number of *rare* items. An immediate observation is that the second part of the formula dominates the first with the logarithm ranging on any positive integer while the fraction (first component) ranges to [0,1]. However, the first part is intended to cover the (many) cases of users which have just a handful of items available at their profile. Because of the limited information available on their profiles, the ratio of rare items is the major identifying factor, instead of the possible combination of their (few) items. Another observation is that our metric may discourage early adopters, as the normalization parameters would be misleading if the database is almost empty (i.e., a fresh system). However, we argue that if a provider would publish a sanitized database for research purposes it would make sure that it offers some utility to researchers, something that an empty database is lacking. Hence, even in the case where early adopters receive incorrect privacy scores, as the database gets populated, their system would be able to correctly calculate their privacy level.

3.2.2 Algorithms. Our system consists of two components: (a) The *database instance* located at the provider and (b) the *user instance* located on candidate clients' devices.

Database instance. We represent a database of microdata as N x M matrix. We assume that every row in the database is a profile $p \in P$, where P denotes the collection of users, while every column denotes an item.

The algorithm running on the provider is presented in Algorithm 1. The provider first calculates the popularity of all items (line 2), based on how many users have interacted with them, and groups the items into two major categories, "popular" and "rare", in line 3.

After creating the two categories based on items' popularities, the provider starts calculating the privacy score according to Equation (1) for all existing clients in his database, as seen in line 4. For every user, our tool finds the total amount of items and the amount

Algorithm 1 Provider's algorithm

```
1:  procedure PRIVACY_GROUPS
2:      find_item_popularities()
3:      group_items_to_bins()                    ▷ based on their popularity
4:      for p in P do
5:          all ← number_of_all_user_items()
6:          popular ← number_of_popular_user_items()
7:          privacy[p] ← (all − popular)/all + log(all)
8:      most_prv← max(privacy)                   ▷ most private user
9:      least_prv ← min(privacy)                 ▷ least private user
10:     for p in P do                            ▷ normalize score
11:         privacy[p] ← 1 − (p − least_prv)/(most_prv − least_prv)
12:     (centroids,clusters) ← run_clustering(P)
13:     centroids ← reduce_noise_in_clusters()
14:     publish_item_popularities()
15:     publish_centroids_of_privacy_groups()
16:     publish_norm_parameters() ▷ max and min privacy scores
```

of items categorized as popular (lines 5 and 6). In line 7, the privacy score for each user is calculated. The lowest and highest scores of all users are found in lines 8 and 9. Then, these scores are used to normalize every user's score to the interval [0,1] (lines 10 and 11). In the remainder of this paper, we will refer to those two values as the "normalization parameters". A score of 1 indicates the most private user.

In lines 12 and 13, we cluster all users according to their privacy score. For the clustering process, we use the X-means algorithm [13], which is an extended k-means algorithm that automatically determines the optimal number of clusters based on the Bayesian Information Criterion (BIC) scores. Even though we use an automated process to define the number of clusters (hence the number of the privacy groups), they can also be selected by the provider by substituting the X-means algorithm with any other clustering method. Noise reduction is performed on the clusters by filtering out the top and bottom 5% of every group's users, leading to higher intra cluster similarity without the effect of outliers (line 13).

In lines 14 - 16, the popularity of all items, the normalization parameters, as well as the centroids of each cluster (privacy groups) are published. The popularity can be published in a variety of ways depending on the provider's policy. Those include, charts with the items popularity, colored indicators next to each item name or, as common in many websites, quotes that refer to the number of customers that interacted with the item. We propose the provider to publish popularity information regarding all items, and not just the required ones to new customers, in order to avoid inference attacks based on the information provided. The provider will periodically run the above procedure due to the dynamic nature of the service, as new items and clients are added on a frequent basis.

User instance. The tool on the clients' side locally computes the privacy score of the user, hereby using the published information from the provider. Following Algorithm 2, it receives from the provider the popularity of all items in the database, the normalization parameters, and the centroids of the privacy groups (lines 2 - 4). The tool then computes the privacy score, based on the items the user wants to disclose to the provider in line 5, using Equation (1), and then normalizes the score in line 6 using the normalization parameters. Using the privacy score and the centroids of each cluster, in line 7, the tool classifies the user in one of the existing privacy groups. The visual result is finally displayed to the client, as seen in line 8.

Algorithm 2 User's algorithm

```
1:  procedure FIND_PRIVACY_SCORE
2:      receive_item_popularities()
3:      receive_centroids()
4:      (least_prv, most_prv) ← receive_norm_parameters()
5:      privacy_score← (all − popular)/all + log(all)
6:      privacy_score← 1 − (privacy_score − least_prv)/(most_prv − least_prv)
7:      find_privacy_group(privacy_score, centroids)
8:      present_visual_result(privacy_score)
```

4 EVALUATION

4.1 Datasets

We evaluate our tool using real datasets, in order to capture realistic profile conditions that an artificial dataset would not provide. We use the following datasets in our evaluation:

Netflix. [1] Netflix is an entertainment company providing video on demand. However, during the time of this dataset collection Netflix was mostly active in online DVD rental. The dataset was released for research purposes and more precisely, to support participants in the Netflix Prize contest. It is a fraction of the original dataset and it contains ratings between December 1999 and December 2005 from a huge community, numbering more than 480 thousand subscribers, 100 million ratings and over 17 thousand items.

Yahoo! Music.[2] This dataset represents a snapshot of the Yahoo! Music community's preferences for various songs. The dataset contains over 717 million ratings of 136 thousand songs given by 1.8 million users of Yahoo! Music services. The data was collected between 2002 and 2006.

Yahoo! Movies.[2] This dataset contains a small sample of the Yahoo! Movies community's preferences for various movies. Users are represented by numerical pseudonyms, so that no identifying information is revealed. User ratings are on a scale from A+ to F.

Due to increased execution time, we randomly sampled the original vast datasets, and work with subsets. However, this does not affect the quality of our experiments: there is theoretical evidence that random sampling of a database has the same sparseness as the original [11]. Hence, sparseness, crucial for privacy in microdata publication, is not affected due to our sampling method. For the

[1] http://www.netflixprize.com/faq.html
[2] https://webscope.sandbox.yahoo.com

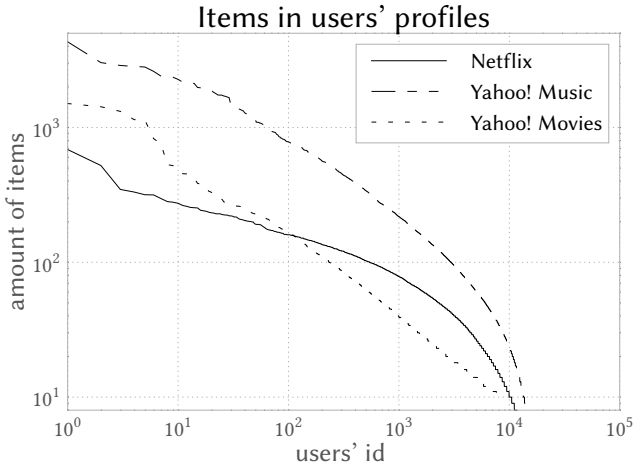

Figure 2: Number of items in user profiles. Users are represented by consecutive ids on the x axis. They are sorted descending, according to the amount of items in their profiles. The number of items is displayed on the y axis. Both axes are in logarithmic scale.

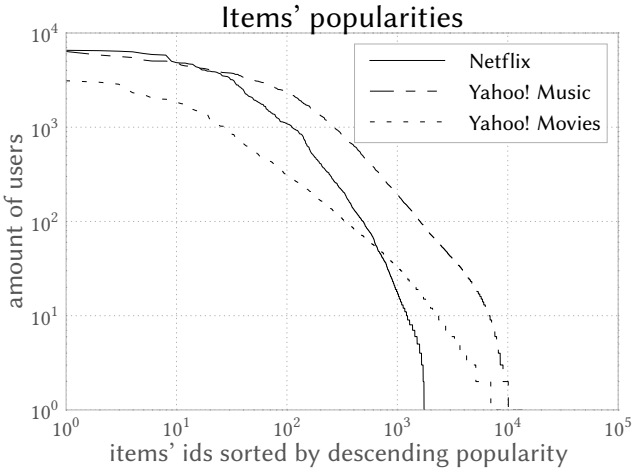

Figure 3: Items' popularities. Items are represented by consecutive ids on the x axis. They are sorted descending, according to how many users have interacted with every item. On the y axis, the amount of users is displayed. Both axes are in logarithmic scale.

	Users	Ratings	Items	Items/User
Netflix	9,938	426,639	1,763	38.5
Yahoo! Music	13,265	1,190,496	13,630	79.9
Yahoo! Movies	7,626	201,579	11,916	26.6

Table 1: Statistics of datasets. The datasets presented are randomly sampled from the original datasets and without outliers.

remainder of the paper, whenever a dataset is mentioned with its name, it will refer to the subset, created for the experiments.

For the evaluation we first sampled the original datasets and then excluded users which were considered as outliers (incomplete profiles with less that 8 items or users with more than 5 thousands items). In Table 1, we summarize the datasets used in this paper. In Figure 2 we plot the amount of items per user profile for all datasets. On the x axis, the users are sorted descending according to the amount of items they have in their profile, while the y axis indicates the amount of items. Both axes are in logarithmic scale. In Figure 3, we plot the amount of users that have interacted with each item. The items for each dataset are sorted descending on the x axis, while the y axis indicates the number of users. Again, both axes are in logarithmic scale. It is clear from Table 1 and Figures 2 and 3, that the datasets are significantly different. Yahoo!-Music users have on average more items per profile, than the rest of the datasets. Yahoo! Movies has both the lowest average amount of items per profile and the lowest ratio of items per profile against all the items in a dataset, which is 0.2%. For comparison, the ratio for Netflix is 2.18% and 0.5% for Yahoo!-Music. Regarding the item's popularities in Figure 3, we observe that the Yahoo! Movies dataset has 50 items that are preferred by more than 1000 users. The Yahoo! Music dataset however, has an order of magnitude more items (500), that are preferred by more than 1000 users.

4.2 Designing the experiments

In order to establish a ground truth to validate our metric, we perform simulations of data breaches on all datasets. For this reason, we assume a realistic attacker that has auxiliary information for every user and tries to reidentify them given only the stored microdata. The Anonymity Set (AS) of a user consists of all profiles that match the adversary's auxiliary information regarding that user. After each deanonymization attempt, every user has an anonymity set $1 \leq |AS| \leq |P|$, where $|AS| = 1$ suggests that the attack was successful. In case where $|AS| \neq 1$, the adversary cannot select the correct profile from the AS with probability higher than $1/|AS|$. Hence, bigger anonymity sets mean that the user is more private. The average size of the AS is a good estimate of the users' privacy, since profiles that are extremely common or identical with many others, do not enhance the adversary's knowledge. To use the minimum or maximum $|AS|$, instead of the average size of the AS, would be too optimistic or pessimistic, respectively, because those events are seldom.

4.2.1 Adversarial Model. In our effort to simulate a realistic adversary, we assume that he has information regarding the existence of several items in every user's profile. In this paper, we consider adversaries that are able to find information about individuals and then try to enrich their knowledge about them. The auxiliary information (the adversary's information) corresponds to 5%-20% of randomly selected items per profile. The adversary attempts to find all profiles in the sanitized database that match his auxiliary information, hoping that only one profile will be returned, the one that he was looking for.

	Spearman's Correlation	Kendall's Correlation
Netflix	0.83	0.65
Yahoo!-Music	0.91	0.73
Yahoo!-Movies	0.84	0.64

Table 2: Correlation coefficients for Spearman's and Kendall's correlation. The compared variables are the average anonymity set for each user and his privacy score based on our metric.

4.2.2 Methodology. We perform Monte Carlo simulation of $n = 10,000$ deanonymization attacks for all users in the datasets. For every user in every round, we select randomly a number between 5-20, denoting the percentage of the adversary's auxiliary information regarding that user. Then, we randomly sample the user's profile for that amount of items. Using this information, we query the database and save the amount of the profiles returned, which is the anonymity set for that user in the specific round. We then average the results of all rounds for each user, in order to estimate the average $|AS|$.

As mentioned in Section 3.2.2, we separate the items in two popularity categories. For every profile in all datasets, we calculate the privacy score based on our metric. Following the algorithm on the provider's side (Algorithm 1), we cluster the profiles based on their privacy score and perform the noise reduction technique.

4.3 Validity of the model

In Figures 4 to 6, we present the correlation between our privacy metric and $|AS|$, as well as the clustering results for all datasets. On the x axis we present the average anonymity set for every user, while on the y axis the score of our privacy metric. We plot with different colors and markers the resulting privacy groups. For all three datasets, our algorithm returned three privacy groups. Those are the "not safe", "medium risk" and "safe". The dashed horizontal lines are the borders between the privacy groups. The lines correspond to the lowest and highest score of each cluster. A black star in the central left part of every cluster denotes the cluster's centroid (calculated solely based only on the privacy score, see lines 11-12 in Alg. 1).

On all datasets, our metric has a strong positive correlation with the anonymity set. More precisely, we measure the correlation using Spearman's rank correlation, as well as with Kendall's rank correlation. The Spearman's correlation is a nonparametric measure of the rank correlation between two variables (in our case the average $|AS|$ per user and his privacy score). It estimates how well the relationship of the two variables is described by a monotonic function. Kendall's correlation is also a non parametric measure, and assesses the ordinal association between two variables. We present the results in Table 2. It is worth mentioning that for all correlations, the p-values of the null hypothesis were zero. This is known to happen in big datasets, with sizes bigger than a few hundred entries. However, a visual examination of the scatter plots (Figures 4 to 6) illustrates that the data are not random, but follow a pattern. Thus, the correlation coefficient scores, as well as the p-values, are trustworthy.

Figures 4 to 6 clearly show that our tool correctly separates users into privacy groups based on their privacy score. The smaller $|AS|$ for the "safe" category range from 50 (Yahoo! Movies) to 200 (Netflix). According to our metric, the scores of the least private users in this category had a score of 0.9 (Yahoo! Movies) and 0.83 (Netflix). The "not safe" group on all datasets includes users who have anonymity sets ranging from 1 (successful reidentification) to 50 (2% chance for successful reidentification) and privacy scores of less than 0.7. However, the anonymity set for the "safe" group on all datasets, ranges in value from 50 (worst case in our experiments) to a few thousands. This happens because users with just a few items (lower risk of reidentification) represent a big part of the database population. Those users, depending on the amount of rare movies they have, have anonymity sets that are significantly different. Users that are in the "safe" group, have low reidentification chances ranging from 2% ($|AS| = 50$) to 0.045% ($|AS| = 2200$).

On all three privacy groups we observe some small overlaps, with regard to the anonymity set, with the next group. This happens because we evaluate real datasets and we rely on probabilistic attacks to create the anonymity sets. However, the majority of the points lay in different AS ranges, depending the privacy group. For example, on the Yahoo! Music dataset (Figure 5), most of the profiles characterized as "safe" are in the AS range of 180 to 1400. Even though a few of the safe profiles are in the range of 180 to 230, where many profiles of the "medium-risk" group are located, they are outliers and do not represent the main mass of the group. In addition, the overlapping ranges are always on the "more private" side of the anonymity set (bigger anonymity sets) where one group fades out and the next one starts.

It becomes clear from Figures 4 to 6 and Table 2, that our privacy tool is able to estimate the privacy levels of individuals as good as the average anonymity set. However, instead of using the whole database, it uses minimal information from the provider's side (published statistics, three centroids and two normalization parameters).

5 PRIVACY THERMOMETER

In this work we develop a tool that is user centric, hence we focus on the usability of our tool. For this reason, we choose to display the privacy levels to the user in a visual, easily comprehensible way. Our choice is based on the notion that privacy is something that everyone should have, not only those with technical background. Hence, complicated scores and definitions should be avoided.

In our attempt to find a way to visualize the resulting privacy level, we identify the following requirements:

- The presentation should avoid the use of numerical scores as they might be confusing,
- preferably it should use colored indicators because they are easily comprehensible and
- ideally, it should relate to a visualization that consumers are already familiar with.

For these reasons we decided to borrow the successful design of the European energy efficiency label (Figure 7), which was introduced by the EU Directive 92/75/EC in 1992.

This label belongs in the family of comparative labels [15], that allow users to compare devices based on a common characteristic

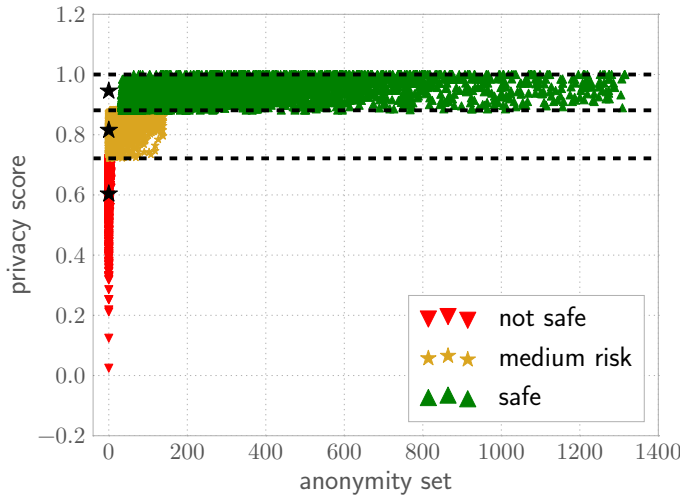

Figure 4: Scatter plot portraying the correlation between the average anonymity set and the privacy score for Yahoo! Movies. The x axis displays the average $|AS|$ for every user, while the y axis the privacy score returned from our metric. Different colors and markers denote the privacy groups returned from our tool. The black stars on the left of the figure are the centroids of the privacy groups. The dashed black lines separate the privacy groups.

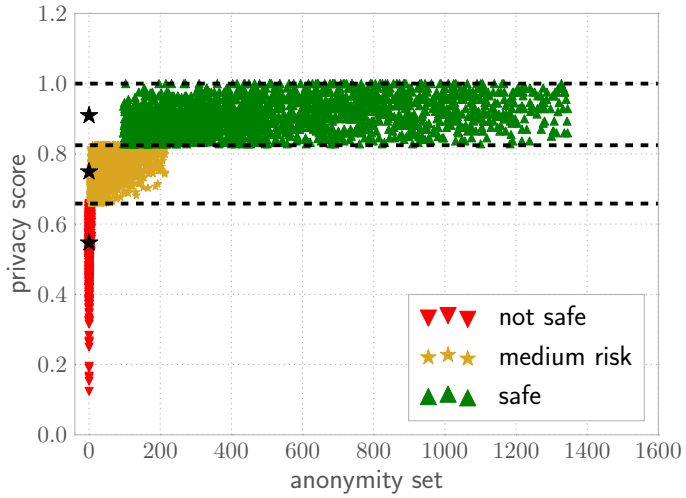

Figure 5: Similar to Figure 4, this scatter plot is showing the correlation between the average anonymity set and our privacy score for the Yahoo! Music dataset, as well as the resulting privacy groups.

(in this work users would compare with other users based on their privacy levels). The privacy levels consists of colored indicators in

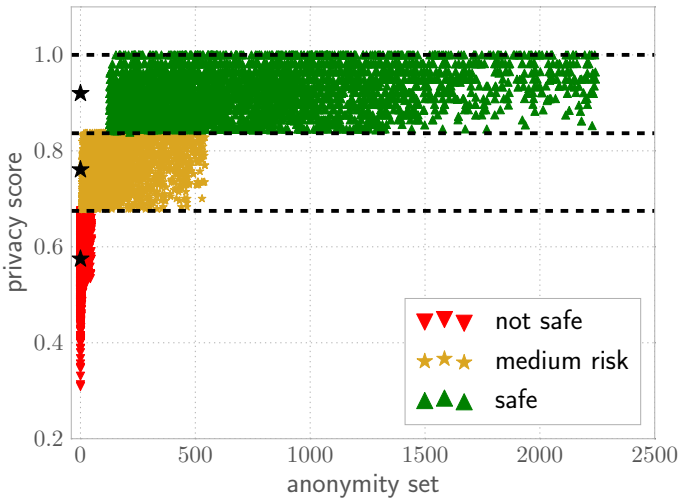

Figure 6: Scatter plot displaying the correlation between the average anonymity set and our privacy score for Netflix, similar to Figures 4 and 5.

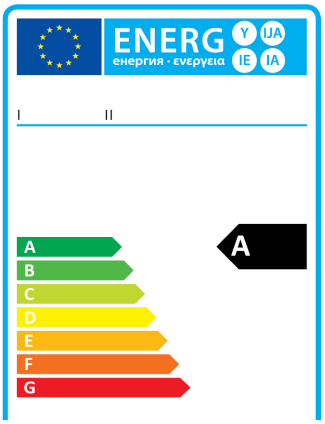

Figure 7: EU energy efficiency label.

the form of bars. The bars correspond to the privacy groups created on the provider's side. The colors range from red to dark green, with red denoting the worst privacy group while dark green is the best. The bars have different lengths with the longest one being the red bar, denoting the highest reidentification risk. All of the bars in our label are used to represent privacy groups (closed alphabetic scale) and there are no unused bars left for future groups (open alphabetic scale). According to research, open alphabetic labels, which would allow for future bars (i.e., more energy efficient devices or in our case, more privacy groups), are harder for consumers to understand than closed alphabetic scales [5]. In addition, our decision to display verbal indicators to each bar instead of numerical has a two fold advantage. According to [5], consumers understand better alphabetic scales than numeric ones. Furthermore, in one of the earliest experiments regarding energy labels, Chestnut [4]

Figure 8: Example of privacy label for the Netflix dataset.

found that verbal ratings are effective for long-term storage of consumer information. Hence, this design enhances our effort in raising privacy awareness, as consumers will be more likely to remember their privacy scores long after they have been displayed to them.

In Figure 8, we give an example of how a privacy label would look for the Netflix dataset. In order to provide more granular information to clients we split the "medium safe" and "safe" group into two categories, depending on the privacy score. Hence, the privacy label for Netflix has five privacy bars denoting each one of the privacy groups.

6 CONCLUSION

Microdata publication presents a significant privacy threat even when data is published sanitized. Motivated by recent deanonymization attacks and focusing primarily on user's privacy, we address the issue of estimating privacy levels with minimal information from the provider's side. More precisely, we develop a tool that runs on the provider's side as well as on clients' devices, that classifies existing users into privacy groups and publishes popularity statistics regarding all items in the database. In order to estimate the privacy of specific individuals using such minimal information, we develop a new lightweight privacy metric. The metric is based on published statistics for an individual's microdata. The result of our tool is a visual estimation of the user's privacy level. Our metric is based on the two fundamental factors, with regard to the user's microdata, that lead to successful reidentification attacks, namely: the total amount of items, as well as the amount of rare items. We evaluate our tool on three real world datasets that are significantly different. We perform reidentification attacks based on random auxiliary information for every user in our datasets. By estimating the average anonymity set of each user and comparing it to our metric's score, we could demonstrate strong correlations between the two on all datasets. The experiments show that our tool is able to estimate the privacy users will have a-priori sharing their data comparatively well. We present the privacy result to clients using a visual representation. In this way, they are more likely to remember their privacy level long after it was displayed to them, hence be more privacy aware. Interesting future work would be a usability evaluation of both the tool and the proposed visual method on real application scenarios and actual users.

ACKNOWLEDGMENT

This work has been funded by the DFG as part of project A1 within the RTG 2050 "Privacy and Trust for Mobile Users".

REFERENCES

[1] Charu C Aggarwal. 2005. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st international conference on Very large data bases*. VLDB Endowment, 901–909.
[2] Dalal Al-Azizy, David Millard, Iraklis Symeonidis, Kieron O'Hara, and Nigel Shadbolt. 2015. A literature survey and classifications on data deanonymisation. In *International Conference on Risks and Security of Internet and Systems*. Springer, 36–51.
[3] Bee-Chung Chen, Daniel Kifer, Kristen LeFevre, Ashwin Machanavajjhala, and others. 2009. Privacy-preserving data publishing. *Foundations and Trends® in Databases* 2, 1–2 (2009), 1–167.
[4] Robert W. Chestnut. 1976. The Impact of Energy-Efficiency Ratings: Selective vs. Elaborative Encoding. *Purdue Papers in Consumer Psychology* 160 (1976).
[5] London Economics. 2014. Study on the impact of the energy label-and potential changes to it-on consumer understanding and on purchase decisions. (2014).
[6] Dan Frankowski, Dan Cosley, Shilad Sen, Loren Terveen, and John Riedl. 2006. You are what you say: privacy risks of public mentions. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 565–572.
[7] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5802–5805.
[8] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 106–115.
[9] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007), 3.
[10] Martin M Merener. 2012. Theoretical results on de-anonymization via linkage attacks. *Transactions on Data Privacy* 5, 2 (2012), 377–402.
[11] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. IEEE Symposium on*. IEEE, 111–125.
[12] Javier Parra-Arnau, David Rebollo-Monedero, and Jordi Forné. 2014. Measuring the privacy of user profiles in personalized information systems. *Future Generation Computer Systems* 33 (2014), 53–63.
[13] Dan Pelleg, Andrew W Moore, and others. 2000. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *ICML*, Vol. 1. 727–734.
[14] Naren Ramakrishnan, Benjamin J Keller, Batul J Mirza, Ananth Y Grama, and George Karypis. 2001. Privacy risks in recommender systems. *IEEE Internet Computing* 5, 6 (2001), 54.
[15] Moritz Rohling and Renate Schubert. 2013. Energy labels for household appliances and their disclosure format: A literature review. *Institute for Environmental Decisions (IED), ETH Zurich, www. ied. ethz. ch/pub/pdf/IED_WP21_Rohling_Schubert. pdf* (2013).
[16] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570.
[17] Isabel Wagner and David Eckhoff. 2015. Technical privacy metrics: a systematic survey. *arXiv preprint arXiv:1512.00327* (2015).