Towards Long-Term Privacy Bounds in Open Data Publishing

Clément Massart
Université catholique de Louvain
ICTEAM Institute, Group Crypto
Louvain-la-Neuve, Belgium
clement.massart@uclouvain.be

François-Xavier Standaert
Université catholique de Louvain
ICTEAM Institute, Group Crypto
Louvain-la-Neuve, Belgium
fstandae@uclouvain.be

***Abstract*—Previous works showed that privacy-preserving open data publishing is a challenging (if achievable at all) goal. Risks are in general hard to quantify and may in particular vary significantly in case databases are extended over time with new data (or are merged). In this paper, we show that the risks of re-identification due to the predictive power of the data can be bounded under reasonable assumptions, thanks to recently introduced information theoretic tools. We illustrate our methodology on a Netflix dataset that was shown to raise privacy issues by Narayanan and Shmatikov (S&P 2008) and motivate a simple protection mechanism based on data swappings as an insufficient but utility-preserving improvement.**

I. Introduction

The General Data Protection Regulation (GDPR) became effective the 25^{th} of May 2018. At the high level, it formalizes the privacy requirements that companies must enforce when manipulating data. For this purpose, the GDPR provides general directions. Yet, it sill lacks systematic evaluation / certification tools to quantify how these recommendations translate into a concrete level of privacy (see for example articles 25.3, 35, 42.1, 43.9, 57.1 of the GDPR).

In the context of open data publishing (discussed in article 89.2 and rule 157 of the GDPR), the only restriction imposed so far is the "anonymization" of the data (see https://www.europeandataportal.eu/en/highlights/protecting-data-and-opening-data). As overviewed by Fung et al. in their survey of recent developments in privacy-preserving data publishing, various metrics can be used to quantify the risks of re-identification when an adversary obtains "internal leakages" allowing to connect a line of the database to a particular target user [5]. One of the easiest to understand metrics is the k-Anonymity introduced by Sweeney [10]. It gives an intuitive (yet limited) privacy measurement based on the similarity between users. Various refinements exist (see for example the aforementioned survey).

In this paper, we are concerned with the complementary issue that an adversary may also have access to "external leakages". That is, fresh observations collected for a user (presumably in a database) can also lead to re-identifications in case the collected data is sufficiently predictive. The risks of such re-identifications exploiting external leakages are harder to bound, since the predictive power of a database typically increases with the amount of collected data.

We mitigate this issue by showing how to leverage recent results / bounds in the field of secure cryptographic implementations [1]. Precisely, we show that the risks of re-identifications with external leakages can be bounded with information theoretic metrics that can be efficiently computed from a database's content (a similar application of these tools to location privacy can be found in [6]). We also show that the bound becomes tighter as the size of the database increases. Since the risks of re-identification with internal leakages also decrease with this size, it implies that database holders have no incentive to hide data in such privacy assessments.

We finally illustrate the application of these tools to the case of the Netflix prize for improving recommendation system that was launched in 2006. In a work from S&P 2008, it was shown that Netflix pseudonyms can be linked to IMDb public accounts based on internal leakages [7]. (The same authors extended their work to social networks in [8]). We show how our tools can be used to quantify the privacy risks in this case study, and how simple manipulations (i.e., data swappings [9]) can reduce these risks at a limited utility cost.

II. Background and notations

We consider a context where users utilize a service and the service provider can collect information such as the users' name, address and their activity while using the service.

A. Data specification

We first define the set of users as:

$$\mathcal{U} = \{u_1, u_2 \ldots, u_{n_u}\},$$

with n_u the number of users. We then define observations $\boldsymbol{o}_{ij}$ that correspond to the j^{th} record for user u_i:

$$\boldsymbol{o}_{ij} = \{o_{ij}^1, \ldots, o_{ij}^{N_c}\},$$

with N_c a number of characteristics. Taking the example of the Netflix database, observations correspond to identifiers, movies, grades, ..., and are reported as:

$$\boldsymbol{o}_{\text{Alice}} = \{\text{Alice}, \text{2019-04-03}, \text{Pulp Fiction}, 5\}.$$

We will assume that the o's are discrete.

A user u_i theoretically follows an unknown distribution which we formalize with the Probability Mass Function (PMF)

$\mathsf{g}(\boldsymbol{o}|u_i)$. In open data publishing, a set of sample observations following this distribution are collected for each user as:

$$\mathcal{D}_i \xleftarrow{N_{\boldsymbol{o}}^i} \mathsf{g}(\boldsymbol{o}|u_i),$$

with $N_{\boldsymbol{o}}^i$ the total number of observation from user u_i.

B. Types of estimations

Given the set of observations $\mathcal{D}_i$'s, recommendation systems will generally try to model the true distributions $\mathsf{g}(\boldsymbol{o}|u_i)$. We consider two types of estimations for this purpose, namely direct estimation (di) and cross-validated estimation (cv). We call direct estimation a modeling process using directly all the available data. By contrast, in the cross-validated estimation the datasset is split in K subsets: $K-1$ are used for model estimation, the last one for model testing (and the model estimation and testing are repeated K times). We denote the K subsets as $\mathcal{D}_i^{(k)}$, $k = 1, \ldots, K$, such that $\bigcup_{k=1}^{K} \mathcal{D}_i^{(k)} = \mathcal{D}_i$ and $\mathcal{D}_i^{(k_1)} \bigcap \mathcal{D}_i^{(k_2)} = \emptyset$, for all $k_1 \neq k_2$. The estimated models are written with a tilde symbol for direct estimation:

$$\tilde{\mathsf{g}}(\boldsymbol{o}|u_i) \xleftarrow{\text{di}} \mathcal{D}_i,$$

and with a hat symbol for cross-validation estimation:

$$\left\{\hat{\mathsf{g}}^{(1:k)}(\boldsymbol{o}|u_i), \mathcal{D}_i^{(1:k)}\right\} \xleftarrow{\text{cv}} \mathcal{D}_i.$$

C. Estimation tools

In order to estimate the true distributions $\mathsf{g}(\boldsymbol{o}|u_i)$ we also need to define statistical tools. The choice of a (e.g., parametric or non-parametric) statistical tool directly impacts the speed of convergence and accuracy of the models, so the closeness between $\mathsf{g}(\boldsymbol{o}|u_i)$ and $\tilde{\mathsf{g}}(\boldsymbol{o}|u_i)$ or $\hat{\mathsf{g}}(\boldsymbol{o}|u_i)$. We consider two simple options for this purpose.

On the one hand, we use a first-order model which treats the characteristics of each observation independently. Concretely, the estimated models are then computed as follows:

$$\tilde{\mathsf{g}}_1\left(\boldsymbol{o}(c) \mid u_i\right) = \frac{1}{N_{\boldsymbol{o}}^i} \sum_{\boldsymbol{o}' \in \mathcal{D}_i} \boldsymbol{o}'(c),$$

$$\hat{\mathsf{g}}_1^{(k)}\left(\boldsymbol{o}(c) \mid u_i\right) = \frac{1}{N_{\boldsymbol{o}}^i - |\mathcal{D}_i^{(k)}|} \sum_{\boldsymbol{o}' \in \mathcal{D}_i \setminus \mathcal{D}_i^{(k)}} \boldsymbol{o}'(c),$$

for any characteristic c with value $\boldsymbol{o}(c)$. For all c's, it counts the number of times a value appears. The resulting model is an histogram of which the size depends on N_c and the cardinality of c (i.e., the range of values the characteristics can take).

Since characteristics can be correlated, we also consider an exhaustive model, which we denote as $\tilde{\mathsf{g}}_{ex}$ or $\hat{\mathsf{g}}_{ex}^{(k)}$ and that directly estimates a histogram for all possible observations. This process can model any type of correlation (i.e., any possible combination of characteristics) but it is naturally much more expensive to estimate. This is reflected by the size of the histograms. Taking the example of our following data where we have 27 categories and each category can come with 5 scores, the exhaustive histogram has 5^{27} possible bins while the first-order one only has $5 \cdot 27$.

Note that intermediate models capturing correlations up to a certain order could also be considered.

Based on these models, the conditional probabilities $\tilde{\Pr}[u_i|\boldsymbol{o}]$ and $\hat{\Pr}[u_i|\boldsymbol{o}]$ that we will need to estimate our metrics can be directly derived thanks to Bayes, assuming an a priori uniform distribution for the users:

$$\tilde{\Pr}[u_i|\boldsymbol{o}] = \frac{\tilde{\mathsf{g}}(\boldsymbol{o}|u_i)}{\sum_{j=1}^{n_u} \tilde{\mathsf{g}}(\boldsymbol{o}|u_j)},$$

$$\hat{\Pr}[u_i|\boldsymbol{o}] = \frac{\hat{\mathsf{g}}(\boldsymbol{o}|u_i)}{\sum_{j=1}^{n_u} \hat{\mathsf{g}}(\boldsymbol{o}|u_j)}.$$

III. Threat model and metrics

A. Threat model

We consider an adversary who aims at re-identifying users in a pseudonymized database, as recommended by the EU in the GDPR. The resulting threat model is depicted in Figure 1. In such a scenario the adversary must link true user identities to pseudonyms thanks to some information that we call leakages. We consider two types of leakages: internal leakages which correspond to a couple $(u, \boldsymbol{o})$ such that both the user and the observation are in the database; external leakages where the user is (assumed to be) in the database but the observation is not (i.e., it is a fresh one). In this second case, the attack crucially relies on the predictive power of the model (as evaluated thanks to cross-validation).

The adversary's success in this threat model depends on two main quantities: the size of the database, measured thanks to the number of observations per user N_o^i (which we assume to be equal for all users $N_o^i = N_o$), and the number of leakages per user M_o^i (where we assume the same $M_o^i = M_o$). Concretely, N_o primarily affects the accuracy of the model, while M_o improves the re-identification rate.

B. Metrics

We use two main metrics to evaluate our threat model: the Perceived Information (PI) and the Hypothetical Information (HI). They provide (on average) a lower and an upper bound for the Mutual Information (MI) that we cannot directly compute in the absence of an exact knowledge of the users' true distributions [1]. As discussed in [6], the HI and PI can be related to the number of (internal and external) leakages needed to re-identify a user.

As will be illustrated next, the fact that the HI upper bounds the PI is handy in a privacy setting, since it implies that the risks of improved (more predictive) models allowing better attacks with external leakages can be bounded based on a database's content. The bound becomes tight as the number of observations in the database increases.[1]

The Hypothetical Information is expressed as,

$$\tilde{\mathrm{HI}}(U;O) = H[U] + \sum_{u_i \in \mathcal{U}} \Pr[u_i] \cdot \sum_{\boldsymbol{o} \in \mathcal{D}_i} \tilde{\mathsf{f}}(\boldsymbol{o}|u_i) \cdot \log_2 \tilde{\Pr}[u_i|\boldsymbol{o}].$$

[1] If the users' distributions are stationary – if they don't the bound is not tight but the risks are also reduced since the models become less predictive.

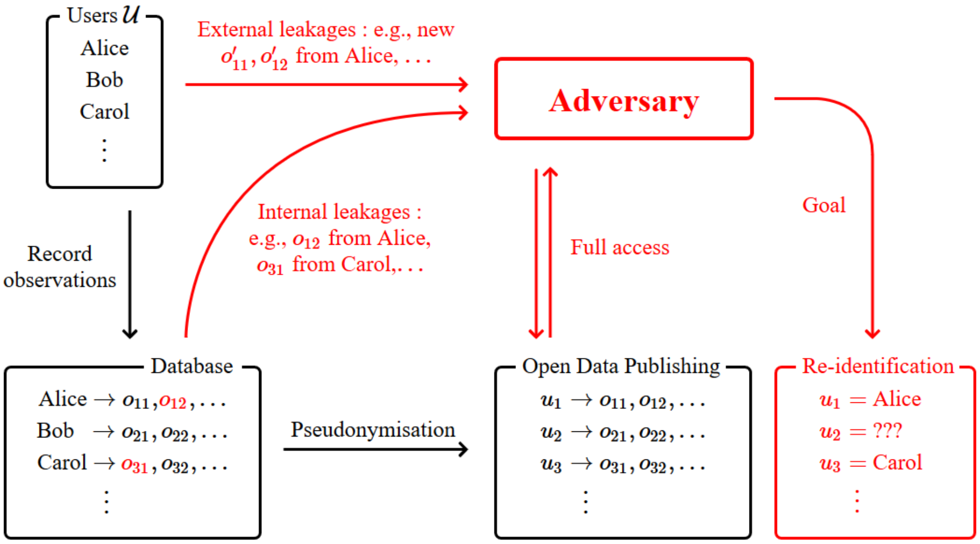
Users $\mathcal{U}$
Alice
Bob
Carol
⋮
External leakages : e.g., new o'_{11}, o'_{12} from Alice, . . .
Adversary
Goal
Internal leakages : e.g., o_{12} from Alice, o_{31} from Carol,. . .
Full access
Record observations
Database
Alice → o_{11}, o_{12}, . . .
Bob → o_{21}, o_{22}, . . .
Carol → o_{31}, o_{32}, . . .
⋮
Pseudonymisation
Open Data Publishing
$u_1 \rightarrow o_{11}, o_{12}, \ldots$
$u_2 \rightarrow o_{21}, o_{22}, \ldots$
$u_3 \rightarrow o_{31}, o_{32}, \ldots$
⋮
Re-identification
u_1 = Alice
u_2 = ???
u_3 = Carol
⋮

Fig. 1. Re-identification threat model.

Intuitively, it depends on the number of collisions among different users' observations. This will lead to a decreasing trend as the size of the database increases in our experiments. It typically means that (on average), users in smaller datasets are easier to re-identify with internal leakages.

The Perceived Information computed for one cross-validation subset can be expressed as:

$$\hat{\mathrm{PI}}^{(j)}(U;O) = H[U] + \sum_{u_i\in\mathcal{U}} \Pr[u_i]\cdot \sum_{\boldsymbol{o}\in\mathcal{D}_i^{(j)}} \frac{1}{|\mathcal{D}_i^{(j)}|}\cdot \log_2 \hat{\Pr}^{(j)}[u_i|\boldsymbol{o}],$$

and a better estimate is then obtained by averaging the K different $\hat{\mathrm{PI}}^{(j)}(U;O)$ values. Thanks to cross-validation, this metric captures the predictive power of the model, which naturally increases (on average) with larger datasets.

Note that estimating the PI requires to deal with outliers, since negligible probabilities for correct users can lead to negative PI values (intuitively reflecting a non-predictive model). We deal with such outliers as in [6] and always reflected the proportion of outliers as f_o in our experiments. (This quantity decreases as the size of the profiling set increases).

In some cases, we will also consider the probability of successful re-identification as an alternative metric. It is more difficult to estimate since it depends jointly on N_o and M_o while the HI and PI metrics depend only on N_o. The number of (internal or external) leakages needed to reach a high probability of success is (inversely) proportional to the HI and PI metrics – so information theoretic metrics are our preferred ones for the evaluation of re-identification attacks. Yet, the success rate sometimes delivers additional intuition. We will in particular consider different levels of success such that a level-l success corresponds to a case where the user to re-identify is among the first l candidates suggested by the attack.

IV. Data description

We next apply our methodology to the Netflix dataset. It was originally published in order to enhance their recommendation system, which led to the privacy issues discussed in [7].

The available dataset regroups 480,189 users who evaluated at least one movie among 17,770 possible ones, between October 1998 and December 2005. The observations contain 4 characteristics: the movie ID, the user ID, the grade and the date of rating. The grades scale from 1 up to 5 and the date is in (year,month,day) format. It corresponds to one eighth of the full Netflix database at the end of 2005.

Considering the full granularity of the data, it turns out that all the observations are unique (or close to be), making any discussion of privacy a bit futile: there are billions of possible observations while the dataset only contains ≈ 100 millions. As a result, our following experiments consider a more optimistic setting from the privacy viewpoint where the (granularity of the) data is reduced in different ways.

First, we removed the time component and the movie ID component which are quite meaningless in the analysis of external leakages. By definition, these quantities are past ones. This is obvious for the time component. For the movie ID one, it relates to the assumption that a single user is unlikely to rate the same movie multiple times. As a result, we decided to report the movies in our database as a combination of categories. Those categories were found thanks to IMDb database. We ended up with a total of 27 categories: Action, Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, Film-Noir, Game-Show, History, Horror, Music, Musical, Mystery, News, Reality-TV, Romance, Sci-Fi, Short, Sport, Talk-Show, Thriller, War, Western. We associated movies and categories with an automated script that checks similarities between Netflix movies and IMDb ones.[2]

Next, we removed all the users with less than 1000 observations and kept this same amount of observations from each of the $13,141$ remaining users. With this reduction we ended up with only 1211 possible combinations of movie types.

Finally, we will analyze different types grades: the "one to five star(s)" one of the original data and a simpler "like-dislike" one. For this second option, we define a like as a grade of 4 or more and a dislike as a grade of 3 or less. We will also consider a "no-grade" case, where users profiles only depend on type of movies they watch (without grades).

V. Experimental results

We now quantify the privacy of users in our modified Netflix dataset and the risks that adversaries can re-identify users.

A. Information theoretic analysis

The HI and PI metrics estimated from the modified Netflix database are reported in Figure 2. As theoretically expected, the (easier-to-estimate) Hypothetical Information (HI) metric is always higher than the Perceived Information (PI) [1]. Hence, it can serve as a bound for the risks of re-identification with external leakages. Since it is monotonously decreasing, this bound becomes tighter as N_o increases.

For the exhaustive model we see a large difference between both metrics, while this difference is much smaller for the first-order model. This is due to the more complex estimation of the exhaustive model. For the first-order model, the HI and PI values for the maximum $N_o = 1800$ are very close, suggesting that this model has converged towards its most informative level (i.e., more observations would not lead better re-identification since all the model parameters are well estimated). Note that the total number of observation is 1800 in this case, since each observation has been split into several independent ones (leading to more "simplified" observations per user). By contrast, such a convergence of the HI and PI curves does not (yet) take place for the exhaustive model which would require much more observations to be perfectly estimated, leading to a much less tight (worst-case) bound.

Concretely, the plots imply that an adversary exploiting a first-order model would be able to extract an amount of information bounded by the HI (i.e., 0.34), and for the amount of observations collected a PI of 0.19 can already be extracted. Given that our experiments include 13,141 users (with $\log_2(13,141) \approx 13.68$), it implies that re-identification could in the worst-case (i.e., for perfectly estimated models) take place after $c \cdot 13.68/0.346 \approx c \cdot 40$ leakage traces (and given the amount of collected observations, $c \cdot 13.68/0.19 \approx c \cdot 72$ leakages are already sufficient), with c a constant depending on the target success rate (e.g., $c = 13.68$ approximately corresponds to an 80% success rate [2], [3]). For the exhaustive model, things get even worse and $c{\cdot}13.68/1.56 \approx c{\cdot}9$ leakages would be sufficient according to the (non tight) bound, due to both the limited number of collected observations and the asymptotically most informative (exhaustive) model.

Note that the previous analysis is an average one, but as suggested by the surfaces of Figure 2, the variability among users is not negligible: some users are (much) more easily re-identifiable than others.

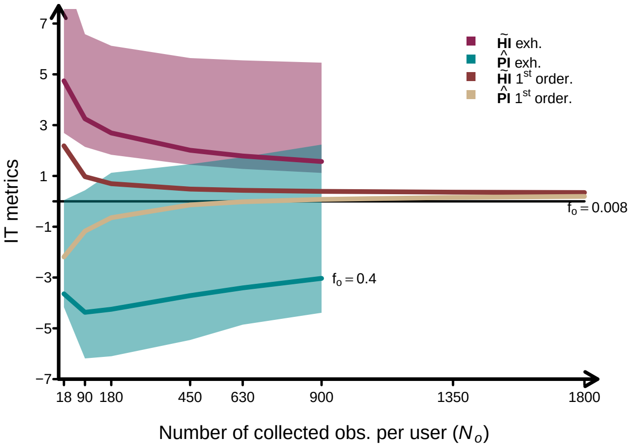

Fig. 2. Information theoretic analysis (II).

B. Impact of granularity

The previous results are extended to the aforementioned levels of granularity for the observations in Table I, where the first-order and exhaustive HI and PI values are provided without grades, with like-dislike grades and with 5-star grades.

There are two opposite effects happening when reducing the granularity. First, it reduces the information available, as reflected by a reduced HI. From the no-grade analysis to the 5-star one, it is constantly increasing. Such a trend is also observed for the PI of the first-order model when using a profiling set of size $1,800$. It means that the model is then able to extract most of the available information. By contrast, this is not the case for the exhaustive model, which has an opposite (decreasing) behavior for the PI. The latter suggests that the model is (much) more complex to estimate and would require (much) more observations to become informative.

This analysis confirms that adding features to estimate in a model implies an increase of the risks of re-identification with internal leakages, while the impact of this addition is contrasted for the PI: if a sufficient number of observations are available, it may improve the asymptotic value of the PI (if the new features capture new details of the true distributions), if not it may reduce the concretely reachable PI.

C. Additional security analysis

Given the variability of users illustrated by Figure 2, one additional question regarding our experiments is whether re-

[2] We therefore cannot pretend that this classification is perfect, but checked that most of the movies were well assigned.

TABLE I
INFORMATION THEORETIC ANALYSIS (II).

N_o		No-grades	Like-dislike	5-star grades
900	$\tilde{\mathrm{HI}}$ (exh.)	0.586	1.001	1.568
	$\hat{\mathrm{PI}}$ (exh.)	-1.782	-2.286	-3.033
	$\tilde{\mathrm{HI}}$ (1st-order)	0.069	0.203	0.392
	$\hat{\mathrm{PI}}$ (1st-order)	0.019	0.090	0.077
1800	$\tilde{\mathrm{HI}}$ (1st-order)	0.061	0.185	0.346
	$\hat{\mathrm{PI}}$ (1st-order)	0.038	0.133	0.194

identification would be significantly easier for certain groups of users having similar behaviors.

We answer this question by investigating the l-order success rate for a re-identification attack with $M_o = 1$ leakage, which is illustrated in Figure 3 for the 5-grade case. Meaningful groups of users would typically be illustrated by a stepped curve for the success rate, which is not observed. We therefore conclude that that grouping users by similar profiles is not helping the re-identification.

Note that the success rate curves would gradually get away from the random line as M_o increases. Note also that the success rate curve for the exhaustive model is below the one of the first-order model, which is expected based on the previous information theoretic analysis.

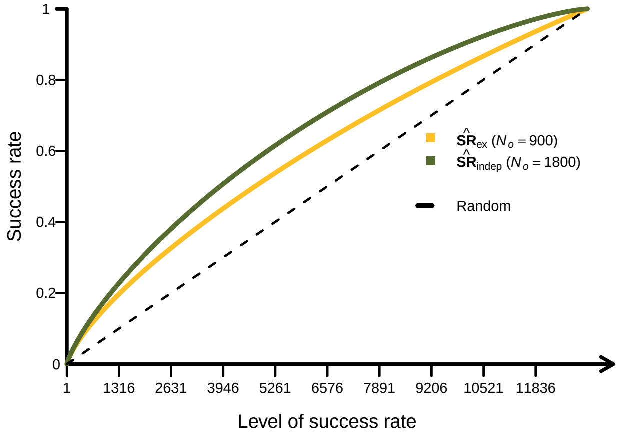

Fig. 3. Re-identification success rate with 5-star grades ($M_o = 1$).

VI. UTILITY-PRESERVING PRIVACY IMPROVEMENT

We conclude the paper by proposing a utility-preserving privacy improvement for the investigated data set.

We use a utility notion inspired from [6] for this purpose. Precisely, we measure utility as the probability to predict the categories of the films watched by the users. Such a utility metric can exploit the same cross-validated estimations as the re-identification attacks with external leakages. The only differences are that it is a single-shot game, meaning that it can be directly evaluated based on the success rate metric, and it aims at predicting categories rather than user IDs (so it essentially exploits the other term of Bayes' law).

Such a utility analysis is represented in Figure 4. We observe that predictions are significantly better than random ones, suggesting that they could be used to guide a recommendation system. The only puzzling fact is the less smooth aspect of the curve corresponding to the exhaustive model. It starts lower than the first-order curve, then rapidly exceeds it until approximately $l = 20$ before running behind it again.

Our tentative explanation for this fact derives from Figure 5, where the categories' distribution are plotted. It shows that categories are very concentrated among a few combinations for the exhaustive representation (meaning that these few categories have a higher chance to be correct). By contrast, the density of this distribution decreases more gradually for the first-order model. As a result, it is natural that the success rate curve is increased for lower-level successes in the case of the exhaustive model (compared to the first-order one).

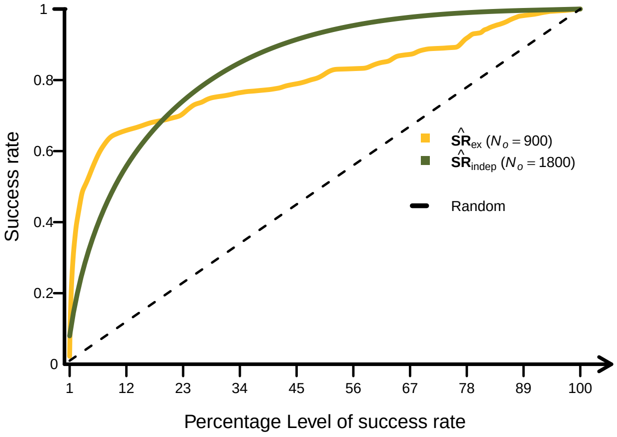

Fig. 4. Success rate of (category) prediction attacks.

Based on the previous experiments we can conclude that the exhaustive model increases the risks of re-identification significantly, while not leading to a comparatively improved utility. As a result, a natural proposal for privacy enhancement is to pre-process the data such that the very possibility to characterize higher-order moments of the observations' distribution vanishes. A very simple solution for this purpose is to break all the observations with combined categories into independent ones, and to exploit data swapping as suggested in [9]. Precisely, for two observations $\boldsymbol{o}_1$ and $\boldsymbol{o}_2$, permuting $\boldsymbol{o}_1(c)$ with $\boldsymbol{o}_2(c)$, for a characteristic c will not affect the first-order modeling, yet it will break any correlation between the characteristics within the observations. As a result, the exhaustive model will not bring any improvement of the re-identification attacks anymore and, as previously mentioned, this will not have any significant utility cost in our case study.

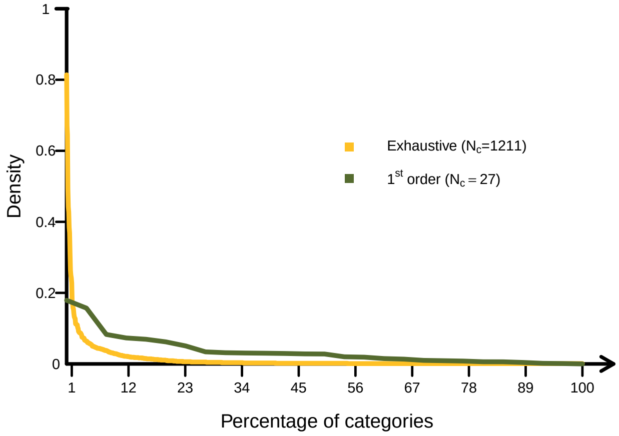

Fig. 5. Density of the categories sorted by cardinality.

Note that the interest of such a data swapping crucially relies on the fact that first-order models are sufficiently useful.

VII. Conclusion

The results in this paper provide tools to bound the risks of re-identification attacks with external leakages in the context of privacy-preserving open data publishing, and to anticipate the impact of extended data collection. However, the concrete values obtained for the HI bound indicate that an accumulation of such leakages may rapidly allow adversaries to re-identify users within databases, with significant variability between users (i.e., some users are much easier to re-identify than others). In the frequent case where simple (e.g., first-order) models are sufficient to maintain the utility of the data, simple data swapping tools can be used to push back the privacy risks, yet in a limited manner. In general and in view of these results, privacy-preserving open data publishing is likely to require strong restrictions of the data. For example, the suppression of any – even pseudonymized – identity in the observations is a good candidate solution in contexts where utility only requires user-independent statistics. Alternatively, pseudonymized data can only be hoped to remain anonymous up to a certain amount of leakages, in which case the number of tolerated leakages could be used as a (weaker) privacy metric by policy makers. In case these options are not applicable / sufficient, replacing the ability to access the full database by the ability to query it as in the differential privacy setting is the only known solution with strong theoretical guarantees [4].

Acknowledgments. François-Xavier Standaert is a senior associate researcher of the Belgian Fund for Scientific Research (FNRS-F.R.S.). This work has been funded in parts by the ERC project 724725 (acronym SWORD).

References

[1] Olivier Bronchain, Julien M. Hendrickx, Clément Massart, Alex Olshevsky, and François-Xavier Standaert. Leakage certification revisited: Bounding model errors in side-channel security evaluations. *IACR ePrint Archive*, 2019:132, 2019.

[2] Eloi de Cherisey, Sylvain Guilley, Olivier Rioul, and Pablo Piantanida. Best information is most successful. Cryptology ePrint Archive, Report 2019/491, 2019. https://eprint.iacr.org/2019/491.

[3] Alexandre Duc, Sebastian Faust, and François-Xavier Standaert. Making masking security proofs concrete - or how to evaluate the security of any leaking device. In Elisabeth Oswald and Marc Fischlin, editors, *Advances in Cryptology - EUROCRYPT 2015 - 34th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Sofia, Bulgaria, April 26-30, 2015, Proceedings, Part I*, volume 9056 of *Lecture Notes in Computer Science*, pages 401–429. Springer, 2015.

[4] Cynthia Dwork. Differential privacy: A survey of results. In Manindra Agrawal, Ding-Zhu Du, Zhenhua Duan, and Angsheng Li, editors, *Theory and Applications of Models of Computation, 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings*, volume 4978 of *Lecture Notes in Computer Science*, pages 1–19. Springer, 2008.

[5] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4):14:1–14:53, 2010.

[6] Clément Massart and François-Xavier Standaert. Revisiting location privacy from a side-channel analysis viewpoint (extended version). *IACR Cryptology ePrint Archive*, 2019:467, 2019.

[7] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (S&P 2008), 18-21 May 2008, Oakland, California, USA*, pages 111–125. IEEE Computer Society, 2008.

[8] Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In *30th IEEE Symposium on Security and Privacy (S&P 2009), 17-20 May 2009, Oakland, California, USA*, pages 173–187. IEEE Computer Society, 2009.

[9] Steven P. Reiss. Practical data-swapping: The first steps. *ACM Trans. Database Syst.*, 9(1):20–37, 1984.

[10] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.