View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by INRIA a CCSD electronic archive server

Probabilistic k^{m}-anonymity

Acs Gergely, Jagdish Prasad Achara, Claude Castelluccia

▶ To cite this version:

Acs Gergely, Jagdish Prasad Achara, Claude Castelluccia. Probabilistic k^{m}-anonymity: Efficient Anonymization of Large Set-Valued Datasets. IEEE International Conference on Big Data (BigData) 2015, Oct 2015, Santa Clara, United States. hal-01205533

HAL Id: hal-01205533

https://hal.inria.fr/hal-01205533

Submitted on 25 Sep 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Probabilistic k^m-anonymity
Efficient Anonymization of Large Set-Valued Datasets

Gergely Acs
INRIA
Email: gergely.acs@inria.fr

Jagdish Prasad Achara
INRIA
Email: jagdish.achara@inria.fr

Claude Castelluccia
INRIA
Email: claude.castelluccia@inria.fr

***Abstract*—Set-valued dataset contains different types of items/values per individual, for example, visited locations, purchased goods, watched movies, or search queries. As it is relatively easy to re-identify individuals in such datasets, their release poses significant privacy threats. Hence, organizations aiming to share such datasets must adhere to personal data regulations. In order to get rid of these regulations and also to benefit from sharing, these datasets should be anonymized before their release.**

In this paper, we revisit the problem of anonymizing set-valued data. We argue that anonymization techniques targeting traditional k^m-anonymity model, which limits the adversarial background knowledge to at most m items per individual, are impractical for large real-world datasets. Hence, we propose a probabilistic relaxation of k^m-anonymity and present an anonymization technique to achieve it. This relaxation also improves the utility of the anonymized data. We also demonstrate the effectiveness of our scalable anonymization technique on a real-world location dataset consisting of more than 4 million subscribers of a large European telecom operator. We believe that our technique can be very appealing for practitioners willing to share such large datasets.

I. Introduction

Today digital data about individuals are being collected on a large scale by different actors. This vast amount of data, termed as *big data* in the literature, could be of great use for social or economical development if shared. Set-valued data is a renowned form of big data, which contains a set of items/values per individual, for example, visited locations, purchased items, watched movies, or search queries. Although set-valued data typically do not include any direct personal identifiers, their release can still lead to a privacy breach if an adversary learns a subset of items of an individual from other sources. In particular, the adversary might be able to re-identify individuals if the known subset of items is unique or not shared by many people. For instance, de Montjoye *et al.* have shown that four spatio-temporal positions are enough to uniquely identify an individual 95% of the times in a dataset of one and a half million users [7]. Similar re-identification attacks have been demonstrated on credit-card metadata [8] and movie ratings [20]. These attacks pose significant privacy threats, as the adversary can learn about *all* items of individuals after identifying their records, which may uncover potential sensitive information such as their health and sex life or political beliefs [20].

In the literature, different privacy models are proposed for privacy-preserving data release. One of the first of these models was k-anonymity [22], [23], which was originally proposed for relational databases. However, k-anonymity is not a meaningful privacy model for set-valued data, where attributes cannot be separated into quasi-identifiers (which the adversary might already know from external sources about an individual) and sensitive attributes (which the adversary intends to learn). Rather, every item can equally serve as a quasi-identifier as well as a sensitive attribute. As a straw man approach of k-anonymization, one can consider all items of an individual as quasi-identifiers. However, it is unreasonable to assume that the adversary knows all the items of an individual, especially if there are many of them. Moreover, k-anonymity suffers from the curse of dimensionality [3], i.e., datasets with many attributes require excessive modification in order to satisfy k-anonymity. As set-valued data are typically large-dimensional and sparse, this would render most of such data practically useless.

As a result, k^m-anonymity has been proposed in [24], which has a weaker but more reasonable guarantee than k-anonymity; any subset of m items must be shared by 0 or at least k individuals in the dataset. In [24], Manolis *et al.* also proposed algorithms to achieve k^m-anonymity but these solutions are not applicable for large datasets. That is, the running time of their apriori-based anonymization remains exponential in m in the worst case, and therefore impractical for larger values of m.

In this paper, we propose a probabilistic version of k^m-anonymity which is suitable for big data. This comes at the cost of a bit relaxation of privacy requirements as it provides a probabilistic guarantee on k^m-anonymity. Roughly saying, our model guarantees that any m items of an individual are shared by at least $k - 1$ other individuals *with a certain probability* σ. Here, the value of σ is typically very close or even equal to 1. If $\sigma = 1$ our model boils down to the standard k^m-anonymity model [24] providing the same privacy guarantee. We propose an anonymization technique based on random sampling to achieve probabilistic k^m-anonymity with a worst-case complexity that is linear in the number of individuals and quadratic in m. Therefore, our solution becomes practical for *any* values of m, and it is also sufficiently flexible to provide different guarantees for different values of m. Indeed, it is rational to have stronger privacy guarantee (i.e., larger k and/or larger σ) for smaller values of m, which are always easier to acquire from external sources than larger number of items. This allows to further improve utility meanwhile providing reasonable privacy guarantee even for large values of m.

Moreover, our model perfectly complies with current personal data regulations and therefore can be a compelling choice for the purpose of anonymization in practice. In particular, the central concept of such regulations is re-identifiability of individuals likewise in our model. For example, if re-identifiability rate can be proved to fall below a particular threshold (see the United States HIPAA Privacy Rule Safe Harbor de-identification standard[1]), then the data is not considered to be personal any more and the regulation does not apply. Similarly, the European Data Protection law (Directive 95/46/EC) [1] defines personal data as "any information relating to an identified or identifiable natural person". In determining whether a person is identifiable "account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person" [1]. Our model provides a straightforward interpretation of these privacy requirements through the fine-tuning of its parameters (i.e., m, k, and σ).

Contributions: More specifically, the contributions of this paper are as follows:

1) We propose a probabilistic relaxation of k^m-anonymity in order to anonymize large set-valued datasets efficiently. The parameters of our model can be fine-tuned to the desired privacy requirements.
2) We design an anonymization technique which uses generalization to achieve our probabilistic privacy guarantees. The running time of our approach is linear in the number of records, the maximum number of generalizations, and quadratic in m. Hence, it remains scalable even for big data regardless what generalization method is used. Our scheme is based on the random sampling of different subsets of items and can optimize the utility with respect to any error function. To this end, we use a Markov Chain Monte Carlo sampling technique, whose worst-case running time (aka., mixing time) is known.
3) We evaluate our anonymization solution on the location data of more than 4 million individuals in a large European city. This location data is derived from their call data records, and consists of the set of visited cell towers per individual. We show that the localization error due to anonymization with reasonable anonymity guarantees falls below a few hundred meters, and the resolution ranges from 100 to 400 partitions of the city which has a total area of 128 km^2. This makes an average localization accuracy between 0.3 and 1 km^2 depending on the applied privacy parameters.

II. Related work

In this section, we briefly review the related work on the anonymization of set-valued data (i.e., excluding the problem of attribute inference). Many works divide items into sensitive attributes and quasi-identifiers, and assume that the adversary's background knowledge is confined to quasi-identifiers [11], [12], [4], [18], [26], [27]. Similarly to [24], [25], our approach does not have such a limitation as all items can equally be sensitive and also quasi-identifier in our model.

In [24], [25], Terrovitis *et al.* propose the model of k^m-anonymity as well as different anonymization algorithms using global and local recoding. However, these algorithms do not scale well for large datasets as their computational cost is exponential in m in the worst case (see Section V-C for details). k^m-anonymity has also been applied on trajectory data in [21], but this solution suffers from the same drawbacks as [25]. Another approach of anonymization of set-valued data were proposed in [17], where the privacy constraints need to be specified in advance in the form of specific itemsets which must satisfy k-anonymity. However, the proposed solutions have a cost which is exponential in m if the privacy constraints are composed of all subsets of items with size m.

He *et al.* propose a top-down generalization algorithm in [13] to provide k-anonymity for set-valued data. However, as [25] also points out, this approach underperforms the anonymization techniques proposed for k^m-anonymity in [25], if m is less than the average record size. Also, k-anonymity has an unnecessarily strong assumption for set-valued data as the adversary is unlikely to know all the items of an individual.

Set-valued data have also been sanitized under differential privacy. Instead of releasing microdata, these solutions publish the noisy occurrence counts of certain items [15] or itemsets [5] in the dataset. Due to the injected noise, these techniques do not preserve data truthfulness, and hence cannot be used in several applications [9], [17]. Moreover, the solution proposed in [5] releases the noisy occurrence counts of complete records which provides weak utility if the records are unique in the dataset (which is the case for most real-world large datasets like in Section VI-A).

Finally, similarly to [21], our scheme does not require a pre-computed generalization hierarchy for anonymization but rather a set of more general constraints describing the desired output. This provides a wide applicability to our approach which is also demonstrated in Section VI.

[1] http://aspe.hhs.gov/admnsimp/final/PvcFR06.txt#

Rec#	Items
1	{LA}
2	{LA, Seattle}
3	{New York, Boston}
4	{New York, Boston}
5	{LA, Seattle, New York}
6	{LA, Seattle, New York}
7	{LA, Seattle, New York, Boston}

Table I: Example for a set-valued dataset, where $\mathbb{I} = \{\text{LA}, \text{Boston}, \text{New York}, \text{Seattle}\}$.

III. Model

Let $\mathbb{I}$ denote the universe of all items (e.g., set of visited locations, purchased items, etc.). A dataset $D \subseteq 2^{\mathbb{I}} \setminus \{\}$ is the ensemble of all items of some set of individuals, where $|D|$ denotes the number of individuals in D. A record D_u, which is a non-empty subset of $\mathbb{I}$, refers to all items of an individual u in D. A set of items with cardinality m is shortly called as an m-itemset henceforth. The set of all m-itemsets over $\mathbb{I}$ is denoted as $\mathbb{I}^m$. A set-valued dataset containing the set of visited cities per individual is illustrated in Table I.

The goal of the adversary is to re-identify a targeted user u in D such that at most m items of u (i.e, a single m-itemset from u's record) are known to the adversary. We assume that the adversary has no other available background knowledge about u.

An itemset is k-anonym in D, if the number of records containing that itemset is either 0 or at least k. k^m-anonymity [24] guarantees that *all* m-itemsets are k-anonym. In other words, if the adversary knows any m-itemset which occurs in at least one record of D, then there are at least $k-1$ other records which also contain that itemset. For example, the dataset in Table I is not 2^2-anonym, as $\{\text{LA}, \text{Boston}\}$ and $\{\text{Seattle}, \text{Boston}\}$ only occur in the last record.

In the following definition, we relax k^m-anonymity and require that any itemset, up to size m, which is known by the adversary must be k-anonym with large probability.

Definition 1 (σ-k^m-anonymity) *Let $supp(x, D)$ denote the support of $x \in \mathbb{I}^m$ in D, i.e., the number of records in D which contain x. Let Ω^ℓ denote the set of all ℓ-itemsets which occur in at least one record in D, i.e., $\Omega^\ell = \{x : x \in \mathbb{I}^\ell \wedge supp(x, D) \geq 1\}$, and let $\mathcal{B}_\ell$ denote a random variable describing a probability distribution over Ω^ℓ. A dataset D is σ-k^m-anonym with respect to the ensemble of $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_m$, if $Pr[supp(\mathcal{B}_\ell, D) \geq k] \geq \sigma$ for all $1 \leq \ell \leq m$.*

σ-k^m-anonymity requires that any ℓ-itemset ($\ell \leq m$) chosen from the distribution of $\mathcal{B}_\ell$ is k-anonym with probability at least σ. In particular, the adversary has a probability distribution of $\mathcal{B}_\ell$ over all ℓ-itemsets, which represents the likelihood that a particular ℓ-itemset is learned by the adversary from external sources. Then, the adversary picks one itemset x from this distribution, and the attack is successful if x is not k-anonym (i.e., $supp(x, D) < k$). In Definition 1, $1 - \sigma$ measures the success probability of this attack.

In order to ease presentation, Definition 1 requires the same privacy guarantee (i.e., identical values of σ and k) for all ℓ-itemsets where $\ell \leq m$. However, we stress that the privacy guarantee may depend on the size of itemsets in reality. That is, larger values of ℓ may need less privacy protection and hence smaller values of σ and k, as the adversary may be less capable to acquire large number of items from external sources. Therefore, in a more general sense, σ and k can be vectors with size m, i.e., $(\sigma_1, \sigma_2, \ldots, \sigma_m)$ and $(k_1, k_2, \ldots, k_m)$ requiring an ℓ-itemset drawn from $\mathcal{B}_\ell$ to be k_ℓ-anonym with probability at least σ_ℓ.

Finally, we note that the k-anonymity of m-itemsets does *not* imply that of shorter itemsets in general, and hence the identical and explicit privacy requirement for all ℓ-itemsets, where $\ell < m$, in Definition 1. Indeed, there is no guarantee that records with size shorter than m also occur in longer records.

IV. Sampling for anonymity

The basic idea of our anonymization approach is to sample a set S of ℓ-itemsets from the distribution of $\mathcal{B}_\ell$ for all $1 \leq \ell \leq m$, and perform anonymization only if any itemset in S for any $\ell \leq m$ violates k-anonymity. It follows from standard sampling complexity bounds that if all itemsets in S are k-anonym and $|S|$ is $O(\ln(1/\delta)/\varepsilon^2)$ then *all* ℓ-itemsets are k-anonym in the *entire* dataset as well with an error of ε and confidence $1 - \delta$. Importantly, the number of required samples is independent of the size of Ω^m, which is exponential in m in the worst case.

A. Sufficient requirement of σ-k^m-anonymity

Theorem 1 *D is k^m-anonym with probability at least σ with respect to $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_m$, if, for each $1 \leq \ell \leq m$, there exists $S \subseteq \Omega^\ell$ such that*

1) *each element of S is independently drawn from the distribution of $\mathcal{B}_\ell$,*
2) *each element of S is k-anonym in D,*
3) *$|S| \geq \frac{\ln(1/x)}{2}\left(1 - \frac{\sigma}{1-x}\right)^{-2}$, where x satisfies $x^2 + 2x\sigma\ln(x) + (\sigma - 2)x - \sigma + 1 = 0$ and $\sigma \geq 0.5$*

Proof: Let $\hat{H}_k$ denote the relative frequency of ℓ-itemsets in S which do *not* satisfy k-anonymity in D. Furthermore, let $H_k = Pr[supp(\mathcal{B}_\ell, D) < k]$. As each element of S is drawn independently from the distribution of $\mathcal{B}_\ell$ (Condition 1), $\hat{H}_k$ is an unbiased estimator of H_k, i.e., $E[\hat{H}_k] = H_k$. It follows from the Chernoff-Hoeffding inequality [14] that $Pr\left[\hat{H}_k - H_k \geq \varepsilon\right] \leq e^{-2|S|\varepsilon^2}$, or equivalently, $Pr\left[\hat{H}_k - H_k < \varepsilon\right] \geq 1 - e^{-2|S|\varepsilon^2}$. Since $\hat{H}_k = 0$ in our case (Condition 2), we obtain that

$$Pr\left[H_k < \varepsilon\right] \geq 1 - e^{-2|S|\varepsilon^2}$$

where $\delta = e^{-2|S|\varepsilon^2}$. Therefore, D is k^m-anonym with probability at least σ, if

$$|S| \geq \frac{\ln(1/\delta)}{2\varepsilon^2}$$

and

$$(1-\varepsilon)(1-\delta) \geq \sigma \tag{1}$$

Our goal is to minimize the number of samples $|S|$ meanwhile satisfying the constraint in Inequality (1). Therefore, we can formulate the following simple optimization problem

$$\begin{aligned} \underset{\varepsilon,\delta}{\text{minimize}} \quad & f(\varepsilon,\delta) = \frac{\ln(1/\delta)}{2\varepsilon^2} \\ \text{subject to} \quad & (1-\varepsilon)(1-\delta) \geq \sigma, \\ & 0 < \varepsilon, \delta \leq 1 \end{aligned}$$

This is a convex non-linear optimization problem, and therefore the Karush-Kuhn-Tucker (KKT) conditions are necessary for a solution to be optimal. Indeed, if $0 < \delta \leq 0.5$, the leading principal minors of the Hessian matrix H of f is strictly positive on $\varepsilon \in (0,1]$ which means that H is positive definite. Hence, f is convex when $\varepsilon \in (0,1]$ and $\delta \in (0,0.5]$. Solving the KKT-conditions yields Condition 3 of the theorem. ■

In the rest of the paper, we assume that $\mathcal{B}_\ell$ is the uniform distribution over all ℓ-itemsets in D (i.e., the adversary can learn any ℓ-itemset in the dataset with equal probability). Although the adversary can always learn certain ℓ-itemsets of some users with larger probability, the uniform distribution over all possible ℓ-itemsets of any user is the most general assumption in practice.

B. Uniform sampling of ℓ-itemsets

As $\mathcal{B}_\ell$ has uniform distribution over Ω_ℓ, our task is to sample an element from Ω^ℓ uniformly at random for any $1 \leq \ell \leq m$. A first (naive) approach could be to use rejection sampling, i.e., sample a candidate ℓ-itemset from $\mathbb{I}^\ell$ uniformly at random, and then accept this candidate as a valid sample only if it also occurs in D. Otherwise, repeat the process until a candidate is accepted. Although sampling a candidate from $\mathbb{I}^\ell$ is straightforward, it is very likely to be non-existent in D (especially if ℓ is large), and hence its running complexity is $O(|\mathbb{I}|^\ell)$ in the worst case. An alternative approach could be to enumerate Ω^ℓ, and choosing one element directly from Ω^ℓ uniformly at random. However, the complexity of this approach is still $O(|D|(\max_u |D_u|)^\ell/\ell!)$ (recall that D_u denotes the record of user u in D). Unfortunately, these naive methods provide acceptable performance only if ℓ is small.

We instead use a sampling technique from [2] based on the Metropolis-Hastings algorithm [19], [6], which is a Markov Chain Monte Carlo (MCMC) method. This technique has a worst-case complexity which is roughly linear in the number of records and ℓ, and hence it remains reasonably fast even for larger values of ℓ.

In particular, an ergodic Markov chain, which is denoted by $\mathcal{M}$ and detailed in Algorithm 1, is constructed such that its stationary distribution π is the uniform distribution over Ω^ℓ that we want to sample from. Each ℓ-itemset in Ω^ℓ corresponds to a state of $\mathcal{M}$, and we simulate $\mathcal{M}$ until it gets close to π, at which point the current state of $\mathcal{M}$ can be considered as a sample from π.

Specifically, at each state transition, $\mathcal{M}$ picks a candidate next state C (i.e., an ℓ-itemset) independently of the current state S (in Line 6-7 of Algorithm 1). In Line 8, the candidate is either accepted (and $\mathcal{M}$ moves to C) or rejected with certain probability (in which case the candidate state is discarded, and $\mathcal{M}$ stays at S). The main idea is that, at each state, we use a fast but biased sampling mechanism to propose a candidate C (in Line 6-7); we first sample a record uniformly at random in D, and then an ℓ-itemset from this record also uniformly at random. This sampling is more likely to select any ℓ-itemset which occurs in multiple records (assuming records have similar sizes). We correct this bias by adjusting the acceptance/rejection probability (in Line 8) accordingly; $\mathcal{M}$ is more likely to accept such states which are less likely to be proposed in Line 6-7.

Algorithm 1 MCMC sampling ($\mathcal{M}$)

```
1: **Input:** Dataset $D$, $\ell$, # of iterations $t$
2: **Output:** A sample $S \in \Omega^\ell$
3: Let $U := \{D_u : |D_u| \geq \ell \wedge D_u \in D\}$
4: Let $S$ be an arbitrary $\ell$-itemset in $\Omega^\ell$
5: **for** $k = 1$ **to** $t$ **do**
6:     Select a record $r \in U$ uniformly at random
7:     Select a subset of items $C \subseteq r$ uniformly at random such
       that $|C| = \ell$
8:     Let $S := C$ with probability $\min(1, q(S)/q(C))$, where
       $q(x) = \sum_{\forall u: D_u \supseteq x} \prod_{i=1}^{\ell} \frac{1}{|D_u|-\ell+i}$
9: **return** $S$
```

Theorem 2 ([2]) *$\mathcal{M}$ in Algorithm 1 is an ergodic Markov chain whose unique stationary distribution is the uniform distribution over Ω^ℓ for any ℓ.*

Convergence of $\mathcal{M}$: To compute t in Algorithm 1, we need to know how many transitions $\mathcal{M}$ should do in order to "forget" its initial state and get "close enough" to its stationary distribution, i.e., the uniform distribution over Ω^ℓ. The time that $\mathcal{M}$ takes to converge to its stationary distribution π is known as the *mixing time* of $\mathcal{M}$, and is measured in terms of the total variation distance between the distribution at time t and π.

Definition 2 (Mixing time) *For $\xi > 0$, the mixing time $\tau_{\mathcal{M}}(\xi)$ of Markov chain $\mathcal{M}$ is*

$$\tau_{\mathcal{M}}(\xi) = \min\{t' : ||P^t_{\mathcal{M}} - \pi||_{tv} \leq \xi, \forall t \geq t'\}$$

where $||P^t_{\mathcal{M}} - \pi||_{tv} = \max_{x\in\Omega^\ell} \frac{1}{2}\sum_{y\in\Omega} |P^t_{\mathcal{M}}(x,y) - \pi(y)|$

defines the total variation distance. $P^t_{\mathcal{M}}(x, y)$ denotes the t-step probability of going from state x to y, and $P^t_{\mathcal{M}}$ denotes the t-step probability distribution over all states.

The next theorem shows that $\tau_{\mathcal{M}}(\xi)$ is $O(|D| \log(1/\xi)/R^*_\ell)$, where $|D|$ is the dataset size and R^*_ℓ is the uniqueness of ℓ-itemsets from the largest record of D, i.e., the probability that an ℓ-itemset selected from the largest record uniformly at random occurs only once in D. As the uniqueness of ℓ-itemsets is usually large in practice, especially if ℓ is large, $\mathcal{M}$ is fast-mixing in general[2]. Indeed, in our dataset detailed in Section VI-A, $R^*_\ell = 0.07$ when $\ell = 5$, and it increases to 0.75 ($\ell = 11$).

Theorem 3 (Mixing time of $\mathcal{M}$ [2]) *Let R^*_ℓ denote the probability that a set of ℓ items selected from the largest record (i.e., having the most items) in D uniformly at random is unique. Then, $\tau_{\mathcal{M}}(\xi) \leq |D| \ln(1/\xi)/R^*_\ell$ for any ℓ.*

Notice that q in Algorithm 1 can be computed rapidly in practice by precomputing a table T, where each row corresponds to an item in D, and row i contains the sorted list of all records which have item i. Hence, the set of records which have a common specific ℓ-itemset can be computed by taking the intersection of the corresponding rows in T in time $O(\ell|i_{\max}|)$, where $|i_{\max}|$ is the maximum row size in T. Fast implementations of the intersection of sorted integers are described in [16].

Supposing that $\ln(1/\xi) \ll i_{\max} \ll |D|$ and $1/R^*_\ell$ is a constant close to 1, the total running complexity of $\mathcal{M}$ is roughly $O(\ell|D|)$. Therefore, a "good" uniform sample from Ω^ℓ can be obtained roughly after $O(\ell|D|)$ iterations in most practical cases.

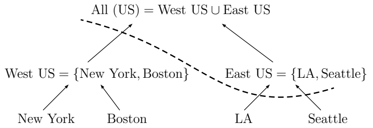
V. Anonymization algorithm

We perform anonymization through the generalization of items in D. Our goal is to find a generalization function $\mathfrak{R} : \mathbb{I} \to 2^{\mathbb{I}} \setminus \{\}$, which maps an item $I \in \mathbb{I}$ to a non-empty subset of $\mathbb{I}$. An anonymized dataset $\langle \mathfrak{R}, D \rangle$ is obtained by applying $\mathfrak{R}$ on the original dataset D, i.e., each item i occurring in D is replaced with $\mathfrak{R}(i)$ (i.e., global recoding is applied on D). The idea is that if multiple items are mapped to the same subset of $\mathbb{I}$, then the records containing these items become indistinguishable w.r.t these items after generalization. A possible (deterministic) anonymization of the example dataset from Table I is illustrated in Figure 1.

In this paper, we assume that the image of $\mathfrak{R}$ must correspond to a partitioning of $\mathbb{I}$. That is, for any $i, j \in \mathbb{I}$, either $\mathfrak{R}(i) = \mathfrak{R}(j)$ or $\mathfrak{R}(i) \cap \mathfrak{R}(j) = \varnothing$, and $\bigcup_{\forall i} \mathfrak{R}(i) = \mathbb{I}$. In order to find a partitioning of $\mathbb{I}$, which results in σ-k^m-anonymity meanwhile maximizes utility, the algorithm needs to be provided with the set of allowed generalization functions, i.e., a subset $\mathbb{C}$ of possible partitionings. In general, $\mathbb{C}$ can be succinctly described in the form of constraints. For example, in Figure 1, the constraint is specified in the form of a generalization hierarchy, i.e., $Image(\mathfrak{R})$ must be a partitioning of $\mathbb{I}$ which corresponds to a horizontal cut of the specified hierarchy.

We define the hamming distance on the set of generalizations as $dist_h(\mathfrak{R}_1, \mathfrak{R}_2) = |\{i \in \mathbb{I} : \mathfrak{R}_1(i) \neq \mathfrak{R}_2(i)\}|$, i.e., the number of items on which $\mathfrak{R}_1$ and $\mathfrak{R}_2$ differ. Two generalization functions $\mathfrak{R}_1, \mathfrak{R}_2 \in \mathbb{C}$ are neighbors at item $i \in \mathbb{I}$, if $\mathfrak{R}_1(i) \neq \mathfrak{R}_2(i)$ and $dist_h(\mathfrak{R}_1, \mathfrak{R}_2) = \min_{\mathfrak{R}' \in \mathbb{C}} dist_h(\mathfrak{R}_1, \mathfrak{R}') = \min_{\mathfrak{R}' \in \mathbb{C}} dist_h(\mathfrak{R}', \mathfrak{R}_2)$. For example, function $\mathfrak{R}$ in Figure 1 is the neighbor of function $\mathfrak{R}'$ at item "New York" and also at item "Boston", where $\mathfrak{R}'$ maps each city to itself (i.e., $\mathfrak{R}'(i) = \{i\}$ for all $i \in \mathbb{I}$)[3].

Rec#	Items
1	{LA}
2	{LA, Seattle}
3	{West US}
4	{West US}
5	{LA, Seattle, West US}
6	{LA, Seattle, West US}
7	{LA, Seattle, West US}



$\mathfrak{R}$(New York) = West US
$\mathfrak{R}$(Boston) = West US
$\mathfrak{R}$(LA) = LA
$\mathfrak{R}$(Seattle) = Seattle

Figure 1: Anonymized dataset (left) generated from Table I in order to guarantee 2^2-anonymity. $\mathfrak{R}$ (right) corresponds to a horizontal cut of the above generalization hierarchy (denoted by dashed line).

Furthermore, $err : 2^{\mathbb{I}} \times 2^{2^{\mathbb{I}}} \to \mathbb{R}$ denotes an error function measuring some distance between the original dataset D and its anonymized version $\langle \mathfrak{R}, D \rangle$. err represents the utility loss due to applying $\mathfrak{R}$ on D. For example, in Figure 1, a possible error function is the average distance between cities and the centroid of their aggregate regions (i.e., the centroid of West US for Boston and New York).

Given an input dataset D and constraint $\mathbb{C}$, our goal is to find $\mathfrak{R} \in \mathbb{C}$ (i.e., a partitioning of $\mathbb{I}$) such that $err(D, \langle \mathfrak{R}, D \rangle)$ is minimized and $\langle \mathfrak{R}, D \rangle$ satisfies σ-k^m-anonymity.

A. Operation

Our proposal, denoted by $\mathcal{A}$, is detailed in Algorithm 2. We perform a randomized greedy search on the set of generalizations in $\mathbb{C}$; it starts from the identity mapping of all items (i.e., when there is no generalization) and it always proceeds to a random neighboring generalization by merging partitions following the given constraint. The search stops as soon as we find a generalization $\mathfrak{R}$ such that $\langle \mathfrak{R}, D \rangle$ satisfies σ-k^m-anonymity, or no more generalization can be done because $Image(\mathfrak{R})$ is a singleton.

[2] The uniqueness of ℓ-itemsets from a single record can easily be approximated with the Chernoff-Hoeffding inequality using uniform samples over all ℓ-itemsets from the record. This sampling is straightforward to implement by choosing ℓ items from the record without replacement.

[3] Recall that a function $\mathfrak{R}$, where $\mathfrak{R}$(Boston) = West US but $\mathfrak{R}$(New York) = New York, is not a valid generalization function as $Image(\mathfrak{R})$ is not a partitioning of $\mathbb{I}$.

In particular, $\mathcal{A}$ first generalizes D by considering shorter itemsets and then proceeds with larger itemsets till the size of m. $\mathcal{A}$ maintains a generalization function $\mathfrak{R}$ which is initialized to the identity mapping of all items in $\mathbb{I}$ (in Line 4), and updates $\mathfrak{R}$ if a sampled ℓ-itemset violates k-anonymity (Line 12-14). Specifically, $\mathcal{A}$ picks an ℓ-itemset s from the original dataset D uniformly at random using $\mathcal{M}$ (in Line 8). If the generalized ℓ-itemset s' (Line 10) occurs at least k times in the anonymized dataset $\langle \mathfrak{R}, D \rangle$, then it is added to the set S of found k-anonym samples (Line 11) and $\mathfrak{R}$ is left unchanged. Otherwise, all generalizations $\mathfrak{R}' \in \mathbb{C}$ are identified such that (1) $\mathfrak{R}'$ is the neighbor of the current generalization $\mathfrak{R}$ at an item $j \in s$, i.e., there exists $j \in s$ where they differ, and (2) $|\mathit{Image}(\mathfrak{R}')| < |\mathit{Image}(\mathfrak{R})|$, i.e., $\mathfrak{R}'$ is obtained from $\mathfrak{R}$ by merging the partition of an item $j \in s$ with another partition in $\mathit{Image}(\mathfrak{R})$ specified by the constraint. Then, we update $\mathfrak{R}$ to the generalization which minimizes the error and satisfies both requirements. Notice that Requirement (2) guarantees that the algorithm will terminate after finite number of steps due to the monotonicity property of k-anonymity [24].

If at least γ samples[4] of ℓ-itemsets satisfy k-anonymity, or no more generalizations can be applied because only a single set of items occurs in each record (i.e., $\mathit{Image}(\mathfrak{R})$ is a singleton), $\mathcal{A}$ stops (Line 7) and proceeds to itemsets with size $\ell+1$ as long as $\ell \leq m-1$ (Line 5). Since γ is adjusted according to Theorem 1, and S contains itemsets selected by $\mathcal{M}$ uniformly at random, it follows from Theorem 1 and 2 that the output of $\mathcal{A}$ satisfies σ-k^m-anonymity.

Notice that in Line 8 of Algorithm 2, the uniform samples are always taken from the original dataset D and *not* from its anonymized version $\langle \mathfrak{R}, D \rangle$. Recall that, in Definition 1, the adversary's background knowledge is assumed to be at most m items from $\mathbb{I}$ (and not from $\mathit{Image}(\mathfrak{R})$).

Example 1 *Consider the anonymization of the dataset from Table I with $\sigma = 0.99$, $k = 2$, $m = 2$, err is the geographic distance, and each valid generalization in $\mathbb{C}$ corresponds to a horizontal cut of the hierarchy in Figure 1. First, $\mathcal{A}$ checks all 1-itemsets. As all cities are present in at least 2 records, $\mathcal{A}$ proceeds to 2-itemsets. Suppose that $\mathcal{A}$ samples $s = \{LA, Boston\}$, which fails 2-anonymity. Hence, $\mathcal{A}$ considers all possible neighboring generalizations of $\mathfrak{R}$, where $\mathfrak{R}$ is the identity mapping of cities, and finds two valid neighbors; $\mathfrak{R}_1$ at item Boston (i.e., Boston is merged with New York), and $\mathfrak{R}_2$ at item LA (i.e., LA is merged with Seattle). In particular, $\mathfrak{R}_1(New\ York) = \mathfrak{R}_1(Boston) = West$, $\mathfrak{R}_1(LA) = LA$, $\mathfrak{R}_1(Seattle) = Seattle$, and $\mathfrak{R}_2(Seattle) = \mathfrak{R}_2(LA) = East$, $\mathfrak{R}_2(New\ York) = New\ York$, $\mathfrak{R}_2(Boston) = Boston$. Since Boston is closer to New York than LA to Seattle, $\mathcal{A}$ chooses $\mathfrak{R}_1$ which results in the least error. The resulting dataset $\langle \mathfrak{R}_1, D \rangle$, which is also shown in Figure 1, is 2^2-anonym and hence returned by $\mathcal{A}$.*

B. Amplifying utility

It can be shown that finding the optimal generalization function $\mathfrak{R}$, which results in the least error and also provides k^m-anonymity, is NP-hard [24]. Although $\mathcal{A}$ tends to output datasets with low error, it is not guaranteed to find a dataset which is "close" to the optimal solution. However, as $\mathcal{A}$ is randomized, we can approach such a solution if $\mathcal{A}$ is executed multiple times.

Algorithm 2 Anonymization $\mathcal{A}$

```
1: Input: $D$, $\mathbb{C}$, $m$, $k$, $\sigma$, $t$
2: Output: Anonymized dataset $\langle \mathfrak{R}, D \rangle$
3: $\gamma := \frac{\ln(1/x)}{2}\left(1 - \frac{\sigma}{1-x}\right)^{-2}$, where $x$ satisfies
   $x^2 + 2x\sigma\ln(x) + (\sigma - 2)x - \sigma + 1 = 0$
4: $\mathfrak{R}(i) := \{i\}$ for all $i \in \mathbb{I}$
5: for $\ell = 1$ to $m$ do
6:     $S := \{\}$ // set of found $k$-anonym itemsets
7:     while $|S| \leq \gamma \wedge |\mathit{Image}(\mathfrak{R})| > 1$ do
8:         $s = \mathcal{M}(D, \ell, t)$ // see Algorithm 1
9:         $s' = \{\mathfrak{R}(i) \mid \forall i \in s\}$
10:        if $\mathit{supp}(s', \langle \mathfrak{R}, D \rangle) \geq k$ then
11:            $S := S \cup \{s\}$
12:        else
13:            $\mathbb{V} := \{\mathfrak{R}' \in \mathbb{C} : \mathfrak{R}'$ is neighbor of $\mathfrak{R}$ at an item $j \in$
               $s$, and $|\mathit{Image}(\mathfrak{R}')| < |\mathit{Image}(\mathfrak{R})|\}$
14:            $\mathfrak{R} := \arg\min_{\mathfrak{R}' \in \mathbb{V}} \mathit{err}(D, \langle \mathfrak{R}', D \rangle)$
15:            $S := \{\}$
16: return $\langle \mathfrak{R}, D \rangle$
```

In particular, we amplify the utility of $\mathcal{A}$ by multiple independent trials. Consider $\mathcal{A}^{(r)}$ which executes $\mathcal{A}$ independently r times, and selects the output among the r executions which has the best utility (i.e., the least error). Notice that multiple executions do not deteriorate anonymity, but rather improve utility. Indeed, $\mathcal{A}^{(r)}$ can find a dataset which is closer to the optimal solution if r is sufficiently large. Moreover, this approach is independent of the error function err.

Theorem 4 *$\mathcal{A}^{(r)}$ achieves σ-k^m-anonymity in $O\left(rm^2|D||\mathbb{I}|(1-\sigma)^{-2}\ln\left(\frac{1}{1-\sigma}\right)\right)$ steps.*

Proof: $\mathcal{A}$ picks γ samples maximum $|\mathbb{I}| - 1$ times in the worst case for any $\mathbb{C}$. In addition, $\gamma = O\left((1-\sigma)^{-2}\ln\left(\frac{1}{1-\sigma}\right)\right)$. Indeed, $\gamma = O\left(\ln(1/\delta)/\varepsilon^2\right)$, where $(1-\varepsilon)(1-\delta) \geq \sigma$. This means that $\varepsilon = (1-\sigma)/2$ implies $\delta = (1-\sigma)/(1+\sigma)$ which results in $\gamma < \ln\left(\frac{4}{1-\sigma}\right)(1-\sigma)^{-2}$.

As the complexity of $\mathcal{M}$ is $O(\ell|D|)$, we obtain that the complexity of $\mathcal{A}^{(r)}$ is $O\left(rm^2|D||\mathbb{I}|(1-\sigma)^{-2}\ln\left(\frac{1}{1-\sigma}\right)\right)$. ■

[4] γ can be efficiently computed with any numerical root-finding method (e.g., Newton's method)

C. Comparison to the deterministic Apriori-based anonymization

In [24], the authors proposed a deterministic anonymization technique to achieve traditional k^m-anonymity (i.e, $\sigma = 1$) based on the apriori-principle. However, the running time of their apriori-based anonymization (AA) remains exponential in m in the worst case. In particular, AA first generalizes all 1-itemsets in D and obtains a new generalized dataset D_1. Then all 2-itemsets of D_1 are generalized in order to get D_2, and so on until D_m is obtained and released. At step ℓ ($\ell \leq m$), AA requires to enumerate and store all ℓ-itemsets of each record in D_ℓ, which has a cost of $O\left(|D| \cdot \binom{t_\ell}{\ell}\right)$, where t_ℓ is the size of the longest record in D_ℓ. Although $t_1 \geq t_2 \geq \ldots \geq t_m$, there is no guarantee that $\binom{t_m}{m}$ is sufficiently small in practice (which is the case for our dataset in Section VI) as its magnitude depends on the applied generalization hierarchy and dataset D.

VI. Evaluation

A. Dataset characteristics

We use a CDR (Call Data Record) dataset provided by a cell phone operator in Europe, where $\mathbb{I}$ represents the set of cell towers of the operator in a large European city. A cell tower $T \in \mathbb{I}$ is visited by an individual, if the operator has a recorded event at tower T related to the individual over the observed period (01/09/2007 - 17/10/2007). An event can be an incoming/outgoing call or message to/from the individual. The dataset contains the events of $4,427,486$ users at $|\mathbb{I}| = 1303$ towers within the administrative region of the city, where the GPS coordinates of all the towers are also available. In our dataset D, a record contains *only* the set of towers that are visited by a user over six weeks, i.e., it does *not* contain the time of visits. Indeed, including time of visits would enable the adversary to easily compute the frequency of each tower per user which allows to deduce the top locations (e.g., home and working place) of *each* individual in D. Top locations of individuals are highly unique [28] even in large populations, which makes the anonymization of time-stamped location data very challenging.[5]

The average number of individuals per tower over this period was 38817 with a standard deviation of 50911. The total area of the city which is covered by cell towers is 128.1 km^2. The main characteristics of our dataset are summarized in Table II. The towers are shown in Figure 2.

B. Uniqueness

In what follows, we illustrate the potential privacy risks of releasing our dataset without any anonymization. The uniqueness of a set of items (i.e., towers) of different sizes is shown in Figure 2. Recall that the uniqueness is defined as the relative frequency of m-itemsets with respect to a single occurrence in D, i.e., $\frac{|\{x: x\in\mathbb{I}^m \wedge supp(x,D)=1\}|}{|\{x: x\in\mathbb{I}^m \wedge supp(x,D)\geq 1\}|}$. As the denominator is infeasible to compute in our case, we use sampling to approximate the uniqueness. In particular, according to the Chernoff-Hoeffding bound, we pick $2\ln(2/\delta)/\varepsilon^2$ subsets of towers with the given size uniformly at random using $\mathcal{M}$, which is described in Section IV-B, and approximate the real uniqueness with the uniqueness of the sample set. We note that, unlike in [7], this technique provides an *unbiased* estimation of the real unicity. In particular, in [7], the authors used a biased sampling technique to estimate unicity; they first selected a record uniformly at random, and then a subset of towers with the given size also uniformly at random. However, their approach is more likely to select subsets which occur in multiple records, and hence underestimates the real unicity in the dataset.

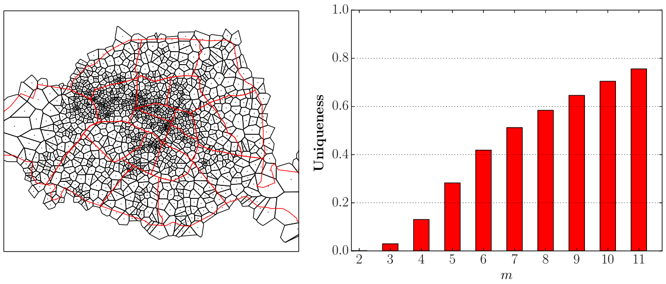


Figure 2: Voronoi-tessellation of cell towers (left). Red lines denote the boundaries of districts. Uniqueness depending on m (right)

Dataset size $\|D\|$	$4,427,486$
# of all towers $\|\mathbb{I}\|$	1303
Maximum record size $\max_u \|D_u\|$	422
Minimum record size $\min_u \|D_u\|$	1
Average record size	11.42
Std.dev of record size	17.23
Total area of all cells	128.1 km^2

Table II: Characteristics of our dataset D

For the sampling algorithm $\mathcal{M}$, we emphasize that the bound in Theorem 3 is a worst-case bound, and the real convergence time can be much smaller depending on the dataset D as well as the starting state of the chain. In order to speed up computations even more, we detected the convergence of $\mathcal{M}$ using the Geweke diagnostic [10] in the rest of the simulations (see the appendix).

For our measurements, we used $\varepsilon = 0.01$ and $\sigma = 0.99$ which requires at least 26492 samples from D (with replacement). This guarantees that the sample uniqueness is within $\pm 1\%$ error of the real uniqueness with probability at least 0.99. As Figure 2 shows, the uniqueness of 5 towers already reaches 0.2 and it increases to 0.78 when the number of towers is 11. These large values of uniqueness even for small number of towers indicates the serious privacy threats of releasing CDR data even if time information is withheld.

[5] Notice that, in our simplified dataset, even if the adversary knows the top locations of a single user, there can be multiple other users who visited each of these locations at least once. Without time information, the adversary does not know whether these are also top locations of other users or not.

C. Anonymization of CDR data

We use $\mathcal{A}$ in Algorithm 2 to anonymize our CDR dataset D with the following constraint. A partition of cell towers, which is always represented by a contiguous region of the city, can only be merged with a neighboring partition in the generalization process. In other words, $\mathbb{C}$ contains all possible partitionings of the towers such that the voronoi polygons of all towers within a partition must constitute a single contiguous polygon. Notice that this requirement cannot be specified in the form of a single generalization hierarchy of cell towers and hence would not be achievable with most previous anonymization techniques.

Initially, $\mathfrak{R}$ is the identity mapping, i.e., every partition is a singleton containing a cell tower from D. Then, at each iteration, when a partition needs to be further generalized (Line 13-14 of Algorithm 2), it is merged with a neighboring partition[6] such that the following error function is minimized.

Let D' denote the anonymized dataset $\langle \mathfrak{R}, D \rangle$. The error function is defined as the average geographical approximation error due to generalization in the anonymized dataset D'. That is,

$$err(D, D') = \frac{\sum_{u \in D_u} \sum_{i \in D'_u} dist(i, \mathfrak{R}(i))}{\sum_{u \in D_u} |D_u|}$$

where $dist$ denotes the geographical distance between a tower and the center of its partition $\mathfrak{R}(i)$ in D'[7]. In other words, the average error is the weighted average of the approximation error of all towers, where the weight of a tower i is the number of occurrences of i in D.

As in Algorithm 2, the anonymization stops when there is a single all-inclusive partition, or all ℓ-itemsets sampled from D satisfy k-anonymity in $\langle \mathfrak{R}, D \rangle$ for $1 \leq \ell \leq m$. In order to amplify utility, we execute the above algorithm 10 times independently, i.e., $r = 10$, and select the output which has the least error.

D. Privacy guarantee

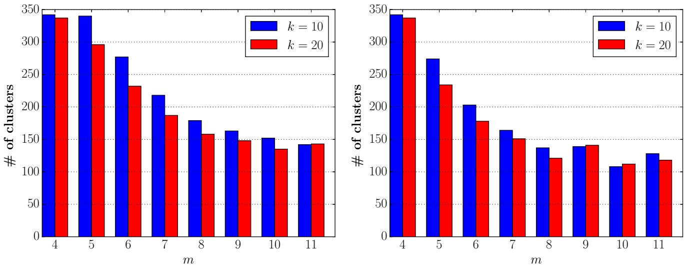
We anonymized our dataset with the following privacy guarantees. For $1 \leq \ell \leq 4$, we guaranteed k^m-anonymity with probability 1, i.e., we used a deterministic apriori-based anonymization (see Section V-C). That is, we replaced the random sampling step of Algorithm 2 with the deterministic enumeration of all ℓ-itemsets. However, for $\ell \geq 5$, this technique turns out to be very expensive, and we rather employed our probabilistic solution described in Section VI-C. Indeed, the size of the longest record after generalizing all 4-itemsets is till more than 300.

In fact, the above approach provides reasonable anonymity guarantee in practice; the adversary is more likely to collect a few, but usually no more than 5 locations of multiple individuals. For example, the adversary can crawl an online social network (OSN) where many individuals publish a few of their visited locations (such as home or working place), and try to re-identify the OSN users in our dataset. However, if $m \geq 5$, the adversary is less likely to collect at least 5 locations of 100 (or say 1000 if $\sigma = 0.999$) OSN users, and hence a probabilistic guarantee can make more sense in practice. That is, as our dataset covers a large proportion of all the citizens, a probabilistic guarantee with $\sigma = 0.99$ results in the privacy breach of 1 out of 100 users on average (assuming that all of them publish at least five of their locations in OSN and are also present in our dataset).

E. Results

As the average record size is 11, we anonymized itemsets up to size 11. In particular, we considered two different values of k; $k = 10$ and $k = 20$, and two different values of σ; 0.99 and 0.999. These probabilistic guarantees require $45,845$ and $5,866,617$ samples of ℓ-itemsets, respectively (see Condition 3 of Theorem 1). For a particular value of m, we used the same value of σ if $5 \leq \ell \leq m$, and $\sigma = 1$ if $1 \leq \ell \leq 4$. In addition, we used the same value of k for all $1 \leq \ell \leq m$.

The average number of released partitions is depicted in Figure 4. Recall that the maximum number of partitions is the original number of towers which is 1303. The number of partitions decreases from 350 ($m = 5$) to 150 ($m = 11$) for $\sigma = 0.99$, and from 260 ($m = 5$) to 120 ($m = 11$) for $\sigma = 0.999$. Interestingly, increasing k from 10 to 20 does not significantly change the results. Although the number of partitions seem to be too limited at first sight, Figure 3 illustrates that they still provide a meaningful partitioning of the city. Indeed, as Figure 5 also shows, the average error ranges from 110 to 350 meters. Specifically, for $\sigma = 0.99$, the error changes from 120 meters ($m = 5$) to 280 meters ($m = 11$), and, for $\sigma = 0.999$, from 160 meters ($m = 5$) to 350 meters ($m = 11$).

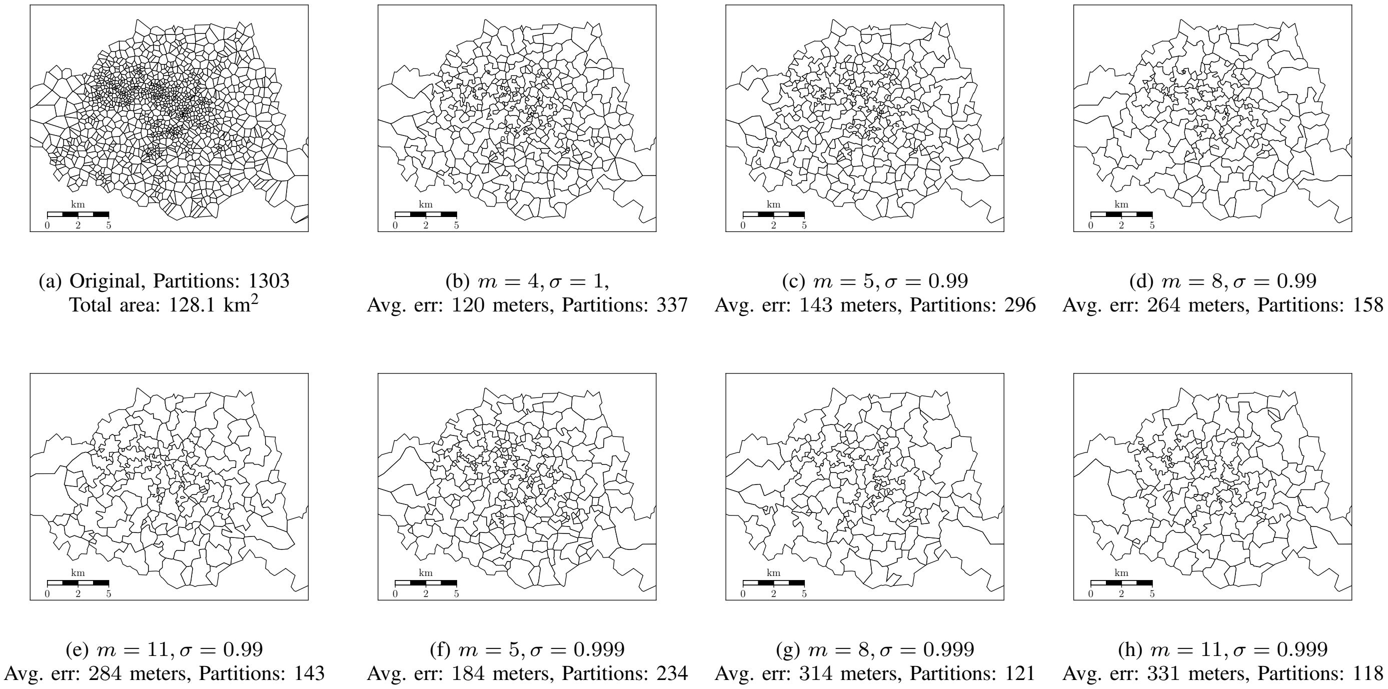


(a) $\sigma = 0.99$ (b) $\sigma = 0.999$

Figure 4: Number of partitions (more are better) ($r = 10$).

[6] Two polygons are immediate neighbors if their boundaries have at least one common point.

[7] The center of a partition is the centroid of the polygon which is obtained by merging all the voronoi polygons of the cell towers belonging to the partition.



(a) Original, Partitions: 1303
Total area: 128.1 km^2

(b) $m = 4, \sigma = 1$,
Avg. err: 120 meters, Partitions: 337

(c) $m = 5, \sigma = 0.99$
Avg. err: 143 meters, Partitions: 296

(d) $m = 8, \sigma = 0.99$
Avg. err: 264 meters, Partitions: 158

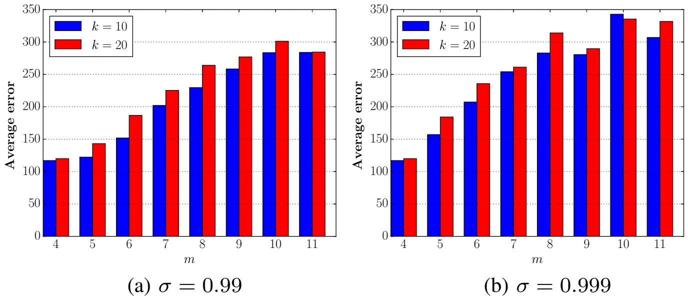
(e) $m = 11, \sigma = 0.99$
Avg. err: 284 meters, Partitions: 143

(f) $m = 5, \sigma = 0.999$
Avg. err: 184 meters, Partitions: 234

(g) $m = 8, \sigma = 0.999$
Avg. err: 314 meters, Partitions: 121

(h) $m = 11, \sigma = 0.999$
Avg. err: 331 meters, Partitions: 118

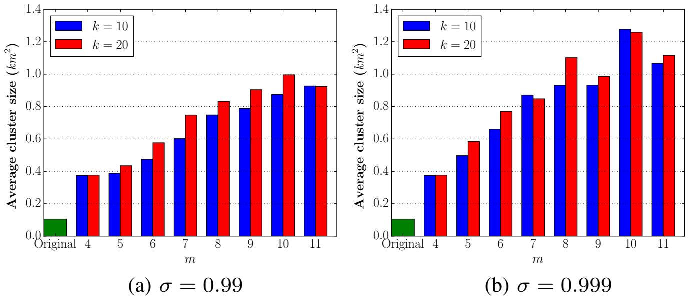
Figure 3: Partitions of cell towers depending on m and σ. $k = 20$ and $r = 10$ in all settings.



(a) $\sigma = 0.99$

(b) $\sigma = 0.999$

Figure 5: Average error (smaller is better) ($r = 10$).



(a) $\sigma = 0.99$

(b) $\sigma = 0.999$

Figure 6: Average partition size (smaller is better) ($r = 10$).

We also computed the average partition size, where the size of each partition is weighted with the number of individuals within the partition. The results are shown in Figure 6. The average partition size in the original dataset is 0.1 km^2, which increases to 1 km^2 when $\sigma = 0.99$, and to 1.3 km^2 when $\sigma = 0.999$. Again, the results are only slightly influenced by the value of k.

VII. Conclusion

We proposed a probabilistic relaxation of k^m-anonymity for the purpose of anonymizing large set-valued datasets. This relaxation is important to achieve scalability and to improve the utility of the anonymized data. We believe that our privacy guarantees are sufficient for most real-world adversaries and also to get rid of personal data regulations if a dataset needs to be shared. We also presented an anonymization technique to achieve this relaxation. Our technique does not rely on a pre-defined generalization hierarchy of the set of items but rather on more general constraints describing the desired output. Moreover, it can optimize the utility of the anonymized dataset against privacy constraints with respect to any error function providing wide applicability to our solution. We evaluated our technique on a real-world large dataset and found that the anonymized dataset is still reasonably accurate. We also studied the effect of the privacy parameters m, σ and k on the utility. According to our measurements, the adversarial background knowledge, which is measured by m, has the largest impact on the utility in general followed by the confidence σ of the privacy guarantee.

Acknowledgements

This work was funded by the PIA (projet investissements d'avenir) XData Project (`http://xdata.fr`).

References

[1] EU Directive 95/46/EC - The Data Protection Directive , 1995. https://www.dataprotection.ie/docs/EU-Directive-95-46-EC-Chapter-1/92.htm.

[2] J. P. Achara, G. Acs, and C. Castelluccia. On the unicity of smartphone applications. In *ACM Workshop on Privacy in the Electronic Society (WPES)*, 2015.

[3] C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In *VLDB*, pages 901–909, 2005.

[4] J. Cao, P. Karras, C. Raïssi, and K.-L. Tan. ρ-uncertainty: Inference-Proof Transaction Anonymization. *VLDB Endow.*, 3(1), September 2010.

[5] R. Chen, B. C. Desai, N. Mohammed, L. Xiong, and B. C. M. Fung. Publishing set-valued data via differential privacy. In *VLDB*, 2011.

[6] S. Chib and E. Greenberg. Understanding the metropolis-hastings algorithm, 1995.

[7] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports, Nature*, March 2013.

[8] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. Pentland. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221), January 2015.

[9] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 2010.

[10] J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics*, pages 169–193. University Press, 1992.

[11] G. Ghinita, Y. Tao, and P. Kalnis. On the anonymization of sparse high-dimensional data. In *ICDE*, 2008.

[12] G. Ghinita, Y. Tao, and P. Kalnis. Anonymous publication of sensitive transactional data. *IEEE TKDE*, 23(2), Feb. 2011.

[13] Y. He and J. F. Naughton. Anonymization of set-valued data via top-down, local generalization. *VLDB Endow.*, 2(1), Aug. 2009.

[14] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[15] A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas. Releasing search queries and clicks privately. In *WWW*, 2009.

[16] D. Lemire, L. Boytsov, and N. Kurz. SIMD compression and the intersection of sorted integers. *CoRR*, abs/1401.6399, 2014.

[17] G. Loukides, A. Gkoulalas-Divanis, and B. Malin. Coat: Constraint-based anonymization of transactions. *Knowledge and Information Systems*, 28(2), 2011.

[18] G. Loukides, A. Gkoulalas-Divaniss, and J. Shao. Anonymizing transaction data to eliminate sensitive inferences. In *DEXA*, 2010.

[19] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, (6):1087–1092.

[20] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE S&P*, pages 111–125, 2008.

[21] G. Poulis, S. Skiadopoulos, G. Loukides, and A. Gkoulalas-Divanis. Distance-based k^m-anonymization of trajectory data. *IEEE MDM*, 2:57–62, 2013.

[22] P. Samarati. Protecting respondents' identities in microdata release. *IEEE TKDE*, 13(6), Nov. 2001.

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

[24] M. Terrovitis, N. Mamoulis, and P. Kalnis. Privacy-preserving anonymization of set-valued data. *VLDB Endow.*, 1(1), 2008.

[25] M. Terrovitis, N. Mamoulis, and P. Kalnis. Local and global recoding methods for anonymizing set-valued data. *VLDB Journal*, 20(1), 2011.

[26] Y. Xu, B. C. M. Fung, K. Wang, and A. W. C. Fu. Publishing sensitive transactions for itemset utility. In *ICDM*, 2008.

[27] Y. Xu, K. Wang, A. W.-C. Fu, and P. S. Yu. Anonymizing transaction databases for publication. In *ACM KDD*, 2008.

[28] H. Zang and J. Bolot. Anonymization of location data does not work: A large-scale measurement study. In *MobiCom*, 2011.

Appendix

Geweke convergence diagnostic: if X_t denotes a random variable describing the number of records in D containing the current state of $\mathcal{M}$ at time t, and $\mathbf{X}_t = (X_1, X_2, \ldots X_t)$, then we compute the z-score $z = \frac{E[\mathbf{X}_a] - E[\mathbf{X}_b]}{\sqrt{Var(\mathbf{X}_a) + Var(\mathbf{X}_b)}}$, where $\mathbf{X}_a$ is the prefix of $\mathbf{X}_t$ (first 10%), and $\mathbf{X}_b$ is the suffix of $\mathbf{X}_t$ (last 50%). We declare convergence when the z-score falls within $[-1, 1]$. Indeed, if $\mathbf{X}_a$ and $\mathbf{X}_b$ become identically distributed (i.e., $\mathbf{X}_a$ and $\mathbf{X}_b$ appears to be uncorrelated), the z values become normally distributed with mean 0 and variance 1 according to the law of large numbers.