Relative privacy risks and learning from anonymized data

Privacy e learning in dati anonimizzati

Michele Boreale and Fabio Corradi

Abstract We consider group-based anonymized tables, a popular approach to data publishing. This approach aims at protecting privacy of the involved individuals, by releasing an *obfuscated* version of the original data, where the exact correspondence between individuals and attribute values is hidden. When publishing data about individuals, one must typically balance the *learner*'s utility against the risk posed by an *attacker*, potentially targeting individuals in the dataset. Accordingly, we propose a MCMC based methodology to learn the population parameters from a given anonymized table and to analyze the risk for any individual in the dataset to be linked to a specific sensitive value when the attacker has got to know the individual's nonsensitive attributes. We call this *relative risk* analysis. Finally, we illustrate results obtained by the proposed methodology on a real dataset.

Abstract *Nel lavoro consideriamo tabelle anonimizzate realizzate per rendere disponibili informazioni sulla popolazione, nascondendo però l'attribuzione dei dati sensibili ai singoli rispondenti. Si valuta l'informazione sulla popolazione che rimane disponibile e il rischio di violare i la privacy dei rispondenti, fornendo diverse forme di apprendimento e di valutazione. Vengono riportati i risultati di un esperimento condotto su un dataset reale.*

Key words: Privacy, anonymization, k-anonymity, MCMC methods.

Michele Boreale
Università di Firenze, Dipartimento di Statistica, Informatica, Applicazioni. e-mail: michele.boreale@unifi.it

Fabio Corradi
Università di Firenze, Dipartimento di Statistica, Informatica, Applicazioni. e-mail: corradi@disia.unifi.it

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

Table 1 A table (left) anonymized according to Local recoding (center) and Anatomy (right).

ID	Nat.	ZIP	Dis.
1	Malaysia	45501	Heart
2	Japan	45502	Flu
3	Japan	55503	Flu
4	Japan	55504	Stomach
5	China	66601	HIV
6	Japan	66601	Diabetes
7	India	77701	Flu
8	Malaysia	77701	Heart

Original table

ID	Nat.	ZIP	Dis.
1	{M,J}	4550*	Heart
2	{M,J}	4550*	Flu
3	Japan	5550*	Flu
4	Japan	5550*	Stomach
5	{C,J}	66601	HIV
6	{C,J}	66601	Diabetes
7	{I,M}	77701	Flu
8	{I,M}	77701	Heart

Local recoding

GID	Nat.	ZIP	Dis.
1	Japan	45502	Heart
1	Malaysia	45501	Flu
2	Japan	55504	Flu
2	Japan	55503	Stomach
3	Japan	66601	HIV
3	China	66601	Diabetes
4	Malaysia	77701	Flu
4	India	77701	Heart

Anatomy

1 Introduction

It is a common practice to release datasets involving individuals in some anonymized form. The goal is to enable the computation of population characteristics with reasonable accuracy, at the same time preventing leakage of sensitive information about individuals in the dataset. We are interested in *group-based* techniques, put forward in Computer Science in the last 15 years or so: k-anonimity [5] and its variants, like ℓ-diversity [2], and Anatomy [8]. Despite their weakness against attackers with strong background knowledge, these techniques are a common choice when it comes to table publishing [3]. In group-based methods, the anonymized or obfuscated version of a table is obtained by partitioning records in groups enjoying certain properties (see Section 2). Generally speaking, even knowing that an individual belongs to a group of the anonymized table, it will not be possible for an attacker to link an individual to a specific sensitive value in the group. Two examples of group based anonymization are in Table 1, adapted from [7]. The original table collects medical data from eight individuals; here *Disease* is considered as the only sensitive attribute. The central table is a 2-anonymous table, obtained by *local recoding*: within each group, the nonsensitive attributes are generalized so as to make them indistinguishable. This is an example of *horizontal* scheme. Generally speaking, each group in a k-anonymous table consists of *at least k* records, which are indistinguishable as far as the nonsensitive part is concerned. Finally there is an example of application of *Anatomy*: within each group, the nonsensitive part of the rows are *vertically* randomly permuted, thus breaking the link between sensitive and nonsensitive values.

We put forward a probabilistic model to reason about the *relative risk* posed by the release of anonymized datasets (Section 2), i.e. the leakage of sensitive information for an individual in the table, *beyond* what is implied for the general population. To see what is at stake here, consider the central table of Fig. 1. An adversary may reason that, with the exception of the first group, a Japanese is never connected to Heart Disease. This hint can become a strong evidence in a larger, real-world table. Suppose now that the attacker's target, a Malaysian living at ZIP code 4550*, is known to belong to the table, so he must be in the first group. On the basis of the

evidence about Japanese not suffering from Heart Disease, the attacker can then link with high probability his target to Heart Disease. Here, the attacker combines knowledge learned from the anonymized table and about his victim with the group structure of the table itself. To formally reason about this phenomenon, we will define the *relative* privacy risks by comparing two conditional probability distributions, encoding respectively: what can be learned about the population from the anonymized table; and what can be learned about a the victim, given knowledge of her/his nonsensitive attributes *and* presence in the table (Sections 3). Generalizing Kifer [1] and Wong et al. [7], we propose a MCMC to learn both the parameter's population and the attacker's probability distribution from the anonymized data (Section 4). We finally illustrate the results of an experiment on a real-world dataset (Section 5).

2 Group based anonymization schemes and the probabilistic model

Given a dataset of N individuals, let $\mathscr{R}$ and $\mathscr{S}$, ranged over by r and s, be finite nonempty sets of *nonsensitive* and *sensitive* values. A *row* is a pair $(s,r) \in \mathscr{S} \times \mathscr{R}$.

In a group based scheme a cleartext *table* is an arrangement of a multiset of N rows, say $d = (s_1,r_1),...,(s_N,r_N)$, into a sequence of *groups*, $t = g_1,...,g_k$, where each group is a sequence $g_j = (s_{j_1},r_{j_1}),...,(s_{j_{n_j}},r_{j_{n_j}})$. Given a generic group g, its *obfuscation* is a pair $g^* = (l,m)$, where $m = s_1,s_2,...$ is the sequence of sensitive values occurring in g, and l, called *generalized nonsensitive value*, is:

- a *superset* of g's nonsensitive values for *horizontal* schemes (e.g. k-anonymity);
- the *multiset* of g's nonsensitive values $\{|\, r_1,r_2,...\,|\}$, for *vertical* schemes.

Given a table $t = g_1,...,g_k$, an obfuscated table is a $t^* = g_1^*,...,g_k^*$, such that each g_j^* is an obfuscation of the corresponding group g_j. An *anonymization algorithm* $\mathscr{A}$ is a – possibly probabilistic – mechanism that maps collections of N rows, d, into obfuscated tables, t^*.

Our model consists of the following random variables with the associated meaning.

- Π, taking values in the set of full support probability distributions $\mathscr{D}$ over $\mathscr{S} \times \mathscr{R}$: the (unknown) joint probability distribution of the population.
- $T = G_1,...,G_k$, taking values in the set of tables. Each group G_j is in turn a sequence of n_j consecutive rows in T, $G_j = (S_{j_1},R_{j_1}),...,(S_{j_1+n_j},R_{j_1+n_j})$; the number k of groups is not fixed, but depends itself on the rows S_j,R_j;
- $T^* = G_1^*,...,G_k^*$, taking values in the set of obfuscated tables.

We assume that the above three random variables form a Markov chain $\Pi \longrightarrow T \longrightarrow T^*$. In other words, the joint probability density f of these variables can be factorized as:

$$f(\pi,t,t^*) = f(\pi)f(t|\pi)f(t^*|t)\,. \tag{1}$$

We also assume the following.

- $\pi \in \mathscr{D}$ is encoded as a pair of $(\pi_S, \pi_{R|S})$ such that $f(s,r|\pi) \propto f(s|\pi_S) f(r|\pi_{r|s})$. Here, each π_S is a distribution over $\mathscr{S}$, and each $\pi_{R|S}$ is viewed as a collection of distributions over $\mathscr{R}$, $\pi_{R|S} = (\pi_{R|s})_{s\in\mathscr{S}}$. We posit that the π_S and the $\pi_{R|s}$s are chosen independently, according to Dirichlet distributions of hyperparameters $\alpha = (\alpha_1, ..., \alpha_{|\mathscr{S}|})$ and $\beta^s = (\beta^s_1, ..., \beta^s_{|\mathscr{R}|})$, respectively. In other words

$$f(\pi) = \mathrm{Dir}(\pi_S \,|\, \alpha) \cdot \prod_{s\in\mathscr{S}} \mathrm{Dir}(\pi_{R|s} \,|\, \beta^s)\,. \tag{2}$$

- The N individual rows composing the table t, $(s_1, r_1), ..., (s_N, r_N)$ are assumed to be drawn i.i.d. conditionally to Π. This amounts to positing that:

$$f(t|\pi) \propto f(s_1, r_1|\pi) \cdots f(s_N, r_N|\pi)\,. \tag{3}$$

3 The honest learner, the attacker and measures of relative risk

A *honest learner* is someone who, after observing $T^* = t^*$, updates his knowledge on the population parameters π. In addition an *attacker* also knows the nonsensitive value r_v of a victim in T. In what follows we shall fix once and for all t^* and r_v such that $f(r_\mathrm{v}, t^*) \stackrel{\triangle}{=} f(r_\mathrm{v} \text{ occurs in } T\,, T^* = t^*) > 0$. Let $p_\mathrm{L}(s,r)$ be the joint probability distribution on the population that can be learned given from t^*. Formally, for each (s,r)

$$p_\mathrm{L}(s,r|t^*) \stackrel{\triangle}{=} E_{\pi\sim f(\pi|t^*)}[f(s,r|\pi)] \;=\; \int_{\pi\in\mathscr{D}} f(s,r|\pi) f(\pi|t^*)\,\mathrm{d}\pi\,. \tag{4}$$

Of course, we can condition p_L on any given r so also the victim's nonsensitive attribute r_v and obtain the corresponding distribution on $\mathscr{S}$.

$$p_\mathrm{L}(s|r_\mathrm{v}, t^*) \stackrel{\triangle}{=} E_{\pi\sim f(\pi|t^*)}[f(s|r_\mathrm{v},\pi)] \;=\; \int_{\pi\in\mathscr{D}} f(s|r_\mathrm{v},\pi) f(\pi|t^*)\,\mathrm{d}\pi\,. \tag{5}$$

Given knowledge of r_v *and* knowledge that the victim is in T, we can define the attacker's distribution on $\mathscr{S}$ as follows. Let us introduce a random variable V, identifying the victim as one of the individuals in T. In other words, V is is an index, which we posit is a priori uniformly distributed on $1..N$, and independent from Π, T. Recalling that each row (S_j, R_j) is identified by a unique index j, we can define the attacker's probability distribution on $\mathscr{S}$, after seeing t^* and r_v, as:

$$p_\mathrm{A}(s|r_\mathrm{v}, t^*) \stackrel{\triangle}{=} f(S_V = s \mid R_V = r_\mathrm{v}, t^*)\,. \tag{6}$$

Theorem 1 provides $p_\mathrm{A}(s|r_\mathrm{v}, t^*)$ only based on the marginals R_j given t^*.

Theorem 1. *Let $T = (S_j, R_j)_{j\in 1..N}$ and s_j the sensitive value in the row j of t^*. Then*

$$p_{\mathrm{A}}(s|r_{\mathrm{v}},t^*) \propto \sum_{j:s_j=s} f(R_j = r_{\mathrm{v}}\,|\,t^*)\,. \tag{7}$$

We now define some measures of relative privacy risk to be put at work in Section 5.

Definition 1 (risk measures). Let p a full support distribution on $\mathscr{S}$ and (s,r) a row in t. We say this row is *at risk under* p if $p(s) = \max_{s'} p(s')$, and that its *risk level under* p is $p(s)$. For an individual row (s,r) in t, which is at risk under $p_{\mathrm{A}}(\cdot|r,t^*)$, its *relative risk level* is $\mathbf{R}(s,r,t,t^*) \stackrel{\triangle}{=} \frac{p_{\mathrm{A}}(s|r,t^*)}{p_{\mathrm{L}}(s|r,t^*)}$. For $\ell \in \{\mathrm{L},\mathrm{A}\}$, let us define (using the multiset notation $\{|\cdots|\}$) $N_\ell(t,t^*) \stackrel{\triangle}{=} |\{|\,(s,r) \in t : (s,r) \text{ is at risk under } p_\ell(\cdot|r,t^*)\,|\}|$. The *global relative risk* of t given t^* is: $\mathbf{GR}(t,t^*) \stackrel{\triangle}{=} \max\left\{0, \frac{N_{\mathrm{A}}(t,t^*)-N_{\mathrm{L}}(t,t^*)}{N}\right\}$.

4 Gibbs sampling

For real world datasets, none of the distributions (4), (5) or (7) will be computable analytically. Nonetheless, we can build accurate estimations of these distributions from samples of the marginals of the density $f(\boldsymbol{\pi},t|t^*)$, with $t = g_1,...,g_k$ (note that here the sensitive values s_i are actually fixed and known from t^*). This can done using a Gibbs sampler, provided we can effectively sample from the full conditionals of $\boldsymbol{\pi}$ and g_j, for $1 \le j \le N$. This is discussed below.

The Gibb's chain state sequence $(\boldsymbol{\pi}^i,t^i)$, $i = 0,1,...$, is defined in the usual way, starting from an initial state $x_0 = (\boldsymbol{\pi}^0,t^0)$ and sampling in turn $\boldsymbol{\pi}^i$ and each of the groups of $t^i = g_1^i,...,g_k^i$ separately, from the respective full conditionals. From equations (1), (2) and (3), it is easy to check that:

$$f(\boldsymbol{\pi}|t,t^*) = f(\boldsymbol{\pi}|t) \tag{8}$$

$$f(g_j|t_{-j},\boldsymbol{\pi},t^*) \propto f(g_j|\boldsymbol{\pi})f(g_j^*|t) \quad (1 \le j \le k). \tag{9}$$

Each of the above two relations enables sampling from the corresponding full conditional on the left-hand side. Indeed, (8) is a posterior Dirichlet distribution, from which effective sampling can be easily performed. Denote by $\boldsymbol{\gamma}(t) = (\gamma_1,...,\gamma_{|\mathscr{S}|})$ the vector of the frequency counts γ_i of each s_i in t. Similarly, given s, denote by $\boldsymbol{\delta}^s(t) = (\delta_1^s,...,\delta_{|\mathscr{R}|}^s)$ the vector of the frequency counts δ_i^s of the pairs (r_i,s), for each r_i, in t. Then, for each $\boldsymbol{\pi} = (\pi_S,\pi_{R|S})$, we have

$$f(\boldsymbol{\pi}|t) = \mathrm{Dir}(\pi_S\,|\,\boldsymbol{\alpha}+\boldsymbol{\gamma}(t)) \cdot \prod_{s\in\mathscr{S}} \mathrm{Dir}(\pi_{R|s}\,|\,\boldsymbol{\beta}^s+\boldsymbol{\delta}^s(t))\,.$$

Let us discuss now (9). Here we will confine ourselves to the important case when the following conditions are satisfied: (a) the obfuscation function is deterministic, so that $f(g_j^*|t)$ equals 0 or 1; (b) the set $\mathscr{G}_j$ of the g_j's such that $f(g_j^*|g_j,t_{-j}) = 1$ depends solely on $g_j^* = (l_j,m_j)$, and is given by

$$\mathscr{G}_j = \begin{cases} \{g = (s_1,r_1),...,(s_n,r_n) : r_\ell \in l_j \text{ for } 1 \le \ell \le n\} & \text{(horizontal schemes)} \\ \{g = (s_1,r_{i_1}),...,(s_n,r_{i_n}) : \text{ for } r_{i_1},...,r_{i_n} \text{ a permutation of } m_j\} & \text{(vertical schemes).} \end{cases} \tag{10}$$

This assumption is exact in many important cases (e.g. Anatomy) and reasonable in the remaining ones. Under assumptions (a), (b) and (10) above, sampling from (9) amounts to drawing an element $g_j \in \mathscr{G}_j$ with probability $\propto f(g_j|\boldsymbol{\pi})$. This can be achieved via different techniques in each of the two cases of interest, horizontal and vertical; the details are omitted here due to lack of space.

5 Experiments

We have put a proof-of-concept implementation of our methodology at work on a subset of the Adult dataset from the UCI machine learning repository [6]. The considered subset consists of 5692 rows, with the following categorical attributes: *sex*, *race*, *marital status*, *education*, *native country*, *workclass*, *salary class*, *occupation*, with *occupation* considered as the only sensitive attribute. Using the ARX anonymization tool [3], we have obtained three different anonymized versions of the considered dataset, enjoying k-anonymity for, respectively: $k = 4$, $k = 5$ and $k = 10$. The average size of the groups varied from 38 rows (for $k = 4$) to 355 rows (for $k = 10$). We run the Gibbs sampler on each of these three anonymized datasets. We obtained the following figures for the global relative risks (cf. Def. 1) of the three datasets: $\mathbf{GR}_1 = 3.98\%$, $\mathbf{GR}_2 = 1.7\%$ and $\mathbf{GR}_3 = 1.86\%$. In absolute terms, the fraction of rows of t^* correctly classified by the attacker ranged from 27.3% to 29.4%. The *maximal* relative risk level $\mathbf{R}$ ranged from about 1.9 to 3.93.

All in all, these results indicate that, in each case the considered anonymized datasets imply a significant relative privacy risk, for an appreciable fraction of the rows.

References

1. D. Kifer. Attacks on privacy and deFinetti's theorem. *SIGMOD 2009 Conference*: 127-138, 2009.
2. A. Machanavajjhala, J., Gehrke, D., and Kifer. ℓ-diversity: privacy beyond k-anonymity. In *ICDE'06*: 24, 2006.
3. F. Prasser, F. Kohlmayer. Putting Statistical Disclosure Control Into Practice: The ARX Data Anonymization Tool. In: Gkoulalas-Divanis, Aris, Loukides, Grigorios (Eds.): *Medical Data Privacy Handbook*, Springer, November 2015. ISBN: 978-3-319-23632-2.
4. C.P. Robert, G. Casella. *Monte Carlo Statistical Methods*. 2/e, Springer, 2004.
5. L. Sweeney. k-anonymity: a model for protecting privacy. *International journal on uncertainty, Fuzziness and knowldege based systems* 10(5), 557-570, 2002.
6. UCI Machine Learning repository, Adult dataset. `https://archive.ics.uci.edu/ml/datasets/Adult`, 1996
7. R. Chi-Wing Wong, A. Wai-Chee Fu, Ke Wang, Ph. S. Yu, J. Pei. Can the Utility of Anonymized Data be Used for Privacy Breaches? In *TKDD'11* 5(3): 16:1-16:24, 2011.
8. X. Xiao, and T. Tao. Anatomy: Simple and effective privacy preservation. In *VLDB'06*: 139Ű150, 2006.