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.

Key words: Privacy, anonymization, k-anonymity, MCMC methods.
### Table 1 A table (left) anonymized according to Local recoding (center) and Anatomy (right).

<table>
<thead>
<tr>
<th>ID</th>
<th>Nat.</th>
<th>ZIP</th>
<th>Dis.</th>
<th>GID</th>
<th>Nat.</th>
<th>ZIP</th>
<th>Dis.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Malaysia</td>
<td>45501</td>
<td>Heart</td>
<td>1</td>
<td>Japan</td>
<td>45502</td>
<td>Heart</td>
</tr>
<tr>
<td>2</td>
<td>Japan</td>
<td>45502</td>
<td>Flu</td>
<td>2</td>
<td>Malaysia</td>
<td>45501</td>
<td>Flu</td>
</tr>
<tr>
<td>3</td>
<td>Japan</td>
<td>55503</td>
<td>Flu</td>
<td>3</td>
<td>Japan</td>
<td>55504</td>
<td>Stomach</td>
</tr>
<tr>
<td>4</td>
<td>Japan</td>
<td>55504</td>
<td>Stomach</td>
<td>4</td>
<td>Japan</td>
<td>55503</td>
<td>Flu</td>
</tr>
<tr>
<td>5</td>
<td>China</td>
<td>66601</td>
<td>HIV</td>
<td>5</td>
<td>{C, J}</td>
<td>66601</td>
<td>HIV</td>
</tr>
<tr>
<td>6</td>
<td>Japan</td>
<td>66601</td>
<td>Diabetes</td>
<td>6</td>
<td>{C, J}</td>
<td>66601</td>
<td>Diabetes</td>
</tr>
<tr>
<td>7</td>
<td>India</td>
<td>77701</td>
<td>Flu</td>
<td>7</td>
<td>{I, M}</td>
<td>77701</td>
<td>Flu</td>
</tr>
<tr>
<td>8</td>
<td>Malaysia</td>
<td>77701</td>
<td>Heart</td>
<td>8</td>
<td>{I, M}</td>
<td>77701</td>
<td>Heart</td>
</tr>
</tbody>
</table>

Original table  Local recoding  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-anonymity [5] and its variants, like $\ell$-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 the victim, given knowledge of her/his non-sensitive 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 \(\mathcal{R}\) and \(\mathcal{S}\), ranged over by \(r\) and \(s\), be finite nonempty sets of nonsensitive and sensitive values. A row is a pair \((s, r) \in \mathcal{S} \times \mathcal{R}\).

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 \(\mathcal{A}\) 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.

- \(\Pi\), taking values in the set of full support probability distributions \(\mathcal{D}\) over \(\mathcal{S} \times \mathcal{R}\): 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 \(T\), \(G_j = (S_{j_1}, R_{j_1}), ..., (S_{j_{n_j}}, R_{j_{n_j}})\); 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 \(\Pi \rightarrow T \rightarrow T^*\). 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.

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

\[
f(\pi) = \text{Dir}(\pi_S | \alpha) \cdot \prod_{s \in \mathcal{S}} \text{Dir}(\pi_{R|s} | \beta^s).
\]

- 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 \( \Pi \). This amounts to positing that:

\[
f(t|\pi) \propto f(s_1, r_1 | \pi) \cdots f(s_N, r_N | \pi).
\]

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 \( \pi \). 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 \( A \) and \( T^* \), and independent, according to Dirichlet distributions of hyperparameters \( \alpha = (\alpha_1, ..., \alpha_{|\mathcal{S}|}) \) and \( \beta^s = (\beta^s_1, ..., \beta^s_{|\mathcal{R}|}) \), respectively. In other words

\[
f(s|\pi) = \int_{\mathcal{S}} f(s, r|\pi) f(r) \, dr.
\]

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 \( \mathcal{S} \).

\[
p_L(s|r_v, t^*) \triangleq \mathbb{E}_{\pi \sim f(\pi|t^*)} [f(s|r_v, \pi)] = \int_{\pi \in \mathcal{D}} f(s|r_v, \pi) f(\pi) \, d\pi.
\]

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

\[
p_A(s|r_v, t^*) \triangleq f(S_V = s | R_V = r_v, t^*).
\]

Theorem 1 provides \( p_A(s|r_v, t^*) \) 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
Relative privacy risks and learning from anonymized data

\[ p_A(s|r, t^*) = \sum_{j:y_j=s} f(R_j = r_i | t^*). \tag{7} \]

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 \( \mathcal{S} \) 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 \( p_A(s | r, t^*) \), its relative risk level is \( R(s, r, t, t^*) \triangleq \frac{p_A(s | r, t^*)}{p_A(s | r)} \). For \( \ell \in \{L, A\} \), let us define (using the multiset notation \( \{\cdot \} \)) \( N_\ell(t, t^*) \triangleq \{\{s, r\} \in t : (s, r) \) is at risk under \( p_\ell(\cdot | r, t^*) \}\}.

The global relative risk of \( t \) given \( t^* \) is: \( \text{GR}(t, t^*) \triangleq \max \left\{ 0, \frac{N_A(t, t^*) - N_L(t, t^*)}{N} \right\} \).

## 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 \( f(\pi, t | t^*) \), with \( t = g_1, \ldots, g_k \) (note that here the sensitive values \( s_i \) are actually fixed and known from \( t^* \)). This can be done using a Gibbs sampler, provided we can effectively sample from the full conditionals of \( \pi \) and \( g_j \), for \( 1 \leq j \leq N \). This is discussed below.

The Gibb’s chain state sequence \((\pi^i, t^i)\), \( i = 0, 1, \ldots \), is defined in the usual way, starting from an initial state \( x_0 = (\pi_0, t_0) \) and sampling in turn \( t^i \) and each of the groups of \( t^i = g_1^i, \ldots, g_k^i \) separately, from the respective full conditionals. From equations (1), (2) and (3), it is easy to check that:

\[
f(\pi^i | t^i, t^*) = f(\pi^i | t^i) \tag{8}
f(g_j | t_j^i, \pi_t^i) = f(g_j | \pi_t^j) f(g_j^* | t^i) \tag{9} \quad (1 \leq j \leq k).
\]

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 \( \gamma(t) = (\gamma_1, \ldots, \gamma_s) \) the vector of the frequency counts \( \gamma_i \) of each \( g_i \) in \( t \). Similarly, given \( s \), denote by \( \delta^s(t) = (\delta_1^s, \ldots, \delta_s^s) \) the vector of the frequency counts \( \delta_i^s \) of the pairs \((r_i, s)\), for each \( r_i \), in \( t \). Then, for each \( \pi = (\pi_S, \pi_{R|S}) \), we have

\[ f(\pi | t) = \text{Dir}(\pi_S | \alpha + \gamma(t)) \cdot \prod_{s \in \mathcal{S}} \text{Dir}(\pi_{R|s} | \beta^s + \delta^s(t)). \]

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 \( f(g_j^* | t^i) \) equals 0 or 1; (b) the set \( \mathcal{G}_j \) of the \( g_j \)'s such that \( f(g_j^* | g_j, t_{-j}^i) \) depends solely on \( g_j^* = (l_j, m_j) \), and is given by

\[
\mathcal{G}_j = \begin{cases} 
\{g = (s_1, r_1), \ldots, (s_n, r_n) : r_i \in l_j \text{ for } 1 \leq i \leq n \} & \text{(horizontal schemes)} \\
\{g = (s_1, r_1), \ldots, (s_n, r_n) : \text{ for } \{r_1, \ldots, r_n\} \text{ a permutation of } m_j \} & \text{(vertical schemes)}
\end{cases}
\]

(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 $g_j \in G_j$ with probability $\propto f(g_j | \pi)$. 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: $GR_1 = 3.98\%$, $GR_2 = 1.7\%$ and $GR_3 = 1.86\%$. 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 $R$ 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