Joint UNECE/Eurostat work session on statistical data confidentiality
(Ottawa, Canada, 28-30 October 2013)

Topic (ii): New methods for protection of microdata

Evaluating the Potential of Differential Privacy Mechanisms for Census Data

Prepared by Jordi Soria-Comas, Universitat Rovira i Virgili, Catalonia, Spain and Jörg Drechsler, Institute for Employment Research, Germany
Evaluating the Potential of Differential Privacy Mechanisms for Census Data

Jordi Soria-Comas*, Jörg Drechsler**

* Universitat Rovira i Virgili. Dept. of Computer Engineering and Mathematics, UNESCO Chair in Data Privacy, Av. Pasos Catalans 26, E-43007 Tarragona, Catalonia, jordi.soria@urv.cat

** Institute for Employment Research, Regensburger Str. 104, 90478 Nürnberg, Germany, joerg.drechsler@iab.de

Abstract. Despite its undeniable attractiveness as the only data protection mechanism with formal privacy guarantees, the concept of differential privacy has been repeatedly criticized because of the deteriorating effects of currently available differential privacy mechanisms. Due to the strong assumptions regarding the knowledge of a potential data intruder, the amount of noise that needs to be added to sufficiently protect the data is often so large that any inference based on the perturbed data will basically be considered useless. This is especially true if the micro-data should be released to the public. However, we argue in this paper that the situation might be different for Census data. The large number of available records coupled with a limited set of only a few (often categorical) variables will ensure that most of the cells defined by cross-classifying the different attributes still contain an ample number of records. Thus, the noise that needs to be added to fulfill the differential privacy requirements might have only minor effects on data quality. To enable the release of detailed geographical information we propose a differentially private procedure based on a micro-aggregation algorithm with a fixed minimal cluster size. We evaluate whether meaningful results can be obtained with this approach using administrative data gathered by the German Federal Employment Agency. Detailed geocoding information has been added to this database recently and plans call for making this valuable source of information available to the scientific community. We expect that the proposed micro-aggregation algorithm will enable us to release detailed geocoding information while offering strong differential privacy guarantees.

1 Introduction

In 2006, Dwork (2006) and Dwork et al. (2006) introduced the concept of differential privacy that offered strong formal guarantees for disclosure limitation for the first time. It was developed for an interactive query-response mechanism: the database user is allowed to query the database, but queries are caught by a sanitizing mechanism that masks the response prior to release. A query-response mechanism enables
researchers to customize their analysis, while keeping the data confidential. Despite its obvious merits differential privacy has been repeatedly criticized for the distorting effects the sanitizing mechanism has on the query results (Fienberg et al., 2010; Charest, 2012; Muralidhar and Sarathy, 2010; Sarathy and Muralidhar, 2010, 2011). In many cases the amount of noise that needs to be added to fulfill the differential privacy requirements is so large that the answers the user gets from his or her queries are basically useless from a data utility point of view. However, the situation might be different when it comes to Census data. Three important features that are common to most Censuses around the world make the noise based approach for differential privacy a potentially viable solution in this case: First, Census data usually mostly consist of categorical variables. Second, the number of records is large and finally, since Censuses usually are mandatory, the information that Alice is in the database is not a sensitive information that needs to be protected. To the contrary it could be argued that the fact that Bob is not in the database might be considered sensitive information since it would reveal that he did not follow his legal obligation to participate in the Census.

If the final goal is to release the actual micro-data, differentially private micro-data is usually generated by approximating the distribution of the data: the domain is partitioned into several prefixed regions and the number of individuals within each of those regions is counted. As differential privacy is agnostic of the original data, the partitioning of the data domain has to be decided prior to accessing the data, which may lead to sparsely populated or even empty partition sets. Such partition sets usually occur when dealing with variables that are unevenly distributed over their domain. They are problematic, as even a small noise can change the counts in those sets substantially.

To allow for the release of very detailed geographical information without sacrificing too much data utility, we propose a procedure that allows to incorporate prior knowledge (which either is already publicly available or is not considered to be disclosive). In particular, we assume that the marginal distribution of the geographical information is known. This is not an unrealistic assumption. The number of inhabitants is often released on a very detailed level. Furthermore, as stated above, the fact that somebody is in the Census database is in itself not a sensitive information. Thus, even if very detailed geocoding information would be released, which would allow to uniquely identify most individuals in the database this would not be problematic as long as the algorithm guarantees that the amount of information that can be learned regarding the relationship between the geographical information and the other variables in the data set is limited by differential privacy.

The evaluation of the approach was motivated by a recent project at the Institute for Employment Research (IAB) that linked the rich source of administrative data gathered by the Federal Employment Agency with detailed geocoding information for each record in the database. Plans call for making this valuable source of inform-
formation available to the scientific community. Given the amount of detail contained in the data set, it is obvious that standard methods like coarsening variables or top- or bottom coding will either not sufficiently protect the data or will destroy the value added by linking the data sources. We expect that the proposed micro-aggregation algorithm combined with noise addition will enable us to release detailed geocoding information while offering strong differential privacy guarantees.

The remainder of the paper is organized as follows. In Section 2 we review the concept of differential privacy and some of the literature on the generation of differentially private micro-data. Section 3 describes the introduction of prior knowledge. Section 4 provides some analytical results regarding the expected impact of the proposed approach on data utility. In Section 5 we introduce the data set that we use to empirically evaluate the approach, discuss how prior knowledge can be incorporated for this data set and present some evaluation results. The paper concludes with some final remarks.

2 Differential Privacy

Differential privacy guarantees that the influence of any single individual in the query response is limited. Thus, the disclosure risk for individuals is also limited. This is formalized by requiring the probability distribution of the response to be similar regardless of whether any specific individual is included in the data set or not.

Definition 1 (ε-differential privacy, Dwork (2006)). A randomized function \( \kappa \) gives \( \varepsilon \)-differential privacy if, for all data sets \( D_1, D_2 \) such that one can be obtained from the other by adding or removing a single record, and all \( S \subset \text{Range}(\kappa) \)

\[
P(\kappa(D_1) \in S) \leq \exp(\varepsilon)P(\kappa(D_2) \in S). \tag{1}
\]

The factor \( \exp(\varepsilon) \) is known as the knowledge gain or leakage.

To provide a differentially private response to a query \( f \), the database access mechanism needs to come up with \( \kappa_f \), an approximation to \( f \) that satisfies the requirements of Definition 1. To satisfy Inequality 1, uncertainty needs to be introduced into \( \kappa_f \). This uncertainty is usually achieved by adding random noise. With noise addition, the task is to determine the minimal amount of noise required to satisfy differential privacy. The addition of a Laplace distributed noise, whose scale parameter depends on the variability of the query function \( f \), is a common approach to noise calibration. In our case we are only concerned with counting the number of individuals that belong to each of the components of a partition of the data domain. As the output of such queries are natural numbers, we employ a two-sided geometric distribution (Ghosh et al., 2009); a discrete version of the Laplace distribution with the following distribution:

\[
P(\text{Noise} = i) = p_0 \times \exp(-\varepsilon \cdot |i|), \forall i \in \mathbb{Z},
\]
where $p_0$ depends on the desired level of differential privacy,

$$p_0 = \frac{1 - \exp(-\epsilon)}{1 + \exp(-\epsilon)}.$$

For the generation of differentially private micro-data, the usual approach computes the frequency distribution of the original data and then a masked version of that distribution satisfying differential privacy is published. To compute the frequency distribution of the original data, the data domain is partitioned into several disjoint regions and the number of individuals that belong to each of those regions is counted. For a data set composed of multiple attributes, each attribute is partitioned individually, and combined to form the data set partition. Thus, if the data set only consists of categorical variables a straightforward approach would be to simply use the full cross classification of all variables as the partition of the data set domain. However, if the data set consists of a large number of variables, continuous variables are involved, or the number of categories for some of the variables is large, finding a useful partition of the domain is not so easy. As the amount of noise added to the count of each of the partition sets is independent of the actual count, the greater the count the smaller the relative error. Thus, there is a trade-off between the partition granularity and the accuracy of the approximation. This trade-off becomes a fundamental limitation for the case of a data set with a large number of attributes (as we may be forced to select a coarse partition for each of the attributes) and for the sparsely populated areas in the original data (as we get less accuracy in those areas). Several methods for the generation of differentially private micro-data have been proposed. Each of them taking a different approach towards the trade-off between granularity and accuracy. In Machanavajjhala et al. (2008) a data set containing commuting patterns is to be published. Because of the large number of commuting destinations, this is a very sparse data set. The authors deal with this problem by introducing the notion of $(\epsilon, \delta)$-probabilistic differential privacy, where $\epsilon$-differential privacy is enforced except for a small probability $\delta$. In Korolova et al. (2009) the aim is to publish query logs. To avoid queries with small counts, only queries with counts over a predefined threshold are considered. In Qardaji et al. (2013) the authors deal with sparse areas by considering an adaptive partitioning that is finer in more populated areas. Other approaches for the generation of differentially private micro-data apply functional transformations to the original data (such as a wavelet (Xiao et al., 2011) or a Fourier (Barak et al., 2007) transformations) and then apply the random noise in the frequency domain.

### 3 Exploiting Prior Knowledge

We use the term prior knowledge to refer to any kind of knowledge about the data that is available before the data release takes place. Because the prior knowledge is assumed to be shared by everybody there is no need to mask it in the differentially
private data set. By introducing prior knowledge we restrict the data sets $D_1$, $D_2$ that need to be considered in Definition 1 to those that are consistent with the prior knowledge. Apart from publicly available information, we can also consider any kind of information that is not considered to be disclosive as prior knowledge, even if it is not publicly available. For example, for a data set with two attributes, if the attributes are not considered disclosive on their own, the marginal distributions can be taken as prior knowledge.

4 Some Analytical Results Regarding Data Utility

To obtain differential privacy, we add an independent noise to each of the sets in which the data domain has been partitioned. Under the geometric noise distribution described in Section 2 the expected error – in absolute terms – for a single partition set is:

$$\sum_{i \in Z} |i|P(N = i) = \frac{2}{\exp(\varepsilon) - \exp(-\varepsilon)}.$$

For a region that comprises $m$ of the partition sets $m$ independent noises are added. Therefore, the expected error is multiplied by $m$. However, the real impact of the expected error depends on the actual number of records contained in the area under consideration. The larger the actual count, the less impact the expected error has. To account for this, we compute the relative error ($RE$); the percentage of records that differ between the original and protected data set:

$$RE = \frac{2m}{n(\exp(\varepsilon) - \exp(-\varepsilon))},$$

where $n$ is the number of records of the original data contained in the region under consideration. Note that in the previous formula we can express $\varepsilon$ in terms of $RE$, $m$ and $n$.

The evaluation of $RE$ requires that $n$ is known. In the basic approach (without the assumption of any prior knowledge) $n$ is unknown, and the computation of $RE$ must rely on an estimation. If the actual $n$ is small, a small error in the estimation results in a large error in $RE$. As we will see later, the approach that we used to protect the very detailed geographical information in our data set will allow us to exactly determine $RE$ for each area under consideration.

5 Application to the Georeferenced Integrated Employment Biographies

In this section we provide an empirical evaluation of the approach described above on the data set that motivated this research: the georeferenced Integrated Employment Biographies. We first describe the data set, provide some information how
we used the idea of incorporating prior knowledge to protect the detailed geocoding information, give some analytical results regarding the relative error in our case and finally evaluate the utility of the protected data empirically.

5.1 The georeferenced Integrated Employment Biographies
The Integrated Employment Biographies (IEB) integrate five different sources of information collected by the Federal Employment Agency through different administrative procedures: the Employment History, the Benefit Recipient History, the Participants-in-Measures History, the Unemployment Benefit II Recipient History, and the Jobseeker History. We refer to Jacobebinghaus and Seth (2010) for a detailed description of the different data sources and of the IEB. Information available in the IEB include among other things: beginning and ending date of every employment, date of birth, gender and nationality, education and health status, employment status, monthly wages, working place and place of residence on the zip code level.

Recently, exact geocoding information has been added for individuals and establishments included in the IEB at the reference date June 30, 2009. The geocoding information was obtained from a georeferenced address database for Germany provided by the Federal Agency for Cartography and Geodesy which contains approximately 22 million addresses of German buildings and their corresponding geographic coordinates. Based on exact matches regarding the address information it was possible to obtain geocoding information for 94.6% of the 36.2 million individuals and 93.2% of the 2.5 million establishments contained in the IEB on the reference date (see Scholz et al. (2012) for more details regarding the matching of the two data sources).

To be able to provide access to the data for external researchers the IAB is looking for innovative ways to generate a sufficiently protected version of the linked data that still contains useful information at least on some of the variables. It was decided to start with a very limited set of variables initially and extend the set in the future if reasonable results, both in terms of disclosure risk and data utility could be achieved in the first round. The variables that are currently included in this preliminary data set are listed in Table 1.

Since the data set only contains a small number of categorical variables and information is available for the full population of individuals that were in contact with the German Social Security System on or before June 30, 2009, the characteristics of this data set should be comparable to the data characteristics typically found in Census data and our findings can be seen more broadly as a general evaluation of the feasibility of the noise addition approach for differential privacy in the Census data context.

5.2 Incorporating prior knowledge for the data set
As stated above, we consider the geocoding information to be prior knowledge for our data set. By taking the geocoding information as prior knowledge, we restrict
Table 1: Variables included in the data set used for the evaluations

<table>
<thead>
<tr>
<th>variable</th>
<th>characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>exact geocoding information</td>
<td>recorded as distance in meters from the point 52 northern latitude, 10 eastern longitude</td>
</tr>
<tr>
<td>sex</td>
<td>male/female</td>
</tr>
<tr>
<td>foreign</td>
<td>yes/no</td>
</tr>
<tr>
<td>employed</td>
<td>yes/no</td>
</tr>
<tr>
<td>unemployment benefits</td>
<td>yes/no</td>
</tr>
<tr>
<td>skills</td>
<td>low/medium/high</td>
</tr>
<tr>
<td>wage</td>
<td>low/medium/high</td>
</tr>
<tr>
<td>distance to work</td>
<td>5 categories (&lt;=1, 1–5, 5–10, 10–20, &gt;20 km)</td>
</tr>
<tr>
<td>zip-code</td>
<td></td>
</tr>
</tbody>
</table>

the data sets under consideration in Definition 1 to those with that exact geocoding data. Now we can use the actual geocoding data to partition the geocoding attribute domain, thus being able to adjust the size of the partitions to the density of the population. In particular the partitioning of the geocoding information is done by means of a micro-aggregation algorithm. Micro-aggregation algorithms group the data in clusters of \( k \) records (or more). In the construction of each cluster we seek to maximize the homogeneity of the records it contains. The partitioning that we get from the micro-aggregation offers a big advantage over the traditional partitioning based on a prefixed grid; it allows us to completely avoid the presence of empty partition sets or sets that only contain a small number of records. For the evaluations we use the MDAV micro-aggregation algorithm (Domingo-Ferrer and Mateo-Sanz, 2002) and replace the individual information with the centroid of the cluster. A downside of this approach is that it will not be possible for the user to determine which geographical area is represented by each cluster. If the cluster size is small enough this might not be a problem. Otherwise, a solution could be to include an additional variable in the released data set that contains random samples of the exact geocoding within each cluster from all individuals belonging to this cluster. Since this random sampling will insure that the provided geocodes are independent from any of the other attributes in the data set, no additional information except for the marginal distribution of the geocodes (which is assumed to be non-sensitive) is released. However, this information still provides useful information for the user, since it helps to determine the geographical area that formed the basis for the cluster. At the same time it also protects individuals that are not in the data set even though participation is mandatory: Since the geocodes are only samples from the original geocodes, the user never knows whether the fact that a specific individual is not in the database is because this individual was not in the original database or because he or she was not sampled. Of course if this idea will be incorporated in the final data product it is mandatory to clearly communicate to the users that the exact geocoding information should never be used for substantive analyses since any results based on this variable would be meaningless.

To further increase the utility of the released data set, we also force the clustering algorithm to find clusters only within zip-code areas, i.e., no clusters can be formed
Table 2: Evaluation of the relative error within a cluster for several cluster size and $\varepsilon$ values.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\varepsilon = 0.1$</th>
<th>$\varepsilon = 1$</th>
<th>$\varepsilon = 4.6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>37.53</td>
<td>3.19</td>
<td>0.08</td>
</tr>
<tr>
<td>100</td>
<td>18.76</td>
<td>1.60</td>
<td>0.04</td>
</tr>
<tr>
<td>200</td>
<td>9.38</td>
<td>0.80</td>
<td>0.02</td>
</tr>
</tbody>
</table>

that contain records with two different zip-codes. This will ensure that the relationship between the variables on the zip-code level is not further distorted beyond the noise that is introduced.

Apart from the geocoding data, we can take advantage of other pieces of prior knowledge. The data set contains several categorical variables, and there are some combinations of categories that do not make sense (such as receiving unemployment benefits while being employed). Since any user can determine these unsensible combinations, assuming these logical constraints to be prior knowledge will not disclose any sensitive information.

5.3 Analytical results regarding data utility

As we have partitioned the geospatial information in clusters of size $k$, we can exactly determine the relative error ($RE$) as defined in (2) for each of the clusters. The fact that the relative error is equal for each of the clusters is an important improvement over the basic approach, where sparse areas are subject to large errors.

According to Table 1, without taking into account the geocoding information, a fully cross classified table of all the variables (excluding impossible combinations as described above) contains 188 cells. Table 2 evaluates the relative error for several cluster size values ($n=50$, 100, and 200), and several values of $\varepsilon$ ($\varepsilon=0.1$, 1, and 4.6). We include the results for $\varepsilon = 4.6$ as this value was considered acceptable in other applications of the $\varepsilon$-differential privacy approach (Machanavajjhala et al., 2008).

It is obvious that choosing a small $\varepsilon$ is not feasible for our data set. For $\varepsilon = 0.1$ and 1, the relative error is large. However, the results seem acceptable for $\varepsilon = 4.6$: The relative error is around 8% for a cluster size of 50 and around 2% when the cluster contains 200 individuals.

5.4 Empirical Evaluations

Since the implemented protection algorithm already offers strict formal privacy guarantees, the main focus of this section is on data utility. To avoid lengthy evaluation runs all simulations are based on a small geographical subset of the original data (the data contains all individuals living in a 110 km x 50 km rectangular around the city of Bayreuth). The subset was chosen to contain a medium sized city as well as very rural areas in an attempt to make the subset as representative as possible for the entire German population.

Coming up with a meaningful global utility measure for this data set is difficult. In the following we will provide simple summaries of the data to get a first impression of the impact of the protection methods on data utility. The two parameters that can be varied to balance the risk and utility are the cluster size and the value of
As mentioned before, $\varepsilon = 4.6$ was considered acceptable in other applications of the $\varepsilon$-differential privacy approach. Thus, we stick with this value for now and only vary the cluster size.

Our approach to evaluating the utility of the protected data set is to compare relative frequencies for various cross tabulations of the variables contained in the original data and the protected data. We compare the relative frequencies on the zip-code level since this is the lowest level of geographical information available in the data set besides the individual geocoding information. Specifically, for each zip-code we first compute the relative frequencies for each cell entry for all possible cross classifications of all variables, i.e., all marginal distributions, two-way interactions, three-way interactions, etc. Then we evaluate how much these relative frequencies differ between the original data and the protected data.

For a formal definition of the utility measure we need some additional notation. Let $s = 1, \ldots, p$ be the different interaction levels, i.e., $s = 1$ are the marginal distributions, $s = 2$ are two-way interactions, etc. Since we have seven variables in our data set (excluding the geographic information) $p = 7$ in our case. Let $m_s = \binom{p}{s}$ be the number of different possible cross tabulations between all variables for interaction level $s$. For example, for our data set we can specify up to 21 different tables if we look at all two-way interactions, i.e., $m_2 = \binom{7}{2} = 21$. Finally, let $n^s_k$ be the number of cells in table $k^s$, where $k^s = 1, \ldots, m_s$. Using these definitions our utility measure can simply be expressed as:

$$\delta_{i,j,s,z} = f^{\text{org}}_{i,j,s,z} - f^{\text{prot}}_{i,j,s,z}, \quad \text{with} \quad f^{\text{org/prot}}_{i,j,s,z} = \frac{F^{\text{org/prot}}_{i,j,s,z}}{N_{j,s,z}} \cdot 100,$$

where $F^{\text{org/prot}}_{i,j,s,z}$ is the number of records in cell $\{i,j,s,z\}$ computed from the original/protected data set, with $i = 1, \ldots, n^s_k$, $j = 1, \ldots, m_s$, $s = 1, \ldots, p$, and $z = 1, \ldots, n_z$, where $z$ identifies the different zip-code areas with $n_z$ being the total number of zip-code areas in the data set. Finally, $N_{j,s,z} = \sum_{i} F^{\text{org/prot}}_{i,j,s,z}$ is the total number of records in each table.

Figure 1 contains boxplots for $\delta_{i,j,s,z}$ for three different cluster sizes: 50, 100, and 200 records. In all cases $\varepsilon = 4.6$. Different interaction levels are presented separately. We excluded cell entries with less than 50 observations.

Only for a very small number of cells the relative frequencies differ by more than two percentage points. More than 58% of the cells differ by less than 0.5 percentage points for all interaction levels for a cluster size of 50 records. As expected the analytical validity increases with increasing cluster size. For example, the percentage of cells that differ less than 0.5 percentage points increases to 75.3% and 92.3% for cluster sizes of 100 and 200 records respectively. However, this increased validity at the zip-code level comes at the price of less flexibility for the analyst. The algorithm will never produce meaningful results below the cluster level. Thus, it
Figure 1: Differences in percentage points for relative frequencies between the original and the protected data for different cluster sizes. All results are based on $\epsilon = 4.6$. The numbers above each boxplot show which percentage of the cell entries differ by less than 0.5 or 0.1 percentage points, respectively.

only makes sense to analyze the data for areas that are aggregates of the clusters. If for example a major is interested in the employment situation in certain areas of his or her town, this area can be better approximated if the cluster size selected for generating the data was small. This is the direct implication of the granularity-accuracy-trade-off described in Section 2. Thus, the data generating agency will have to balance carefully between offering a high level of flexibility for the analyst in terms of selecting the geographical area of interest and the general analytical validity of the protected data, since more noise needs to be added if the cluster sizes are small.

6 Conclusions
We have proposed to incorporate available prior knowledge when constructing differentially private micro-data. By introducing prior knowledge, we restrict the data set that must be considered to those that are consistent with the prior knowledge. We have evaluated the proposed approach with a census-like data set containing several
categorical attributes and detailed geospatial information. Having the geospatial data as prior knowledge, we are able to adjust the partitioning of the geospatial attribute domain to the actual data. By taking larger partition sets in sparse areas, we are able to provide the same relative error for both sparse and populated areas. The first results presented in this paper look promising but the limitations of the approach should not be ignored. To achieve useful results the partition granularity must be large enough that all partitions contain a sufficient number of records. As the number of variables in the data set grows this becomes unfeasible quickly. Furthermore, all continuous variables have to be descretized. Besides, compared to the approach taken by Machanavajjhala et al. (2008) that uses equidistant grids to partition the data, the approach has the disadvantage that the analyst can no longer directly identify which geographical area is covered by the points belonging to the same cluster. As we discussed, a possible solution to this problem could be that individual georeferences are sampled within each cluster and this information is released together with the data. However, it is not clear whether data releasing agencies would be willing to follow this approach. Still we believe, the advantage of distributing the amount of noise that is added equally over the clusters outweighs this downside of the clustering approach. Generally, further research including a detailed disclosure risk evaluation is needed to fully evaluate the potential and limitations of noise based differential privacy approaches for Census data.

Acknowledgements
We thank Theresa Scholz for her great support in all the data preprocessing steps and her useful data utility evaluation routines that helped us detect problems with earlier versions of the implemented algorithm.

References


