Partially Synthetic Dataset Generated for the Testing Purposes on the Basis of Available Public Use Anonymized Microdata

MARIO MILICEVIC, KRUNOSLAV ZUBRINIC, TOMO SJEKAVICA
Department of Electrical Engineering and Computing
University of Dubrovnik
Cira Carica 4, Dubrovnik
CROATIA
{mario.milicevic, krunoslav.zubrinic, tomo.sjekavica}@unidu.hr

Abstract: - Governments and organizations increasingly recognize huge opportunities in sharing and distribution of collected data, and research community must provide methods and algorithms for privacy-preserving data publishing. Without access to the original microdata it is impossible to estimate the quality of developed anonymization methods or to compare the classification accuracy and the computational time of various algorithms applied both on anonymized and original datasets. We propose another high-quality microdata source for testing purposes - partially synthetic dataset generated on the basis of actual public use anonymized microdata set. The original distribution of the data should be simulated in a significant extent, as well as attribute value correlations or functional dependencies. Since the synthesized data are based on published microdata sets, it is expected that hidden complex patterns within a dataset can be also preserved.

Key-Words: - Synthetic data, Confidentiality, Disclosure, Microdata, PPDP

1 Introduction

More than 10 years ago, IT community talked about the knowledge gap - the simple idea that the amount of data we are creating is surpassing our ability to analyze that data. But nowadays knowledge gap generates yet another cause - large amounts of data cannot be made available to all interested parties that could be involved in their analysis.

Current information technologies enable organizations to collect, store and use various types of information about individuals and population. Governments and organizations increasingly recognize tremendous opportunities in sharing such as wealth of information for research and knowledge-based decision making. On the basis of mutual benefit or regulations requiring that certain information must be disclosed, there is a demand for the exchange and dissemination of data among different parties. However, data in its original form frequently contains sensitive information about individuals or other entities. Although the problem of privacy-preserving data publishing has received a lot of attention in recent years, famous illustration of this problem dates back to 1997 when L.Sweeney [1] re-identificated Massachusetts Governor William Weld's medical data. The Group Insurance Commission (GIC) was responsible for purchasing health insurance for state employees. Because the data were believed to be anonymous, GIC gave a copy of the data to researchers and sold a copy to industry. Also, it was possible to buy the voter registration list for Cambridge Massachusetts and this information can be linked using ZIP code, birth date and gender to the medical information, thereby linking diagnosis, procedures, and medications to particularly named individuals. For example, William Weld was governor of Massachusetts at that time and his medical records were in the GIC data. Governor Weld lived in Cambridge Massachusetts. According to the Cambridge Voter list, six people had his particular birth date; only three of them were men; and, he was the only one in his 5-digit ZIP code.

2 Problem Formulation

Publishing of detailed person-specific data in its original form immediately violates individual privacy. The current practice primarily relies on policies and guidelines to restrict the types of publishable data and on agreements on the use and storage of sensitive data. The limitation of this approach is that it either distorts data excessively or requires a trust level that is impractically high in many data-sharing scenarios [2]. A task of the

ISBN: 978-960-474-304-9

utmost importance was to develop methods and tools for publishing data in a more hostile environment, so that the published data remains practically useful while individual privacy is preserved.

Privacy-preserving data publishing (PPDP) provides methods and tools for publishing useful information while preserving data privacy. Recently, PPDP has received considerable attention in research communities, and many approaches have been proposed for different data publishing scenarios. Even if identifiers such as names and social security numbers have been removed, the adversary can use linking, homogeneity and background attacks to re-identify individual data records or sensitive information of individuals [3].

2.1 Related work

To prevent the re-identification attacks, ***k***-anonymity was proposed [4–6]. Specifically, a dataset is said to be k-anonymous ($k \geq 1$) if, on the quasi-identifier (QID) attributes each record is identical with at least k - 1 other records. QID is a subset of attributes that can indirectly reveal private information, possibly by joining with other data.

The larger the value of k, the better the privacy is protected. Several algorithms are proposed to enforce this principle [7-12]. Machanavajjhala et al. [13] showed that a k-anonymous table may lack diversity in the sensitive attributes. To overcome this weakness, they propose the ***ℓ***-diversity. However, even ***ℓ***-diversity is insufficient to prevent attribute disclosure due to the skewness and the similarity attack. To amend this problem, ***t***-closeness [14] was proposed to solve the attribute disclosure vulnerabilities inherent to previous models.

In [2] authors systematically summarize and evaluate different PPDP approaches, recently developed in several directions: generalization, suppression, anatomization, permutation and perturbation. As mentioned above, generalization and suppression replace values of specific description, typically the QID attributes, with less specific description. Anatomization and permutation disassociate the correlation between QID and sensitive attributes by grouping and shuffling sensitive values in a QID group. Perturbation distorts the data by adding noise, aggregating values, swapping values, or generating synthetic data based on some statistical properties of the original data.

Many statistical disclosure control methods [15] use synthetic data generation to preserve record owners' privacy and retain useful statistical information. The general idea is to build a statistical model from the data and then to sample points from the model. These sampled points form the synthetic data for data publication instead of the original data. An alternative synthetic data generation approach is condensation [16,17].

Synthetic data generation also addresses a major challenge for researchers who develop PPDP methods - the availability of real microdata in their original form before QID attributes are suppressed or generalized. The effectiveness of methods and algorithms is difficult to estimate on small artificial datasets, while the actual microdata are available only to a small number of researchers (who are mostly employees of state institutions). For others, the sixth United Nations Fundamental Principle of Official Statistics [18] is very clear on statistical confidentiality: "Individual data collected by statistical agencies for statistical compilation, whether or not they refer to natural or legal persons are to be strictly confidential and used exclusively for statistical purposes."

Due to the aforementioned reasons, researchers can use anonymized microdata files (Public Use Files - PUFs). However, the trend is to reduce the amount of data available in public use files and to put more reliance on licensed anonymized microdata files. This is an arrangement where specific users are authorized or licensed to use anonymized microdata files after making a relevant undertaking or contract. Although these files have been anonymized and individuals cannot be identified from these microdata files in isolation, it may be possible to do so by (statistical) matching with other files, hence the need for a license.

But regardless of the delivered anonymized microdata format, without original microdata researchers also cannot estimate the quality of resulting anonymized data, because information loss is an inherent feature of anonymization [19]. A recent illustration of this problem is the analysis done by Alexander et al. [20], who pointed out that there were substantial discrepancies between analyses (i.e. number of men and women at each individual age) done with disclosure-protected, public use census microdata samples and those done with actual census data. Also, in [21] the authors analyze recent cases when microdata perturbation gone wrong. This is another proof that anonymization algorithms researchers must have access to the original data, or at least its simulation.

Likewise, when applying data mining algorithms, without original microdata it is impossible to compare the classification accuracy

and the computational time of various algorithms applied both on anonymized and original datasets.

3 Experimental study

Aforementioned reasons motivated us to propose another high-quality microdata source for testing purposes - partially synthetic dataset generated on the basis of actual public use anonymized microdata set. The original distribution of the data should be simulated in a significant extent, as well as attribute value correlations or functional dependencies. Since the synthesized data are based on published microdata sets, it is expected that hidden complex patterns within a dataset can be preserved.

As a source of data, we chose public use micro-data files of the 2010 National Hospital Ambulatory Medical Care Survey (NHAMCS) [22]. NHAMCS is a national probability sample survey of visits to hospital outpatient and emergency departments (ED), conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention. A sample of 449 emergency services areas (ESAs) was selected from the EDs. Of these, 427 responded fully or adequately by providing. In all, 34936 Patient Record forms (PRFs) were submitted during reporting period. Based on the processed PRFs, each visit to ED in the end results with 417 attributes included in the published microdata. Range of attributes is very wide - from patient's age, sex, race, ethnicity, height and weight, reason for visit, cause of injury, diagnosis, ambulatory procedures and treatments, medications, hospital and ED characteristics, etc. - to census contextual variables about population in patient's ZIP code. Of course, the original PRFs contain detailed data that are generalized or suppressed in published microdata - for example the patient's name, date of birth and ZIP code, date and time of the visit and discharge.

2.1 Synthesis of generalized and suppressed data

Used method will be explained in a few examples - from simple to complex. Thus, generation of a random date of birth is trivial - it is based on the disclosed patient's age and on the assumption that there is no significant correlation, nor the functional dependencies between the exact date of birth (not the patient's age) and other attributes.

The entire patient's ZIP code is suppressed in published microdata, and since it is a potentially important attribute for a different analysis, special attention was paid to generate a realistic synthetic ZIP code. Due to the potential risk that someone can "link" an imaginary patient (based on the generated random date of birth and ZIP code) with a real person, it is also desirable that the generated ZIP codes do not exist in the U.S.

After analyzing the published public use microdata, we have found the functional dependence (1).

```
URBANRUR, PCTPOVR, HINCOMER, PBAMORER →
PAT_ZIP_CODE                                (1)
```

URBANRUR, HINCOMER, PCTPOVR and PBAMORER are census contextual variables: urban-rural classification of patient's ZIP code grouped into quartiles (URBANRUR), household income in patient's ZIP code grouped into quartiles (HINCOMER), percent impoverished in patient's ZIP code grouped into quartiles (PCTPOVR), and percent with bachelor's degree or higher grouped into quartiles (PBAMORER).

We use group by clause (GROUP BY URBANRUR, HINCOMER, PCTPOVR, PBAMORER) to detect how many ZIP code candidates exist in published microdata. After exclusion of instances with incomplete or missing values of these attributes, there are 241 different combinations, and thus at least as many ZIP codes. There are certainly more ZIP codes, particularly associated with smaller settlements in the sample. Therefore, the next step could be randomly dividing the rows which are related to the attribute value URBANRUR = 4 (small metro) into more than one ZIP code. But as the goal is to create data for testing, not for credible statistical analysis, this step is not essential.

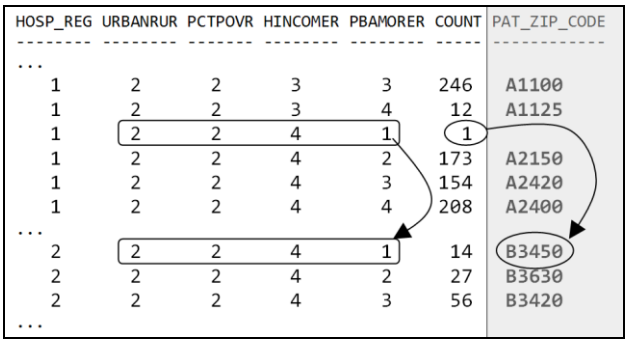
In the published microdata there are no more patient's attributes suitable for more accurately synthesizing of ZIP codes. However, it is indicated that the hospital belongs to one of the 4 U.S. regions (Northeast, Midwest, South and the West). As NHAMCS is related to EDs, it is reasonable to assume that in most cases patients seek help in one of the local hospitals. Therefore - with a certain error - the functional dependence (1) can be extended to (2).

```
HOSP_REGION, URBANRUR, PCTPOVR, HINCOMER,
PBAMORER → PAT_ZIP_CODE                     (2)
```

The error is due to the fact that some patients use EDs outside their region - for example when traveling. Because of this, a correction is made, so in the case that the grouping results in 1 or 2 patients, we assumed that these patients are from

ISBN: 978-960-474-304-9

other regions (Fig.1), i.e. from the city from another region which has the same attribute values (URBANRUR, PCTPOVR, HINCOMER, PBAMORER).



HOSP_REG	URBANRUR	PCTPOVR	HINCOMER	PBAMORER	COUNT	PAT_ZIP_CODE
...						
1	2	2	3	3	246	A1100
1	2	2	3	4	12	A1125
1	2	2	4	1	1	
1	2	2	4	2	173	A2150
1	2	2	4	3	154	A2420
1	2	2	4	4	208	A2400
...						
2	2	2	4	1	14	B3450
2	2	2	4	2	27	B3630
2	2	2	4	3	56	B3420
...						

Fig.1: Patient ZIP code generation

After application of the grouping according to (2), and elimination of the rows with 1 or 2 patients (Fig.1), 564 candidates remain for ZIP codes generation. Given the 33364 instances (ED patient visits), we concluded that 564 synthesized ZIP codes well represent the sample distribution.

In further procedure, artificial ZIP codes are generated within the region as pseudo-random numbers, taking into account the size of the area based on the number of patients and, what is more important, proximity to the regional center. A similar logic is generally embedded in ZIP codes worldwide.

Simple algorithm determining the neighboring areas will be explained using the first line in Fig.1. This area contributes with 246 patient visits, and is determined by the attributes HOSP_REGION, URBANRUR, PCTPOVR, HINCOMER, PBAMORER (hereinafter HUPHP) and their values 1-2-2-3-3 (respectively). Zip code A1100 is generated for this area. On Fig.2 is shown which hospitals are receiving patients from one area.

HOSPCODE	COUNT
242	86
284	20
091	19
040	17
087	14
334	12
166	11
125	9
310	7
262	7
320	5
...	

Fig.2: Distribution of patients from one ZIP code across hospitals

It is noticeable that 86 patients (from ZIP code A1100) visited the hospital 242, 20 patients visited the hospital 284 etc. Similarly, for these hospitals, we can analyze areas where live patients who visited them (Fig.3).

HOSP_REG	URBANRUR	PCTPOVR	HINCOMER	PBAMORER	HOSPCODE	COUNT
1	2	2	3	3	242	86
1	2	1	3	4	242	14
1	2	1	4	4	242	2
1	2	2	4	4	242	2
...						

HOSP_REG	URBANRUR	PCTPOVR	HINCOMER	PBAMORER	HOSPCODE	COUNT
1	2	2	3	3	284	20
1	2	1	4	4	284	16
1	2	2	3	2	284	8
1	2	1	4	3	284	6
1	1	3	3	2	284	3
...						

Fig.3: Areas from which are the patients visited hospitals 242 and 284

It is obvious that the areas with values of HUPHP attributes 1-2-1-3-4, 1-2-1-4-4, 1-2-2-3-2, etc. are adjacent areas to the ZIP code A1100, so the ZIP codes for these areas can be determined accordingly.

Median household income in patient's ZIP code is generalized and grouped into quartiles. The quartiles are identified by values of 1 to 4, indicating the poorest to wealthiest populations. According to U.S. Census Bureau [23] household income includes the income of the householder and all other individuals 15 years old and over in the household. They also claim that median income for households, families, and individuals is computed on the basis of a standard distribution.

In order to synthesize the value of median household income for previously generated ZIP codes, it is necessary to analyze in detail the features of the median household income distribution. The URBANRUR attribute values indicate that the NHAMCS 2010 uses old median household income data (probably from the year 2000). Meanwhile, the median and mean values were increased, but the main features of the distribution are not substantially changed, so we used data from the year 2010.

On the base of available data, we analyzed the distribution of 32486 ZIP codes by median household income. We didn't take into consideration 150 ZIP codes with a population of less than 20 (Fig.4).

ISBN: 978-960-474-304-9

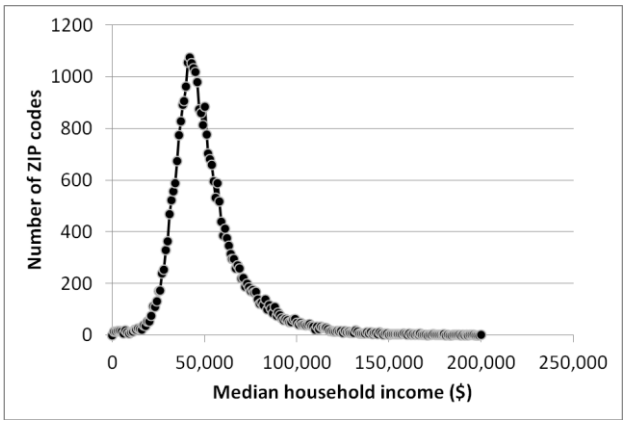


Fig.4: Number of ZIP codes accross median household income in patient's ZIP code

Minimal median household income in ZIP code was $316 (population 60), and maximum value was $223106 (population 2513). Distribution is right skewed (skewness=1.8, kurtosis=6.1), but if we assume that the maximum median household income is $150000 (thus excluding only 1% of ZIP codes), distribution is closer to normal (skewness=1.4, kurtosis=3.1) and skewed normal distribution offers satisfactory approximation. During generation of household income in patient's ZIP code within a given quartile (HINCOMER), we used pseudo-random algorithm that takes into account (in a certain extent) the ZIP code population. The reason is explained in Fig.5.

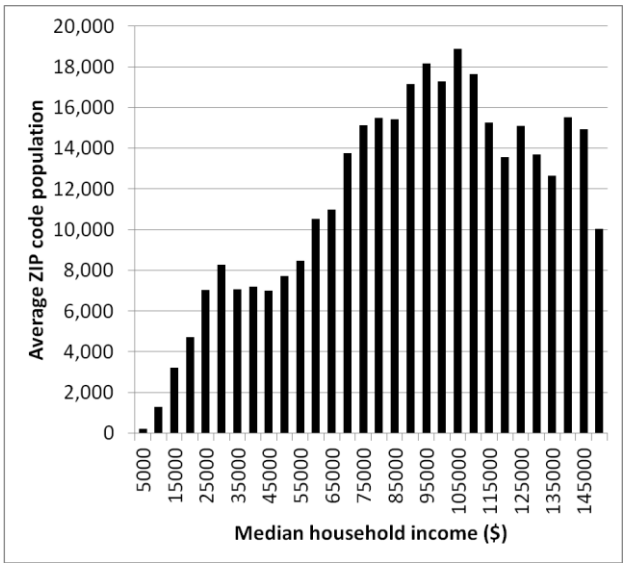


Fig.5: Average ZIP code population accross median household income in patient's ZIP code

It is obvious that the extreme median household income values (especially those close to the minimum) occur more frequently in ZIP codes with a smaller population.

4 Conclusion and future work

Lack of realistic data for anonymization algorithms and methods testing motivated us to develop a framework for the generation of synthetic data that can facilitate the development and testing of PPDP tools. Since the synthesized data are based on published microdata sets, it is expected that hidden complex patterns within a dataset can be preserved. Information that is the basis for the synthesis of the generalized or suppressed data is publicly available, and the resultant dataset cannot be used to identify individuals even though it contains their actual attributes.

Initial experiments conducted with synthesized dataset show the viability of using this approach for testing the existing and development of new PPDP techniques. In the continuation of this research, a framework can be extended so that, in addition to generating generalized or suppressed attributes for existing examples, it also allows the generation of completely new instances that are compatible with existing data.

References:

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

[2] B. C. M. Fung, K.Wang, R.Chen and P.S. Yu, Privacy-preserving data publishing: A survey of recent developments, *ACM Comput. Surv.* 42, 4, Article 14 (June 2010).

[3] X. Sun and L. Sun. Privacy Preserving Large-Scale Rating Data Publishing, *EAI Endorsed Transactions on Scalable Information Systems*, 13 (01-03), ICST. 2013.

[4] P. Samarati and L. Sweeney, Generalizing data to provide anonymity when disclosing Information, *PODS* 1998.

[5] P. Samarati and L. Sweeney, Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, *Technical Report SRI-CSL-98-04*, SRI Computer Science Laboratory, 1998.

[6] P. Samarati, Protecting respondents' identities in microdata release, *IEEE Transactions on Knowledge and Data Engineering*, 13(6): pp: 1010-1027. 2001.

[7] C. Aggarwal, On k-anonymity and the curse of dimensionality, *VLDB* 2005, pp.901-909

[8] R. J. Bayardo and R. Agrawal, Data privacy through optimal k-anonymisation, *ICDE* 2005, pp.217-228

[9] B. C. Fung, K. Wang and P. S. Yu, Top-down specialization for information and privacy preservation, *ICDE* 2005, pp.205-216

[10] K. LeFevre, D. DeWitt and R. Ramakrishnan, Incognito: efficient full-domain k-anonymity, *SIGMOD* 2005, pp.49-60

[11] K. LeFevre, D. DeWitt and R. Ramakrishnan, Mondrian multidimensional k-anonymity, *ICDE* 2006.

[12] J. Li, Y. Tao and X. Xiao, Preservation of Proximity Privacy in Publishing Numerical Sensitive Data, *ACM Conference on Management of Data (SIGMOD),* 2008. pp.473-486

[13] A. Machanavajjhala, J. Gehrke, D. Kifer and M.Venkitasubramaniam, l-Diversity: Privacy beyond k-anonymity, *ICDE* 2006.

[14] N. Li, T. Li and S. Venkatasubramanian, t-Closeness:Privacy Beyond k-anonymity and l-diversity, *ICDE* 2007, pp.106-115

[15] D.B. Rubin, Discussion statistical disclosure limitation, *Journal of Official Statistics*, Vol.9, No.2, 1993. pp.461–468

[16] C.C. Aggarwal and P.S. Yu, A framework for condensation-based anonymization of string data, *Data Mining and Knowedge. Discovery* 13, 3, pp.251–275

[17] C.C. Aggarwal and P.S. Yu, On static and dynamic methods for condensation-based privacypreserving data mining, ACM Trans. Datab. Syst. . Vol. 33, Issue 1, March 2008..

[18] United Nations Statistical Commission, "Fundamental Principles of Official Statistics", April 1994., available at: http://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx (accessed 20.4.2013).

[19] K. Purdam and M J. Elliot, A Case Study of the Impact of Statistical Disclosure Control on a Data Quality in the Individual UK Samples of Anonymised Records, *Environment and Planning A*(2007) , pp.1101-1118

[20] J.T. Alexander, M. Davern, and B. Stevenson, Inaccurate Age and Sex Data in the United States Census PUMS Files: Evidence and Implications, *Public Opinion Quarterly* (2010)74(3). pp.551-569

[21] L. Cleveland, R. McCaa, S. Ruggles and M. Sobek, When Excessive Perturbation Goes Wrong and Why IPUMS-International Relies Instead on Sampling, Suppression, Swapping, and Other Minimally Harmful Methods to Protect Privacy of Census Microdata, *Privacy in Statistical Databases* 2012, pp.179-187

[22] National Center for Health Statistics, Description of the National Hospital Ambulatory Medical Care Survey (NHAMCS), available at: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHAMCS/ (accessed 15.1.2013).

[23] U.S. Census Bureau. 2010 Household Income Table of Contents, available at: http://www.census.gov/hhes/www/cpstables/032011/hhinc/toc.htm (accessed 2.3.2013).