EGOV-CeDEM-ePart

Detecting and Generalizing Quasi-Identifiers by Affecting *Singletons*

Matteo Pastore*, Maria Angela Pellegrino**, Vittorio Scarano***

**Dipartimento di Informatica, Università degli Studi di Salerno, Italy, m.pastore48@studenti.unisa.it, {mapellegrino, vitsca}@unisa.it*
***Dipartimento di Informatica, Università degli Studi di Salerno, Italy, mapellegrino@unisa.it*
****Dipartimento di Informatica, Università degli Studi di Salerno, Italy, vitsca@unisa.it*

Abstract: In order to adhere to Open Government doctrine, Public Administrations (PAs) are requested to publish Open Data while preventing the disclosure of personal information of their citizens. Therefore, it is crucial for PAs to employ methods that ensure Privacy-preserving data publishing by distributing useful data while protecting individual privacy. In this paper, we study this problem by providing a two phases approach. First, we detect privacy issues by recognizing the minimum number of attributes that expose the highest number of unique values (that will be referred to as singletons) as Quasi-Identifier. We test our approach on real datasets openly published by the Italian government, and we discover that the quasi-identifier (year_of_birth, sex, ZIP_of_residence) discloses up to 2% unique values in already anonymized datasets. Once accomplished the detection phase, we propose an anonymization approach to limit the privacy leakage. We investigate which combination of attributes must be generalized to achieve the minimum number of singletons while minimising the amount of modified and removed rows. We tested our approach on real datasets as in the previous phase, and we noticed that by generalizing only rows corresponding to the singletons, we achieve nearly no singletons while affecting only the 2% of rows.

Keywords: Privacy, Quasi-Identifies, Anonymization, Generalization

1. Introduction

Data owners are spur in opening up their data to enable informed decision making, ensure transparency, audience engagement, and release social and commercial value (OKF). *Open data* (OD) is any information that people are free to use, re-use, and redistribute - without any legal, technological, or social restriction (OKF). Unfortunately, data in their raw and original form could contain personal and sensitive information about individuals. Publishing such data violate individual privacy (GDPR, 2016). Thus, data providers should perform *Privacy-preserving data publishing* (PPDP) (Chen, 2009) to provide useful data without violating individuals' privacy. According to the PPDP principles, data publishers have tables containing *(Identifiers, Quasi-*

Identifiers, Sensitive Attributes, Non-Sensitive Attributes) where Identifiers (IDs) are a set of attributes that identifies record owners; Quasi-Identifier (QID) is a set of attributes that could potentially identify record owners; Sensitive Attributes are person-specific information, such as, diseases, salary, religion, political party (GDPR, 2016); while all the remaining attributes are Non-Sensitive.

Data owners, such as public administrations (PAs), health care, and financial institutions, may release the data they collect by de-identifying them, i.e., by masking, generalizing, or deleting IDs. However, even anonymized public information may be re-identified by exploiting other pieces of available data. A 2002 study found that 87% of the U.S. population can be identified using gender, birthdate, and zip code as QID by matching anonymized hospital visit records and voting lists (Sweeney, 2002). These data are not problematic if isolated but lead to the re-identification of individuals by exploiting additional information. The individuals whose data are re-identified risk of having their private information, such as their finances, health or preferences, and their identity, sold to organizations without user conses (Porter, 2008) or disclosed to undesired end-users (Porter, 2008), or even it can cause the refusal of an insurance provision (Sweeney, 2002).

The choice of the QID is an open question (Fung, 2010) since it depends on attributes that the attacker can exploit to link actual data to any external source.

In this article, we aim to prove that PAs current anonymization practices are not as waterproof as first perceived. We propose an approach to identify the *best* QID by counting the number of uniquely occurrences of a combination of values in a dataset. These unique occurrences will be referred to as *singletons*. We define the *best* QID as the *minimum* number of columns/attributes that leads to the disclosure of the *highest* number of singletons. By recalling that PPDP is interested in minimising both the information loss and the privacy leakage (Fung, 2010), we interpreted privacy leakage as the number of occurred singletons, and we estimate the information loss as the number of suppressed and modified rows, decreasing thereby the overall dataset quality. We focus on the following Research Questions (RQs):

- RQ1: What is the best QID according to our approach, and how many singletons does it disclose?
- RQ2: The generalization of which attribute or set of attributes among date of birth, ZIP and sex (one of the most common QID in the literature) leads to the minimum number of singletons while minimising the affected rows?

The main contributions of this article can be summarized as follows:

- We designed and implemented an algorithm that classifies the minimum set of attributes that exposes the greatest number of singletons as the best QID where singletons are the unique occurrence of a combination of values for a set of attributes. Its implementation is freely available on GitHub (https://github.com/isislab-unisa/qid_identifier_and_anonymizer);
- We observed that by generalizing only singletons detected by the QID (sex, ZIP, date_of_birth), we obtain nearly no singletons, and we just affected up to 2% of rows. The experiments have been performed on real open datasets released by the Italian Ministry of Infrastructure and Transport (freely available on GitHub https://github.com/isislab-unisa/driver-license-datasets).

The rest of the article is structured as follows: in Section 2, we overview alternative approaches proposed in the literature to detect QID and perform anonymization; in Section 3, we present our approach to detect privacy issues by the number of occurred singletons, and we propose an anonymization approach based on the generalization and suppression; in Section 4, we report and discuss the performed evaluation; then, we conclude with some final considerations and future research directions.

2. Related Work

In this section, we aim to overview the alternatives in detecting QIDs and corrective actions to sanitise a dataset. According to the PPDP principles, we focus on microdata, i.e., data records about individuals instead of data mining results.

2.1. Detect Privacy Leakage Algorithms

Braghin et al. (2016) defined an approach to detect QIDs by considering both single columns and their collections and counting the unique occurrence of values. All the QIDs are returned. Our approach works similarly, but it elects the *best* QID as the smallest combination of columns enabling the most extensive identification of singletons.

Motwani and Xu (2007) exploit the *separation* and the *distinct ratio* to quantitatively describe the ability of attributes to distinguish an individual from another. While the distinct ratio measures the percentage of distinct values, the separation ratio measures the proportion of tuple pairs that can be uniquely distinguished. We aim to find the QID by the separation ratio while keeping the distinct ratio as a metric that does not contribute to the best QID election.

2.2. Anonymization Approaches

The *anonymization* approaches explicitly removes IDs and hide the sensitive information assuming that the latter should not be used in data mining algorithms. ID removal might not be enough: it is still possible to recognise individuals by QIDs. To prevent linking attacks, the datasets must be *sanitized* (Samarati, 2001) by applying *anonymization operations* among generalization, suppression, anatomization, permutation, and perturbation (Federal committee of statistical methodology, 2005).

Among the most famous privacy models, we cite *k-Anonymity* (Ciriani, 2007) that is based on the fundamental concept that if a record has a particular value for a QID, at least other k-1 records will have the same value for that QID. Multiple versions of k-Anonymity have been proposed to overcome some of its limitations. For instance, (X-Y)-anonymity (Wang, 2006) face the case in which multiple rows of the dataset are related to the same individual; MultiRelational k-Anonymity (Nergiz, 2009) focuses on the anonymization of multiple relational tables; l-diversity (Machanavajjhala, 2007) guarantees that each equivalence class has at least *l* well-represented values for each sensitive attribute overcoming the risk of the homogeneity attack; (α, k)-anonymity (Wong, 2006) experiences local recording by reducing data distortion; t-closeness (Li, 2007) requires that the distance between the distribution of a sensitive attribute and the distribution of the attribute in the overall table should be no more than *t*.

Our approach is a modified version of k-Anonymity where:

- k is at least equal to 2;
- we discourage suppression in favour of generalization;
- k-Anonymity operates at a global level, we can also locally work.

3. Our Approach for Detecting and Solving Privacy Issues

Our approach is based on privacy issues detection, followed by an anonymization approach based on generalization and suppression. The workflow starts from the human provision of the dataset to test, and it automatically returns the best QID. If the best QID matches (year, municipality, gender), it also provides the corresponding generalization.

3.1. Detection of Privacy Issues

We interpreted the detection of privacy issues as the occurrence of unique values by considering a single column or a combination of them. These unique values are referred to as *singletons*. We elect IDs and QIDs by the number of occurred singletons. The implemented pseudo-code follows:

```
def detect_ID_and_QID(dataset):
    identifiers = []
    stats = {}
    for size in range(1, num_columns):
        # all the dataset column subsets of "subset_size"
        subsets = get_subsets(columns_to_check, size)
        # IDs: columns containing all distinct values
        temp_IDs, still_to_check = get_IDs(dataset, subsets)
        IDs += temp_IDs
        # for each subset, it stores the singletons number
        stats.update(get_stats(dataset, still_to_check))
        columns_to_check = list(set(still_to_check))
    # best QID: the smallest subset of columns exposing the greatest number of singletons
    best_QID = get_best_QID(stats)
```

In the best QID election, we first consider the number of singletons detected by each set of attributes. If more than one set of columns share the same number of singletons, we elect as *best* QID the minimum set of columns.

3.2. Anonymity Based on Generalization and Suppression

To ameliorate the detected privacy issues, we propose an anonymization approach based on generalisation and suppression. We suppress incomplete rows at the beginning of our technique; then, only the generalization is permitted. We favor the generalization to the suppression since we prefer to publish *incomplete* information rather than preventing the publication at all.

We focus on the QID (date_of_birth, ZIP, sex) since it is a well-known QID. In particular, we focus on a slightly simplified version of this QID, where we only have access to the year_of_birth. It is simplification without loss of generality since it can be easily generalized to the entire date. However, by only considering years, we can straightforwardly generalize dates by the mean value

of a range. Moreover, in Italy, there is a two-way correspondence between ZIP codes and Municipalities. Therefore, they can be used interchangeably. We aim to detect which column or combination of columns is *worth* to generalize to achieve the minimum number of singletons by modifying the minimum amount of rows.

For the numerical attribute (i.e., year_of_birth), we consider the standard approach of substituting values by intervals. Therefore, we sort rows by year_of_birth, split all the rows into groups of at least k values. If we split two rows containing the same year in creating groups, we iteratively merge those rows in the same group until the cut splits rows with different years. Each year is substituted with the interval [*min_year, max_year*). We also apply a second strategy where the average value of the interval replaces each interval. We hypothesize that if current years mainly correspond to the mean value of the range, fewer rows will be modified while still reducing the number of singletons.

About the sex column, we replace *male* and *female* values by *any gender*. In this case, the generalization plays the same role as cell suppression. About the municipality column, we exploit the hierarchy induced by our national administrative levels: municipalities are generalized by provinces, provinces by regions, regions by states. In our experiments, we only consider the first level of this hierarchy by generalizing municipalities by provinces. For categorical attributes (i.e., sex and municipality), we apply both a global and a local recording. While in the global recording we affect the entire dataset, at a local level, we only modify rows related to the singletons disclosed by the best QID. We hypothesize that the local recording can introduce a sufficient level of privacy protection while affecting a minimal number of rows (i.e., slightly decreasing the overall dataset quality). The implemented approach can be resumed as follows:

- the rows containing empty values are dropped out, and the removed rows alter the counter of affected rows. The full version of the dataset (i.e., only rows without any empty cell) is considered in the following steps;
- the following generalizations are performed:
  - all the municipalities are generalized by the corresponding province;
  - only the municipalities of the rows corresponding to the singletons detected by the best QID are generalized by the corresponding province;
  - all the values of the sex column are generalized by any gender;
  - only the values of the sex columns corresponding to the singletons detected by the best QID are generalized by any gender;
  - all the birth_years are generalized by the corresponding intervals;
  - all the birth_years are generalized by the mean value of the intervals generated as before;
  - all the attributes are generalized by combining every pair of the generalizations described so far and by generalizing all fields at once;
- for each performed generalization, we compute the number of singletons, the percentage of singletons, the number of distinct values, the number of modified and removed rows;
- we elect the best generalization by considering the one that achieves the minimum number of singletons while affecting the minimum number of rows.

4. Evaluation

We tested our approaches on real datasets released by the Italian Ministry of Infrastructure and Transport. These (anonymized) datasets contain information related to the driver licenses of all the Italian regions. We downloaded them in October, 2019 from the official site (http://dati.mit.gov.it/catalog/dataset/patenti). In January 2020, the Ministry updated the online version by significantly reducing the (already minimal) available content. The tested datasets (in their original form) are available on the GitHub (https://github.com/isislab-unisa/driver-license-datasets).

4.1. Detecting Privacy Issues

We selected only the columns related to personal information (i.e., the municipality and the province reported as the driver residence, the year of birth, and the sex), while we discarded all the Non-sensitive information, such as the driver's license details, and the numerical ID. The algorithm is linearly correlated to the dataset size. It takes 0.0152 seconds for processing datasets with 6M rows.

Results of the QID identifier. The results of our privacy issue detecting approach are available on GitHub, here omitted due to the lack of space. [Birth_year, ZIP, Sex] is reported as *best* QID in all the regions. Even if datasets are anonymized, our approach highlights the possibility of distinguishing up to 2% (1.93%) of singletons uniquely. If 2% seems to be a negligible amount of disclosed identities, it is worth noticing that the maximum number of disclosed singletons is more than 25K. It implies that removing IDs is not enough, and further anonymization actions must be performed to publish sanitized datasets. These results reply to **RQ1**.

4.2. Anonymization Approach Based on Generalization

We considered three datasets used in the previous analysis, heterogeneous in the disclosed percentage of singletons. Results are provided in Table 1.

Year range VS Year mean value. By comparing the year generalization by intervals (row **Y_ran** of Table 1) and by mean values (row **Y_avg** of Table 1), we observed that the mean value gains 95% fewer singletons while affecting 7% rows less than the range approach.

Global VS Local recording. By comparing the global generalization of the Municipality (**M** row of Table 1) and its local recording (**M_loc** row of Table 1), we observed the local recording achieves results close to 0, while the global recording succeeds in completely avoiding the disclosure of singletons. On the other side, the global recording affects the entire dataset, significantly decreasing the dataset quality, while the local recording affects only the 2% of the rows. We consider a good thread-off between privacy-preserving and data quality the generalization of the municipality (and the sex) only of singletons disclosed by the QID (year_of_birth, sex, ZIP) (**RQ2**).

Comparison with k-Anonymity. We used the *Valle d'Aosta* dataset to compare our approach and k-Anonymity (*https://dzone.com/articles/an-easy-way-to-privacy-protect-a-dataset-using-pyt*). We run k-Anonymity by generalizing Municipalities by their Provinces, Sex by any gender, and the Year by intervals of width 4. We consider the generalization of any set of attributes. We allowed suppression

equals to 0.01 in all cases, but in the sex, generalization is set to 0.05. At the global level, both our approach and k-Anonymity modify almost the entire dataset. While we obtain a generalized version of the dataset information, k-Anonymity removes many rows. When a small number of singletons occurs, k-Anonymity drops the corresponding rows. In all the other cases, it drops at least 200 rows and modifies all the other ones. Thanks to the local recording performed by our approach, we can obtain a minimum number of singletons (near to 0) while affecting a small portion of the dataset (up to 2%). Concluding, we achieve the same results in privacy-preserving, while our results lead to better data quality thanks to the local recording.

Table 1: ***Results of the anonymization approach****.* ***#S*** *is the number of singletons, and* ***%S*** *its percentage,* ***size*** *is the full dataset size,* ***DV*** *stands for Distinct values. The first row of the table reports the number of singletons disclosed by the best QID (year_of_birth, sex, ZIP). Then, we report the effect produced by our generalization algorithm obtained by affecting the columns reported in the first column of this table. The rows entitled with the suffix* ***loc*** *are related to a local recording approach, while the others correspond to a global one.* ***S*** *is related to the sex column;* ***M*** *to the Municipality column;* ***Y*** *to the Year_of_birth column. While* ***Y_ran*** *represents the generalization of years by range,* ***Y_avg*** *exploits range mean value.*

	Valle d'Aosta				**Molise**				**Umbria**			
	#S	**%S**	**Size**	**DV**	**#S**	**%S**	**Size**	**DV**	**#S**	**%S**	**Size**	**DV**
	1,684	1.93	87,464	9,174	869	1.30	597,243	13,391	2,569	0.15	198,312	16,628
Cols	**#S**	**%S**	**MR**	**DV**	**#S**	**%S**	**MR**	**DV**	**#S**	**%S**	**MR**	**DV**
S_loc	1,264	1.45	1,684	8,964	745	0.12	869	13,329	2,219	1.07	2,569	16,408
M_loc	4	~0	1,679	7,501	7	~0	860	12,539	7	~0	2,556	14,078
S, M_loc	1	~0	1,684	7,785	32	~0	860	12,680	3	~0	2,569	14,446
S	621	0.71	87,464	5,166	295	0.05	597,243	7,101	909	0.46	198,312	9,490
M	4	~0	87,464	167	127	~0	414,584	338	7	~0	198,312	324
Y_ran	198	0.002	87,464	2,739	1	~0	556,875	3,641	325	~0	198,312	4,806
Y_avg	9	0.001	81,476	607	1	~0	597,243	736	3	~0	198,312	1,089
S, M	1	~0	87,464	85	48	~0	597,243	171	7	~0	198,312	324
S, Y_ran	70	~0	87,464	1,442	0	0	597,243	1,895	116	~0	198,312	2,606
S, Y_avg	5	~0	87,464	308	0	0	597,243	368	1	~0	198,312	545
M, Y_ran	0	0	87,464	43	0	0	597,243	44	0	0	198,312	4
M, Y_avg	0	0	85,948	8,817	0	0	584,794	8	0	0	194,932	16
S,M,Y_ran	0	0	87,464	22	0	0	597,243	44	0	0	198,312	43
S,M,Y_avg	0	0	87,464	4	0	0	597,243	8	0	0	198,312	8

5. Conclusions and Future Work

In this article, we propose to detect QID by counting *singletons* in a dataset. We observed that the best QID (date_of_birth, sex, ZIP) discloses up to 2% (and up to 25K) of singletons in already anonymized datasets. When a privacy leakage is reported, PAs usually react by closing data or publish poorly informative datasets. As an example, the datasets we analyzed were substituted with a version with significantly lower informativeness, as only the province of residence, driving license category, and release date are provided. These datasets, in our opinion, are reduced to *pointless Open Data*. Instead of making data useless, we suggest investing in further sanitation actions. In this article, we observed that we achieve the minimum number of modified rows (up to 2% of affected rows) while obtaining the number of singletons close to 0 thanks to a local recording. We aim to provide our proposal as a framework to support PAs in publising datasets significantly more informative than the currently available on their websites while preserving citizens privacy.

References

Braghin, S., Gkoulalas-Divanis, A., Wurst, M. (2016) *Detecting quasi-identifiers in datasets* (US Patent 15 193 536)

Chen, B.C., Kifer, D., LeFevre, K., Machanavajjhala, A. (2009) *Privacy-preserving data publishing*. Foundations and Trends in Databases (2), pp. 1–167

Ciriani, V., di Vimercati, S.D.C., Foresti, S., Samarati, P. (2007) *k-anonymity*. In: Secure Data Management in Decentralized Systems, pp. 323–353

European Regulation 2016/679 of the European Parliament and of the Council (2016) *GDPR*, https://eur-lex.europa.eu/eli/reg/2016/679/oj, last access March, 2020

Federal Committee on Statistical Methodology (2005) *Statistical policy working paper 22*. Report on Statistical Disclosure Limitation Methodology

Fung, B.C.M., Wang, K., Chen, R., Yu, P.S. (2010) *Privacy-preserving data publishing: A survey of recent developments.* ACM Computing Surveys 42 (4), 14:1–14:53

Li, N., Li, T., Venkatasubramanian, S. (2007) *t-closeness: Privacy beyond k-anonymity and l-diversity*. In: IEEE 23rd Inter. Conf. on Data Engineering. pp. 106–115

Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M. (2007) *l-diversity: Privacy beyond k-anonymity*. ACM Trans. Knowledge Discovery Data 1 (1), pp. 3

Motwani, R., Xu, Y. (2007) *Efficient algorithms for masking and finding quasi-identifiers*. In: Proceedings of the Conference on Very Large Data Bases. pp. 83–93

Nergiz, M.E., Clifton, C., Nergiz, A.E. (2009) *Multirelational k-anonymity.* IEEE Trans-action on Knowledge and Data Engineering 21 (8), pp. 1104–1117

OKF: *Open data*, https://okfn.org/opendata/, [last access May, 2020]

Porter, C.C. (2008) *De-identified data and third party data mining: the risk of re-identification of personal information*. Journal of Law, Commerce Technology (5), pp. 1

Samarati, P. (2001) *Protecting respondents' identities in microdata release*. IEEE Trans-actions on Knowledge and Data Engineering 13 (6), pp. 1010–1027

Sweeney, L. (2002) *Achieving k-anonymity privacy protection using generalization and suppression.* International Journal Uncertainty Fuzziness Knowledge -Based System 10 (5), pp. 571–588

Wang, K., Fung, B.C.M. (2006) *Anonymizing sequential releases*. In: 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. pp. 414–423

Wong, R.C.W., Li, J., Fu, A.W.C., Wang, K. (2006) *(a,k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing.* In: ACM SIGKDD on Knowledge Discovery and Data Mining. pp. 754-759.

About the Authors

Matteo Pastore

Matteo Pastore was Born in Salerno in 1998. He will graduate at the University of Salerno in 2020. He will continue his studies with a Master's degree in Computer Science. He is interested in Cloud Computing, Cybersecurity, Open Data, Networks.

Maria Angela Pellegrino

Maria Angela Pellegrino, born in 1994, graduated in Computer Science at Master Level in 2018 at the University of Salerno. From 2018, she is a Ph.D. student in Computer Science at the University of Salerno under the supervision of Professor Vittorio Scarano. Her studies focus on how to improve the quality of (Linked) Open Data while preserving the privacy of individuals. Research interests include data quality, privacy, Open Data, and Semantic Web.

Vittorio Scarano

Vittorio Scarano is a Full Professor of Computer Science at the University of Salerno (Italy). Since 1996, he (with Alberto Negro) funded and co-directs the ISISLab laboratory within the Department. ISISLab has been hosting, until now, the research activity of 20 PhD students, more than 20 collaborators (grants, fellowships) and provided support for more than 120 theses (Bachelor and Master level) with computing facilities such that a 38-nodes IBM cluster, file and application servers, teaching lab, a stereoscopic screen, augmented reality devices etc.He is co-author of more than 140 papers in internationally refereed journals and conferences of IEEE, ACM, etc. and he has been the PhD supervisor of several PhD students of the Computer Science PhD program at the University of Salerno. In 2000, he has been awarded (with his co-authors) the ”Best poster award" at the 9th World Wide Web Conference (WWW9). In 2008, he has been awarded by IBM with the International IBM Jazz Innovation Award as a grant of 25.000$. He coordinated the European funded research H2020 project ROUTE-TO-PA "Raising Open and User-friendly Transparency-Enabling Technologies for Public Administrations" (grant agreement No 645860) with 12 partners and a budget above 3M€. He has participated and coordinated local units in European, national, and regional funded research and innovation projects. Research interests include Distributed Systems, Collaborative Systems and Open Data, and Enhanced (Virtual/Augmented) Reality.