ISBN 978-981-11-0008-6
Proceedings of 2016 6th International Workshop on Computer Science and Engineering (WCSE 2016)
Tokyo, Japan, 17-19 June, 2016, pp. 594-598
doi: 10.18178/wcse.2016.06.102

ODD Visualizer: Scalable Open Data De-identification Visualizer

Chiun-How Kao [1], Chih-Hung Hsieh [1+], Chien-Lung Hsu [2], Yu-Feng Chu [1] and Yu-Ting Kuang [1]

[1] Institute for Information Industry, Taipei, Taiwan

[2] Department of Information Management, Chang-Gung University, Taoyuan, Taiwan

Abstract. Due to the significant values it can derive, large-scaled open data analysis (or big data analysis) attracts lots of attentions from various domains researchers and experts. However, the progresses of data releasing for open usages are still slow in the latest decade. Only about 10% amount of datasets owned by worldwide governments have been released, and the main reason of that is due to concern for "privacy preserving'. According to previous real case studies, even though the personally identifiable information have been de-identified, sensitive personal information still could be uncovered by heterogeneous or cross-domain data joining operation. This kind of privacy re-identification are usually too complicated or obscure to be realized by data owner, not to mention that this problem will be more severe as the scale of data goes large. To our best knowledge so far, none of existent research work leverages data visualization approach to provide direct and clear manner detecting information re-identification problem. In this project, we aim to propose a method for scalable open data de-identification visualization consisting of: 1) platform for scalable storing and computation for de-identification measuring and 2) novel data visualization technique depicting distribution of de-identification robustness in a global view. It was demonstrated that our work not only provides efficient estimation and visualization for data de-identification but also a useful guideline helping users determine which parts of data should be released or not.

Keywords: data de-identification, data visualization, privacy preserving, personally identifiable information, sensitive personal information.

1. Introduction

In recent decades, for the sake of rising of big data analytics, open data applications create almost 3,000-billion market size cross various domains [1]. Governments and companies compete to pay more attention on heterogeneous or cross-domain data analysis year by year. As reported by statistics from World Wide Web Foundation and Open Data Institute, there are totally 77 countries dedicating opening governments' data [2][3]. Yet, only 10% amount of datasets owned by worldwide governments have been released, and the main reason of that is due to concern for "privacy preserving' such that there are still plenty space for improving [2][3].

For problem of Personally Identifiable Information (PII) or Sensitive Personal Information (SPI) re-identification, most of previous studies usually adopt two strategy to de-identify data which owner want to release: 1) generalization [4][5] and 2) suppression [4][6][7]. Both of them sacrifice information and corresponding value, that dataset originally can provide, to some extent. Besides, it has been shown that even after dataset being de-identified, there still chances that PII may be revealed by cross joining different datasets [8][9]. To quantify the likelihood where PII or SPI being re-identified, the *k*-anonymity model was proposed and can be used to measure how well the upcoming released data being de-identified [5]. It was also known that time complexity to calculate the optimal *k*-anonymity value for a given dataset is Non-deterministic Polynomial-time Hard (NP-hard) level [10] thus researchers devoted themselves to develop scalable *k*-anonymity computation platform [11][12][13]. However, when data scale goes larger to big-data

+ Corresponding author. Tel.: + 886-2-6607-2076 fax: +886-2-6607-2026.
E-mail address: chhsieh@iii.org.tw.

level, it is not only need a scalable approach to estimate *k*-anonymity of large-scaled data but also an efficient approach to provide user a direct and clear picture telling the robustness for data against PII re-identification. To our best knowledge so far, there are no existent research focusing on making de-identification degree of sensitive data being easily readable and understandable by human being.

In this paper, the effectiveness of data visualization analysis, which is one of rising data mining topics and proven to be useful on various domains [14][15], was leveraged to build a scalable Open Data De-identification Visualizer (ODD Visualizer). The ODD Visualizer is inspired and modified from Matrix Visualization (MV) approach[16], as well as benefited from a novel Hierarchical Analysis and Clustering Tree (HACT). The variant of matrix visualization depicts a brief distribution among different *k*-anonymity estimation given different attribute subsets of data, while HACT considers this *k*-anonymity distribution as input to reason out an optimal alignment of attributes giving users the most robust attribute subset against PII re-identification, under various de-identification thresholds. The merits of the proposed ODD Visualizer are threefold. 1) A scalable database and computation platform were implemented in ODD Visualizer such that *k*-anonymity of each attribute subset of data can be rapidly and efficiently estimated. 2) Users can easily get a whole picture depicting the *k*-anonymity distribution among different attribute subset combination. Hence, it can be known where the weakness of current dataset against PII re-identification is. 3) Based on the optimal alignment sorted by HACT, users get suggestion to decide which attribute subsets can be released or not.

Details about the definition of PII re-identification and *k*-anonymity model discussed in this paper can be found in Section 2. The architecture and implementation of proposed ODD Visualizer are all mentioned in Section 3. Sections 4 is responsible for demonstration effectiveness of our proposed method using benchmark of employee information. At last, section 5 concludes this paper as well as gives some possible directions for further researches.

2. Problem Definition-Record Linkage Attack and *k*-Anonymity Model

As what being told in related surveys and researches [17][18], for any given dataset table ,***T***, and one record, ***D***, attributes = {V_1, V_2, …} contained in ***D*** can be partitioned into different groups according their roles during progress of de-identification or re-identification, as followings.

D (**Explicit Identifier**, **Quasi Identifier**, **Sensitive Attributes**, **Non-Sensitive Attributes**), where:

1) **Explicit Identifier** (**EID**): is a set of attributes, such as name and social security number (SSN), containing information that explicitly identifies record owners.
2) **Quasi Identifier** (**QID**): is a set of attributes that could potentially identify record owners. Note that the value of **QID** in this study is represented as ***qid***.
3) **Sensitive Attributes** (**SA**): Sensitive Attributes consists of sensitive personal-specific information such as disease, salary, and disability status.
4) **Non-Sensitive Attributes** (**NSA**): Non-Sensitive Attributes contains all attributes that do not fall into above three categories.

The discussed record linkage attack and *k*-anonymity model can be illustrated using following example as shown in table 1 [17]. Given that a hospital is going to release patient's information (as patient table in table 1(a)). Once a PII hacker having privilege to access another public external dataset (as external table in table 1(b)), and the hacker also know that instances in patient table and external table came from the same population. In the above condition, hacker then do have chance to identify the **SA** of "Diagnosis" via **QID** = {Job, Gender, Age}. For example, privacy leakage crisis is on Chris, cause he is the only one whose value of **QID** (named as ***qid***) being {Salesman, Male, 31}. Using such ***qid*** value to link two tables, the sensitive information that Chris's diagnosis result will leak out. On the contrary, PII for Bob and Kevin is much safer, as they share the same ***qid*** of {Military, Male, 27}. As a consequence, the *k*-anonymity model can be used to measure the likelihood of this sensitive information leakage.

k*-anonymity model**: for a given ***T, assume ***qid*** is one existent value of one possible **QID** combination. For any ***qid*** of each **QID** existing in ***T***, if there are at least *k* records sharing the same ***qid***, then such ***T*** satisfying this requirement is called *k*-anonymous. The probability of linking a victim to a specific record

through **QID** is at most 1/*k*.

Table 1. Examples for illustrating record linkage re-identification and *k*-anonymity model

(a) Patient table

Job	Gender	Age	Diagnosis
Engineer	Male	32	Hepatitis
Engineer	Female	25	Flu
Military	Male	27	Diabetes
Military	Male	35	Hepatitis
Salesman	Male	31	HIV
Artists	Female	28	Flu
Artists	Female	35	Diabetes

(b) External table

Name	Job	Gender	Age
Smith	Engineer	Male	32
Bob	Military	Male	27
Louis	Artists	Female	35
Chris	Salesman	Male	31
John	Military	Male	35
Amy	Artists	Female	28
Nancy	Engineer	Female	25
Eddie	Engineer	Male	32
Kevin	Military	Male	27

3. Proposed Method

The proposed ODD Visualizer is designed under two concerns when dealing dataset de-identification evaluation. One is for scalable de-identification measuring, and the other is to make the measuring result being direct, clear and easily understandable. The Elasticsearch [19] database and corresponding API of version 1.7.2 provides scalable data storing, loading, and querying functionalities to implement *k*-anonymity computation in ODD Visualizer. On the other hand, the following two subsections discuss the two major components in the ODD Visualizer: 1) matrix visualization and 2) hierarchical analysis and clustering tree.

3.1. Matrix visualization for depicting *k*-anonymity distribution

The matrix visualization (MV) technique used in this paper, focusing on depicting a brief distribution among different *k*-anonymity estimation given combinations of any two attributes, as well as a measurement of robustness against PII re-identification in terms of *k*-anonymity value for whole dataset. Figure 1(a) shows the MV in the proposed ODD Visualizer. The color spectrum in the bottom of figure 1(a) indicates different *k*-anonymity values from lowest to highest. For matrix itself, the row and column indexes lists the attributes (i.e., V_1, V_2, …) constituting input dataset, and the color shown in position (i, j) represents the *k*-anonymity values for **QID** = $\{V_i, V_j\}$. The diagonal position stands for only single variable. Moreover, the color of matrix borderline is for *k*-anonymity value of entire dataset considering all attributes. It should be note that the attributes imported into the matrix are only **QID**, **SA**, and **NSA**. **EID** is excluded from MV, because **EID** always produce *k*-anonymity of 1 due to its definition. All *k*-anonymity computation use scalable Elasticsearch API functionalities.

3.2. Hierarchical Analysis and Clustering Tree for sorting and grouping attributes

This paper also proposes a novel Hierarchical Analysis and Clustering Tree (HACT) to analysis the distribution after MV. As in the figure 1(b), The HACT first sorts all of *k*-anonymity values for each combination of two variables and uses greedy strategy to merge two variables (attributes) who result in maximum *k*-anonymity value. The two selected variables, V_i and V_j, will form a new cluster and be treated as one "complex" variable, $\{V_i, V_j\}$. The clustering operation will iterate until all attributes merged as one tree. The "uncle flipping" mechanism [20] is then adopted to determine the attribute order inside each clustered subtree. Because HACT keeps merging, aligning, and grouping variables to show which subsets of data having most robustness against re-identification. After HACT process, given any *k*-anonymity threshold = δ, users get suggestion and guideline to select which attribute subsets can be released or not based on the optimal alignment sorted by HACT (figure 1(c)). Followings are the pseudo code of HACT process:

Step 1. Excluding diagonal, pick one position (i, j) in MV who produces largest *k*-anonymity value.

Step 2. Merge the two picked variables, V_i and V_j, as new one "complex" variable, $\{V_i, V_j\}$, and form a new clustering tree.

Step 3. Replace index V_i and V_j, with $\{V_i, V_j\}$, and update the corresponding *k*-anonymity values related to V_i and V_j.

Step 4. Go back to Step 1，until all subtree being merged.

Step 5. Alignment all clustered variables based on "uncle flipping" order determining mechanism.

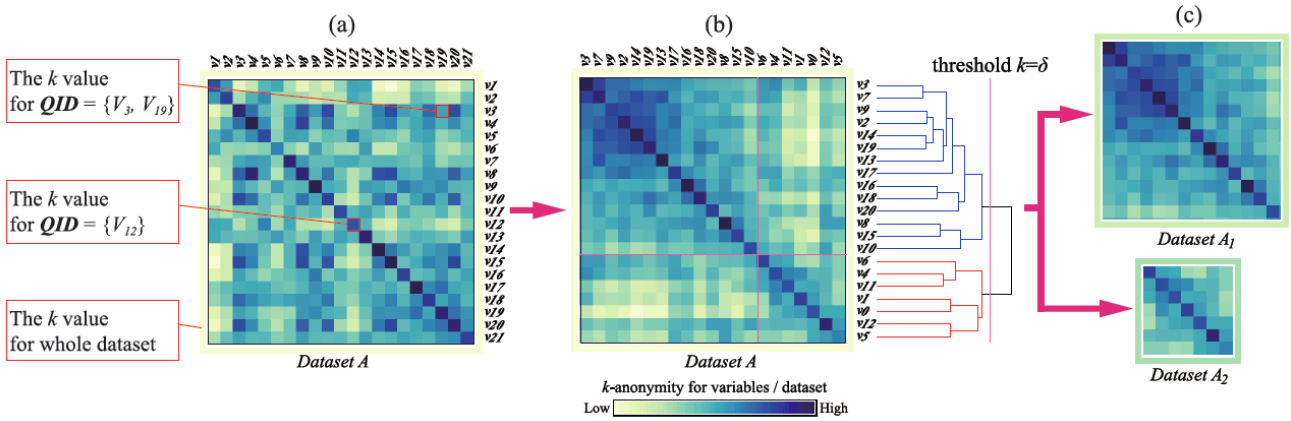
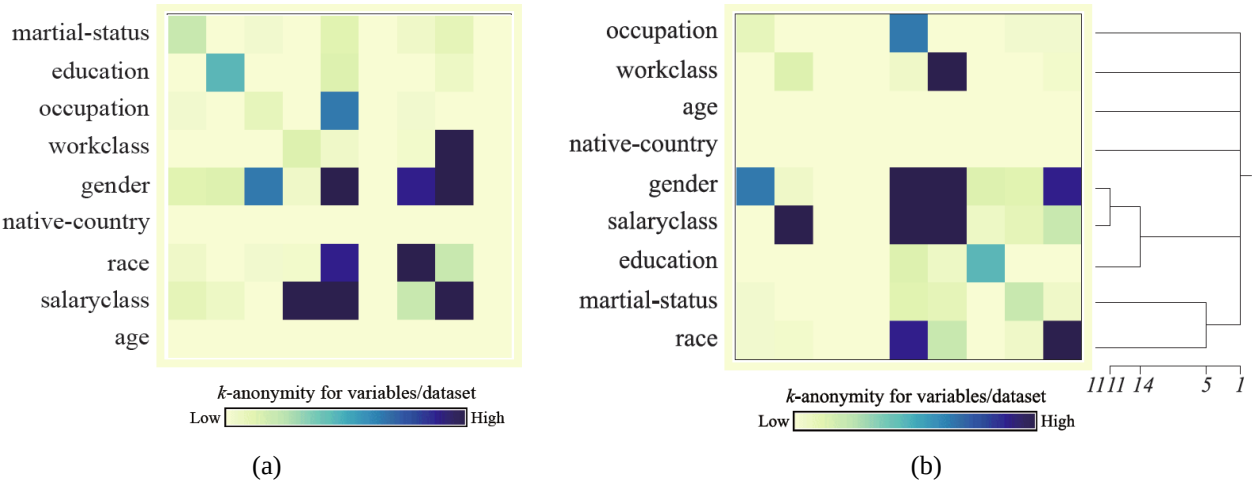


Fig. 1: The modified matrix visualization to display *k*-anonymity distribution.

4. Demonstration

For demonstration the effectiveness of the proposed ODD Visualizer, an employee dataset from [13] was download as benchmark. This data contains totally 10 attributes, one of them is employee ID, used as **EID**, was excluded without importing into ODD Visualizer. And the results of using ODD Visualizer to display, to analysis, and to group *k*-anonymity values of this benchmark are all shown in Figure 2. Figure 2(a) displays the *k*-anonymity values not only about the whole employee dataset as well as considering its single attribute or attributes combination. For instance, attributes: gender, race, and salary class have the highest *k*-anonymity values, while native-country has the lowest one. Combination of {salaryclass, gender} provides highest *k*-anonymity value, however PII may be re-identified via combination of {salaryclass, age} whose corresponding *k*-anonymity value is too low. It also can be observed that the original matrix visualization was shown in a multi-fragment manner. After HACT processing, figure 2(b) shows the aligned matrix result where attributes are grouped into several blocks. The resulted cluster representing {gender, salaryclass, education martial-status, race} has more resistance against re-identification than other attributes. The clustering tree in the right side of figure 2(b) indicates the order for those block being merged. Moreover, the tree also reflect different de-identification robustness after each merging from *k* =1111 to *k* =1. This example successfully demonstrate that the users can take advantage of this visualization as a guideline deciding to share which parts of dataset is appropriate or not.



(a) (b)

Fig. 2: A demonstration for how ODD Visualizer works.

5. Conclusion

Open data analysis creates huge amount revenue for various domain and its related application. However, the current progresses of worldwide data releasing is still too slow to catch up the growing applications

because of the “privacy preserving” concern. To our best knowledge so far, this paper first propose a scalable open data de-identification measuring and visualization mechanism, named as ODD Visualizer. The proposed method depicts the distribution of *k*-anonymity values among various dataset attributes in a direct and clear manner. Besides, it also provides a useful guideline to help user decide which parts of dataset can be released. Other various kinds of anonymity models and their corresponding specific visualization way are the future works for further research.

6. References

[1] “Open data: Unlocking innovation and performance with liquid information | McKinsey & Company,” 06-Aug-2015.[Online].Available: http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information. [Accessed: 06-Aug-2015].

[2] “New Report Highlights Successes and Challenges of Worldwide Open Data Policies,” *TechPresident*, 06-Aug-2015. [Online]. Available: http://techpresident.com/news/wegov/24480/new-report-highlights-successes-and-challenges-worldwide-open-data-policies. [Accessed: 06-Aug-2015].

[3] “PM speech at Open Government Partnership 2013 - Speeches - GOV.UK,” 06-Aug-2015. [Online]. Available: https://www.gov.uk/government/speeches/pm-speech-at-open-government-partnership-2013. [Accessed: 06-Aug-2015].

[4] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Incognito: Efficient Full-domain K-anonymity,” in *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, 2005, pp. 49–60.

[5] L. Sweeney, “k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY,” *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 05, pp. 557–570, Oct. 2002.

[6] P. Samarati, “Protecting respondents identities in microdata release,” *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 6, pp. 1010–1027, Nov. 2001.

[7] L. Sweeney, “ACHIEVING k-ANONYMITY PRIVACY PROTECTION USING GENERALIZATION AND SUPPRESSION,” *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 05, pp. 571–588, Oct. 2002.

[8] L. Sweeney, “Simple Demographics Often Identify People Uniquely,” 2000. .

[9] “Harvard Professor Re-Identifies Anonymous Volunteers In DNA Study,” *Forbes*. [Online]. Available: http://www.forbes.com/sites/adamtanner/2013/04/25/harvard-professor-re-identifies-anonymous-volunteers-in-dna-study/. [Accessed: 08-Mar-2016].

[10] A. Meyerson and R. Williams, “On the Complexity of Optimal K-anonymity,” in *Proceedings of the Twenty-third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, New York, NY, USA, 2004, pp. 223–228.

[11] F. Liu, X. Shu, D. Yao, and A. R. Butt, “Privacy-Preserving Scanning of Big Content for Sensitive Data Exposure with MapReduce,” in *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*, New York, NY, USA, 2015, pp. 195–206.

[12] X. Zhang, L. T. Yang, C. Liu, and J. Chen, “A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 2, pp. 363–373, Feb. 2014.

[13] F. Prasser and F. Kohlmayer, “Putting Statistical Disclosure Control into Practice: The ARX Data Anonymization Tool,” in *Medical Data Privacy Handbook*, A. Gkoulalas-Divanis and G. Loukides, Eds. Springer International Publishing, 2015, pp. 111–148.

[14] Y.-J. Tien, Y.-S. Lee, H.-M. Wu, and C.-H. Chen, “Methods for simultaneously identifying coherent local clusters with smooth global patterns in gene expression profiles,” *BMC Bioinformatics*, vol. 9, p. 155, 2008.

[15] J.-K. Chou and C.-K. Yang, “PaperVis: Literature Review Made Easy,” *Comput. Graph. Forum*, vol. 30, no. 3, pp. 721–730, Jun. 2011.

[16] H.-M. Wu, S. Tzeng, and C. Chen, “Matrix Visualization,” in *Handbook of Data Visualization*, Springer Berlin Heidelberg, 2008, pp. 681–708.

[17] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, “Privacy-preserving Data Publishing: A Survey of Recent Developments,” *ACM Comput Surv*, vol. 42, no. 4, pp. 14:1–14:53, Jun. 2010.

[18] B. L, B.-S. K, P. Al, and A. H, “The ‘GeneTrustee’: a universal identification system that ensures privacy and confidentiality for human genetic databases.,” *J. Law Med.*, vol. 10, no. 4, pp. 506–513, May 2003.

[19] “Elasticsearch.” [Online]. Available: https://www.elastic.co/products/elasticsearch. [Accessed: 10-Mar-2016].

[20] H.-M. Wu, Y.-J. Tien, and C. Chen, “GAP: A graphical environment for matrix visualization and cluster analysis,” *Comput. Stat. Data Anal.*, vol. 54, no. 3, pp. 767–778, 2010.