Clustering Heuristics for Efficient *t*-closeness Anonymisation

Anne V.D.M. Kayem(✉) and Christoph Meinel

Faculty of Digital Engineering, Hasso-Plattner-Institute for Digital Engineering
GmbH, University of Potsdam, Prof.-Dr.-Helmert Str. 2-3, 14440 Potsdam, Germany
anne@mykayem.org

Abstract. Anonymisation based on *t*-closeness is a privacy-preserving method of publishing micro-data that is safe from skewness, and similarity attacks. The *t*-closeness privacy requirement for publishing micro-data requires that the distance between the distribution of a sensitive attribute in an equivalence class, and the distribution of sensitive attributes in the whole micro-data set, be no greater than a threshold value of t. An equivalence class is a set records that are similar with respect to certain identifying attributes (quasi-identifiers), and a micro-data set is said to be *t*-close when all such equivalence classes satisfy *t*-closeness. However, the *t*-closeness anonymisation problem is NP-Hard. As a performance efficient alternative, we propose a *t*-clustering algorithm with an average time complexity of $O(m^2 \log n)$ where n and m are the number of tuples and attributes, respectively. We address privacy disclosures by using heuristics based on noise additions to distort the anonymised datasets, while minimising information loss. Our experiments indicate that our proposed algorithm is time efficient and practically scalable.

Keywords: Anonymisation · *t*-closeness · Privacy · Performance

1 Introduction

Published data, such as medical and crime data facilitates data analytics for improved service delivery. Such data releases must be protected to prevent sensitive personal information disclosures due to privacy subversion attacks. There are two categories of privacy subversion attacks, namely, identity and attribute disclosures. Identity disclosures, occur when a person can be uniquely linked to a specific data item in the released dataset. While attribute disclosures occur when the released dataset is combined with other publicly released data sources to uniquely identify individuals based on specific attribute values. Frequently, such vulnerabilities can be exploited to trigger a chain reaction of privacy disclosure attacks. For example, Machanavajjhala et al. [1,2] demonstrated that in a released medical data set, knowledge of the fact that heart attacks are rare among the Japanese could be used to reduce the range of sensitive attributes required to infer a patient's disease.

D. Benslimane et al. (Eds.): DEXA 2017, Part II, LNCS 10439, pp. 27–34, 2017.
DOI: 10.1007/978-3-319-64471-4_3

In t-closeness anonymisation, data privacy is ensured by building on previous anonymisation algorithms namely, k-anonymisation [4–7], and l-diversity [1,8,9] to maximise data utility and at the same time protecting against skewness and similarity attacks [3]. t-closeness addresses both attacks by taking into account global background knowledge as well as possible disclosure levels, based on the distribution of sensitive attributes in both the equivalence classes and the entire dataset. However, as Liang and Yuan [22] have shown, the t-closeness problem is NP-Hard. Having a more efficient solution is practical for real world applications involving large datasets, and low-powered, low-processing devices. Application scenarios emerge on lossy networks and the Internet-of-things (such as opportunistic, and Fog computing networks), where personal data is collected over several devices, may need to be anonymised before it is transferred to a forwarding device.

In this paper, we propose a performance efficient algorithm based on clustering as a classification heuristic to ensure that the distance between sensitive attributes and the cluster centroid is no more than a threshold value of t. The degree of similarity between a cluster and a sensitive attribute is computed by using a combination of severity rankings (cost to privacy due to attribute exposure), the Jaccard coefficient for categorical attributes, and a Euclidean distance for numerical attributes in the quasi-identifier. A high degree of similarity, is captured by a smaller distance from the cluster centroid, and the reverse is true for a low similarity degree. Using these criteria minimises the amount of information that an observer can infer from the published data, based on background knowledge. Our proposed algorithm has an average time complexity of $O(m^2 \log n)$ where n and m are the number of tuples and attributes, respectively.

The rest of the paper is structured as follows. In Sect. 2 we present related work on the general topic of syntactic anonymisation approaches. We proceed in Sect. 3 with a description of our proposed approach to clustering supported t-closeness anonymisation. In Sect. 4, we discuss the performance complexity of our proposed t-closeness clustering algorithm. We follow this in Sect. 5, with some experimental results based on the UCI Adult dataset. In Sect. 6, we offer concluding remarks.

2 Related Work

Sweeney's work [4] on sharing personal data without revealing sensitive information, by k-anonymising the data prior to publication, has triggered a plethora of algorithms aimed at circumventing deanonymisation attacks while at the same time ensuring that the data remains usable for operations such as querying [1–3,5–24]. Privacy preserving data publishing algorithms can basically be classified into two categories namely, syntactic and semantic approaches. Syntactic approaches work well with both categorical and numerical data, and have a well defined data output format. This property allows for confirmation of privacy traits of the data by visual inspections. Adversarial models for constructing deanonymisation attacks are based for the most part on generally available information and inferences drawn from the syntactic and semantic meaning of the underlying data. Examples of algorithms that fall under this category include, k-anonymity [4], l-diversity [1], and t-closeness [3].

The t-closeness algorithm was proposed to alleviate vulnerabilities to skewness and similarity attacks [3] to which both k-anonymisation and l-diversity algorithms are vulnerable. In t-closeness the idea is to structure the anonymised dataset to ensure that the distance between the distribution of a sensitive attribute in a given equivalence class, and the distribution of sensitive attributes in the entire dataset is no more than a threshold value of t [3]. This approach to anonymisation overcomes the limitations of l-diversity in preventing attribute disclosure, and those of k-anonymisation by preventing inference of sensitive attributes, in addition to protecting against background attacks. However, t-closeness anonymisation is performance intensive in the average performance case, and as is the case with k-anonymisation [5,11,19], and l-diversity [20], achieving optimal t-closeness is an NP-Hard problem [22]. Using heuristics, is one method of obtaining near-optimal results.

Anonymisation by clustering has been studied as an approach to improving the performance of k-anonymisation by alleviating the cost of information loss [23,24]. The idea behind these clustering schemes is to cluster quasi-identifiers in equivalence classes of size k, and to avoid using generalisation hierarchies when this impacts negatively on information loss. This property of clustering lends itself well to t-closeness anonymisation as an approach to alleviating the performance demands of anonymising large datasets, particularly when this is done on low-powered, low-processing devices. In the next section we describe our proposed clustering algorithm.

3 t-closeness Clustering

Before we discuss our t-clustering algorithm we first consider aspects such as information loss and sensitive attribute severity weightings which are important in achieving a tradeoff between data utility and privacy. In order to determine information loss, we use a generalisation hierarchy denoted $T(a)$, where $T(a)_{max}$ is the root node or maximum numerical value for an attribute, and $T(a)_{min}$ a leaf node or minimum numerical value. In $T(a)$, P is the set of parent nodes, $T(a)_p$ is the subtree rooted at node $p \in P$, and $T(a)_{tot,p}$ is total number of leaf nodes in the subtree rooted at node $p \in P$. We handle NULL values by classifying them as categorical values.

To calculate information loss for categorical attributes, we consider the proportion of leaf nodes that are transformed to the parent node in the the subtree rooted at p in comparison to the total number of parent nodes P in $T(a)$ excluding the root node. Information loss as $IL(a)$ for categorical attributes is computed with

$$IL(a) = \frac{T(a)_{tot,p} - 1}{P - 1}.$$

and for numerical attributes

$$IL(a) = \frac{T(a)_{max,p} - T(a)_{min,p}}{T(a)_{max} - T(a)_{min}}$$

is used to compare the loss incurred within the subtree in which the value falls, to maximum and minimum values both in the subtree and the entire hierarchy, $T(a)$. Finally, we express the combined information loss over both categorical and numerical attributes for the entire dataset is computed using $IL_{tot} = \sum_{t \in D} \sum_{a \in A} IL_(a)$.

We introduce a severity weighting scheme to determine the level of loss of privacy due to classifying a tuple with a given sensitive attribute in one cluster over another. For example, a severe illness like "stomach cancer" carries a higher risk of privacy loss than "flu". We denote the sensitive attribute severity weight as $S(s)$ where $s \in S(a)$ and $S(\cdot)$ maps the sensitive attribute to its weight. In this case, the weight is a guideline for the duration, severity of the illness, and/or the likelihood of stigmatisation in the case of exposure. For instance, on a scale of $1 - 10$, S(Cancer) = 10, while S(Allergy) = 4.

In order to cluster data to ensure t-closeness anonymity with clustering, it is important to determine the minimum size of a cluster required to guarantee a global minimum level of t-closeness that all clusters must adhere to. The clustering algorithm uses a value k_{min} as the minimum cluster size and moves tuples into appropriate clusters, based on both the severity weighting and the distance from the cluster centroid. We define k_{min} as follows: $k_{min} = Max(k_{cons}, min(S_D(\cdot)))$ where k_{cons} is a pre-defined minimum cluster size and $S_D(\cdot))$ represents the set of all sensitive attribute severities for D.

Based on the cluster size, we must determine which tuples to either include or exclude from a cluster. As a first step, we use the relative distance between tuples to decide which tuples to classify in the same cluster. The inter-tuple distance is computed based on both categorical and numerical attributes. The distance between categorical attributes is measured using the Jaccard's coefficient [17], as a similarity measure that is easy to interpret and works well for large datasets with a proportionately small number of NULL or missing values. We define the Jaccard coefficient for our t-clustering algorithm using

$$sim_{t_i,t_j} = \frac{Q_{t_i} \cap Q_{t_j}}{Q_{t_i} \cup Q_{t_j}}$$

where Q_{t_i} and Q_{t_i} are the quasi-identifiers for t_i and t_j, respectively. t_j is the centroid of the cluster that t_i is classified in. The value of sim_{t_i,t_j} varies between 0 and 1, 1 indicates a strong similarity between the tuples, and 0 a strong dissimilarity, based on the quasi-identifier attributes.

To reduce the rate of information loss due to tuple suppressions, we also compute the Euclidean distance between numerical attributes with an n-dimensional space function which is represented as follows:

$$Dist(t_i, t_j) = \sqrt{\left(t_i\,(a_1) - t_j\,(a_1))^2 + ... + (t_i\,(a_m) - t_j\,(a_m))^2\right)}$$

where a_i is an attribute in Q. Tuples separated by a small Euclidean distance are classified in the same cluster.

Next we consider the sensitive attribute severity weightings and compute the average severity weighting AS_D for D as well as the average severity weighting AS_e for e for a given cluster (equivalence class). The AS_e serves to evaluate the distribution of sensitive attributes in e, while AS_D does this for the entire dataset D, which is similar to how t-closeness decides on tuple classifications based on statistical distributions of sensitive attributes, and also to prevent skewness as well as similarity attacks. The AS_D is used to start the anonymisation process and is computed as follows:

$$AS_D = \frac{\sum S_{t_i}(a)}{\|D\|}$$

where $S_{t_i}(a)$ is the severity weight of sensitive attribute $a \in t_i$. A high AS_D indicates a high level of diversity in the entire dataset. In a similar manner, we compute AS_e as follows:

$$AS_e = \frac{\sum S_{t_i}(a)}{\|e\|}$$

In line with using the t parameter in the t-closeness scheme as a method of optimising dataset utility, we evaluate the level of loss of privacy with respect to information loss in forming the clusters, using a fitness function that is expressed as follows:

$$t = \frac{1}{Max\,(AS_D, IL_{tot})}.$$

Expressing the fitness function in this way captures the fact that when t is low a high degree of loss of either privacy or information is likely to occur, while a high value indicates a good balance between privacy and data utility.

Finally, the Kullback-Leibler distance between AS_e and AS_D is used to determine the level of diversity of sensitive attributes in e with respect to D. Using the Kullback-Leibler distance ($\sum AS_e \log \frac{AS_e}{AS_D}$) serves as an entropy-based measure to quantify the distribution of sensitive attributes both in e and D; and is computed as follows: $Dist\,(AS_e, AS_D) = \sum AS_e \log \frac{AS_e}{AS_D} \leq t$ where $\sum AS_e \log \frac{AS_e}{AS_D} = H(AS_e) - H(AS_e, AS_D)$ such that $H(AS_e) = \sum AS_e \log AS_e$ is the entropy of AS_e and $H(AS_e, AS_D)$ is the cross entropy of AS_e as well as AS_D. When $Dist\,(AS_e, AS_D) \leq t$ the anonymised dataset mimics t-closeness by ensuring that sensitive attributes are classified according to severity of exposure. When $Dist\,(AS_e, AS_D) \nleq t$, we must rerun the whole algorithm to re-compute cluster structures to ensure privacy. In the next section we provide a complexity analysis of the average case running time for our proposed scheme.

4 Complexity Analysis

We know from Liang and Yuan's work [22] that the t-closeness anonymization problem is NP-Hard. With respect to t-clustering, we know that clustering problems are in general NP-Hard. However, with our heuristics we are able to drop the performance cost to $O(n^2 \log m)$ where n and m represent the tuples and attributes in the dataset D. We achieve this by dividing up D into at most n clusters, computations required for classification are in $O(n)$ and the fraction of attributes that are critical for classification are in $O(\log m)$, which results in a total time complexity of $O(n^2 \log m)$.

5 Experiments and Results

In this section we present some results of experiments that we conducted to evaluate the performance of our proposed t-clustering anonymisation scheme. We applied our proposed scheme to the Adult Database from the UCI Machine Learning Repository [25]. We modified the table to include 12 attributes namely: *Age, Race, Gender, Salary, Marital Status, Occupation, Education, Employer, Number of Years of Education, Workclass, Relationship, Native Country.* We included 3 quasi-identifiers, and 2 sensitive attributes. From the base original table (45222 tuples), we extracted dataset sizes to experiment with, and randomly generated an additional 20000 tuples to observe the behaviour of the proposed scheme on larger dataset sizes. With respect to the anonymisation process, we used the following parameters - cluster size: 2, 3, 9, 10, 11, 15, 16, 17, 18; maximum suppression allowed: 0%, 1%; $0.017 \leq t \leq 0.2$. From the table above, we observe that the Kullback-Leibler distance ($Dist\,(AS_e, AS_D)$) between the severity weightings both within the clusters and the dataset are relatively low which indicates a high level of privacy in terms of protection against background knowledge attacks such as skewness and similarity attacks. In this way our proposed scheme inherits the privacy properties of the t-closeness anonymisation algorithm. In terms of performance of our proposed scheme, in line with the theoretical performance discussed in Sect. 4, we note that the time required for clustering grows linearly with the size of the dataset. Finally, the percentage information loss falls between 9% and 25% depending on the number of clusters formed, the cluster size, and the dataset size. Lower information loss percentages occur when smaller and more clusters are formed for a dataset and the reverse happens when larger clusters are formed. The trade-off however, is that smaller clusters result in a higher risk of privacy loss while larger clusters reduce the privacy risk (Table 1).

Table 1. Classification time with respect to dataset size

Dataset size	Cluster size	$Dist\,(AS_e, AS_D)$	Time (ms)
30000	9	0,0015	74
35000	17	0,00117	83
40000	16	0,0015	92
45000	16	0,00035	90
50000	16	0,002	98

6 Conclusion

In this paper we presented a clustering scheme to alleviate the performance cost of t-closeness anonymisation. Basically, what we do is to rank sensitive attributes by a severity weighting and classify tuples to minimise the risk of privacy disclosure of tuples containing high severity weight sensitive attributes. Clustering has the advantage of reducing the need for extensive attribute generalisation in order to classify tuples based on similarity. This is good, in addition, because it reduces the cost of information loss. As we have mentioned earlier, high levels of information loss make datasets unusable in practical situations. By considering severity weightings both for individual clusters and the entire dataset, we mimic the t-closeness principle, of seeking to distribute tuples in ways that ensure that the difference in distributions both within the equivalence classes and the entire dataset, does not surpass a threshold value of t. In this way, our proposed scheme also offers protection against skewness and similarity attacks. Finally, a further benefit of our scheme is that because it is not performance intensive, it can be used on low-powered, low-processing networks for guaranteeing privacy of data under data forwarding schemes.

References

1. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-diversity: privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data **1**(1), 1–52 (2007). Article 3
2. Kifer, D., Machanavajjhala, A.: No free lunch in data privacy. In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, SIGMOD 2011, pp. 193–204. ACM, New York (2011)
3. Li, N., Li, T., Venkitasubramaniam, S.: t-closeness: privacy beyond k-anonymity and l-diversity. In: Proceedings of the 23rd International Conference on Data Engineering, pp. 106–115 (2007)
4. Sweeney, L.: K-anonymity: a model for protecting privacy. Int. J. Uncertainty Fuzziness Knowl. Based Syst. **10**(5), 557–570 (2002)
5. Aggarwal, C.: On k-anonymity and the curse of dimensionality. In: Proceedings of the 31st International Conference on Very Large Databases, VLDB 2005, pp. 901–909. VLDB Endowment (2005)
6. Bayardo, R.J., Agrawal, R.: Data privacy through optimal k-anonymization. In: Proceedings of the 21st International Conference on Data Engineering, ICDE 2005, pp. 217–228. IEEE (2005)
7. Liu, K., Giannella, C., Kargupta, H.: A survey of attack techniques on privacy-preserving data perturbation methods. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy-Preserving Data Mining. Advances in Database Systems, vol. 34, pp. 359–381. Springer, Boston (2008). doi:10.1007/978-0-387-70992-5_15
8. Shmueli, E., Tassa, T.: Privacy by diversity in sequential releases of databases. Inf. Sci. **298**, 344–372 (2015)
9. Xiao, X., Yi, K., Tao., Y.: The hardness of approximation algorithms for l-diversity. In: Proceedings of the 13th International Conference on Extending Database Technology, EDBT 2010, pp. 135–146. ACM, New York (2010)

10. Iyengar, V.S.: Transforming data to satisfy privacy constraints. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2002, pp. 279–288. ACM, New York (2002)
11. Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., Zhu, A.: Anonymizing tables. In: Eiter, T., Libkin, L. (eds.) ICDT 2005. LNCS, vol. 3363, pp. 246–258. Springer, Heidelberg (2004). doi:10.1007/978-3-540-30570-5_17
12. Ciriani, V., Tassa, T., De Capitani Di Vimercati, S., Foresti, S., Samarati, P.: Privacy by diversity in sequential releases of databases. Inf. Sci. **298**, 344–372 (2015)
13. Aggarwal, C.C.: On unifying privacy and uncertain data models. In: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, ICDE 2008, pp. 386–395. IEEE, Washingtion, D.C. (2008)
14. Aggarwal, C.C., Yu, P.S.: Privacy-Preserving Data Mining: Models and Algorithms, 1st edn. Springer Publishing Company Incorporated, New York (2008)
15. Lin, J.-L., Wei, M.-C.: Genetic algorithm-based clustering approach for k-anonymization. Expert Syst. Appl. **36**(6), 9784–9792 (2009)
16. Shmueli, E., Tassa, T., Wasserstein, R., Shapira, B., Rokach, L.: Limiting disclosure of sensitive data in sequential releases of databases. Inf. Sci. **191**, 98–127 (2012)
17. Aggarwal, C.C.: Data Mining: The Textbook. Springer, Cham (2015)
18. Xiao, Q., Reiter, K., Zhang, Y.: Mitigating storage side channels using statistical privacy mechanisms. In: Proceedings of 22nd ACM SIGSAC Conference on Computer Communications Security, CCS 2015, pp. 1582–1594. ACM, New York (2015)
19. Meyerson, A., Williams, R.: On the complexity of optimal k-anonymity. In: Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems, PODS 2004, pp. 223–228. ACM, New York (2004)
20. Dondi, R., Mauri, G., Zoppis, I.: On the complexity of the l-diversity problem. In: Murlak, F., Sankowski, P. (eds.) MFCS 2011. LNCS, vol. 6907, pp. 266–277. Springer, Heidelberg (2011). doi:10.1007/978-3-642-22993-0_26
21. Ciglic, M., Eder, J., Koncilia, C.: k-anonymity of microdata with NULL values. In: Decker, H., Lhotská, L., Link, S., Spies, M., Wagner, R.R. (eds.) DEXA 2014. LNCS, vol. 8644, pp. 328–342. Springer, Cham (2014). doi:10.1007/978-3-319-10073-9_27
22. Liang, H., Yuan, H.: On the complexity of t-closeness anonymization and related problems. In: Meng, W., Feng, L., Bressan, S., Winiwarter, W., Song, W. (eds.) DASFAA 2013. LNCS, vol. 7825, pp. 331–345. Springer, Heidelberg (2013). doi:10.1007/978-3-642-37487-6_26
23. Kabir, M.E., Wang, H., Bertino, E., Chi, Y.: Systematic clustering method for l-diversity model. In: Proceedings of the Twenty-First Australasian Conference on Database Technologies, ADC 2010, Brisbane, Australia, vol. 104, pp. 93–102 (2010)
24. Aggarwal, G., Panigrahy, R., Feder, T., Thomas, D., Kenthapadi, K., Khuller, S., Zhu, A.: Achieving anonymity via clustering. ACM Trans. Algorithms **6**(3), 1–19 (2010). ACM, New York
25. Frank, A., Asuncion, A.: UCI machine learning repository (2010). http://archive.ics.uci.edu/ml