Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 63 (2015) 348 – 355

Procedia Computer Science

The 5th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2015)

Systematic Literature Review on the Anonymization of High Dimensional Streaming Datasets for Health Data Sharing

Benjamin Eze[a], Liam Peyton[a]*

[a]*University of Ottawa, Ottawa, ON, K1N 6N5, Canada*

Abstract

One of the biggest challenges to health data sharing is regulations that prohibit the transmission and distribution of Personal Health Information (PHI) even among collaborating organizations. This impedes research and reduces the utility of these datasets. Anonymization can address this issue by hiding PHI while maintaining the analytical utility of the data. Much research has focused on data that is static, independent and complete. Unfortunately, this is not typical of health data. Instead of static, independent tables, health data is in relational databases with multiple high-dimensional tables that are transactional and constantly changing. Data recipients usually receive multiple versions of the database over time. This study reviews literature on anonymization methodologies for large and fast changing high-dimensional datasets, especially health data. Relevant papers are analyzed, categorized and compared in terms of scope, and contributions. Finally, we used the extracted details from our analysis to outline possible research direction for developing a realistic anonymization framework for health data sharing.

Keywords: privacy; personal health information; high-dimensional datasets; streaming databases; data sharing

1. Introduction

Anonymization of datasets requires a thorough understanding of various risk models that simulate the behaviours of de-anonymization adversaries[1]. The complexity of high-dimensional data makes it a particularly hard problem in the data anonymization space. Applicable anonymization methodologies for such datasets are as complex as the

* Corresponding author. Tel.: +1-613-562-5800 x2122; fax: +1-613-562-5664.
E-mail address: beze080@uottawa.ca, lpeyton@uottawa.ca

doi:10.1016/j.procs.2015.08.353

various attack scenarios they represent. Therefore the appropriate anonymization model needs to adapt to the diverse nature of these datasets. This is especially important for health data.

Most anonymization techniques[2–4] fail with high-dimensional health data because they try to group actors so they look similar to an adversary trying to re-identify them[5]. Exact equivalence that is required for such grouping results in excessive information loss, virtually wiping off the analytical utility of the dataset[3]. Another issue that comes up often is associated with the state of the dataset. Data sharing in a streaming model[6] assumes that health data being shared is constantly changing. Privacy protection needs to be consistent even within such streaming models. You cannot suppress the attribute of an actor in one stream and reveal it with another.

We conducted a systematic literature review to investigate, document and analyze research on anonymizing high dimensional streaming datasets, particularly as it relates to health data. The methodologies used for querying, filtering and selecting the relevant academic papers from various academic publication databases are documented. The selected papers are analyzed against their relevance to our defined research question. The results of this analysis are then documented and discussed. Information extracted from these papers is analyzed, categorized and compared in terms of scope, and contributions. Finally, the extracted details from our analysis are used to outline possible research directions for developing a realistic anonymization framework for health data sharing.

2. Background

For health data, Personal Health Information (PHI) attributes are categorized under one of three categories: Direct Identifier attributes singularly identify an individual in the dataset. These attributes include names, and given identifiers like social insurance numbers, social security numbers and driver license identifiers[7]. Quasi Identifier attributes, on their own, cannot identify an individual[7]. However, when quasi-identifiers (QI) are combined, they behave like direct identifiers. Most re-identification efforts link QI values to publicly available data repositories to re-identify individuals in an anonymized dataset. Finally, Sensitive Attributes are not usually public data but are sensitive if associated with an individual. Procedure and drug codes, as well as disease conditions are some of the popular sensitive attributes in the healthcare domain.

In this study, we only consider research efforts aimed at anonymizing quasi-identifiers and sensitive attributes. The target of anonymization is to ensure that both identifying and sensitive attributes are protected from a de-anonymization adversary. There are three types of disclosures: identity, attribute and membership disclosures. Identity disclosure occurs when the record of an individual in a dataset is re-identifiable. It is also called record linkage since a record in a dataset can be linked to the individual[8]. Attribute disclosure occurs when new information can be gained on sensitive attributes by an attacker or adversary. It is important to note that attribute disclosure is often a consequence of identity disclosure. Membership disclosure on the other hand is a probabilistic measure of the presence or absence of an individual in a dataset. This knowledge changes the behavior of an adversary towards de-anonymization. Both identity and attribute disclosures are important in this study.

Various research efforts are geared at creating privacy models to ensure proper anonymization of one or more target datasets. *k*-Anonymity is the most popular measure of anonymity. Introduced by Samarati and Sweeney[7] in 1998, it ensures that each record in a dataset has at least *k*-1 indistinguishable records. *k*-Anonymity protects against identity disclosure but cannot guarantee protection against attribute disclosure of sensitive attributes[10,11]. A re-identification adversary can discover the values of sensitive attributes when their diversity is low after anonymization. Basically, even when an equivalence class satisfies *k*-anonymity, it may not satisfy *l*-diversity for the sensitive attributes if they are not diverse enough[10,11]. *l*-diversity protects against sensitive attribute disclosure by requiring every equivalence class to have at least *l* well-represented values for each sensitive attribute. *t*-Closeness[10,12] takes this a little further by requiring that the distribution of these sensitive attributes in each equivalence class be close to the distribution of these attributes in the entire dataset. Table linkage represents those scenarios where the adversary is able to infer the presence or absence of an individual's record in a dataset or table. *d*-Presence is the privacy model that protects against such adversarial knowledge. *e*-Differential privacy is the privacy model that focuses on how an adversary would change the probabilistic belief on the sensitivity information of a victim after accessing the published anonymized data.

These four basic models form the basis for most anonymization algorithms and frameworks. Generalization is the most popular anonymization technique[10,13] for satisfying *k*-Anonymity. However, real datasets tend to produce very unique groups, making most of the equivalence classes or quasi-identifier groups non *k*-Anonymous. Anonymization processors usually require a generalization tree or generalization hierarchies[13] for each quasi-identifier. These hierarchies then get combined to form a generalization lattice. It is usually necessary to search this lattice for an optimal generalization set for all quasi-identifiers that satisfy *k*-Anonymity and sensitive attributes satisfying *l*-Diversity, while providing the least information loss and the greatest data utility. This process of finding the optimal solution from a generalization lattice is NP Hard. There are many algorithms for searching the lattice more efficiently for an optimal solution, notably among them are Samarati's k-minimal generalization algorithm[7], Sweeney's Datafly[14], Kirsten's Incognito[15], El Emam's Optimal Lattice Anonymization (OLA)[2], and Flash from Kohlmayer et al.[4].

Most generalization algorithms use global recoding of attribute values. That means the same transformation is applied to each quasi-identifier value. However, there are efforts to develop algorithms that perform this action locally for each equivalence class through local recoding[16,17]. Local recoding increases data utility but also drastically increases the complexity of the generalization process. Data suppression is used to strike the balance between utility and availability. Equivalence classes whose sizes are less than *k* need to be suppressed to keep the entire dataset *k*-Anonymous. Suppression can be done at record or cell level. Record level suppression simply deletes those equivalence classes that are not *k*-Anonymous. Unfortunately, suppressing those records could be overly excessive. Cell level suppression on the other hand determines those QI values that make a tuple identifiable and removes them. When used appropriately, it provides the least information loss for anonymization. There are algorithms that use perturbation to supplement generalization instead of suppressing identifiable attributes. Usually applied to sensitive attributes, perturbation techniques either shuffle at-risk QI values and sensitive attributes or simply replace them with fake masks of the original.

Finally, another major challenge to anonymization comes from the wide variety of datasets out there. Most healthcare databases do not contain single, independent tables. Instead, they are normalized transactional databases with patient identifying and sensitive attributes scattered in multiple tables. Data streams also present potential challenges to anonymization. For example, EMR and Patient Management (PM) databases see patient data growing with each visit to the hospital or clinic. Guo et al.[18] and Cao et al.[19] show that one of the most important characteristics of data streams is the rate of change, which is usually high, and in turn affects the statistical characteristics of the dataset. Periodic releases of anonymized data can potentially reveal identifiable as well as sensitive attributes over time, even though each release is individually anonymous.

3. Problem Description

For effective health data sharing, anonymization must support both the complexity of health data and frequency of change in complex transactional systems. This systematic literature review tries to answer the question – are there anonymization frameworks for high-dimensional streaming health data?

3.1. Research Method

In this paper we followed the guidelines for a systematic literature review in software engineering[20,21]. Our study is summarized in three stages: planning the review, conducting the review, and reporting the review. In planning the review, we explain how we formulated the search keywords. Conducting the review delves into details of where and how we searched for the keywords, and how the results were obtained and filtered. Finally, we present the results of the review by classifying and comparing our gathered knowledge.

3.2. Search Terms and Research Databases

Our search was for academic papers that present or discuss algorithms, methodology or application frameworks for anonymizing high-dimensional datasets. The search also includes work done on streaming datasets or incremental data releases where data is continuously published to the data recipients.

The following terms are considered: anonymization plus synonyms like de-identification and antonyms like re-identification; high-dimensional datasets plus synonyms like streams, streaming, longitudinal events and relational databases; anonymization techniques like generalization, suppression, aggregation, and clustering. We were interested in papers that present models, algorithms and application frameworks. We also required that *k*-Anonymity or *k*-Anonymization be mentioned somewhere in the content. These keywords were used to formulate advanced search queries that were run against some of the most popular academic databases: Google Scholar, Springer Link, Science Direct, PubMed, ACM Digital Library and IEEE Explore.

3.3. Exclusion Criteria

Papers that focus on these domains are excluded: anonymization techniques that solely apply to cross-sectional datasets with the exception of those that investigate clustering for sensitive attributes as well streaming databases or data streams; unstructured data anonymization; geo-location or positioning systems or trajectory based system; biomedical anonymization especially DNA database anonymization.

4. Conducting the review

In conducting the review, we started off constructing advanced queries that create the right filters for each of the academic databases based on search terms in section 3.2. Our final search queries, the electronic repositories of scholar studies, and the number of returned results are summarized above in Table 1. Because of space limitation, we are only able to show the search queries for Science Direct and Google Scholar. However, the other databases were still considered in our analysis.

Table 1. Advanced Search Queries

Query	Database	#Results	Selected
(*anonymiz* AND *identification) AND (stream* OR relational OR (high AND multi AND dimensional) OR longitudinal) AND (generalization OR suppression OR aggregation OR clustering OR cluster) AND (model OR algorithm OR framework)	Science Direct	592	18
(anonymization AND *identification AND privacy) AND ("*k*-anonymization" or "*k*-anonymity" OR "*k*-anonymous") (relational OR datebase OR (high AND multi AND dimensional) OR longitudinal) AND (generalization OR suppression OR aggregation OR cluster*) -dna –gis	Google Scholar	582	21

Initial review of the search results was done by reviewing the title, abstract, keywords and sometimes conclusions to determine if a paper meets the criteria described in section 3.3. For those selected, we then filtered them through the exclusion criteria before making the final selection for each database or repository. Finally, duplicate selections were excluded. At the end of this process, we ended up with 26 papers for data extraction.

Based on the extracted data from each paper, we categorized the papers into three categories to cover our research domains:

CAT1 – Improving high-dimensional data anonymization or datasets with many sensitive attributes.

CAT2 – Improving cross-sectional data anonymization with many quasi-identifier and sensitive attributes.

CAT3 – Streaming database anonymization.

We then did a quality appraisal in terms of contributions with respect to the following areas:

Q1 - Attribute categorization for high-dimensional or streaming datasets.

Q2 - Adversary and privacy Models.

Q3 - Anonymization techniques for high-dimensional datasets.

Q4 - Anonymization techniques for streaming datasets.

We scored the contribution of each paper on a scale of 3. “0” – No contribution, “1” – Some contribution, “2” – Major contribution to the research space. The total score shows how relevant the research is to our study.

5. Data Extraction

In this section we explain how we decided what pieces of information to extract from each paper under study. The result of the extraction process is summarized in Table 2. In summary, only 3 of the papers present research work on high-dimensional, relational databases usually associated with health data. Most studies focus on cross-sectional datasets with multiple sensitive attributes. There are 9 papers in this category and additional 5 that did not meet all our inclusion criteria. For streaming data anonymization, 6 papers were reviewed and an additional 2 also did not meet our inclusion criteria.

Table 2. Selected Papers for Data Extraction

Cat	Paper	Q_1	Q_2	Q_3	Q_4	Total
CAT 2	Abdalaal, et al. (2013)[22]	2	2	1	2	**7**
CAT2	Aggarwal, C. C. (2008)[23]	1	2	2	0	7
CAT 3	Byun, et al. (2006)[24]	2	2	1	2	9
CAT2	El Emam & Dankar (2008)[25]	1	2	1	0	4
CAT3	Fung, et al. (2012)[26]	2	2	0	1	6
CAT2	Gal, et al. (2014)[27]	1	2	2	2	7
CAT1	Ghinita, et al. (2008)[28]	2	2	0	2	6
CAT2	Gkoulalas-Divanis, et al. (2014)[29]	2	2	2	1	8
CAT2	Jändel (2014)[30]	1	2	1	1	5
CAT3	Kim, et al. (2014)[31]	1	1	2	2	8
CAT3	Mohammed, et al (2010)[32]	2	2	1	1	6
CAT2	Mortazavi, et al (2014)[33]	1	1	2	1	5
CAT1	Narayanan, et al (2008)[34]	1	2	2	1	6
CAT1	Nergiz, et al (2009)[35]	0	1	2	2	4
CAT2	Nergiz, et al (2014)[36]	1	2	2	2	7
CAT3	Pei, et al (2007)[37]	0	1	0	1	4
CAT2	Vatsalan, et al (2013)[38]	2	2	1	1	4
CAT3	Yingjie, et al (2009)[39]	1	2	2	2	7

6. Reporting the review

In this section we provide a summary of the contributions and important remarks from the selected papers according to our four quality areas in section 5.

6.1. Attribute categorization for high-dimensional or streaming datasets

The curse of dimensionality is a problem associated with algorithms processing high-dimensional data. In anonymization, it occurs because there are many sensitive attributes making most anonymization algorithms computationally expensive with unacceptable levels of information loss. Aggarwal et al.[23] show the effect of this curse with *k*-Anonymity, *l*-Diversity and other anonymization techniques using various theoretical models. Abdalaal et al.[22] show that little study has been done to anonymize multiple sensitive attributes such as those in opinion polls. Data Mashup techniques like those from Byun et al.[26] require data from multiple providers to be aggregated before anonymization. This process results in an integrated dataset with too many sensitive attributes, making anonymization impossible as a result because of the curse of high dimensionality[23]. These datasets therefore become very vulnerable to attribute linkage attacks. Based on these studies, we conclude that a successful high-dimensional health data anonymization framework needs to be able to handle multiple quasi-identifier and sensitive attributes.

6.2. Adversary and Privacy Models

As pointed out by Jandel[30], modeling the adversary knowledge correctly provides good protection against de-anonymization. Nergiz and Gok[36] identify three classes of adversaries: 1) Classical adversaries that tries to discover sensitive attributes associated with an individual because of the knowledge of the quasi-identifier attributes. 2) Statistical adversaries that analyse the statistical distribution of the original and the anonymized datasets and discover perturbations to the data from variations in the distributions. 3) Algorithm-aware adversaries that know the algorithm used for anonymization and can reverse-engineer the process.

For datasets with lots of quasi-identifiers, it is not realistic for the adversary to have knowledge of all the QI values. The LKC-privacy model proposed by Mohammed et al.[32] bounds the numbers of QI values known to an adversary to at most L values. The general intuition of this algorithm is to ensure that all combinations of L QI values will satisfy *k*-anonymity constraints and that the confidence of inferring any sensitive values in a sensitive attribute is not greater than a constant C. Unfortunately, combinatorics suffer from the curse of dimensionality[23].

Byun et al.[24] pointed out that anonymized datasets are vulnerable to inference attacks when viewed collectively. The research further shows that the timeline between data releases could help an adversary re-identify some patients by analyzing the deltas between each batch and using those to pick-up sensitive attribute on patients from statically anonymized datasets. It is not uncommon for data publishers to anonymize and publish a sample of the original dataset. El Emam et al.[25] show that for such scenarios, re-identification risk is dependent on the sample size. If there is no sampling, the adversary is certain that observed patients are in the published dataset and the maximum re-identification risk is the same as *1/k*. This is referred to as Prosecutor risk. However, if the data is sampled, an adversary only needs to show that there is at least an individual that can be re-identified in the anonymized dataset to discredit the publisher. This was the case with the 2006 Netflix Prize data[34] breach. El Emam et al.[25] call this Journalist risk.

Finally, the systematic literature survey carried by Gkoulalas-Divanis et al.[29] reviewed over 45 algorithms. While only a few of those focus on attribute disclosure, the paper points out that it is not clear if any of those algorithms would work in the healthcare setting because of the increasing complexity of health data.

6.3. Anonymization Techniques for high-dimensional datasets

Abdalaal et al. MSA-diversity[22] is useful for statistical learning. It borrows from Hilbert curve anonymization[40,41] to increase utility of the anonymized dataset. Hilbert curve maps multi-dimensional quasi-identifier attributes to a flattened single dimension. It subsequently sorts the records according to their QI values and constructs sensitive attribute distinct groups of at least *l* tuples. There is need for further investigations on how this algorithm could be adapted to achieve *l*-diversity for datasets in relational form. Algorithms like MiRaCle[35] and Fast Data-Oriented Micro-aggregation (FDM)[33] use clustering techniques that transform records into vector spaces. These vectors are subsequently processed to create *k*-Anonymous cluster groups. Unfortunately, clustering still suffers from high information loss with many quasi-identifiers[23,34] and fails to protect against *l*-diversity.

Gal et al.[27] pointed out that implementing *k*-anonymity, *l*-diversity and *t*-Closeness in the same dataset leads to competing requirements, and subsequently considerable information loss through over generalization, perturbation and suppression. Their micro-aggregation algorithm targets numerical quasi-identifier attributes only in creating *k*-Anonymous clusters while replacing sensitive attributes with masked values. SHARE introduced by Gkoulalas et al.[29] tries to work around these competing requirements by creating two projections, one for the quasi-identifiers and the other for each sensitive attribute. It then identifies a group membership attribute that specifies the association between records in both projections. The problem is that it assumes only one sensitive attribute per dataset. It is also not known if these projections work with multiple sensitive attributes and quasi-identifiers.

One of the important characteristics of high-dimensional datasets is sparseness especially with patients with long profiles. Correlation-aware Anonymization of High-dimensional Data (CAHD) introduced by Ghinita et al.[28] targets relational datasets and take advantage of data sparseness to explore correlation among items. Correlated attributes form anonymous groups, drastically reducing the number of quasi-identifiers and sensitive attributes being considered for anonymization. Unfortunately, the sensitive attributes are anonymized independently usually by seeding masked values.

Narayanan and Shmatikov[34] identify the need for a proper anonymization strategy for sparse high-dimensional dataset like the 2006 Netflix Prize data showing user movie ratings. Their de-anonymization algorithm Scoreboard-RH adversarial attack tool exploits the facts that most records in a sparse database have little or no similarities. It also sees anonymization changes to a dataset as small "errors" or adversarial impression introduced, otherwise the utility of the dataset would be completely lost. This work also showed that Netflix Prize data was highly re-identifiable because the sampling was not very random. They were successful in identifying Netflix records for known users and subsequently learnt about their political preferences and other potentially sensitive data in their profiles. This research shows that without proper anonymization, high-dimensional datasets are very vulnerable to de-anonymization.

6.4. Anonymization Techniques for streaming datasets

Byun et al.[24] show that with streaming anonymization, the dataset is continuously growing. So anonymization using static means is vulnerable to inference attacks. This paper introduces an anonymization algorithm that inserts new records to anonymized dataset while ensuring data quality and protection against inference attacks. Byun et al.[24], and Kim et al.[31] propose an accumulation-based method for data stream anonymization. The frameworks pass each tuple through an aggregation engine that assigns each tuple to a cluster based on the information loss weight of its quasi-identifier. Every new tuple is moved to the cluster that results in the least information loss. Based on the cluster size, a tuple is either released immediately or kept for release at a later time. Clustering only considers the quasi-identifiers. For the sensitive attributes, they are released only after passing an *l*-diversity pre-condition. If they don't pass, perturbation takes place by seeding some randomly generated attributes to the original.

Fung et al.[26] tackles the problem of aggregating and anonymizing data mashups in a Service Oriented Architecture (SOA). This paper considers two approaches to integrating data. In the first approach, data is first mashed up before generalization (mashup-then-generalize) but the provider holding the mashup has access to sensitive attributes from other providers. The second approach is to generalize before mashup (generalize-then-mashup) but this approach fails to guarantee anonymity when quasi-identifier values span multiple tables.

7. Conclusions and Future Work

Based on the papers reviewed, existing algorithms and frameworks touch various aspects of this problem but to claim full anonymization, a framework needs to touch all these aspects in one privacy model. Our findings show that: 1) It is probably impossible for an adversary to know the values of all quasi-identifiers in most high-dimensional datasets. Therefore, adversarial knowledge needs to be bound as shown in LKC privacy[32]. Privacy models should account for these limits. 2) Clustering and micro-aggregation are good techniques for anonymizing high-dimensional datasets. However, they result in excessive information loss when there are too many quasi-identifiers. They also fail to address *l*-Diversity on their own. 3) With high-dimensionality, there is sparseness and this aligns naturally to correlation. With correlation, as shown in Ghinita et al. algorithm[28], one can reduce the number of quasi-identifiers and sensitive attributes that gets fed into an anonymization framework. 4) Aggarwal et al.[23] show that multi-dimensional data in their natural forms suffer from dimensionality curse. Anonymization algorithms are prone to this curse when used in their natural forms for high-dimensional data. One of the recommended solutions is to transform this data to fewer and lower dimensions before anonymization. Algorithms like Hilbert curve [22] can be used for carrying out such transformations. 5) There is little work being done on suppression and local recoding despite their high utility. This is probably because of their complexity but would require further investigations. 6) There are good approaches for anonymizing streaming cross-sectional datasets that need be evaluated on larger datasets. One of the most important contributions from these papers [9,24,37,39] is that anonymization algorithms should use an accumulation based approach to support streaming data. This way, each data release process learns from transformations done in previous releases. 7) Finally, anonymization for health data sharing cannot be solved by technology alone. There is need to investigate regulatory frameworks that can complement technology to achieve the right mix of solution that allows health care organization to share data more efficiently and effectively.

References

1. Tamersoy A, Loukides G, Nergiz ME, Saygin Y, Malin B. Anonymization of longitudinal electronic medical records. IEEE Trans Inf Technol Biomed. 16(3), 2012. pp 413-423
2. El Emam K, Dankar FK, Issa R, et al. A Globally Optimal *k*-Anonymity Method for the De-Identification of Health Data. J Am Med Informatics Assoc. 16(5), 2009. pp 670-682
3. Fung BCM, Wang K, Chen R, Yu PS. Privacy-preserving data publishing. ACM Comput Surv. 42(4), 2010 pp 1-53.
4. Kohlmayer F, Prasser F, Eckert C, Kemper A, Kuhn KA. Flash: Efficient, stable and optimal *k*-anonymity. In: Proceedings ASE/IEEE International Conference on Privacy, Security, Risk and Trust 2012.
5. Hay M, Miklau G, Jensen D, Towsley D, Li C. Resisting structural re-identification in anonymized social networks. VLDB J. 19(6), 2010. pp 797-823

6. Eze B, Kuziemsky C, Peyton L, Middleton G, Mouttham A. Policy-based data integration for e-health monitoring processes in a B2B environment: Experiences from Canada. J Theor Appl Electron Commer Res. 5(1), 2010. pp 56-70.
7. Samarati P, Sweeney L. Protecting Privacy when Disclosing Information: *k*-Anonymity and its Enforcement Through Generalization and Suppresion. Proc IEEE Symp Res Secur Priv. 1998. pp 384-393
8. Fung BCM, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: A Survey of Recent Developments. CSUR. 42(4), 2010. pp 1-53.
9. Kim S, Sung MK, Chung YD. A framework to preserve the privacy of electronic health data streams. J Biomed Inform. 50, 2014. pp:95-106
10. Ninghui L, Tiancheng L, Venkatasubramanian S. t-Closeness: Privacy beyond *k*-anonymity and ℓ-diversity. Proc - Int Conf Data Eng. 2, 2007. pp 106-115
11. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. ℓ-Diversity: Privacy beyond *k*-anonymity. In: Proceedings - International Conference on Data Engineering.Vol 2006.
12. Kohlmayer F, Prasser F, Eckert C, Kuhn K a. A flexible approach to distributed data anonymization. J Biomed Inform. 50, 2013. pp 62-76
13. Sweeney, L. Achieving k-anonymity privacy protection using generalization and suppression. Int J Uncertainty, Fuzziness Knowledge-Based Syst, 10(05), (2002) pp 571-588
14. Sweeney, L. k-anonymity: A model for protecting privacy. Int J Uncertainty, Fuzziness Knowledge-Based Syst, 10(05), (2002) pp 557-570
15. LeFevre K, DeWitt DJDJ, Ramakrishnan R. Incognito: efficient full-domain *k*-anonymity. In: SIGMOD '05: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. 2005. pp 49-60
16. Li J, Wong RCW, Fu AWC, Pei J. Anonymization by local recoding in data with attribute hierarchical taxonomies. IEEE Trans Knowl Data Eng 20(9), 2008. pp:1181-1194
17. Xu J, Wang W, Pei J, Wang X, Shi B, Fu AW-C. Utility-based anonymization using local recoding. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006. pp 785-790
18. Guo K, Zhang Q. Fast clustering-based anonymization approaches with time constraints for data streams. Knowledge-Based Syst.46, 2013. pp 95-108
19. Cao J, Carminati B, Ferrari E, Tan KL. CASTLE: Continuously anonymizing data streams. IEEE Trans Dependable Secur Comput. 8(3), 2011. pp 337-352
20. Keele, Staffs. Guidelines for performing systematic literature reviews in software engineering. Ver. 2.3 EBSE Technical Report. EBSE. 2007.
21. Okoli C, Schabram K. Working Papers on Information Systems A Guide to Conducting a Systematic Literature Review of Information Systems Research. Work Pap Inf Syst. 10(26), 2010. pp 1-51
22. Abdalaal A, Nergiz ME, Saygin Y. Privacy-preserving publishing of opinion polls. Comput Secur. 37, 2013. pp 143-154.
23. Aggarwal CC. Privacy and the Dimensionality Curse. In: Privacy-Preserving Data Mining - Models and Algorithms. 2008. pp 433-460
24. Byun J-W, Sohn Y, Bertino E, Li N. Secure anonymization for incremental datasets. Secur Data Manag. 4165, 2006. pp 48-63
25. El Emam K, Dankar FK. Protecting Privacy Using *k*-Anonymity. J Am Med Informatics Assoc. 15(5), 2008. pp 627-637
26. Fung BCM, Trojer T, Hung PCK, Xiong L, Al-Hussaeni K, Dssouli R. Service-oriented architecture for high-dimensional private data mashup. IEEE Trans Serv Comput. 5(3), 2012. pp:373-386
27. Gal TS, Tucker TC, Gangopadhyay A, Chen Z. A data recipient centered de-identification method to retain statistical attributes. J Biomed Inform. 50, 2014. pp 32-45.
28. Ghinita G, Tao Y, Kalnis P. On the anonymization of sparse high-dimensional data. In: Proceedings - International Conference on Data Engineering.; 2008. pp 715-724.
29. Gkoulalas-Divanis A, Loukides G, Sun J. Publishing data from electronic health records while preserving privacy: A survey of algorithms. J Biomed Inform. 50, 2014. pp 4-19..
30. Jändel M. Decision support for releasing anonymised data. Comput Secur. 46, 2014. pp 48-61
31. Kim S, Sung MK, Chung YD. A framework to preserve the privacy of electronic health data streams. J Biomed Inform. 50, 2014. pp 95-106
32. Mohammed N, Fung BCM, Hung PCK, Lee C-K. Centralized and Distributed Anonymization for High-Dimensional Healthcare Data. ACM Trans Knowl Discov Data. 4(4), 2010. pp 1-33.
33. Mortazavi R, Jalili S. Fast data-oriented microaggregation algorithm for large numerical datasets. Knowledge-Based Syst. 67, 2014. pp 195-205.
34. Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. In: Proceedings - IEEE Symposium on Security and Privacy.; 2008. pp 111-125
35. Nergiz ME, Clifton C, Nergiz AE. Multirelational *k*-anonymity. IEEE Trans Knowl Data Eng. 21(8), 2009. pp 1104-1117
36. Nergiz ME, Gök MZ. Hybrid *k*-Anonymity. Comput Secur. 44, 2014. pp 51-63
37. Pei J, Xu J, Wang Z, Wang W, Wang K. Maintaining *k*-anonymity against incremental updates. In: Proceedings of the International Conference on Scientific and Statistical Database Management, SSDBM.; 2007. doi:10.1109/SSDBM.2007.16.
38. Vatsalan D, Christen P, Verykios VS. A taxonomy of privacy-preserving record linkage techniques. Inf Syst. 38(6), 2013. pp 946-969
39. Yingjie W, Zhihui S, Xiaodong W. Privacy preserving *k*-anonymity for re-publication of incremental datasets. In: 2009 WRI World Congress on Computer Science and Information Engineering, CSIE 2009.Vol 4.; 2009:53-60. doi:10.1109/CSIE.2009.549.
40. Ghinita G, Karras P, Kalnis P, Mamoulis N. Fast data anonymization with low information loss. In: Proceedings of the 33rd International Conference on Very Large Data Bases, 2007. pp 758-769
41. Chung KL, Huang YL, Liu YW. Efficient algorithms for coding Hilbert curve of arbitrary-sized image and application to window query. Inf Sci (Ny). 177(10), 2007. pp 2130-2151