Privacy Preserving Techniques on Centralized, Distributed and Social Network Data – A Review

R. Padmaja
Research Scholar
SCSE, VIT University, Vellore

V. Santhi, PhD
Associate Professor
SCSE, VIT University, Vellore.

ABSTRACT
Privacy Preserving Data Publishing refers publishing data in such a way that the privacy of the individuals are preserved. The Published data can further be used for various **Data Analysis** and **Data Mining tasks**. Techniques used to preserve privacy of individuals before publishing is called Anonymization Techniques. Initially only centralized data need to be published for analysis and Mining. Later with the advent of Internet, it has become necessary to publish Distributed and Social network data. The Anonymization Techniques that are applied on Centralized data can be applied on both Distributed and Social Network data with little modifications. This Paper is to present a brief review of Anonymization Techniques like k-anonymity and l-diversity **on Centralized, Distributed and Social Network Data.**

Keywords
SMC,TTP

1. INTRODUCTION
Generally Organizations data contains personal information of individuals, so before releasing the data, the privacy should be preserved. Techniques that are used for privacy preserving data publishing are called Anonymization Techniques.

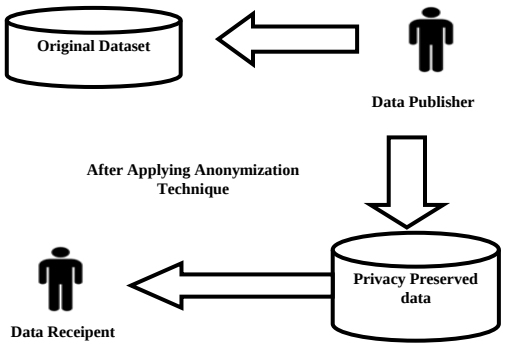

Figure 1: Privacy Preserving Data Publishing

Anonymization can be applied on centralized data, distributed and social network data. The popular anonymization techniques that can be applied on centralized data are k-anonymity and l-diversity. The same with little modifications can be applied on distributed and social network data.

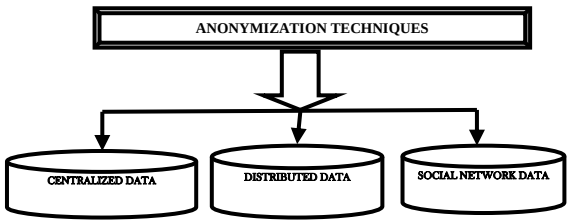

Figure 2: Anonymization Techniques can be applied on different databases

1.1 Relational Data
Organizations often need to publish their data for Research or Mining. Generally such data is stored in a table and each record corresponds to one individual. The attributes of such table are divided into 3 categories. ***Explicit Attributes***, attributes used to identify the tuple. *SSN is an Example of Explicit attribute*. The second category is ***Quasi Identifiers***, whose values collectively used to identify the individual and finally third category is S***ensitive attributes***, whose values are considered sensitive [1].

For Example, medical organizations need to publish their patient data for Medical Research purpose. Since Patient Data contains sensitive information, it should not be published as it is i.e the privacy of the Patients should be preserved before publishing.

Two types of information disclosure are possible [2,3]. Firstly ***Identitity disclosure*** and later ***Attribute disclosure***. ***Identity disclosure,*** occurs when an individual is linked to a particular record in the published table. ***Attribute Disclosure,*** occurs when the published data helps to infer the characteristic of an individual more accurately.

Identity disclosure often leads to attribute disclosure. *Anonymization Techniques* helps to limit such *disclosures*. First step of anonymization is removing *Explicit identifiers* but that is not enough because an adversary can identify an individual from the quasi identifiers.

A common anonymization approach is G*eneralization*, which replaces quasi identifier values into less specific but semantically consistent. As a result, more records with same set of quasi identifier values are retrieved. Identify a set of records whose quasi identifier values are same and make it an equivalence class.

To effectively limit the disclosure, it is necessary to measure the disclosure risk of the anonymized table. Samarati et. al. introduced anonymization technique called k-Anonmity [4,5,6], which only prevents Identity disclosure, but it is not sufficient to prevent Attribute Disclosure, Machanavajjhala et al. [7] introduced a new notion of privacy, called l-diversity**.**

1.2 Distributed Data

There is an increasing need for sharing data repositories containing personal information across multiple distributed, possibly untrusted, and private databases. Such data sharing is subject to constraints imposed by privacy of data.

Government and organizations increasingly recognize the critical value in sharing a wealth of information across multiple distributed, private, and possibly untrusted databases.

An example is the Shared Pathology Informatics Network initiative by the National Cancer Institute that attempts to provide a search interfaces for electronic databases at institutions across the country to locate human specimens and associated clinical and pathologic data needed for cancer research. However, personal health information is protected under regulations such as the Health Insurance Portability and Accountability Act2[8]. In addition, institutions may not want to reveal their private databases to each other for various reasons. i.e. Distributed databases are increasingly used for sharing information in various domains like Health care, Defense etc..

These scenarios can be generalized into the problem of privacy preserving data publishing for multiple distributed databases where multiple data custodians need to publish an anonymized view of the data that does not contain individually identifiable information.

Another example where privacy preserving Distributed Data publishing is needed is in health care domain, where an agenda is needed to develop a Nationwide Health Information Network (NHIN) through which information can be shared among hospitals and also it can be published for research and analysis purpose[9]. Same Anonymization techniques that can be applied on centralized data can be applied on Distributed Data i.e k-anonymity and l-diversity

1.3 Social Network Data

Recently, social networks have received dramatic interest in research and development, partly due to more and more social networks are built online and the fast development of Web 2.0 applications [10]. Social networks model social relationships by graph structures using vertices and edges. Vertices model individual social actors in a network, while edges model relationships between social actors.

Many different kinds of social networks present in our lives such as friendship networks, telephone call networks, and academia co-authorship networks. With the rapid growth of social networks, social network analysis has emerged as a key technique in modern sociology, geography, economics, and information science. The goal of social network analysis is to uncover hidden social patterns.

The power of social network analysis has been shown much stronger than that of traditional methods which focus on analyzing the attributes of individual social actors. social network analysis can serve as a customer relationship management tool for companies selling products and services.

Companies can also use social networks to identify potential customers or recruit candidate employees. For example, according to the statistics published in Time Magazine2, 12% of employers in the United States use popular social networking sites such as MySpace and Facebook to investigate potential employees[10].

Publication of social network data has led to the risk of leakage of personal information of individuals. So, it is necessary to preserve the privacy of individuals before such network data is published by service providers.

Some Companies use social networks to identify potential customers or recruit candidate employees. Neighborhood attack is identified as an essential type of privacy attack. To protect privacy against neighborhood attacks, the conventional *k*-anonymity and *l*-diversity models are extended from relational data to social network data.

2. ANONYMIZATION TECHNIQUES

Techniques that are used to preserve the privacy of individuals before Publishing the Data is called Anonymization Techniques.

2.1 Anonymization Techniques on centralized Data

Most Popular Anonymization Techniques on Centralized Data are k-Anonymity and l-Diversity. Samarati and Sweeney introduced k-anonymity principle to measure the disclosure risk. Initially generalize the table i.e convert the quasi identifier values into less specific but semantically consistent. Later identify those records whose quasi identifier values are same and make it an equivalence class. Then check each equivalence class for k-anonymity principle. *An Equivalence class is said to have k-anonymity if every record in the equivalence class is distinguishable from at least k-1 other records with respect to quasi –identifier attribute[11]*. A table is said to be in k-anonymous if every equivalence class satisfies k-anonymity principle. While k-Anonymity protects identity disclosure it does not protect against identity disclosure so two possible attacks with k-anonymity are Homogeneity and background attack.

To address the above specified limitations of k-anonymity, l–diversity, a stronger notion of Privacy Preserving Principle is introduced. Here the records are not generalized rather the records are divided into equivalence classes in such a way that each equivalence class should satisfy the l-diversity principle. An equivalence class is said to have l-diversity if there are at least l " well-represented" values for the sensitive attribute. A table is said to have l-diversity if every equivalence class satisfies l-diversity principle. While l-diversity protects attribute disclosure, two possible attacks of l-diversity are Similarity attack and skewness attack. These shortcomings are overcome through another anonymization principle called t-closeness[12].

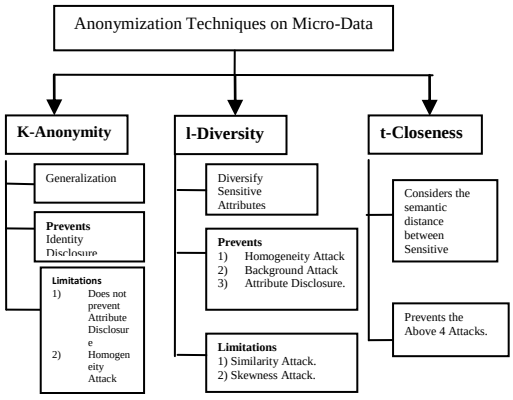

Figure 3: Anonymization Techniques on Micro-Data

2.2 Anonymization Techniques on Distributed Data

Privacy preserving data publishing for multiple distributed databases where multiple data custodians need to publish an anonymized and integrated view of the data that does not contain individually identifiable information.

There are a two approaches one may apply to enable privacy preserving data publishing for distributed databases.

Anonymize–and-Aggregate is one approach, in which each provider should anonymize their data before Integrating. Data Providers can use Anonymization techniques like k-anonymity, l-diversity and t-closeness on their data before aggregating. Here the Data recipients or clients can then query the individual anonymized data or an integrated data. One limitation of this approach is that data is anonymized before the aggregation and hence it does not provide better data utility. In addition, individual databases reveal their ownership of the anonymized data.

A more desirable approach is Aggregate-and-Anonymize or *collaborative data publishing [13]*, in which data from various provides is aggregated and then anonymized as if they would come from one source, using either a trusted third-party (TTP) or Secure Multi-party Computation (SMC) protocols to do computations.

Here the Trusted Third Party has to aggregate the data from various providers and then apply any anonymization techniques like k-anonymity, l-diversity and t-closeness. i.e TTP can use k-anonymity if the third party wants to prevent identity disclosure or use l-diversity to prevent identity and attribute disclosure or use t-closeness to overcome the disadvantages of l-diversity i.e similarity and skewness attack.

This approach assumes an existence of third party that can be trusted by each of the data owners as shown in Figure. 4b. Here, in this approach, data owners send their data to this trusted third party where data integration and anonymization are performed. Then, clients can query the centralized database.

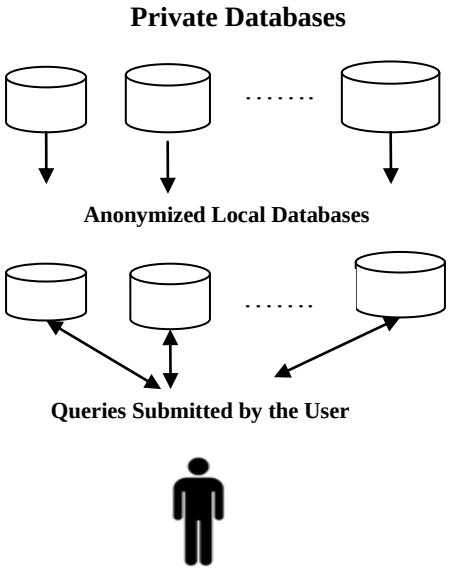

Figure 4a: Anonymization-and-Aggregate

.

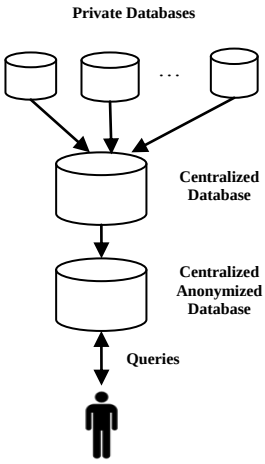

Figure 4b: Aggregate-and-Anonymization

Our goal is to publish an anonymized view of the integrated data such that a data recipient including the data providers will not be able to compromise the privacy of the individual records provided by other parties. Anonymization techniques should prevent the following attacks. 1) Attacks by External Data Recipient Using Anonymized Data 2) Attacks by Data Providers Using Intermediate Results and Their Own Data. 3) Attacks by Data Providers Using Anonymized Data and Their Own Data

2.3 Anonymization technique on Social network

Privacy of individuals may be leaked if a social network data is released improperly to public. In Practice, a systematic approach is needed to anonymize social network data, which is much more challenging than anonymizing the relational data due to the following issues.

The first issue is modeling background knowledge of adversaries. The second issues is measuring information loss in anonymizing social network data and the third one is that devising anonymization methods for social network data. Background knowledge of adversaries may be modeled in various ways.1) identifying attributes of vertices.2) vertex degree Pair of an Edge 3) link relationships.4) neighborhoods 5) embedded subgraphs 6) graph metrics. In social network data publication different methods are proposed to model the different background knowledge, because one method cannot solve all the problems in one shot. The existing anonymization methods on social network data publication is classified into three categories such as 1) Identity preserving methods, 2) Link preserving methods, and 3) Sensitive attribute preserving methods.

To protect privacy against neighborhood attacks, The conventional k-anonymity and l-diversity models can be extended from relational data to social network data.

2.3.1 K-Anonymity in social networks.

An adversary may attack the privacy using the neighborhoods. Let G be a social network and G' an anonymization of G. If G' is k-anonymous, then with the neighborhood background knowledge, any vertex in G cannot be re-identified in G' with confidence larger than 1/k[19].

Given a social network G, the k-anonymity problem is to compute an anonymization G' such that (1) G' is k-anonymous;

(2) each vertex in G is anonymized to a vertex in G' and G'does not contain any fake vertex;(3) every edge in g is retained in G'; and (4) the number of edges to be added is minimized.

A practical method to anonymize a social network data which satisfies k-anonymity requirement has two steps. First, The neighborhoods of all vertices in the network should be extracted. The simple and effective technique for extracting the neighborhoods of all vertices is *neighborhood component coding technique* [19]. This technique can be used to represent the neighborhoods in a concise way and to facilitate the comparisons among neighborhoods of different vertices including the isomorphic tests. In the second step, greedily organize vertices into groups and anonymize the neighborhoods of vertices in the same group. Due to the well-recognized power law distribution of the degrees of vertices in large social networks, start with those vertices of high degrees.

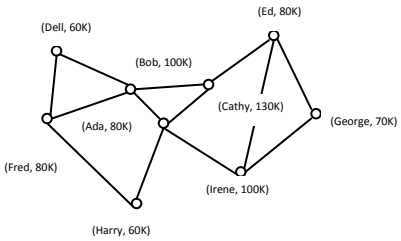

Figure 5a: **A sample social network graph**

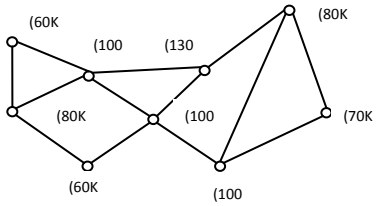

Figure 5b: **A 2-Anonymous Network Still Leak**

This practical solution to the k-anonymity problem can obtain k-anonymous social networks with low information loss.

2.3.2 L-Diversity In Social Networks

As how a k-anonymized relational table may not preserve privacy sufficiently because it lacks diversity in sensitive attributes, k-anonymized social network data still may leak privacy. If an adversary can link a victim to a group of vertices anonymized together all associated with a sensitive attribute value, then the adversary still can link the victim to the sensitive attribute value.

As a concrete example, consider the social network in Figure.5a. Each vertex in the social network carries two labels: the name and a sensitive attribute value Salary. Figure.5b is a 2-anonymous network of Figure.5a. Does Figure.5b preserve the privacy on the sensitive salary information sufficiently?

If an adversary is equipped with the background knowledge of the 1-neighborhood of ada, due to the 2-anonymity, the adversary cannot identify the vertex of ada in Figure.5b. However, since ada, bod and Irene have the isomorphic 1-neighborhood in Figure.5b, and no one else has the same 1-neighborhood, the adversary is sure that ada must be one of the three vertices. Importantly, since ada, bod, and Irene all have salary 100k, the adversary can accurately determine the salary of ada.

The above example clearly demonstrates that a k-anonymized social network may still disclose sensitive information due to the lack of diversity. L-diversity principle overcomes this problem by distributing the sensitive values in each equivalence class sufficiently diverse. Technically, let G be a social network and G' be an anonymization of G[19]. G' is said to be l-diverse if in every equivalence group of vertices, at most 1/l of the vertices are associated with the most frequent sensitive label. As a result, an adversary with the background knowledge of 1-neighborhood structure only can infer the sensitive label for a target victim with the probability not large than 1/l. The larger the value of l, the better privacy is protected.

3. CONCLUSION & FUTURE WORK

It became evident from the literature that privacy of users is the main concern and topic of research now a days. Various Anonymization techniques like k-anonymity and l-diversity are primary techniques that can be applied on tabular microdata. Later many studies proved that same techniques can be applied on distributed data. As social network data is much more complicated than relational data, privacy preserving in social networks is much more challenging and needs many serious efforts in the future. Particularly, modeling adversarial attacks and developing privacy preservation strategies are critical. Privacy preservation in social networks is a relatively new research direction. There is much future work needed to be done. For example, this paper is to review the complete 1-neighborhood attack as the background knowledge. Considering different kinds of background knowledge, the privacy preservation model and methods in social network data can be completely different. Furthermore, there may be various kinds of other privacy attacks in the social network data, thus effective and efficient anonymization methods with respect to different attacks are quite interesting.

4. ACKNOWLEDGEMENT

we would like to thank all our well wishers who has supported us to do this Paper work. Last but not least we would like to thank our Family for their constant support and encouragement.

5. REFERENCES

[1] Li N, Li T, Venkatasubramanian S (2007) t-Closeness: privacy beyond k-anonymity and l-diversity. In: Proceedings of the 23rd international conference on data engineering (ICDE'07), IEEE, pp 106–115.

[2] G. T. Duncan and D. Lambert. Disclosure-limited data dissemination.J. Am. Stat. Assoc., pages 10–28, 1986.

[3] D. Lambert. Measures of disclosure risk and harm. J. OfficialStat., 9:313, 1993.

[4] P. Samarati. Protecting respondent's privacy in microdata release. IEEE T. Knowl. Data En., 13(6):1010–1027, 2001.

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

[6] L. Sweeney. K-anonymity: A model for protecting privacy.Int. J. Uncertain. Fuzz., 10(5):557–570, 2002.

[7] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. InProc. 22nd Intnl. Conf. Data Engg. (ICDE), page 24, 2006.

[8] Pawel Jurczyk, LiXiong, "Privacy Preserving Data Publishing for Horizontally Partitioned Databases", In Proc. Of 17th ACM conference on Information and Knowledge management, ACM, New York, USA, Pages 1321-1322, 2008.

[9] Slawomir Goryczka, Li Xion, Benjamin C.M.Fung, "M-Privacy for Collaborative Data Publishing", In Proc of 7th International Conference on Collaborative Computig:Networking, Applications and Worksharing, Orlando, FL,Pages:1-10,2011.

[10] B. Zhou, Jian Pei, Wo-Shun Luk, "A brief survey on anonymization techniques for privacy preserving publishing of social network data," ACM SIGKDD Explorations Newsletter, Vol. 10, pp. 12-22, 2008.

[11] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.(3,23).

[12] Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," In Proc. of 23rd International Conference on Data Engineering ICDE 2007, IEEE, Istanbul, pp 106-115, 2007.

[13] N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. Lee, "Centralized and distributed anonymization for high-dimensional healthcare data," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 4, no. 4, pp. 18:1–18:33, October 2010.(2,4).

[14] B.C.M. Fung, K. Wang, R. Chen, and P.S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," ACM Computing Surveys, vol. 42, no. 4, pp. 1-53, 2010.(3,11).

[15] B.C.M. Fung, K. Wang, and P.S. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 5, pp. 711-725, May 2007.(3,12).

[16] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In Proceedings of the 22nd International Conference on Data Engineer ing (ICDE'06), page 25, Washington, DC, USA, 2006.IEEE Computer Society.

[17] B. C. M. Fung, K. Wang, and P. S. Yu. Anonymizing classication data for privacy preservation. IEEE Transactions on Knowledge and Data Engineering,19(5):711{725, 2007.

[18] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In Proceedings of the 7th ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems (PODS'98), page 188, New York, NY, USA, 1998. ACM Press.

[19] B. Zhou and J. Pei. The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. 31 May 2010 © Springer-Verlag London Limited 2010

[20] B. Zhou and J. Pei. Preserving privacy in social networks against neighborhood attacks. In Proceedings of the 24th IEEE International Conference on Data Engineering (ICDE'08), pages 506 to 515, Cancun, Mexico, 2008. IEEE Computer Society.