Privacy and Anonymity For Multilayer Networks: A Reflection

Abhishek Santra[1], Kiran Mukunda[2], Sharma Chakravarthy[3]
IT Lab and CSE Department, UT Arlington, Arlington, Texas
abhishek.santra@mavs.uta.edu[1], kiran.mukunda@mavs.uta.edu[2], sharmac@cse.uta.edu[3]

***Abstract*—Privacy of data as well as providing anonymization of data for various kinds of analysis have been addressed in the context of tabular transactional data which was mainstream. With the advent of the Internet and social networks, there is an emphasis on using different kinds of graphs for modeling and analysis. In addition to single graphs, the use of MultiLayer Networks (or MLNs) for modeling and analysis is becoming popular for complex data having multiple types of entities and relationships. They provide a better understanding of data as well as flexibility and efficiency of analysis.**

In this article, we understand the provenance of data privacy and some of the thinking on extending it to graph data models. We will focus on the issues of data privacy for models that are different from traditional data models and discuss alternatives. We will also consider privacy from a visualization perspective as we have developed a community Dashboard for MLN generation, analysis, and visualization based on our research.

I. Introduction

Privacy is not a new problem or research area. It has been researched extensively in the context of traditional transactional data as well as to some extent on non-traditional data, such as graphs, if not explicitly on multilayer networks (MLNs), etc. With the advent of the internet and social media, as data collection and sharing have reached unexpected heights, there is renewed interest in privacy issues associated with unstructured data. Big data analytics certainly add a new urgency and seriousness regarding privacy issues. As big data analysis uses different models, such as graphs, multilayer networks, and others for data representation and analysis, it is imperative that privacy be extended and addressed in this context as well.

Privacy has been broadly defined as "freedom from unauthorized intrusion" [16]. However, in the current age of social media, both data collection and sharing are difficult to understand due to limited guidelines and regulations. In certain domains such as healthcare, there are strict regulations and compliance requirements for data collection, sharing, and privacy preservation. Health Information Exchange (HIE) has to conform to strict rules for information exchange. Violation of privacy can occur due to security breaches as well as sharing data improperly and without appropriate technical precautions.

In this reflection, we will not address security breaches as it is intentional and/or adversarial, in most cases, using security holes or lapses in the system. We reflect on data privacy when it is explicitly shared with others for various purposes. In this case, a data provider needs to provide or use privacy tools or mechanisms, so that users (researchers and other data consumers) can use or consume data in a manner that does not compromise users' privacy. Techniques used for preserving privacy could take many forms including data redaction, data masking (DM), perturbation of data, anonymization, pseudonymization, format-preserving encryption (FPE), and others. Briefly, data masking creates characteristically intact, but inauthentic, replicas of personally identifiable data or other highly sensitive data in order to uphold the complexity and unique characteristics of data. In this way, tests performed on properly masked data will yield the same results as they would on the authentic data set. Format-preserving encryption (FPE) refers to encrypting in such a way that the output (the cipher text) is in the same format as the input (the plain text). The meaning of "format" varies. Encrypting a 16-digit credit card number to a cipher text that is also a 16-digit is an example of FPE. Typically, FPE is reversible. In the anonymization literature, a single table with rows and columns is typically used as an example for discussion. However, our interest is regarding more complex data structures, such as a graph or an MLN created for analysis in any domain, not merely social networks.

The goal of data analysis or knowledge discovery in the presence of sensitive information requires sanitization of data before it is published for public usage. Data sanitization is a complex problem that is essentially a trade-off between hiding private information and the reduced utility of data. The challenge is to remove or perturb *sensitive information* to thwart an adversary from inferring identity or other sensitive information from the published data. The approach depends on the properties of data and the notion of privacy and utility of the data. It may also depend on some of the assumptions made on the type and amount of data needed for inferring or identifying sensitive information.

One way to sanitize data is through anonymization. It is described as a technique that allows "hiding in the crowd" for easy understanding. k-anonymization was introduced in 1988 [13]. Transactional (or tabular) data attributes can be categorized into: i). *Explicit or sensitive identifiers* (e.g., SSN, name, address) which need to be either encrypted or suppressed and ii). *Quasi-identifiers*, whose release needs to be protected or controlled. This is because using other linkage information, the identity of an individual can be discerned using these quasi or non-sensitive identifiers. k-anonymity requirement is that each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to k individuals. The value of *k* needs to be determined. k-anonymity is built on the idea that by combining (linkage) sets of other available data with similar attributes to identify information about any one of the individuals in the released data can be obscured. Individuals' data is pooled in a larger group (determined by k), meaning information in the

group could correspond to any single member, thus masking the identity of the individual or individuals in question.

To overcome some of the subtle issues that can thwart *k*-anonymization, the notion of *l*-diversity has been introduced [10]. This refers to the diversity of values within each of the k blocks used for k-anonymization for a sensitive attribute.

A. Privacy Issues for graphs

With the advent of social networks and sharing of social media data which is modeled as graphs (non-tabular form), the related privacy issues need to be addressed. It is not surprising that both the notions of anonymity and diversity have been extended to graphs [7]. A graph is called (*k, l*) anonymous if for every node in the graph there exists at least *k* other nodes that share at least *l* of its neighbors. It has also been shown that simple graph anonymization that removes the identity of each node in the graph by replacing it with a random identification number instead, is not adequate for preserving the privacy of nodes [4]. Complexity results have also been shown for both strong and weak (*k, l*) anonymization.

The rest of the paper is organized as follows: Section II briefly discusses related work. Section III succinctly introduces various graph types under consideration and MLNs. Section IV discusses various approaches to privacy based on the available products. Section V discusses privacy issues in MLNs along with our thoughts on how they can be addressed at different levels. Conclusions are in Section VII.

II. RELATED WORK

Early work on data privacy focused on obscuring data to prevent inference using and combining other (publicly available or otherwise) data with the released data [16]. For sensitive attributes that can reveal identity immediately, suppression or encryption techniques needed to be used. Even for non-sensitive attributes (identified as quasi-identifiers), it was discovered that inferences can be drawn using other information on the values of quasi-identifiers. This led to the notion of k-anonymization where information is hidden in a crowd of k values [13]. Higher the k, stronger is the anonymization or difficulty in inferring. However, this did not address the diversity (of values) issue which could be used to infer identity. This was later formulated as l-diversity [10]. These two together make anonymization robust.

The same concepts were extended to graphs [7] resulting in (*k, l*)-anonymization of graphs. Mislove et al. [11] addressed the problem of inferring profiles on online social networks based on the attributes given for some fraction of the users in an online social network. This would use the graph as well as attribute information to infer the attributes of other users. In their experiments, using as little as 20% of the users' providing attributes, they could infer the attributes for the remaining users with over 80% accuracy. They have used a modified community detection approach/algorithm for their work. This work implies that making attributes private (using access control mechanisms) is not sufficient to guard privacy in a social network.

More recent work [17] has identified fundamental vulnerability of best established graph anonymization mechanisms, namely k-DA [9] and SalaDP [12] which do not take into account key structural characteristics of a social graph when adding fake edges to it making it possible to identify them. Enhancements have been proposed to existing anonymization techniques to make them more robust.

III. GRAPHS AND MULTILAYER NETWORKS

Graphs capture relationships between entities in application data using nodes and edges. This representation allows us to perform various types of analysis depending on the relationships found in the data. As graph data sets are becoming larger and more complex (in terms of entity and relationship types), we need to address privacy issues associated with data sharing using these models.

A. Graph Types Used as Data Models

A **simple graph** is defined as (V, E) where V is a set of vertices or nodes and E is a set of edges connecting two *distinct* vertices. E is a subset of V $\times$ V. The edges are assumed to be unweighted, either directed or undirected, and loops and multiple edges between nodes are not allowed. Typically, vertices have unique numbers, but labels of nodes and edges need not be unique. This simple graph model is adequate for many purposes and is widely used.

An **attributed graph** (also called a multigraph) is defined as (V, E, ϕ) where V is a set of vertices or nodes, E is a set of edges connecting two distinct vertices, and ϕ is a function mapping of E to $\{\{x, y\} \mid x, y \in V \ and \ x \neq y\}$. If the distinctness of nodes is removed, loops will be allowed as well. The main advantage of a multigraph or attributed graph from a modeling viewpoint is that it captures multiple entities and multiple relationships between entities. Multiple labels can be associated with nodes and entities. With the attributed graph model, it is possible to include relevant information from the data description as labels and hence is more expressive as a model than a simple graph model.

Figure 1 shows different types of graph models that have been considered in this paper. Figure 1 (a) shows a **simple graph** with a single type of nodes and edges, but without any label information. Figure 1 (b) shows an **attributed graph (or multigraph)** that includes multiple node and edge types (illustrated using different colors). It also illustrates the support for multiple edges between two nodes (multiple edges between light green-colored nodes). Figure 1 (c) is a multilayer network which is defined below.

A **multilayer network (or MLN)** is a *network of simple graphs* (or forests). In this model, every layer represents a distinct relationship among entities with respect to a single (or combination of) features. The sets of entities across layers, which may or may not be of the same type, can be related to each other too.

Formally, a *multilayer network*, $MLN(G, X)$, is defined by two sets of graphs: i) The set $G = \{G_1, G_2, \ldots, G_N\}$ contains simple graphs of N individual layers, where $G_i(V_i, E_i)$ is

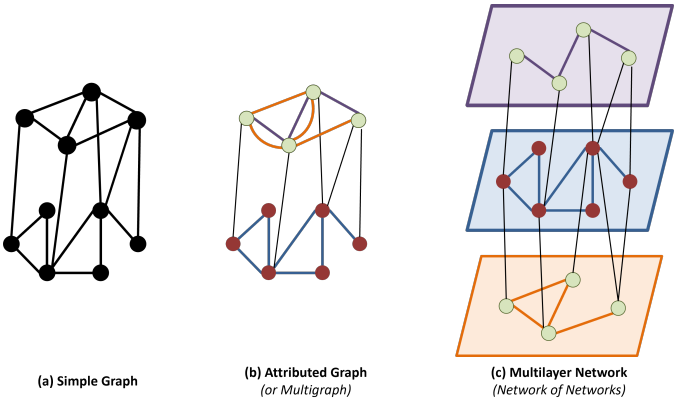


Figure 1: Different Types of Graph Models

defined by a set of vertices, V_i and a set of edges, E_i. An edge $e(v,u) \in E_i$, which is a subset of V_i x V_i, connects vertices v and u, where $v,u \in V_i$ and ii) A set $X = \{X_{1,2}, X_{1,3}, \ldots, X_{N-1,N}\}$ consists of bipartite graphs. Each graph $X_{i,j}(V_i, V_j, L_{i,j})$ is defined by two sets of vertices V_i and V_j, and a set of edges (also called links or inter-layer edges) $L_{i,j}$, such that for every link $l(a,b) \in L_{i,j}$, $a \in V_i$ and $b \in V_j$, where V_i (V_j) is the vertex set of graph G_i (G_j.)

An MLN can be used to separate entities and corresponding relationships from an attributed graph into separate layers where each layer is a simple graph. This provides more clarity in understanding and processing. MLN representations are widely used for modeling complex data sets with multiple types of entities and multiple relationships between the same types of entities. They can also capture relationships between different types of entities.

Based on the type of relationships and entities, multilayer networks can be classified into different types. Layers of a **homogeneous MLN (or HoMLN)** are used to model the diverse relationships that exist among the **same type of entities** like movie actors who are linked based on co-acting (i.e., they act together in a movie) or have similar average rating. Thus, $V_1 = V_2 = \ldots = V_n$ and inter-layer edge sets are empty as no relations across layers are necessary. Relationships among **different types of entities** like researchers (connected by co-authorship), research papers (connected if published in same conference), and year (related by pre-defined ranges/eras) are modeled through **heterogeneous MLN (or HeMLN).** The inter-layer edges represent the relationship across layers like writes, published-in, and active-in. In addition to being collaborators, researchers may be Facebook friends. Thus, to model multi-feature data that capture **multiple relationships within and across different types of entity sets**, a combination of homogeneous and heterogeneous MLNs is used, termed **hybrid MLN (or HyMLN).**

The above three graph types as well as variants of MLNs clearly provide alternatives for modeling for any data set based on entity types and relationships as well as objectives to be explored (in terms of labels retained). It also provides the flexibility of choice as the same information can be represented in attributed graphs and MLNs. Simple graphs and MLNs also provide clarity in understanding the data set. In addition, the availability of algorithms for a specific graph model will also play a key role in the choice of the graph model. As an example, there are not many algorithms available for attributed graphs in contrast to simple graphs. As there is considerable ongoing research in developing algorithms for the multilayer networks [6], [14], [15] and clarity of the model is better, MLNs are preferred for modeling complex data sets.

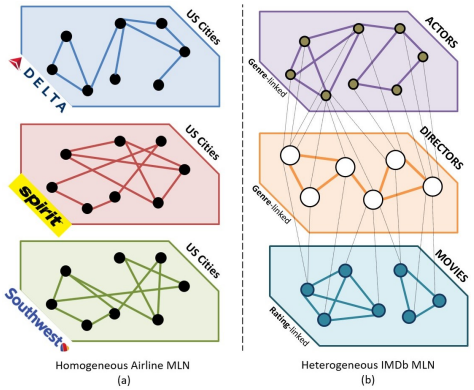


Figure 2: Multilayer Network Types

There exist two distinct types of multilayer networks: Homogeneous and Heterogeneous. When all the layers of MLN have a common set of entities, it is a homogeneous MLN (or HoMLN). For example, the US Airline data set can be modeled using a HoMLN (see Figure 2 (a)), where nodes in each layer represent the cities and edges correspond to the flights between cities of a *specific airline*. The other type of multilayer network is the heterogeneous MLN (or HeMLN), where the set of entities is different across layers and edges exist within and across layers. IMDB (International Movie Database) data set [2] requires HeMLNs (see Figure 2 (b)) to model actors, directors, and movies as different layers along with inter-layer edges to capture relationships across layers (such as, acts-in-a-movie, directs-a-movie, same-rating movies, etc.) Hybrid multilayer networks (HyMLN) are also possible as a combination of the above.

Figure 1 (c) also shows a **multilayer network** (specifically, HyMLN) where the attributed graph of Figure 1 (b) is separated and modeled as different layers such that each layer captures information about a single type of entity and relationship in the form of a simple graph/network. Due to the presence of three types of relationships (shown through orange, blue, and purple colored edges) three separate layers/simple graphs are generated. Also, note that the first and the third layer have the same node types but with different relationship types making it a Hybrid MLN.

We have modeled a number of data sets including IMDb, DBLP (Database bibliography data set), and others. We have also performed various kinds of analysis on the MLN model generated for the data sets. Visualization is important, as the data sets are large and it is difficult to understand the results without visualization. Hence, we have paid special attention

to visualizing data sets as well as analyzing results in various ways.

IV. Product Support for Privacy

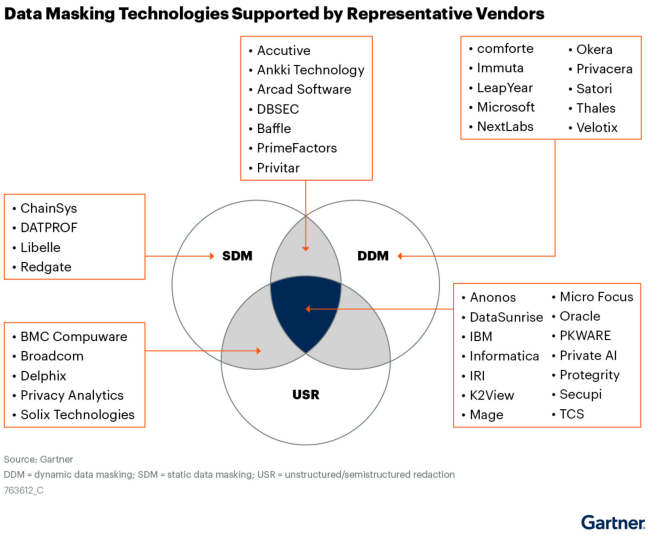


Figure 3: Data Masking Products

The purpose of this section is to highlight or contrast the amount of research on data privacy issues and techniques/algorithms developed with the product support currently available for handling data privacy. A recent (January 2023) Gartner report [5] lists a number of basic data masking products in different categories. Although several algorithms have been developed for k-anonymizaton, l-diversity for both tabular data and graphs, it is unclear whether there are any products based on that research for real-world data.

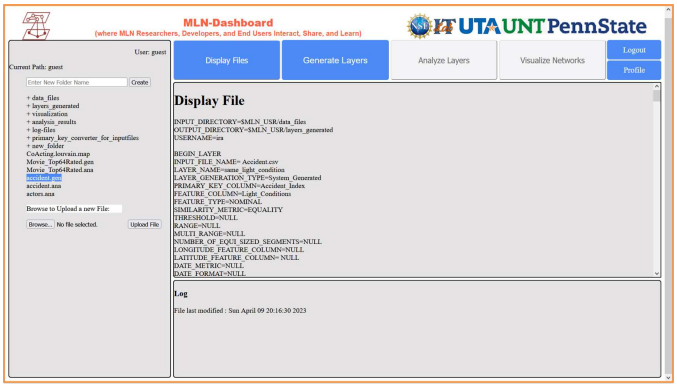
However, there seem to be some data anonymization tools [1] available, such as Clover DX's Anonymization tool, Docbyte's real-time automated anonymization (for images), a programmable g9 anonymizer tool, and others. This situation poses issues for privacy compliance and assurance despite a considerable amount of research results available on this topic.

Figure 3 from [5] shows Data Masking technologies supported by representative vendors (not exhaustive). Figure 3 shows the overlap among dynamic data masking (DDM), static data masking (SDM), and unstructured/semi-structured redaction (USR) tools. The names are self-descriptive.

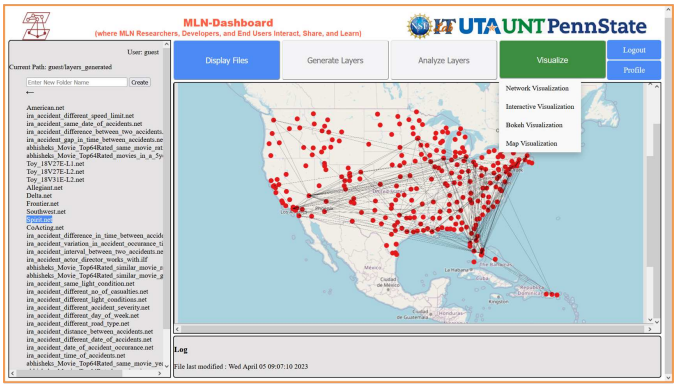
V. Privacy Issues in Graphs and MLNs

Our interest is in adapting privacy tools and techniques for MLNs both for the long term and short term. Long term privacy solutions require research that takes into account MLN characteristics and the analysis algorithms used for them. Mainly, we are considering aggregate computations, such as community detection, computation of various centrality measures (such as degree, betweenness, closeness, etc.), substructure discovery etc. For this, we want to consider both simple graph algorithms and decoupling-based algorithms that have been recently proposed for MLNs [14], [15]. For the short term, we are interested in adapting and integrating publicly available tools into the MLN dashboard that we have developed, as well as implementing simple tools that may be specific to MLNs.

Figure 4 shows the dashboard we have developed, using which MLN layer generation, complex MLN analysis of different types, and visualization of both raw data and analysis results can be done by a user using the web-based graphical interface. Users can upload data files, use configuration files to generate (Figure 4(a) is displaying a layer generation configuration file content) MLN layers, and write complex analysis expressions for community, centrality, and substructure detection on combinations of layers using algorithms developed using the decoupling approach. Depending on the file extension, buttons are highlighted to indicate what can be performed on the file contents. Figure 4(b) shows a visualization of an airline's schedule of flights on a map so it is easy to understand.



(a) Layer Generation Module activated for .gen files



(b) Available Visualization Alternatives

Figure 4: MLN Dashboard Main Window and visualization Display

Most of the work in the literature on graph privacy addresses nodes and edges and there is very little existing work on how to deal with labels, which is an important characteristic of big data analysis. Most of the data sets that we have used include labels – both for nodes and edges – which form an important aspect of the model. These can be seen as the equivalent of attribute values in the tabular context.

Figure 5 shows different aspects of graphs and MLNs that need to be considered for privacy as well as their combinations. This requires adaptation of existing results where possible and development of new ones. We have also used synthetically generated data for our experiments which makes it easier and not worry about privacy issues. We can also generate synthetic graphs with desired characteristics to stress test various hypotheses.

In the short term, we will pursue the following from a pragmatic perspective to make the MLN dashboard as useful for the community as possible.

1) Provide access control mechanisms to users for sharing or selectively sharing data
2) Provide/adapt tools for masking data at different levels as characterized in Figure 5
3) Provide/adapt tools to mask different components of a graph (nodes, edges, labels, or their combinations) as per user request
4) Provide/adapt tools to mask visualization outputs so the results can be useful without having to reveal any data
5) Provide/adapt tools for providing a percentage of data (instead of all data) in conjunction with 3) above based on criteria given by the data producer

In addition, our long-term interest in incorporating privacy to MLN analysis will include additional research on various types of anonymization, different types of masking, and techniques that are specific to MLNs. Interestingly, our earlier work on modeling raw data using the widely used EER (Extended Entity Relationship) model [3] [8] and converting that into MLNs may also provide an opportunity to incorporate privacy at the modeling stage itself.

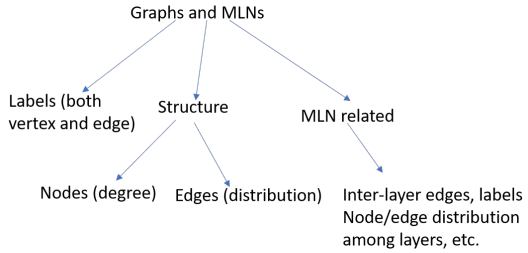


Figure 5: Graph and MLN Characteristics for privacy

VI. Desiderata

Privacy is going to be critical in sharing and analyzing social networks and other types of data. The linkage problem is more complex in the case of graphs and MLNs as compared to tabular data. Also, as graph similarity and isomorphism are complex computations, they may also factor into privacy issues depending on the graph size to be k-anonymized. Certainly, more tools of various kinds are needed to meet user privacy requirements.

VII. Conclusions

This paper reflects upon privacy issues and challenges as it pertains to graphs and MLNs. It also discusses preliminary ideas on how it can be incorporated into graph and MLN analysis at different levels that are based on the structure of the model used. Future work includes fleshing out more details and making it available initially to community users. Compliance and assurance aspects of data are other issues that need to be addressed as well.

Acknowledgments: For this work, Dr. Sharma Chakravarthy was partly supported by NSF Grants CCF-1955798 and CNS-2120393.

References

[1] Data anonymization tools. https://blog.gramener.com/10-best-data-anonymization-tools-and-techniques-to-protect-\\sensitive-information/
[2] The internet movie database. ftp://ftp.fu-berlin.de/pub/misc/movies/database/
[3] From base data to knowledge discovery – a life cycle approach – using multilayer networks. Data Knowledge Engineering **141**, 102058 (2022)
[4] Backstrom, L., Dwork, C., Kleinberg, J.: Wherefore art thou r3579x? anonymized social networks, hidden patterns, and structural steganography. Association for Computing Machinery, New York, NY, USA (2007)
[5] Bales, A., Fritsch, J.: Market guide for data masking, id g00763612-21. In: Gartner Report (January 2023)
[6] Boden, B., Günnemann, S., Hoffmann, H., Seidl, T.: Mining coherent subgraphs in multi-layer graphs with edge labels. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1258–1266. KDD '12, ACM (2012). https://doi.org/10.1145/2339530.2339726, http://doi.acm.org/10.1145/2339530.2339726
[7] Feder, T., Nabar, S., Terzi, E.: Anonymizing graphs. CoRR **abs/0810.5578** (10 2008)
[8] Komar, K.S., Santra, A., Bhowmick, S., Chakravarthy, S.: EER → MLN: EER approach for modeling, mapping, and analyzing complex data using multilayer networks (mlns). In: Dobbie, G., Frank, U., Kappel, G., Liddle, S.W., Mayr, H.C. (eds.) Conceptual Modeling - 39th International Conference, ER 2020, Vienna, Austria, November 3-6, 2020, Proceedings. LNCS, vol. 12400, pp. 555–572. Springer (2020)
[9] Liu, K., Terzi, E.: Towards identity anonymization on graphs. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. p. 93–106. SIGMOD '08, Association for Computing Machinery, New York, NY, USA (2008). https://doi.org/10.1145/1376616.1376629, https://doi.org/10.1145/1376616.1376629
[10] Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: L-diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data **1**(1), 3–es (mar 2007). https://doi.org/10.1145/1217299.1217302, https://doi.org/10.1145/1217299.1217302
[11] Mislove, A., Viswanath, B., Gummadi, K.P., Druschel, P.: You are who you know: Inferring user profiles in online social networks. pp. 251–260 (02 2010). https://doi.org/10.1145/1718487.1718519
[12] Sala, A., Zhao, X., Wilson, C., Zheng, H., Zhao, B.: Sharing graphs using differentially private graph models. Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC (11 2011). https://doi.org/10.1145/2068816.2068825
[13] Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. In: Proceedings of the IEEE Symposium on Research in Security and Privacy (1998), citeseer.ist.psu.edu/samarati98protecting.html
[14] Santra, A., Bhowmick, S., Chakravarthy, S.: Efficient community re-creation in multilayer networks using boolean operations. In: International Conference on Computational Science, ICCS 2017 (2017)
[15] Santra, A., Bhowmick, S., Chakravarthy, S.: Hubify: Efficient estimation of central entities across multiplex layer compositions. In: Gottumukkala, R., Ning, X., Dong, G., Raghavan, V., Aluru, S., Karypis, G., Miele, L., Wu, X. (eds.) 2017 IEEE International Conference on Data Mining Workshops, ICDM Workshops 2017, New Orleans, LA, USA, November 18-21, 2017. pp. 142–149. IEEE Computer Society (2017)
[16] Vaidya, J., Clifton, C.: Privacy-preserving data mining: Why, how, and when. IEEE Security & Privacy **2**(6), 19–27 (2004)
[17] Zhang, Y., Humbert, M., Surma, B., Manoharan, P., Vreeken, J., Backes, M.: Towards plausible graph anonymization (2019)