Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 10, October 2014, pg.910 – 917

SURVEY ARTICLE

A Survey on Optimized Structural Diversity

R.Rajeshkumar[1], S.Saranya[2], S.Shanthi[3]
[1,2]M.E Scholar, Department of Computer Science and Engineering
[3]Assistant Professor, Department of Computer Science and Engineering
Sri Eshwar College of Engineering, Coimbatore
rajeshdreams07@gmail.com [1], saranyacse41@gmail.com [2], shan.sece@gmail.com [3]

Abstract - This paper identifies various concepts involved in social networks for anonymizing the original details of the user. We focus on the various methods that can be applied for applying the anonymization techniques. The methods used are re-identification, k-isomorphism, k-automorphism and k^w-SDA. These methods are used to provide the security and privacy for each user and the community in the social networks. The k^w-SDA method is used to prevent privacy breaches in dynamic networks and minimize graph alterations.

Keywords - Social network, privacy, anonymization, re-identification, k-isomorphism, k-automorphism

I. INTRODUCTION

Social data mining has some privacy and security challenges like stealing the original information of users. To overcome these challenges there are variety of methods have been proposed. Although, anonymization of the information is again the challenging task. For this, the methods such as re-identification, k-isomorphism, k-automorphism and k^w-SDA and R-MAT. In re-identification, privacy risks are associated with releasing network datasets are identified. This method mainly concentrates on risk assessment. K-isomorphism is used for privacy preserving publication against structural attacks. K-anonymity does not protect data against link info attacks. k^w-Structural Diversity Anonymity is used to identify protection in sequential releases of dynamic networks. This technique is used to anonymize the graph based on previous releases and minimize graph alterations. This method mainly used for protecting multicommunity identity. R-MAT generates graph by operating on its adjacency matrix. It is used to generate the graph automatically with communities within communities.

A. R-MAT

The World Wide Web, the Internet topology and Peer-to-Peer networks follows surprising power-laws exhibit strange “bow-tie" or “jellyfish" structures, while still having a small diameter.

Ideally, we would like a generative model with the following properties:

- **Parsimony**: It would have a few only parameters.
- **Realism**: It would only generate graphs that obey the above "laws" and it would match the properties of real graphs (degree exponents, diameters etc.) with the appropriate values of its parameters.
- **Generation speed**: it would generate the graphs quickly, ideally, linearly on the number of nodes and edges.

The "recursive matrix" (R-MAT) model, which can quickly generate realistic graphs [1], capturing the essence of each graph in only a few parameters. Contrary to existing generators, our model can trivially generate weighted, directed and bipartite graphs; it subsumes the celebrated Erdos-Renyi model as a special case; it can match the power law behaviours, as well as the deviations from them. The typical representative here is the Barabasi-Albert (BA) method with the "preferential attachment" idea: keep adding nodes; new nodes prefer to connect to existing nodes with high degrees.[1]

The goals a graph generator should achieve are that the generated graph should:

- match the degree distributions (power laws or not)
- exhibit a "community" structure
- have a small diameter, and match other criteria

The Main Idea here is to provide a method which fits both unimodal and power-law graphs using very few parameters.[1] This method, called Recursive Matrix or R-MAT for short, generates the graph by operating on its adjacency matrix in a recursive manner. The partitions a and d represent separate groups of nodes which correspond to communities. The partitions b and c are the cross-links between these two groups; edges there would denote friends with separate interests. The recursive nature of the partitions means that here automatically get sub-communities within existing communities.

Third point results in "communities within communities". The skew in the distribution of edges between the partitions leads to lognormals and the DGX distribution. Here show experimentally that R-MAT also generates graphs with small diameter and matching other criteria as well.

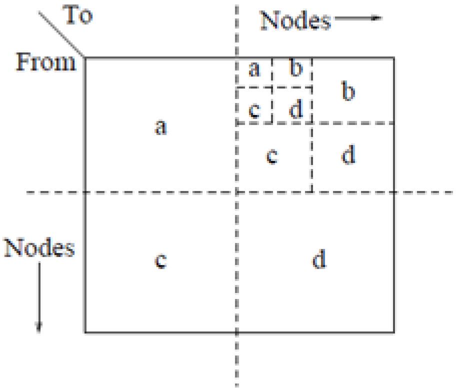


Fig: R-MAT Model

The basic idea behind R-MAT is to recursively subdivide the adjacency matrix into four equal-sized partitions, and distribute edges within these partitions with a unequal probabilities: starting off with an empty adjacency matrix, we "drop" edges into the matrix one at a time. Each edge chooses one of the four partitions with probabilities a; b; c; d respectively (Figure1). The chosen partition is again subdivided into four smaller partitions, and the procedure is repeated until we reach a simple cell. This is the cell of the adjacency matrix occupied by the edge. The number of nodes in the R-MAT graph is set to 2^n.This technique is used for generating "communities" in the graph. The skew in the distribution of edges between the partitions (a >= d) leads to lognormals and the DGX distribution.

R-MAT model can be considered as a binomial cascade in two dimensions. We can calculate the expected number of nodes ck with out-degree k:

$$c_k = \binom{E}{k} \sum_{i=0}^{n} \binom{n}{i} \left[p^{n-i}(1-p)^i\right]^k \left[1 - p^{n-i}(1-p)^i\right]^{E-k}$$

Where 2^n is the number of nodes in the R-MAT graph and p = a + b.

An undirected graph must have a symmetric adjacency matrix. In this the resulting matrix will be symmetric, and hence the corresponding graph will be undirected. For a bipartite graph, the length and height may be different, and the adjacency matrix will be a rectangle instead of a square.

B. K-Isomorphism

The information in social networks becomes an important data source, and sometimes it is necessary or beneficial to release such data to the public. If we represent each user as a node, and create an edge between two nodes when there exists sufficient email correspondence between the two corresponding individuals, then we arrive at a data graph, or a social network. Privacy preservation is about the protection of sensitive information. [2] From the examples of real datasets we identify two main types of sensitive information that a user may want to keep private and which may be under attack in a social network.

NodeInfo:

The first type, which we call *NodeInfo*, is some information that is attached with a vertex.

LinkInfo:

The second type, which we call *LinkInfo*, is the information about the relationships among the individuals, which may also be considered sensitive.

Most of the above works in privacy preserving publishing of social network aim at the issue of node reidentification, [2] which means that in the published data the adversary is not able to link any individual to a node with high confidence. Most of the solutions target at *k*-anonymity. The issue is that if the privacy is some sensitive value linked with some individual, then *k*-anonymity is an overkill.[2] The reason is that if the sensitive values of a set of *k* records are simply separated from the individual records and published as a set (bucket), then privacy can be guaranteed and there is no need to change the raw data related to the other parts of the records.

The important point is that even if a graph is *k*-anonymous or *k*-automorphic, it cannot protect the data from LinkInfo attack. The fact that there are k different ways to map each of 2 individuals in a graph does not protect the linkage of the 2 individuals if all the mappings are identical in terms of the relevant graph structure. A social network is modelled as a simple graph, in which vertices describe entities and edges describe the relationships between entities. A vertex in the graph has an identity and is also associated with some information such as a set of emails. The NodeInfo or LinkInfo[2] of any individual can be inferred only with a probability not higher than a pre-defined threshold, whereas the information loss of the published graph with respect to the original graph is kept small.

The main contributions can be summarized as follows. (1) Here identify two realistic targets of privacy attacks on social network publication, NodeInfo and LinkInfo. Here point out that the popular notion of *k*-anonymity in graph data does not protect the data against LinkInfo attacks. Although some previous works have considered protection of links, there has not been any definition of a quantifiable guarantee in the protection. To our knowledge, the first to define this problem formally. (2) Here prove that this problem is NP-hard. (3) Here propose a solution by *k*-isomorphism anonymization and show that this is the only solution to the problem. (4) Here design a number of techniques to make the anonymization process efficient while maintaining the utility. (5) Here introduce a dynamic release mechanism that has a number of advantages[2] over previous work.

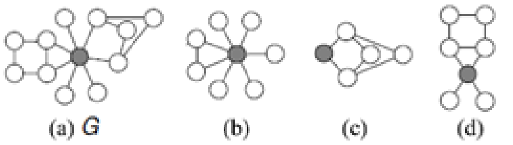
Minimal anonymization cost is given by two conditions:

(1) The difference between the number of edges in *G* and the number of edges in *Gk* is minimized.
(2) Under condition (1), the edit distance *ED* (*G;Gk*) is minimized.

For simplicity we first assume that for the given graph *G = {V; E}*, |*V* | is a multiple of *k*. This assumption can be easily waived by adding no more than *k* − 1 dummy vertices in the graph. The solution relies on the concept of graph isomorphism.

Enforcing *k*-isomorphism could mean that we have reduced the information of the given graph *G* to 1=*k* the original size. In fact, since all the graphs *gi* is isomorphic we may as well publish just one of the sub graphs *gi* in *Gk*, if graph structure is the only interested information. The graph is partitioned into different sub graphs. After the partitioning, the sub graphs are augmented by edge addition and deletion to ensure pair wise graph isomorphism. The frequent sub graphs have a high potential to generate VD-embeddings and large connected sub graphs minimize the edge augmentation needed for graph isomorphism.

The discovery of frequent sub graphs is costly, especially when considering large sub graphs. Also considering large sub graphs may not be useful since they are less likely to be frequent. For better performance we set a threshold *maxPAGsize* on the size of the maximum sub graphs to be considered, where the size is in terms of the number of edges in the sub graph.



Neighbourhood sub graphs as NAGs

C. Resisting Structural Re-identification

A network dataset is a graph representing entities connected by edges representing relations such as friendship, communication, or shared activity. Maintaining privacy when publishing a network dataset is uniquely challenging because an individual's network context can be used to identify them even if other identifying information is removed. Network data can describe a variety of domains: a social network might describe individuals connected by friendships; an information network might describe a set of articles connected by citations; a communication network might describe Internet hosts related by traffic flows. [3]

Anonymization techniques [3] for tabular data do not apply to network data because they fail to account for the interconnectedness of the entities. The network analysis can be performed in the absence of entity identifiers a natural strategy for protecting sensitive information is to replace identifying attributes with synthetic identifiers. We refer to this procedure as naive anonymization. A distinctive threat in network data is that an entity's connections can be distinguishing, and may be used to re-identify an otherwise anonymous individual. Although an adversary may also have information about the attributes of nodes resulting from structural or topological re-identification. [3] The use of attribute knowledge to re-identify individuals in anonymized data has been well-studied, as have techniques for resisting it.

In this work, main contributions are follows:

Adversary Model: here propose a flexible model of external information used by an adversary to attack naively-anonymized networks. The model allows us to evaluate re-identification risk efficiently and for a range of different adversary capabilities. Here also formalize the structural indistinguishability of a node with respect to an adversary with locally-bounded external information.

Empirical Risk Assessment: here evaluate the effectiveness of structural attacks on real and synthetic networks, measuring successful re-identification and edge disclosures. Here find that real networks are diverse in their resistance to attacks. Nevertheless, the results demonstrate that naive anonymization provides insufficient protection, especially if an adversary is capable of gathering knowledge beyond a target's immediate neighbours.

Theoretical Risk Assessment: In addition to the empirical study, we perform a theoretical analysis of random graphs. Here show how properties such as a graph's density and degree distribution affect reidentification risk. A significant finding is that in sufficiently dense graphs, nodes can be re-identified even when the graph is extremely large.

Privacy Definition: here also propose strategies for mitigating re-identification risk. First, here propose a privacy condition, which formally specifies a limit on how much the adversary can learn about a node's identity.

Anonymization Algorithm: Then here propose a novel algorithm to achieve this privacy condition. The algorithm produces a generalized graph, which describes the structure of the original graph in terms of node groups called supernodes. The generalized graph retains key structural properties of the original graph yet ensures anonymity.

Algorithm Evaluation:

This includes a comparison with other state-of-the-art graph anonymization algorithms.

Naive anonymization:

Formally, we model a network as an undirected graph G = (V,E). The naive anonymization of G is an isomorphic graph, Ga = (Va,Ea), defined by a random bijection Π : V → Va. The anonymization mapping Π, also shown, is a random, secret mapping. Naive anonymization prevents re-identification when the adversary has no information about individuals in the original graph. Formally stated, an individual x ∈ V, called the target, has a candidate set, denoted cand(x), which consists of the nodes of Ga that could feasibly correspond to x.

Threats:

The external information may be available through a public source beyond the control of the data owner, or may be obtained by the adversary's malicious actions. Re-identification can lead to additional disclosures under naive anonymization. [3] If an individual is uniquely re-identified, then the entire structure of connections surrounding the individual is revealed. If two individuals are uniquely re-identified, then the presence or absence of an edge between them is revealed directly by the naively anonymized graph. Such an edge disclosure, in which an adversary is able to accurately infer the presence of an edge between two identified individuals, can be a serious privacy threat. In the present work, we consider the general threat of re-identification as well as the more specific threat edge disclosure.

Anonymity through structural similarity:

A strong form of structural similarity between nodes is automorphic equivalence. Two nodes x, y ∈ V are automorphically equivalent (denoted x ≡A y) if there exists an isomorphism from the graph onto itself that maps x to y. Automorphic equivalence induces a partitioning on V into sets whose members have identical structural properties. It follows that an adversary — even with exhaustive knowledge of a target node's structural position —cannot identify an individual beyond the set of entities to which it is automorphically equivalent. Two such nodes are structurally indistinguishable.

Adversary model based on structural signatures:

It is based on a class of knowledge queries, of increasing power, which report on the local structure of the graph around a node. These queries are inspired by iterative vertex refinement, a technique originally developed to efficiently [3] test for the existence of graph isomorphism's. External information about a published social network may be acquired through malicious actions by the adversary or from public information sources. In addition, a participant in the network, with some innate knowledge of entities and their relationships, may be acting as an adversary in an attempt to uncover unknown information. A legitimate privacy objective in some settings is to publish a graph in which participating individuals cannot re-identify themselves.

D. Protection in Dynamic Networks

This paper addresses the privacy risks of identity disclosures in sequential releases of a dynamic network, to prevent privacy breaches, we proposed novel k^w-structural diversity anonymity [4], where k is an appreciated privacy level and w is a time period that an adversary can monitor a victim to collect the attack knowledge. For protecting against such an adversary, here introduce a new dynamic privacy scheme, named dynamic k^w-structural diversity anonymity (k^w-SDA), which ensures that the probability of a vertex identity or a multicommunity identity being revealed is limited to 1k. After that, here develop a scalable heuristic algorithm [4] to provide dynamic k^w-SDA. The proposed algorithm can anonymize the graph based on the previous w - 1 release and minimize the graph alterations.

The contributions of this work are summarized as follows:

1. This paper is the first work that presents a privacy model for protecting multicommunity identity.
2. This paper introduces a new privacy model, dynamic kw-SDA, to ensure the protection for both vertex and multicommunity identities in sequential publications.
3. To achieve kw-SDA, we develop an efficient solution for anonymizing large-scale dynamic networks with limited information distortion.

For better execution efficiency, we propose to construct a table, named the Cluster Sequence Table (CS-Table), to summarize the vertex information of sequential releases and avoid the need of scanning all the releases for anonymization.

Anonymization algorithms:

The problem of dynamic k^w-SDA mainly differs from the other anonymizations for static networks at the consideration of previous releases. The consideration of the previous releases brings significant challenges in minimizing information distortion and improving the computational efficiency of anonymization algorithm [4] since an exponential number of enumerations is possible. Finding optimal solutions on large-scale dynamic networks is thus computationally infeasible. Compared to the anonymizations for static networks, the challenge of the dynamic k^w-SDA is to generate an anonymized release of the current network based on w _ 1 previous release. A naive solution to this problem is first generating a release based on the current network alone and then modifying the current anonymized release based on each previous release to eliminate possible privacy breaches one by one. However, this approach is time consuming because it requires searching all possible privacy breaches through w releases at each time instance of anonymization.

For anonymization, three operations are used to adjust the degree of a vertex.
1) Operation Adding Edge connects two vertices in the same community,
2) Operation Redirecting Edge increases the degree of a vertex v by changing the not-yet-anonymized end-point of a previously added edge to vertex v.
3) Operation Adding Vertex connects a vertex and an additional fake vertex.

E. K-Automorphism

In this work, propose k-automorphism to protect against multiple structural attacks and develop an algorithm (called KM) that ensures k-automorphism [5]. Considering the limitations in previous works, here propose a systematic method for privacy-preserving publishing of social network data. The technique has three advantages.

1) *It can guarantee privacy under any structural attack.* In this method, do not assume one type of attack. Here assume that an adversary can have complete information about the target, such as degree, neighbours, and shortest-distances from hubs and so on. This method can guarantee privacy, even though an adversary can launch multiple and different types of structural attacks.

2) *The released network has no uncertainty.* The released network generated by our method can provide not only a summary of the structural information regarding the whole network, but also structural information about each individual vertex.

3) *It can guarantee privacy under dynamic releases.* Even though an adversary can have historical information about the target, the target cannot be identified in our released networks.

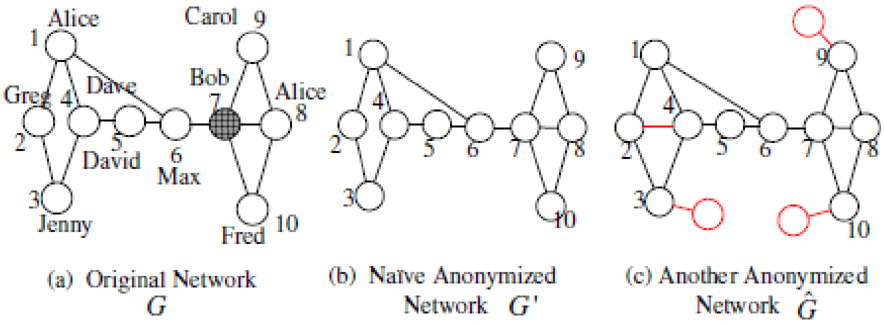


Fig: Anonymized Networks

In summary, we make the following contributions:

1. Here propose a systematic method to protect the released social network data from all structural attacks. Specifically, we propose an algorithm to convert original network *G* into *k*- automorphic network *G*¤, which is then released.

2. Here consider dynamic releases of networks. In order to avoid privacy disclosure in re-publication of networks, here propose vertex ID generation technique.

The Structural Attacks includes *Degree Attack, Sub-graph Attack, 1-Neighbor-Graph Attack and 1-Neighbor-Graph Attack.* [5]

K-Match algorithm:

The KM algorithm [5] is user here to support dynamic social network analysis. The K-Match algorithm start by removing all identity information from the original network *G* resulting in the naive anonymized network *G'.*

S.No	Methods	Usage
1	R-MAT	To generate the graph automatically.
2	k-isomorphism	To protect sensitive information(privacy protection)
3	Structural re-identification	To ensure privacy and security.
4	K^w-SDA	to prevent privacy breaches in dynamic networks and minimize graph alterations.
5	k-automorphism	To protect multiple structural attacks.

Table: comparison of various methods used in anonymization

II. CONCLUSION

The paper describes the comparison and analysis between various anonymization methods involved in the structural diversity in community identification. It also illustrates that there are many techniques that can be followed for anonymization in individual and community identities. This kind of comparison reflects that the efficiency differs from each method. This paper shows the usage of re-identification and k-anonymization methods.

REFERENCES

[1] Deepayan Chakrabarti, Yiping Zhan and Christos Faloutsos, *R-MAT: A Recursive Model for Graph Mining*

[2] James Cheng, Ada Wai-Chee Fu and Jia Liu, *K-Isomorphism: Privacy Preserving Network Publication against Structural Attacks*, SIGMOD'10, June 6–11, 2010, Indianapolis, Indiana, USA.

[3] Michael Hay, Gerome Miklau, David Jensen, Don Towsley and Chao Li, *Resisting Structural Re-identification in Anonymized Social Networks*, VLDB Journal, Volume 19, Number 6, 797-823, doi: 10.1007/s00778-010-0210-x.

[4] Chih-Hua Tai, Peng-Jui Tseng, Philip S. Yu and Ming-Syan Chen, *Identity Protection in Sequential Releases of Dynamic Networks*, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 3, MARCH 2014.

[5] Lei Zou, Lei Chen and M. Tamer Ozsu, *KAutomorphism: A General Framework for Privacy Preserving Network Publication*, VLDB '09, August 2428, 2009, Lyon, France

[6] Chih-Hua Tai, Philip S. Yu, Fellow, IEEE, De-Nian Yang, IEEE, and Ming-Syan Chen, *Structural Diversity for Resisting Community Identification in Published Social Networks*, ieee transactions on knowledge and data engineering, vol. 26, no. 1, january 2014

[7] Aaron Clauset, M. E. J. Newman,and Cristopher Moore, *Finding community structure in very large networks*", arXiv:cond-mat/0408187v2 [cond-mat.stat-mech] 30 Aug 2004

[8] Michael Hay, Chao Li, Gerome Miklau, and David Jensen, *Accurate Estimation of the Degree Distribution of Private Networks*.

[9] Xiaowei Ying and Xintao Wu, *Randomizing Social Networks: a Spectrum Preserving Approach*, U.S. National Science Foundation NSF IIS-0546027.

[10] M.E. Nergiz, M. Atzori, and C. Clifton, *Hiding the Presence of Individuals from Shared Databases*, Proc. ACM SIGMOD Int'l Conf. Management of Data, 2007.

[11] M.E. Nergiz, C. Clifton, and A.E. Nergiz, *Multirelational k-Anonymity*, IEEE Trans. Knowledge & Data Eng., vol. 21, no. 8, pp. 1104-1117, Aug. 2009.

[12] P. Samarati and L. Sweeney, *Generalizing Data to Provide Anonymity When Disclosing Information*, Proc. ACM SIGACTSIGMOD-SIGART Symp. Principles of Database Systems (PODS '98), 1998.