Data Utility Evaluation Framework for Graph Anonymization

Jordi Casas-Roma
Internet Interdisciplinary Institute (IN3)
Universitat Oberta de Catalunya (UOC)
CYBERCAT – Center for Cybersecurity Research of Catalonia
Barcelona, Spain
jcasasr@uoc.edu

***Abstract*—Anonymization of graph-based data is a problem which has been widely studied over the last years and several anonymization methods have been developed. Information loss measures have been carried out to evaluate the noise introduced in the anonymized data. However, there is no consensus about how to evaluate perturbation and data utility in privacy-preserving and anonymization scenarios, where released datasets contain some noise to hinder re-identification processes. Thus, it is quite complex to compare different methods or algorithms in literature. In this paper we propose a framework to evaluate and compare anonymous datasets in a common way, providing an objective score to clearly compare methods and algorithms.**

***Index Terms*—Privacy-preserving, Anonymity, Evaluation framework, Data utility, Social networks, Graphs**

I. Introduction

Currently, data mining processes require large amounts of data, which often contain personal and private information of users and individuals. Although basic processes are performed on data anonymization, such as removing names or other key identifiers, remaining information can still be sensitive as well as useful for an attacker to re-identify users and individuals. To solve this problem, methods which introduce noise to the original data have been developed in order to hinder the subsequent processes of re-identification. However, the noise introduced by the anonymization processes may affect data by reducing its usefulness in subsequent processes of data mining. It is necessary to keep the main properties of data to ensure the data mining process is not altered by the anonymization process.

Anonymization processes should allow the analysis performed in the anonymized data to lead to results as equal as possible to the ones obtained when applying the same analysis to the original data. Nevertheless, data modification is contrary to data utility. The larger data modification, the less data utility. Thus, a good anonymization method hinders the re-identification process while causing minimal distortion to the data.

Owing to what we have mentioned in the previous paragraph, several measures have been designed to evaluate the goodness of the anonymization methods. Generic information loss measures evaluate to what extent the analysis on anonymized data differs from the original data. Each measure focuses on a particular property of the data and it is application-independent. We assume that if these metrics show little variation between original and anonymized data, then the subsequent data mining processes will also show little variation between original and anonymized data. Furthermore, we can also use measures specifically designed to quantify perturbation on real-world specific problems, such as community detection (i. e. clustering) or information flow.

However, there is no standard or common way to evaluate data utility or information loss. Usually, several authors use different generic information loss measures to quantify perturbation on anonymized data. Hence, it is very hard (or even impossible) to compare information loss and data utility among different methods and algorithms, since each work of literature uses specific (and usually different) metrics to evaluate the perturbation or noise in anonymous graphs.

In this paper we propose a common framework to evaluate data utility and information loss on privacy-preserving data publication processes. Specifically, we use generic information loss measures and clustering-specific ones on graph formatted data in order to provide a clear comparison among an original graph and several perturbed (i. e. anonymous) graphs.

A. Notation

Let $G = (V, E)$ be a simple graph, where V is the set of nodes and E the set of edges in G. We use $v_i \in V$ to denote node i and $(v_i, v_j) \in E$ to indicate an edge connecting nodes v_i and v_j. We define $n = |V|$ to denote the number of nodes and $m = |E|$ to denote the number of edges. We use $G = (V, E)$ and $\widetilde{G} = (\widetilde{V}, \widetilde{E})$ to indicate the original and the anonymized graphs, respectively.

B. Roadmap

This paper is organized as follows. We review the state of the art of anonymization in networks in Section II. Our framework is introduced in Section III. Then, we discuss metrics related to generic information loss measures in Section IV and a methodology to compare clustering-specific information loss measures in Section V. Finally, brief examples of our results are presented in Section VI, while Section VII concludes the research and points future work directions.

II. Related work

The two main objectives of an anonymization process are: (1) to preserve the privacy of users or individuals who appear in a dataset, hindering the re-identification processes, and (2) to preserve data utility on anonymized data, i.e., minimizing information loss.

Anonymization methods and graph assessment depend on the type of data they are intended to work with. In this paper, we will work with simple, undirected and unlabelled graphs.

A. Anonymization

We categorize anonymization methods on graph formatted data into three main categories [6]:

- Graph modification approaches: These methods anonymize a graph by modifying (adding and/or deleting) edges or nodes in a graph. There are two basic approaches: (1) The simplest way alters the graph structure by removing and adding edges randomly. It is called randomization or random-based approach. (2) Another way consists of edge addition and deletion to fulfill some desired constraints, i.e. anonymization methods do not modify edges at random, they modify edges to meet some desired constraints. For example, k-anonymity-based approaches modify graph structure (by adding and removing edges) in order to get the k-anonymity value for the graph.
- Generalization approaches (also known as clustering-based approaches): These methods cluster nodes and edges into groups. Then, they anonymize each group into a super-node to publish the aggregate information about structural properties of the nodes [12]. The details about individuals can be hidden properly, but the graph may be shrunk considerably after anonymization, which may not be desirable for analyzing local structures.
- Differentially private approaches: These methods refer to algorithms which guarantee that individuals are protected under the definition of differential privacy [8]. Differential privacy imposes a guarantee on the data release mechanism rather than on the data itself. The goal is to provide statistical information about the data while preserving the privacy of users.

We will focus on graph modification approaches, which preserve local structures and keep the details of the data for clustering processes.

Randomization methods are based on introducing random noise in the original data. For graphs, there are two main approaches: (a) Rand Add/Del that randomly adds and deletes edges from the original graph (this strategy keeps the number of edges) and (b) Rand Switch that exchanges edges between pairs of nodes (this strategy keeps the degree of all nodes and the number of edges). Naturally, edge randomization can also be considered an additive-noise perturbation. There are several works on graph randomization in literature, such as [11], [18], [19], [3].

Another widely adopted strategy for graph modification approaches consists of edge addition and deletion to meet desired constraints, usually to achieve a certain level of privacy. For instance, take the k-anonymity concept that was introduced by Sweeney [17] for the privacy preservation on relational data. It states that an attacker cannot distinguish among k different records although he managed to find a group of quasi-identifiers. Consequently, the attacker cannot re-identify an individual with a probability greater than $\frac{1}{k}$. The k-anonymity model can be applied using different concepts when dealing with networks rather than relational data like in our case. A widely used option is to consider the node degree as a quasi-identifier, which corresponds to k-degree anonymity [14]. It is based on modifying the network structure (by adding and removing edges) to ensure that all nodes satisfy this condition. There are several works in literature on constrained anonymization, for instance [21], [22], [12], [5], among many others.

B. Graph assessment

Several generic measures have been used to quantify the structure's properties in graph formatted data. Authors usually use these measures and compare the values obtained by the original and the anonymized data in order to quantify the noise introduced by the anonymization process. When we quantify the information loss as described above, we talk about generic information loss measure.

Hay et al. [11] utilized five structural properties from graph theory for quantifying network structure. For each node, the authors evaluate closeness centrality, betweenness centrality and path length distribution. For the graph as a whole, they evaluate the degree distribution and the diameter. The objective is to keep these five measures close to their original values, assuming that little distortion is involved in the anonymized data. Ying and Wu [18] and Ying et al. [19] used both real space and spectrum based characteristics to study how the graph is affected by randomization methods. The authors focused on four real space characteristics of the graph and on two important eigenvalues of the graph spectrum. The real space characteristics are: the harmonic mean of the shortest distance, the modularity, the transitivity, and the sub-graph centrality. Since graph spectrum has close relations with many graph characteristics and can provide global measures for some network properties, the authors also consider the following two spectral characteristics: the largest eigenvalue of the adjacency matrix and the second smallest eigenvalue of the Laplacian matrix. Alternatively, Zou et al. [22] defined a simple method for evaluating information loss on undirected and unlabelled graphs. The method is based on the difference between the original and the anonymized graph edges, $cost(G, \widetilde{G}) = (E \cup \widetilde{E}) - (E \cap \widetilde{E})$. Liu and Terzi [14] used clustering coefficient and average path length for the same purpose. Hay et al. [12] examined five properties commonly measured and reported on network data: degree, path length, clustering coefficient, network resilience and infectiousness.

III. Framework

It is important to emphasize that these generic information loss measures only evaluate structural and spectral changes between original and anonymized data. That is, these measures do not evaluate the data mining processes on anonymized data, and as such, they are general or application-independent. The analysis of specific and application-dependent quality measures is an open problem. We consider in this paper the case of an application in clustering.

Although it is possible to compare space and time complexity of these works using the Big $\mathcal{O}$ notation, it is not possible (or at least, not easy) to compare data utility and information loss among relevant works in literature, since they are evaluated using different methodologies and metrics. Our objective is to provide a common framework to quantify information loss on graph perturbation processes. To do so, we will use generic information loss (GIL) measures and

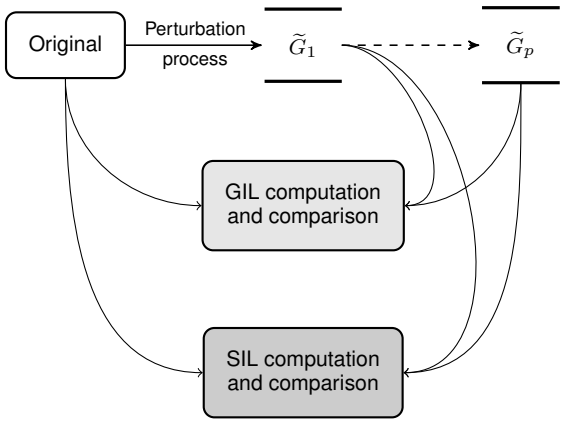

Fig. 1: Experimental framework. Each dataset is anonymized from $\widetilde{G}_1$ to $\widetilde{G}_p$ using some anonymization method. Next, we compare the original and perturbed data using GIL measures in order to quantify the noise introduced on the data. Then, we do the same with real clustering processes and SIL measures.

specific information loss (SIL) measures based on clustering processes.

As we have mentioned previously, in Section II, our framework is focused on graph modification approaches. All aforementioned methods share some properties that we use to evaluate the data utility in our evaluation framework. Specifically, the vertex set remains the same on perturbed graphs, i. e. $V = \widetilde{V}$ and $n = \widetilde{n}$. However, the edge set changes due to the randomization or constraint modification process, i. e. $E \neq \widetilde{E}$, while the number of edges is usually modified during the anonymization process, i. e. $m \neq \widetilde{m}$. The number of edges on perturbed graphs could be greater than the original one, i. e. $m < \widetilde{m}$, or smaller, i. e. $m > \widetilde{m}$. Mainly, it depends on algorithm's edge modification technique, which could be based on edge addition or edge removing.

Our experimental framework is shown in Figure 1. First, we apply perturbation to graph datasets using some graph modification approach. Each dataset is perturbed from $\widetilde{G}_1$ to $\widetilde{G}_p$ of edge set. Then, we evaluate original and perturbed data using GIL measures for quantifying network structure (details in Section IV). Next, we apply the clustering processes both on original and perturbed data and we use clustering-based specific measures (Section V) to evaluate the results.

IV. Generic Information Loss Measures

We use different generic measures for quantifying network structure. These generic measures are used to compare both the original and the anonymized data to quantify the noise introduced in the perturbed data by the anonymization process. These generic measures evaluate some key graph's properties, which are relevant according to [4]. They evaluate the graph structure, so they are general or, in other words, application-independent. Information loss was defined by the discrepancy between the results obtained from the original and the anonymized data. In our experiments we use several graph measures based on structural and spectral properties. In the rest of this section we review the measures used.

Average distance (AD) is defined as the average of the distances between each pair of nodes in the graph. It measures the minimum average number of edges between any pair of nodes. Formally, it is defined as:

$$AD(G) = \frac{\sum_{i,j} d_{ij}}{\binom{n}{2}} \tag{1}$$

where d_{ij} is the length of the shortest geodesic path from v_i to v_j, meaning the number of edges along the path.

Another used measure is **edge intersection** [22], [14] (EI). It is defined as the percentage of original edges which are also in the anonymized graph. Formally:

$$EI(G, \widetilde{G}) = \frac{|E \cap \widetilde{E}|}{max(|E|, |\widetilde{E}|)} \tag{2}$$

Clustering coefficient [14], [12], [10], [7] (C) is a measure widely used in literature. The clustering coefficient of a graph is the average:

$$C(G) = \frac{1}{n} \sum_{i=1}^{n} C(v_i) \tag{3}$$

where $C(v_i)$ is the clustering coefficient for node v_i. The clustering of each node is the fraction of possible triangles that exist. For each node the clustering coefficient is defined by:

$$C(v_i) = \frac{2T(v_i)}{deg(v_i)(deg(v_i) - 1)} \tag{4}$$

where $T(v_i)$ is the number of triangles surrounding node v_i, and $deg(v_i)$ is the degree of v_i.

Transitivity [18], [19], [7] (T) is the fraction of all possible triangles present in the graph. Possible triangles are identified by the number of triads (two edges with a shared node), as we can see in Equation 5.

$$T(G) = \frac{3 \times (number\, of\, triangles)}{(number\, of\, triads)} \tag{5}$$

Betweenness centrality [11] (BC) is a centrality measure, which calculates the fraction of number of the shortest paths that go through each node. This measure indicates the centrality of a node based on the flow among other nodes in the graph. A node with a high value indicates that this node is part of multiple shortest paths in the graph, which will be a key node in the graph structure. We define the betweenness centrality of a node v_i as:

$$BC(v_i) = \frac{1}{n^2} \sum_{s,t} \frac{g_{st}^i}{g_{st}} \tag{6}$$

where g_{st}^i is the number of geodesic paths from v_s to v_t that pass through v_i, and g_{st} is the total number of geodesic paths from v_s to v_t.

The second centrality measure is **closeness centrality** [11] (CC), which is described as the inverse of the average distance to all accessible nodes. Closeness is an inverse measure of centrality in which a larger value indicates a less central node, while a smaller value indicates a more central node. Formally, we define the closeness centrality of a node v_i as:

$$CC(v_i) = \frac{n}{\sum_j d_{ij}} \tag{7}$$

And the last centrality measure is **degree centrality** [11] (DC). It evaluates the centrality of each node associated with its degree. That is, the fraction of nodes connected to it. A higher value indicates greater centrality in the graph. The degree centrality of a node v_i is depicted in Equation 8.

$$DC(v_i) = \frac{deg(v_i)}{m} \tag{8}$$

The last three centrality measures described above evaluate the centrality of each node of the graph from different perspectives. These measures give us a value of centrality for each node. To assess the perturbation introduced in the graph by the anonymization process, we compute the vector of differences for each node between the original and the anonymized graph. Then, we compute the root mean square (*RMS*) to obtain a single value for the whole graph. We calculate the difference of the centrality measures between the original and the anonymized graph as follows:

$$\epsilon(G, \widetilde{G}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(g_i - \widetilde{g}_i)^2} \tag{9}$$

where g_i is the value of the centrality measure for the node v_i of G, and $\widetilde{g}_i$ is the value of the centrality measure for the node v_i of $\widetilde{G}$. In our experiments we use Equation 9 to compute a value representing the error induced in the whole graph by the anonymization process in the centrality measures.

We also focus on the **largest eigenvalue of the adjacency matrix A** (λ_1) [18] where λ_i are the eigenvalues of A and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. The eigenvalues of A encode information about the cycles of a graph as well as its diameter. The spectral decomposition of A is:

$$A = \sum_i \lambda_i e_i e_i^T \tag{10}$$

where e_i is the eigenvector corresponding to λ_i eigenvalue.

The number of nodes, edges and average degree are not considered parameters to assess anonymization process, since anonymization methods analysed in this work keep these values constant.

V. Specific Information Loss Measures

Variations in the generic graph properties are a good way to assess the information loss but they have their limitations because they are just a proxy to the changes in data utility we actually want to measure. For instance, the average distance or the diameter could remain constant while the topology of the network completely changes at the node level. What we are truly interested in is, given a data mining task at hand, quantify the disparity in the results between performing the task on the original network and on the anonymized one. We chose clustering because it is an active field of research, which provides interesting and useful information in community detection for instance. Therefore, the extracted clusters/communities of nodes are the data utility we want to preserve.

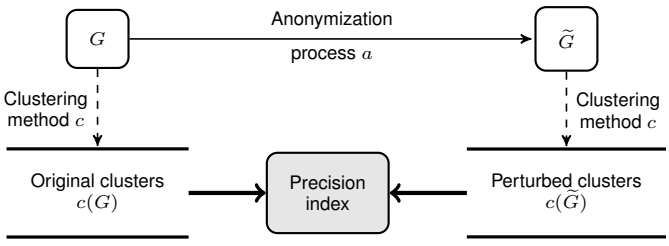

Fig. 2: Framework for evaluating the clustering-specific information loss measure.

A. Clustering

In this work we want to analyse the utility of the perturbed data by evaluating it on different clustering processes. Like generic graph measures, we compare the results obtained both by the original and the perturbed data in order to quantify the level of noise introduced in the perturbed data. This measure is specific and application-dependent, but it is necessary to test the perturbed data in real clustering processes.

We considered the following approach to evaluate the clustering assignment made by a given clustering method c using a particular graph perturbation method a: (1) apply a to the original data G and obtain $\widetilde{G} = a(G)$; (2) apply c to G and $\widetilde{G}$ to obtain the cluster assignments $c(G)$ and $c(\widetilde{G})$; and (3) compare $c(G)$ to $c(\widetilde{G})$, as illustrated in Figure 2. In terms of information loss, it is clear that the more similar $c(\widetilde{G})$ is to $c(G)$, the less information loss. Thus, clustering-specific information loss metrics should measure the divergence between both cluster assignments $c(G)$ and $c(\widetilde{G})$. Ideally, if the anonymization step was lossless in terms of data utility, we should have the same number of clusters with the same elements in each cluster. When the clusters do not match, we need to quantify the divergence.

For this purpose, we used the *precision index* [2]. Assuming we know the true communities of a graph, the precision index can be directly used to evaluate the similarity between two cluster assignments. Given a graph of n nodes and q true communities, we assigned to nodes the same labels $l_{tc}(\cdot)$ as the community they belong to. In our case, the true communities are the ones assigned to the original dataset (i.e. $c(G)$) since we want to obtain communities as close as the ones we would get on non-anonymized data. Assuming the perturbed graph has been divided into clusters (i.e. $c(\widetilde{G})$), then for every cluster, we examine all the nodes within it and assign to them as predicted label $l_{pc}(\cdot)$ the most frequent true label in that cluster (basically the mode). Then, the precision index can be defined as follows:

$$precision\_index(G, \widetilde{G}) = \frac{1}{n}\sum_{v \in G} \mathbb{1}_{l_{tc}(v)=l_{pc}(v)} \tag{11}$$

where $\mathbb{1}$ is the indicator function such that $\mathbb{1}_{x=y}$ equals 1 if $x = y$ and 0 otherwise. Note that the precision index is a value in the range [0,1], which takes value 0 when there is no overlap between the sets and value 1 when the overlap between the sets is complete. To be consistent with the notion of error for the generic graph properties, we report 1 - $precision\_index$ in the results tables so that the lower,

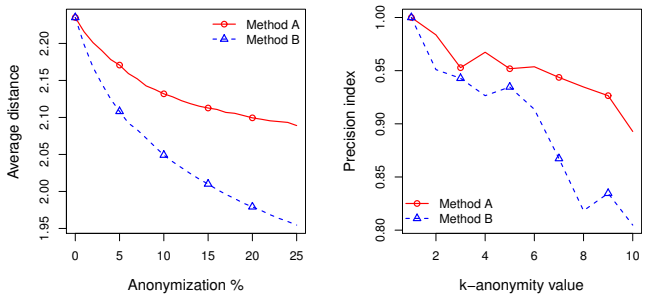

Fig. 3: Examples of our framework results. The horizontal axis presents the anonymization (randomization % or k-anonymity value), while vertical axis indicates the value of the original graph (leftmost point) and the evolution during anonymization processes.

the better.

Regarding the clustering methods c, we propose 4 graph clustering algorithms to evaluate the edge modifications techniques. All of them are unsupervised algorithms based on different concepts and developed for different applications and scopes. An extended revision and comparison of them can be found in Lancichinetti and Fortunato [13] and Zhang et al. [20]. The selected clustering algorithms are:

- *Fastgreedy* (FG) [9], a hierarchical agglomeration algorithm for detecting community structure based on modularity optimization.
- *Walktrap* (WT) [15] that tries to find densely connected sub-graphs, i.e. communities, in a graph via random walks.
- *Infomap* (IM) [16] that optimizes the map equation, which exploits the information-theoretic duality between the problem of compressing data and the problem of detecting significant structures in the graph.
- *Multilevel* (ML) [1], a multi-step technique based on a local optimization of Newman-Girvan modularity in the neighborhood of each node.

Even though some algorithms permit overlapping among different clusters, we did not allow it in our experiments by setting the corresponding parameter to zero, mainly for ease of evaluation.

VI. Application examples

In this section we briefly present some hypothetical results obtained by our experimental framework[1]. The framework expects two input graphs, G and $\widetilde{G}$, and it returns a score error for each GIL metric and a precision score for each selected clustering algorithm.

Usually, researchers are interested in comparing the original graph to a set of anonymous graphs obtained from the original one (i. e. $\widetilde{G}_1, \ldots, \widetilde{G}_p$), by applying different percentages of randomization or different k values in k-anonymity-based algorithms. Not only may this help to understand the behavior of the datasets, but also to choose the best parameter according to privacy and data utility requirements. An example is presented in Figure 3. For instance, the perturbation of average distance on anonymization percentage in range $[0, \ldots, 25\%]$ can be seen in Figure 3a. It is clear to see how this metric evolves during anonymization process. Therefore, it can help us to choose right algorithm's parameters to fulfill data utility and privacy constraints. A similar example is presented in Figure 3b, where the precision score clearly shows the behavior of two different methods over a set of $k \in [1, \ldots, 10]$.

[1]R Source code at: https://bitbucket.org/jcasasr/data-utility-framework/

VII. Conclusions

In recent years several anonymization algorithms have appeared to protect users' privacy. However, it is quite difficult to compare data utility among them, since each work usually uses different measures to compute and evaluate information loss. In this paper we have proposed a framework to evaluate data utility and information loss on privacy-preserving graph data. We claim that some generic information loss measures can be used to compute and evaluate information loss. Nevertheless, metrics related to application-specific real-world problems must be defined and used to compute and compare data utility among methods and algorithms in literature. Our framework provides a standard way to compute both metrics and can be easily used to perform comparisons among graph modification techniques (including random-based and constrained-based methods).

Many interesting directions for future research have been uncovered in this work. It would also be interesting to consider other specific information loss measures, such as those related to information flow or remaining ratio of top influential users. It would be also thought-provoking to extend this analysis to other graph's types (directed or labelled graphs, for instance).

Acknowledgements

This work was partly funded by the Spanish MCYT and the FEDER funds under grant TIN2014-57364-C2-2-R "SMART-GLACIS".

References

[1] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, Vol. 2008(10), P10008.

[2] Cai B-J, Wang H-Y, Zheng H-R and Wang H (2010) Evaluation repeated random walks in community detection of social networks. In: 2010 International Conference on Machine Learning and Cybernetics (ICMLC). IEEE Computer Society, Qingdao, pp 1849-1854

[3] Casas-Roma, J. (2014). Privacy-Preserving on Graphs Using Randomization and Edge-Relevance. In V. Torra (Ed.), International Conference on Modeling Decisions for Artificial Intelligence (MDAI) (pp. 204-216). Tokyo, Japan: Springer International Publishing Switzerland.

[4] Casas-Roma, J., Herrera-joancomartí, J., and Torra, V. (2015). Anonymizing Graphs: Measuring Quality for Clustering. Knowledge and Information Systems (KAIS), Vol. 44(3), pp. 507-528

[5] Casas-Roma, J., Herrera-joancomartí, J., and Torra, V. (2016). k-Degree Anonymity And Edge Selection: Improving Data Utility In Large Networks. Knowledge and Information Systems (KAIS), Vol. 50(2), pp. 447-474

[6] Casas-Roma, J., Herrera-Joancomartí, J., and Torra, V. (2017). A survey of graph-modification techniques for privacy-preserving on networks. Artificial Intelligence Review, 47(3), 341-366. http://doi.org/10.1007/s10462-016-9484-8

[7] Chakrabarti D and Faloutsos C (2006) Graph mining: Laws, generators, and algorithms. ACM Comput Surv 38(1):2:1-2:69

[8] Dwork C (2006) Differential Privacy. In: Proceedings of the 33rd International Conference on Automata, Languages and Programming (ICALP). Springer-Verlag, Berlin, pp 1-12

[9] Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. Physical Review E, Vol. 70(6), 66111.

[10] Girvan M and Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99(12):7821-7826

[11] Hay M, Miklau G, Jensen D, Weis P and Srivastava S (2007) Anonymizing Social Networks. Report, University of Massachusetts Amherst

[12] Hay M, Miklau G, Jensen D, Towsley D and Weis P (2008) Resisting structural re-identification in anonymized social networks. Proc VLDB Endow 1(1):102-114

[13] Lancichinetti, A., and Fortunato, S. (2009). Community detection algorithms: a comparative analysis. Physical Review E, Vol. 80(5), 56117.

[14] Liu K and Terzi E (2008) Towards identity anonymization on graphs. In: Proceedings of the ACM International Conference on Management of Data (SIGMOD). ACM Press, New York, pp 93-106

[15] Pons, P., and Latapy, M. (2005). Computing Communities in Large Networks Using Random Walks. In Computer and Information Sciences (ISCIS), Vol. 10, pp. 284293. Springer Berlin Heidelberg.

[16] Rosvall, M., and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences, Vol. 105(4), pp. 11181123.

[17] Sweeney L (2002) k-anonymity: a model for protecting privacy. Int J Uncertain Fuzziness Knowl Based Syst 10(5):557-570.

[18] Ying X and Wu X (2008) Randomizing Social Networks: a Spectrum Preserving Approach. In: Proceedings of the SIAM International Conference on Data Mining (SDM). SIAM, Atlanta, pp 739-750

[19] Ying X, Pan K, Wu X and Guo L (2009) Comparisons of randomization and k-degree anonymization schemes for privacy preserving social network publishing. In: Proceedings of the 3rd Workshop on Social Network Mining and Analysis (SNA-KDD). ACM Press, New York, pp 10:1-10:10

[20] Zhang, K., Lo, D., Lim, E.-P., and Prasetyo, P. K. (2013). Mining indirect antagonistic communities from social interactions. Knowledge and Information Systems, Vol. 35(3), pp. 553583.

[21] Zhou B and Pei J (2008) Preserving Privacy in Social Networks Against Neighborhood Attacks. In: Proceedings of the 24th International Conference on Data Engineering (ICDE). IEEE Computer Society, Washington, pp 506-515

[22] Zou L, Chen L and Özsu MT (2009) K-Automorphism: A General Framework For Privacy Preserving Network Publication. Proc VLDB Endow 2(1):946-957