Preservation of Structural Properties in Anonymized Social Networks

Traian Marius Truta, Alina Campan
Department of Computer Science
Northern Kentucky University
Highland Heights, KY 41099, USA
{trutat1, campana1}@nku.edu

Anca L. Ralescu
Department of Computer Science
University of Cincinnati
Cincinnati, OH 45221, USA
anca.ralescu@uc.edu

***Abstract*—Social networks such as Facebook, LinkedIn, or Twitter have nowadays a global reach that surpassed all previous expectations. Many social networks gather confidential information of their users, and as a result, the privacy in social networks has become a topic of general interest. To defend against privacy violations, several social network anonymization models were introduced. In this paper, we empirically study how well several structural properties of a social network are preserved through an anonymization process. We first anonymize several real and synthetic social networks using the *k*-anonymous cluster social network model, and then we compare how well structural properties such as diameter, centrality measures, clustering coefficients, and topological indices are preserved between the original networks and their anonymized versions. Our experiments show that there are correlations between the structural properties' values obtained from the original network and from the corresponding anonymized networks. Preserving such graph properties through anonymization might be extremely important / essential for subsequent graph-mining of the anonymized networks.**

***Index Terms*—K-Anonymity, Privacy, Social Networks, Structural Properties.**

I. Introduction and Motivation

Social networks such as Facebook [12], LinkedIn [18], or Twitter [36] have nowadays a global reach that surpassed all previous expectations. Smaller social networks that focus on specialized domains such as sports, games, and technology have also attracted a large number of users in the last years. For instance, FanCru offers sport fans a place to connect and share information [13], Playfire [28] and WeeWorld [39] are social networks that attract online gamers, and Toolbox for IT (Information Technology) is a knowledge-sharing community for IT members [34]. Most Internet users are part of one or more social networks today and they contribute with a wealth of information to these networks.

Many social networks gather confidential information about their users, information that could potentially be misused. For instance, in the healthcare field, PatientsLikeMe [26], a social network with more than 150,000 users as of July 2012, creates communities of patients for various diseases. Due to this amount of sensitive data gathered by social network sites, the privacy in social networks is a concern for many users and the research in this field has flourished in the past several years.

Several research directions in the social networks' privacy field are outlined next.

Backstrom et al. illustrate the shortcomings of the naïve graph anonymization, which replaces the identity of individual nodes by synthetically created identifiers. Two types of attacks, passive and active attacks, are presented in this context [2]. Narayanan and Shmatikov performed a de-anonymization experiment that compromised the privacy of a third of the users who had accounts on both Twitter and Flickr, with a 12% error rate [22].

To defend against privacy attacks, several social network privacy models were introduced. These models can be categorized into graph modification models and clustering-based models.

In the graph modification category, Liu and Terzi's introduced the *k*-degree anonymity model, in which the original social network is modified such that the released social network will have at least *k* nodes with the same degree [19]. Zhou and Pei defined a model called *k*-neighborhood anonymity, in which each node must have *k* others nodes with the same 1-neighborhood characteristics [43]. Edge additions and/or deletions are performed in order to satisfy both *k*-degree anonymity and *k*-neighborhood anonymity. Zou et al. assume a more powerful adversary and their model, titled *k*-automorphism anonymity, requires that each node from the social network is unindistinguisable from other *k*-1 nodes with respect to any subgraph in which the node belongs [45]. Two other models, named *k*-symmetry [8] and *k*-isomorphism [40], are similar to *k*-automorphism. The social networks that satisfy one of these three models are created via a process of both node- and edge- additions / deletions. It is not well understood how the graph structure is preserved during the anonymization process, and this represents a significant limitation of the graph modification techniques.

In the clustering-based category, Campan and Truta introduced the *k*-anonymous clustered social network model, in which nodes are grouped together in clusters and super-nodes and super-edges are created [6]. This clustering-based approach to social network anonymity is briefly presented in Section 2 of this paper. Its full presentation can be found in [6]. Related clustering approaches were presented in [3, 17, 41].

The research in social network privacy extends beyond the presented privacy attacks and defenses. Recent survey of this area can be found in [42, 44].

In this paper we empirically study how well several structural properties of a social network are preserved during an anonymization process. We first anonymize several real and synthetic social networks using the *k*-anonymous cluster social network model, and then we compare how well structural properties such as diameter [16], centrality measures [14], clustering coefficients [37, 38] and topological indices (also known as graph theoretical invariants) [21] are preserved between the original networks and their anonymized version.

There has been some preliminary work in assessing structural property preservation in anonymized social networks. In the only work that considers a *k*-anonymous clustered social network model, the structural properties considered do not include clustering coefficients and topological indices [35]. In addition to analyzing new structural properties, our approach is novel since we do not compute the structural properties values directly on the anonymized graph as in [35] (which have a reduced number of super-nodes and super-edges, thus generating less conclusive results); instead, we generate for each anonymized graph, a subset of possible graphs that match the anonymized graph structure, and we compute the structural properties' values as the average of the corresponding values obtained for each "de-anonymized" graph. A second related work considers most of the structural properties from this paper except the topological indices measures [31]. It is worth mentioning that in the above mentioned paper the authors focus on a graph modification approach, *k*-automorphism, and they do not address any clustering-based anonymization model. Other less related works that analyze a limited number of structural properties were performed only for graph modification approaches such as *k*-isomorphism [8] and *k*-symmetry [40].

The remaining of this paper is structured as follows. Section 2 summarizes the clustering-based social network privacy model. Section 3 presents the structural properties that we study in our experiments. Section 4 describes our experiments, and presents our preliminary findings. The paper ends with future work directions and conclusions.

II. Social Network Anonymization Model

In this section we succinctly present an adaptation of the *k*-anonymous clustered social network model [6]. Since in this paper our focus is on preservation of the social networks' structural properties, we make the additional simplifying assumption that the nodes in the social network do not have quasi-identifier attributes (example of such quasi-identifier attributes are *Age* and *ZipCode*; they may be used to discover the identity of the nodes); accordingly, the anonymization process is based on the social network structure only. The nodes in the social network still have sensitive attribute values that need to be protected from potential intrudes (example of such sensitive attributes are *Diagnosis* and *Income*). For details on how this model approaches the more general problem of anonymizing networks where individual nodes also have quasi-identifier attributes, please refer to [6].

Consider an initial social network modeled as a simple undirected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ is the set of nodes and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ is the set of edges. Only binary relationships are allowed in this model. Additionally, all relationships are of the same type and they are represented as unlabeled undirected edges. These relationships (or at least a subset) are assumed to be known by an intruder, thus they are similar to a quasi-identifier attribute. Using this known graph structure, an intruder is able to identify individuals and to reveal their sensitive information due to the uniqueness of the neighborhoods of various individuals.

Using a grouping strategy, one can partition the nodes from this network into pairwise disjoint clusters. Clusters can then be generalized to super-nodes, which may be connected by super-edges. The goal of this process is to make any two nodes coming from the same cluster indistinguishable based on their relationships. To achieve this objective, Campan and Truta developed intra-cluster and inter-cluster edge generalization techniques that were used for generating super-nodes and super-edges, and so generalizing the social network structure. The definition of such a clustered anonymized graph is presented next [6].

Definition 1. (***anonymized social network***): Given an initial social network, modeled as a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, and a partition $\mathcal{S} = \{c_1, c_2, \ldots, c_v\}$ of the node set $\mathcal{N}$, $\bigcup_{j=1}^{v} c_j = \mathcal{N}$, $c_i \cap c_j = \emptyset$; $i, j = 1..v$, $i \neq j$; the corresponding ***anonymized social network*** $\mathcal{AG}$ is defined as $\mathcal{AG} = (\mathcal{AN}, \mathcal{AE})$, where:

- $\mathcal{AN} = \{C_1, C_2, \ldots, C_v\}$, C_i is a node corresponding to the cluster $c_j \in \mathcal{S}$ and is described by the intra-cluster generalization pair $(|c_j|, |\mathrm{E}_{cj}|)$, where $|c_j|$ is the number of nodes in the cluster C_i and $|\mathrm{E}_{cj}|$ is the number of edges that exist in $\mathcal{G}$ between nodes belonging to c_j;
- $\mathcal{AE} \subseteq \mathcal{AN} \times \mathcal{AN}$; $(C_i, C_j) \in \mathcal{AE}$ iff $C_i, C_j \in \mathcal{AN}$ and $\exists$ $X \in c_j$, $Y \in c_j$, such that $(X, Y) \in \mathcal{E}$. Each generalized edge $(C_i, C_j) \in \mathcal{AE}$ is labeled with the inter-cluster generalization value $|\mathcal{E}_{ci,cj}|$, which represents the number of edges with one end in c_i and the other in c_j.

Based on this definition, all nodes from a cluster c are collapsed into the super-node C and are indistinguishable from each other. To satisfy the *k*-anonymous clustered model – model derived from the well-known *k*-anonymity property for microdata [30, 32], each cluster must have at least *k* nodes.

Definition 2. (***k-anonymous clustered social network***): An anonymized social network $\mathcal{AG} = (\mathcal{AN}, \mathcal{AE})$, where $\mathcal{AN} = \{C_1, C_2, \ldots, C_v\}$, and $C_j = [(|c_j|, |\mathrm{E}_{cj}|)]$, $j = 1, \ldots, v$ is *k*-anonymous iff $|c_j| \geq k$ for all $j = 1, \ldots, v$.

Based on the social network $\mathcal{G}_{ex}$ depicted in Figure 1, we illustrate, in Figure 2, a possible 3-anonymous clustered social network $\mathcal{AG}_{ex}$.

The algorithm used in the anonymization process, called the *SaNGreeA* (Social Network Greedy Anonymization) algorithm, performs a greedy clustering processing of an initial social network in order to generate a *k*-anonymous clustered social network. In this algorithm the nodes that are more similar in

terms of their neighborhood structure are clustered together using a greedy approach. To do so, a measure that quantifies the extent to which the neighborhoods of two nodes are similar with each other is used. Full descriptions of this measure and of the *SaNGreeA* algorithm are presented in [6].

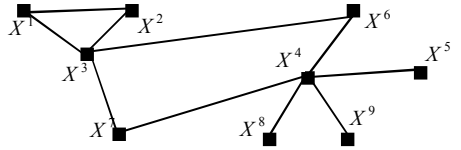

Figure 1. The Social Network $\mathcal{G}_{ex}$.

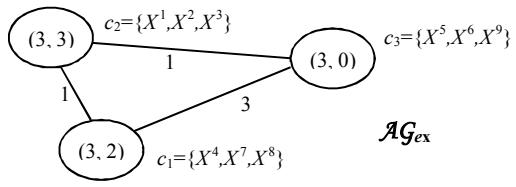

Figure 2. The 3-anonymous clustered social networks $\mathcal{AG}_{ex}$.

III. STRUCTURAL PROPERTIES

There is a wide array of structural properties or measures that characterize the structure of a social network. A good survey that includes most of the structural properties that we considered in this paper (diameter, centrality measures, and clustering coefficients) is in [9]. In addition, the topological indices are summarized in [21]. In this section we briefly present all the structural properties of social networks that we consider in our experiments. We will use the terms "graph" and "social network" interchangeably.

Let $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ be an undirected graph (that represents a social network), where $\mathcal{N}$ is the set of nodes (the cardinality of $\mathcal{N}$, $|\mathcal{N}| = n$) and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ is the set of edges (the cardinality of $\mathcal{E}$, $|\mathcal{E}| = m$).

A. Diameter

In order to define the diameter of a graph we introduce first the concepts of distance between two nodes and eccentricity.

The *distance* between two nodes in a graph is the number of edges in a shortest path connecting them.

The *eccentricity of the node v* is the maximum distance from *v* to any node. That is, $\varepsilon(v) = \max\{d(v, w) \mid w \in \mathcal{N}\}$.

The *diameter of G* is the maximum eccentricity among the nodes of $\mathcal{G}$ (the longest shortest path). That is, *diameter*(G) = $\max\{\varepsilon(v) \mid v \in \mathcal{N}\}$.

B. Centrality Measures

Freeman introduced three centrality measures namely *degree*, *betweenness*, and *closeness centrality* [14]. These measures are computed for each node in a network. For the entire network, Freeman introduced *centrality* (also known as centralization) *measures* for a network that assess how central its most central node is compared to all the other nodes. The network centrality measures calculate the sum of the differences in centrality between the most central node in a network and all other nodes, divided by the theoretically largest such sum of differences in any network with the same number of nodes [14]. We present next the degree, betweenness, and closeness centrality measures for both a node and a network.

The *degree centrality of a node v* is the number of edges adjacent to the node (degree) normalized to the interval [0, 1]. Thus, $C_D(v) = \frac{\deg(v)}{n-1}$. The larger the degree centrality of a node *v*, the stronger its communication potential; alternatively, the lower the degree centrality, the more peripheral the node is perceived.

The *degree centrality of* $\mathcal{G}$ is defined as follows: $C_D(G) = \frac{\sum_{i=1}^{n}[C_D(v*)-C_D(v_i)]}{n-2} = \frac{\sum_{i=1}^{n}[deg(v*)-deg(v_i)]}{(n-1)\cdot(n-2)}$, where $v^* = \text{argmax}\{C_D(v) \mid v \in \mathcal{N}\}$ is the node that has the maximum degree centrality from all nodes from $\mathcal{G}$.

The *betweenness centrality of a node v* is the normalized sum of the number of shortest paths between any pair of nodes (except the considered node) going through the node, divided by the number of shortest paths between any pair of nodes. In other words, $C_B(v) = \frac{2\cdot\sum_{s\neq v\neq t\in N}\frac{\sigma_{st}(v)}{\sigma_{st}}}{(n-1)\cdot(n-2)}$, where $\sigma_{st} = |\{p \mid p$ is a shortest path from *s* to *t*$\}|$, is the number of shortest paths from *s* to *t*, and $\sigma_{st}(v) = |\{p \mid p$ is a shortest path from *s* to *t*, $v \in p\}$ is the number of shortest paths from *s* to *t* that pass through the node *v*. This measure expresses a node's potential for control of communication.

The *betweenness centrality of* $\mathcal{G}$ is defined as follows: $C_B(G) = \frac{\sum_{i=1}^{n}[C_B(v*)-C_B(v_i)]}{n-1}$, where $v^* = \text{argmax}\{C_B(v) \mid v \in \mathcal{N}\}$, is the node of maximum betweenness centrality in $\mathcal{G}$.

The *closeness centrality of a node v* is defined as the inverse of the average of shortest paths' lengths between the node *v* and all other nodes from $\mathcal{G}$, normalized to [0, 1]. That is, $C_C(v) = \frac{n-1}{\sum_{i=1}^{n} d(v_i,v)}$, where $d(v,w)$ is the length of the shortest path from *v* to *w*. This measure assesses the potential for independent communication of a node, i.e. the extent to which the node can avoid the potential control of others.

The *closeness centrality of* $\mathcal{G}$ is defined as follows: $C_C(G) = \frac{\sum_{i=1}^{n}[C_C(v*)-C_C(v_i)]}{(n-1)\cdot(n-2)/(2n-3)}$, where v^* is the node of maximum betweenness centrality in $\mathcal{G}$.

As already mentioned above, for all three network centrality measures of $\mathcal{G}$, the denominators are computed based on the maximum possible sum of differences in the corresponding node centrality for a graph of *n* nodes, that is, $max\sum_{i=1}^{n}[C_X(v^*) - C_X(v_i)]$, where *X* represents degree (*D*), betweenness (*B*), and closeness (*C*). More details about these measures can be found in [14].

C. Clustering Coefficients

Luce and Perry introduced the family of measures called clustering coefficients to describe the likelihood that any node *w* in the neighborhood of node *v* is also adjacent to other nodes in *v*'s neighborhood [20]. The neighborhood of a node *v* represents all nodes that are connected with *v*. We present next two such global clustering coefficients.

Watts and Strogatz define first a local clustering coefficient as follows [38]. The *local clustering coefficient of a node v* ($LCC_1(v)$) is the ratio of actual edges between *v*'s neighbors, and all possible edges between its neighbors. Thus, $LCC_1(v) = \frac{2\cdot|E(G_1(v))|}{\deg(v)\cdot(\deg(v)-1)}$, where $|E(G_1(v))|$ denotes the number of edges between nodes in the 1-neigborhood of node *v*.

The *Watts-Strogatz clustering coefficient* (also known as the *global clustering coefficient*) is defined as the mean of all local clustering coefficients ($WS\_CC(G) = \frac{\sum_{i=1}^{n} LCC_1(v_i)}{n}$).

An alternative clustering coefficient, the *network clustering coefficient* (*N_CC*) or *transitivity* is defined as the number of closed 2-paths divided to the number of 2-paths in the graph [23]. In Figure 1, examples of closed 2-paths are (X^1, X^2, X^3), (X^2, X^1, X^3), and any other combination of these three nodes. (Unclosed) 2-paths examples include (X^3, X^6, X^4) and (X^6, X^4, X^9).

D. Topological Indices

Topological indices are mainly used in chemical graph theory and pharmacology. Good surveys of such indices can be found in [10, 33]. More recently, these topological indices are being used in social networks, in particular for community analysis [1].We briefly describe next four topological indices.

The *Zagreb group index 1* (M_1) is defined as the sum of the squared node degrees [15]. That is, $M_1 = \sum_{i=1}^{n}[\deg(v_i)]^2$.

The *Zagreb group index 2* (M_2) is defined as the sum of the products of the degrees of pairs of adjacent nodes [15]. Thus, $M_2 = \sum_{(i,j)\in E} \deg(v_i)\cdot\deg(v_j)$.

The *Randic connectivity index* (*Xr*) is also defined based on the degree of nodes [29]. It is computed as follows: $Xr = \sum_{(i,j)\in E} \frac{1}{\sqrt{\deg(v_i)\cdot\deg(v_j)}}$.

The *Platt index* (*F*) is computed by summing for each edge the number of its adjacent edges [27]. Thus, $F = \sum_{(i,j)\in E}\big(\deg(v_i) + \deg(v_j) - 2\big)$.

E. An Illustration

Using the social network G_{ex} from Figure 1 we illustrate the values of all structural properties described in this section.

TABLE I. STRUCTURAL PROPERTIES VALUES FOR G_{EX} SOCIAL NETWORK

Structural Property	Value
Diameter	4
Degree Centrality	0.44642857
Betweenness Centrality	0.55133929
Closeness Centrality	0.34419152
Watts-Strogatz Clustering Coefficient	0.36111111
Network Clustering Coefficient	0.15000000
Zagreb Group Index 1	60
Zagreb Group Index 2	71
Randic Connectivity Index	3.88831
Platt Index	40

IV. EXPERIMENTS

We study the above illustrated structural properties on the original and de-anonymized versions of several real and synthetic datasets. These datasets are described next.

The **Enron** dataset is a network of e-mail exchanges available online at [11]. A node in this network represents an e-mail address. An edge exists between two nodes if at least one e-mail was sent from one node to the other node from that edge. This network has 36,692 nodes and 183,831 edges.

The **Random1** and **Random2** datasets are synthetically generated using the Erdos-Renyi random network model [4] from the social network analysis program Pajek [24]. For Random1 we used as input parameters for the social network generator 10,000 nodes and an average vertex degree of 20. Since multiple edges between the same nodes are reduced to single edges for conforming to the social network model assumed by the *k*-anonymous clustered social network model, the final average degree slightly dropped and the resulting social network had 10,000 nodes and 99,945 edges. The Random2 dataset had as input parameters 10,000 nodes and 200 as average node degree. After multiple edges' elimination, the final network had 10,000 nodes and 995,011 edges.

The **ScaleFree** dataset is an undirected network generated based on the scale free model introduced in [25]. This approach models real world social networks that follow a power-law degree distribution [5]. We generated this dataset using Pajek with the following initial parameters: the number of nodes = 10,000, average degree of nodes = 33, the number of nodes in the initial Erdos-Renyi graph = 10, probability of edges in the initial Erdos-Renyi graph $\alpha = 0.2$, $\beta = 0.4$, $\gamma = 0.4$. Details about how this network is generated based on above mentioned parameters can be found in [25]. The generated graph has a significant number of multiple edges (more than 60,000) which are eliminated in a post-processing step. This final scale free network that we used in our experiments has 10,000 nodes and 100,657 edges.

The last dataset, labeled **RMAT**, is based on the R-MAT model introduced in [7]. We implemented an R-MAT graph generator that takes the number of nodes (*n*), the average node degree (*avg_deg*), and four probabilities as input parameters. The location of each edge is determined based on a recursive algorithm that divides the adjacency matrix into 4 equal-sized partitions and the edge location is probabilistically selected in

one of the 4 partitions, based on the four probability parameters (we used the values 0.45, 0.15, 0.15, and 0.25 for RMAT dataset generation). Once a partition is decided, it is again divided into four sub-partitions until there will be only one cell from the adjacency matrix left in the partition. If this cell has value 1 (an edge exists in that location), this procedure is repeated from the beginning (multiple edges between the same pair of nodes are not allowed in our graph model). This approach also models real-world graphs that follow power-law degree distributions [7]. More details about this algorithm can be found in [7].

We depict in Figure 3 the flow of our experiments. This framework consists of five steps.

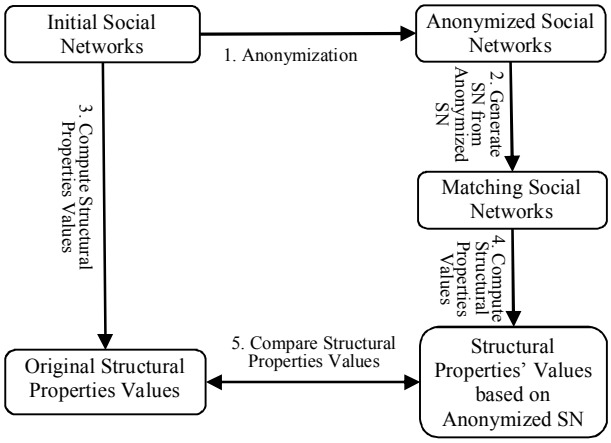

Figure 3. General framework of the experiments.

We start from the initial social networks (Enron, Random1, Random2, ScaleFree, and RMAT) previously described. First, the initial social networks are anonymized into *k*-anonymous clustered social networks as described in Section 2. For each dataset we used the following values for *k*: 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, and 50.

Second, from each *k*-anonymous clustered social network ten possible de-anonymized social networks are generated. In this process we generate edges randomly (uniform probability of selecting any pair of nodes in the super-node) within each cluster until the number of generated edges is equal to the number of edges recorded in the super-node description. This process continues with the generation of edges between super-nodes. These edges are also generated randomly (uniform probability of selecting any pair of nodes that connects two specified super-nodes) until the number of generated edges is equal with the number of edges that describes the corresponding super-edge between the two super-nodes. This process guarantees that each generated "de-anonymized" social network will have the same number of nodes and edges as the original network. We decided to generate ten networks for each anonymized social network to avoid any possible outliers.

We used Java to implement the anonymization algorithm as well as the social network de-anonymization process described above for Step 2.

In Steps 3 and 4, we compute the structural properties' values for the original and de-anonymized social networks. Since we generated ten social networks for each anonymized network, we report the average for each structural property. To compute all structural properties' values we use Pajek [24].

Last, we compare the structural properties values measured for the original social network with the ones obtained from the anonymized networks. The results are shown in Figures 4 – 15. In each figure, the vertical axis shows either the actual values of a structural property, or the ratio between the values measured for the de-anonymized network and the original network. The reason we chose to report the values or the ratio is due to the fact that the values can be very different between the five considered datasets and the representation of all the values is difficult to include in one chart. On the other hand, for some structural properties (such as diameter) reporting the values provides more information. On the horizontal axis we show the various values of *k* (2, 4, 6, 8, 10, 15, 20, 25, and 50) for which we report the graph properties' values and ratios on the *y* axis. *k*=1 represents the original social network.

Figure 4 shows the diameter values for all datasets. We notice that the diameter value is 5 for the original Random1 dataset and all corresponding de-anonymized datasets. In the same way, the diameter value 3 is obtained for all datasets that correspond to the Random2 network. While not identical, the diameter values are also preserved in the same range for ScaleFree and RMAT datasets. For the Enron network, the results are not as easy to interpret. We believe that the spike obtained for $k = 8$ is due to particularities of this real social network. The network contains small well-connected groups of nodes (of average size close to 8) which are only weakly connected to other such groups. The connections between groups are normally realized through very central nodes, with high betweenness centrality. Nodes in well-connected groups are similar and therefore the anonymization algorithm clusters them together in super-nodes; this leads to weakly connected super-nodes. When the de-anonymization algorithm reconnects pairs of nodes from the de-anonymized clusters, it will miss re-connecting the very central nodes that originally connected the clusters. Instead, it has great chances of picking nodes with lower betweenness centrality; this will certainly have a direct effect on the network diameter, which will almost double. Once the super-nodes get larger, they combine several internally-well-connected, but weakly inter-connected groups; the de-anonymization algorithm will however connect pairs of nodes in the super-nodes regardless of their origin (from the same well-connected group, or from two different weakly inter-connected groups).As a result, the diameter value is decreasing.

Figures 5 and 6 show the degree and betweenness centralities ratios. For Random1 and Random2 datasets the values obtained for de-anonymized networks are, in general, slightly greater that the original values. For the other three datasets the centrality values for the de-anonymized networks are less than the original values. As explained before, these results are due to Step 2, where we use a uniform random approach to generate inter-cluster and intra-cluster edges.

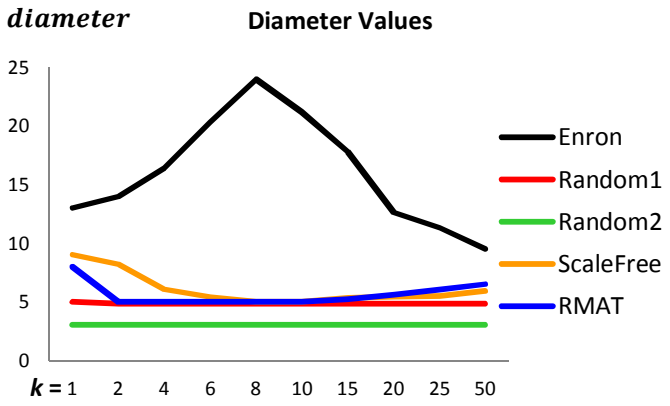

Figure 4. Diameter values.

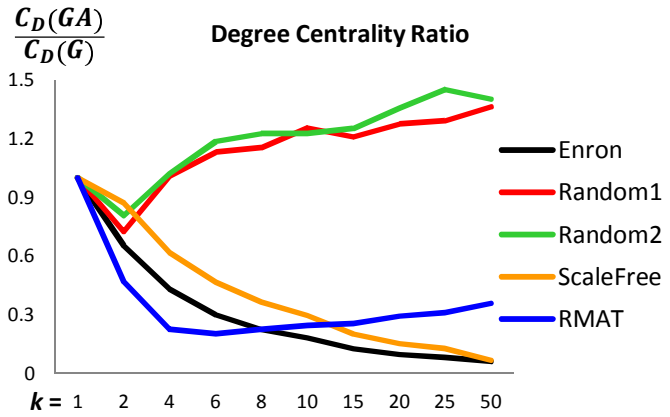

Figure 5. Degree centrality ratio.

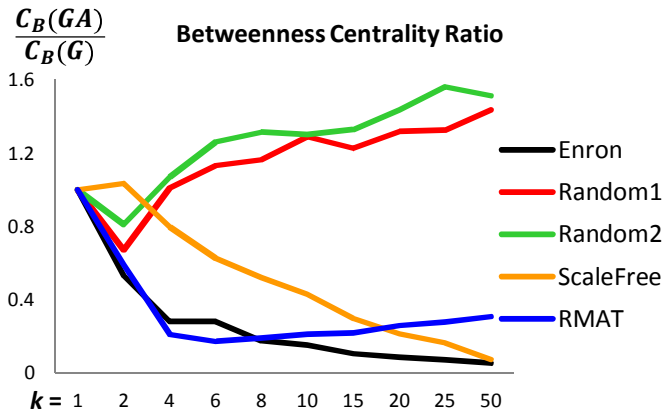

Figure 6. Betweenness centrality ratio.

We report in Figure 7 the closeness centrality values for Random1 and Random2 datasets. These are the only two datasets that are connected and this measure is computed only for connected graphs. Similar to degree and betweenness centrality, the closeness centrality values are well-preserved between the original and de-anonymized versions of these two datasets.

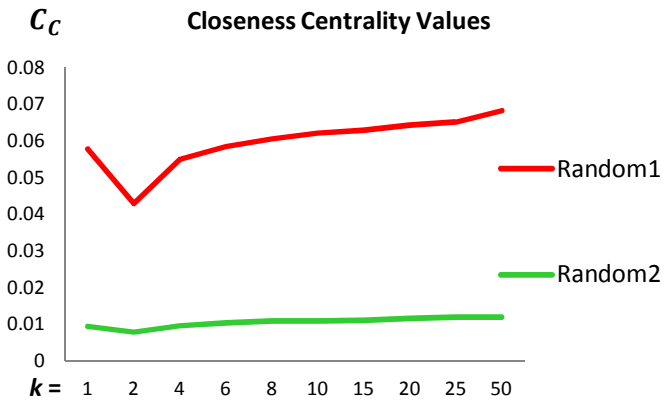

Figure 7. Closeness centrality ratio.

Figures 8-11 show the values and ratios for the Watts-Strogatz clustering coefficient and for the network clustering coefficient. Again we notice that the values are well preserved for both Random1 and Random2 datasets. Enron and ScaleFree networks have smaller values for both clustering coefficients when k increases. This decrease is almost linear with a high slope. For RMAT dataset the decrease is only until k reaches the 4, then the clustering coefficients values are almost the same for k greater than 4.

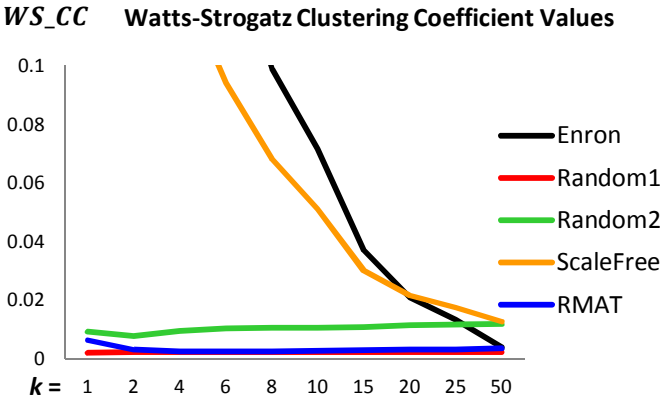

Figure 8. Watts-Strogatz clustering coefficient values.

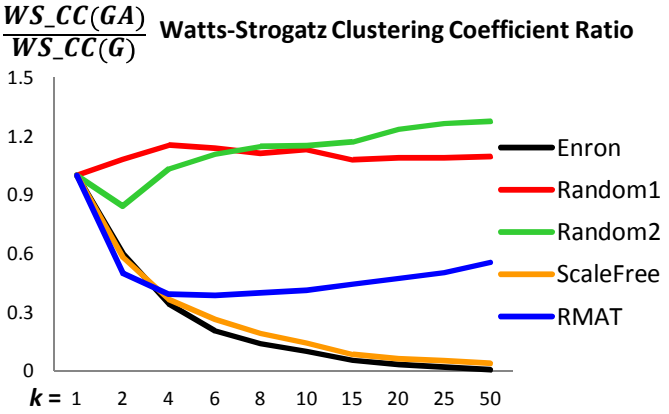

Figure 9. Watts-Strogatz clustering coefficient ratio.

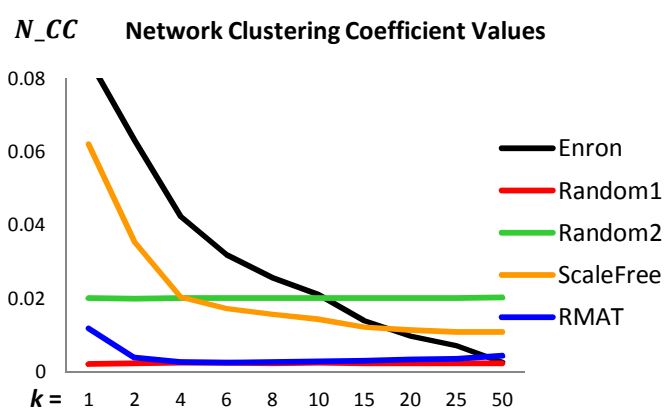

Figure 10. Network clustering coefficient values.

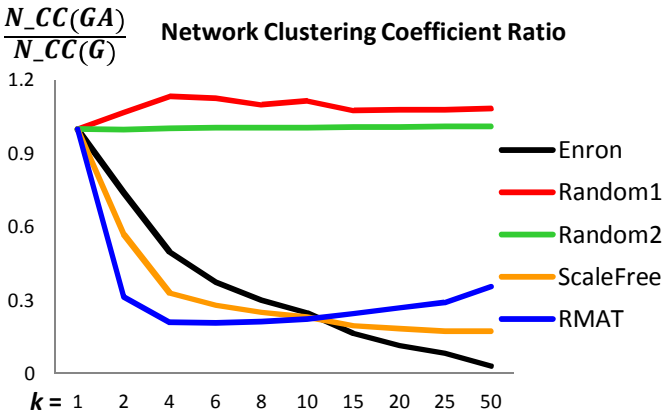

Figure 11. Network clustering coefficient ratio.

Figures 12-15 show the ratios for topological indices. We notice that the results for Zagreb group index 1, Zagreb group index 2, and Platt index are similar with the results obtained for clustering coefficients. These topological indices are well preserved for Random1 and Random2 datasets. For the RMAT dataset there is an initial drop in value (until approximately 0.7 from the original value) followed by a small increase (to approximately 0.8 from the original value). Enron and ScaleFree have a continuous decrease of the topological index values. For Randic connectivity index the results, as expected, are mirrored compared with the other three indices (this is due to the product of the degrees being part of the denominator).

Our experiments show that the structural properties are well preserved for the datasets that were generated using the Erdos-Renyi random network model. We expect this is also true for other random network models. It is worth noting that while the structural properties' values are significantly altered for k-anonymous clustered social networks with large values of k, in particular for Enron and ScaleFree datasets, if a researcher has the additional knowledge of the model that the social network follows, then the original value, or more specifically the range for it, can be estimated with a good probability.

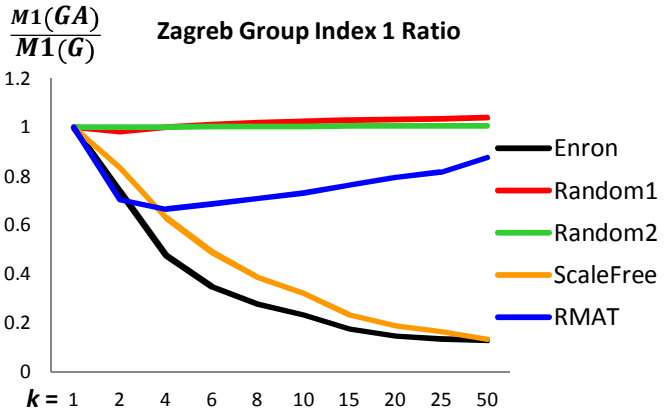

Figure 12. Zagreb group index 1 ratio.

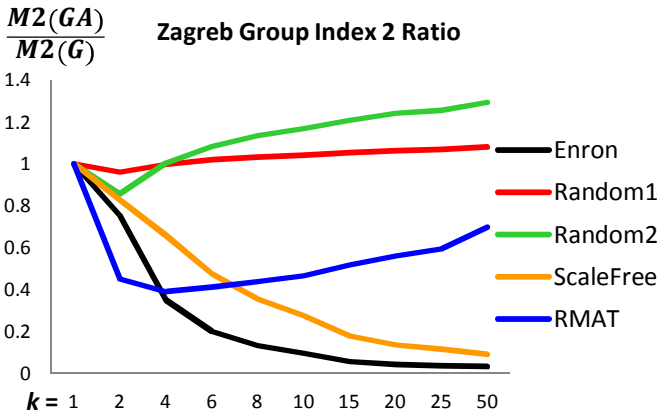

Figure 13. Zagreb group index 2 ratio.

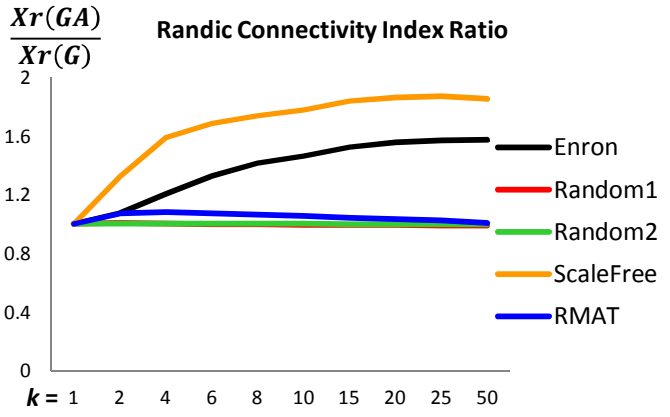

Figure 14. Randic connectivity index ratio.

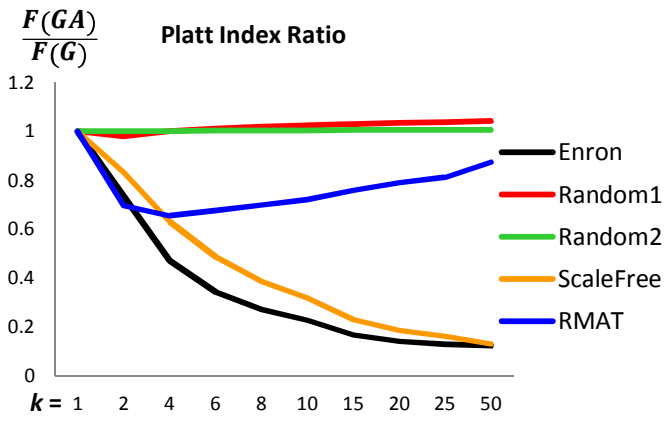

Figure 15. Platt index ratio.

It is also worth comparing our results and conclusions with the authors of [35] and [31]. In [35], the same anonymization model was used and the centrality measures were computed on the anonymized graph directly. Their results are more unpredictable and they concluded that "….experiments show a weak correlation between the anonymization level (the *k* value) of a graph and the centrality measures" [35]. Our results show that such correlations might exist and are based on the network model. The reason of obtaining "better" results than those in [35] is the use of de-anonymized graphs (see Step 2 of our experimental framework) instead of the anonymized graphs, for computing structural properties values. In [31], the anonymization model used was *k*-automorphism. Their conclusion was "This comprehensive set of experiments on graphs from real social networks demonstrates that utility metrics are significantly impacted by *k*-automorphism anonymization" [31]. Our results show significantly better results due to the selected model.

V. Conclusions and Future Work

In this paper we empirically studied how well several structural properties of a social network such as diameter, centrality measures, clustering coefficients, and topological indices are preserved during an anonymization process. Our experiments show that these structural properties are well preserved for datasets that were generated using the Erdos-Renyi random network model. In addition, for networks that follow a power law degree distribution, if a researcher has the additional knowledge of the social network degree distribution, then the original value (more precisely, its range), can be estimated with a high probability. The experiments described in this paper are the first to experimentally show a correlation between the structural properties computed for an original network and its corresponding de-anonymized networks.

There are two directions that we plan to investigate in the future. First, we plan to study how well other anonymization models, in particular graph modification models, preserve social networks structural properties. Second, we plan to explore in depth the conditions on the degree distribution of the original network under which its structural properties are preserved in its de-anonymized networks. It is likely that pursuing these directions will lead to other research avenues that will increase our knowledge regarding the trade-off between privacy and utility in social networks.

Acknowledgment

The authors would like to thank Fang-Yu Rao, Chenyun Dai, and Elisa Bertino for updates in the implementation of the *SaNGreeA* algorithm.

References

[1] A. Arenas, L. Danon, A. Díaz-Guilera, P. M. Gleiser, and R. Guimerá, "Community Analysis in Social Networks," *The European Physical Journal B*, Vol. 38, Number 2, pp. 373-380, 2004.

[2] L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography," in *Proc. WWW'07*, pp. 181-190, 2007.

[3] S. Bhagat, G. Cormode, B. Krishnamurthy, and D. Srivastava, "Class-based Graph Anonymization for Social Network Data," in *Proc. VLDB'09*, pp. 766-777, 2009.

[4] B. Bollobás, *Random Graphs*, 2nd ed., Cambridge University Press, 2001.

[5] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnady, "The Degree Sequence of a Scale-Free Random Graph Process," *Journal of Random Structures and Algorithms*, Volume 18, Issue 3, pp. 279-290, 2001.

[6] A. Campan and T. M. Truta, "Data and Structural K-Anonymity in Social Networks," *Lecture Notes in Computer Science*, Berlin, Germany: Springer, vol. 5456, pp. 33-54, 2009.

[7] D. Chakrabarti, Y. Zhan and C. Faloutsos, "R-MAT: A Recursive Model for Graph Mining," in *Proc. SDM'04*, pp. 442-446, 2004.

[8] J. Cheng, A. W. C. Fu, and J. Liu, "K-Isomorphism: Privacy Preserving Network Publication against Structural Attacks," in *Proc. SIGMOD'10*, pp. 459-470, 2010.

[9] L. Costa, F. Rodrigues, G. Travieso, and P. Boas, "Characterization of Complex Networks: A Survey of Measurements," *Advances in Physics*, vol. 56, no. 1, pp. 167-242, 2007.

[10] J. Devillers and A. T. Balaban (Eds.), *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon &Breach, Amsterdam, 1999.

[11] ENRON Dataset, Available http://snap.stanford.edu/data.

[12] Facebook, Available http://www.facebook.com.

[13] FanCru, Available http://fancru.com.

[14] L. C. Freeman, "Centrality in Social Networks: Conceptual Clarification," *Social Networks*, vol. 1, no. 3, pp. 215-239, 1979.

[15] I. Gutman and N. Trinajstic, "Graph Theory and Molecular Orbitals. Total ¼-Electron Energy of Alternant Hydrocarbons," *Chemical Physics Letter*, Vol. 17, pp. 535–538, 1972.

[16] F. Harary, *Graph Theory*, Addison-Wesley, 1994.

[17] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weiss, "Resisting Structural Re-identification in Anonymized Social Networks," in *Proc. VLDB'08*, pp. 102-114, 2008.

[18] LinkedIn, Available http://linkedin.com.

[19] K. Liu and E. Terzi, "Towards Identity Anonymization on Graphs," in *Proc. SIGMOD'08*, pp. 93-116, 2008.

[20] R. Luce and A. Perry, "A Method of Matrix Analysis of Group Structure," *Psychometrika*, Vol. 14, pp. 95-116, 1949.

[21] I. Lukovits, S. Nikolic, and N. Trinajstic, "On Relationships between Vertex-degrees, Path-numbers and Graph Valence-shells in Trees," *Chemical Physics Letter*, Vol. 354, pp. 417-422, 2002.

[22] A. Narayanan and V. Shmatikov, "De-anonymizing Social Networks," in *Proc. IEEE Security and Privacy*, pp. 173-187, 2009.

[23] M. E. Newman, S. H. Strogatz, and D. J. Watts, "Random Graphs with Arbitrary Degree Distributions and Their Applications," *Physical Review* E, 64:026118, 2001.

[24] W. de Nooy, A. Mrvar, and V. Batagelj, *Exploratory Social Network Analysis with Pajek*, Revised and Expanded Second Edition, Structural Analysis in the Social Sciences, Vol. 34, Cambridge University Press, 2011.

[25] D.M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, and C. L. Giles "Winners don't Take All: Characterizing Competition for Links on the Web," *PNAS*, Vol. 99, No 8, pp. 5207-5211, 2002.

[26] PatientsLikeMe, Available http://www.patientslikeme.com.

[27] J. R. Platt, "Prediction of Isometric Differences in Paraffin Properties," *Journal of Physics Chemistry*, Vol. 56, pp. 328-336, 1952.

[28] Playfire, Available http://playfire.com.

[29] M. Randic, "On Characterization of Molecular Branching," Journal of the American Chemical Society, Vol. 97, pp. 6609–6615, 1975.

[30] P. Samarati, "Protecting Respondents Identities in Microdata Release," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010-1027, 2001.

[31] Y. Song, S. Nobari, X. Lu, P. Karras, and S. Bressan, "On the Privacy and Utility of Anonymized Social Networks," in *Proc. of the iiWAS'11*, Ho Chi Minh City, Vietnam, 2011.

[32] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," *International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems*, vol. 10, no. 5, pp. 557 – 570, 2002.

[33] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000.

[34] Toolbox for IT, Available http://it.toolbox.com.

[35] T. M. Truta, A. Campan, A. Gasmi, N. Cooper, and A. Elstun, "Centrality Preservation in Anonymized Social Networks," in *Proc. of the DMIN'11*, Las Vegas, NE, 2011.

[36] Twitter, Available http://twitter.com.

[37] S. Wasserman and K.Faust, *Social Network Analysis: Methods and Applications*, Cambridge: Cambridge University Press, 1994.

[38] D. J. Watts and S. H. Strogatz, "Collective Dynamics of 'Small-World' Networks," Nature, Vol. 393, pp. 440-442, 1998.

[39] WeeWorld, Available http://weeworld.com.

[40] W. Wu, Y. Xiao, W. Wang, Z. He, and Z. Wang, "K-Symmetry Model for Identity Anonymization in Social Networks," in *Proc. EDBT'10*, pp. 111-122, 2010.

[41] E. Zheleva and L. Getoor, "Preserving the Privacy of Sensitive Relationships in Graph Data," in *Proc. Privacy, Security, and Trust in KDD Workshop*, pp. 153-171, 2007.

[42] E. Zheleva, E. Terzi, and L. Getoor, *Privacy in Social Networks*, Synthesis Lecture on Data Mining Series. Book published by Morgan and Claypool Publishers. 2012.

[43] B. Zhou and J. Pei, "Preserving Privacy in Social Networks against Neighborhood Attacks," in *Proc ICDE'08*, pp. 506-515, 2008.

[44] B. Zhou, J. Pei, and W. S. Luk, "A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data," *SIGKDD Explorations*, vol. 10, no. 2, pp. 12-22, 2008.

[45] L. Zou, L. Chen, and M. T. Ozsu, "K-Automorphism: A General Framework for Privacy Preserving Network Publication," in *Proc. VLDB Endowement*, Vol. 2, Issue 1, pp. 946-957, 2009.