Centrality Preservation in Anonymized Social Networks

Traian Marius Truta, Alina Campan, Ashley Gasmi, Nicholas Cooper, Andrew Elstun

***Abstract*—Social network sites continue to grow in number and size and accumulate information about their members. Among the data provided by members on the social sites they use, there are pieces of sensitive information about themselves. The identity and confidential information about social networks' individual nodes should be protected in all situations, including when the data is made public or released to third parties for analytical tasks. A possible solution to preserve the privacy of individuals is to anonymize the social network data and / or structure, i.e. to modify social network data and structure such that to make several individuals in the network alike, data and neighborhood-wise. Several anonymity definitions and methods to achieve them were introduced in the last few years. Of course, all anonymization approaches aim to preserve as much as possible the data and structural content of the initial social network; the less the inherent informational content is disturbed in the anonymization process, the more accurate are the results obtained by exploring the anonymized social network. Our work aims to study an existing anonymization approach with respect to how it preserves the structural content of the initial social network; specifically, we study how various graph metrics (centrality measures, radius, diameter etc.) change between the initial and the anonymized social network. This study is carried out for a number of synthetic social network datasets.**

I. Introduction and Motivation

The advent of social networks in the last few years created an enormous amount of social network data that could be potentially used for many purposes: for marketing, research, etc. This huge amount of data has created a revitalized interest in social network analysis and mining [1], [18], [20].

Some of the social networks gather individuals' confidential information and/or confidential relationships between individuals. For instance, PatientsLikeMe [24], Rareshare [26], and Daily Strength [12] are social networks in the healthcare field that create communities of patients for various diseases. As a result, privacy in social networks has become a serious concern and the research in this area has flourished in the past few years. Not only the privacy in social networks has become a topic discussed by scientists, but also the large public has shown a vivid interest for this matter. Concerns about privacy with respect to various social networks sites such as Facebook are reported in various media outlets and raise general public awareness about this problem [9]. Yet, the research in the social network privacy area is still very recent, and many problems remain to be solved.

Here are a few research directions in social network privacy.

Attacks in social networks are discussed in several research papers. In one of the early works in this field, Backstrom et al. described two types of attacks: active and passive [2]. An interesting de-anonymization experiment was performed by Narayanan and Shmatikov [23]. They showed that a third of the users who have accounts on both Twitter and Flickr can be re-identified in the anonymous Twitter graph with only a 12% error rate. An inference attack for released social networking data to infer undisclosed private information about individuals is presented in [21].

To defend against privacy attacks, several privacy models, which can be classified as graph modification and clustering-based approaches, were introduced. In the *graph modification approach* category, Liu and Terzi add edges to the original social network so that there are at least *k* nodes with the same degree [22]. Zhou and Pei introduce a stronger requirement: that each vertex must have *k* others with the same *k*-neighborhood characteristics [31]. In order to achieve this property, edge deletions and additions are performed. Other works in this direction include [7], [17], [29]. Unfortunately, it is not clear how well the graph structure is preserved during these graph modification processes, and this represents a major limitation of the graph modification techniques. In the *clustering-based approaches*, vertices and edges are grouped together in clusters and super-nodes and super-edges are created. One clustering-based approach is briefly presented in Section 2, and the full presentation can be found in [5]. Other works in this subarea include [3], [17], [30].

The research in social network privacy extends beyond the privacy attacks and defenses. Anonymization in bipartite graphs is studied in [8]. A relaxation of differential privacy [13] in the context of social networks is presented in [25]. A recent survey of this field can be found in [32].

In this paper the focus is how much data utility is preserved in the anonymized social networks. Specifically, we look at how social networks characteristics such as radius, diameter, and centrality measures [14], [15] are preserved through anonymization.

Other recent papers have also explored utility preservation in anonymized graphs. In our previous work, we introduced a measure of structural information loss that quantifies the probability of error when trying to reconstruct

T. M. Truta (phone: +1-859-572-7551; fax: +1-859-572-5398, e-mail: trutat1@nku.edu), A. Campan (e-mail: campana1@nku.edu), N. Cooper (e-mail: coopern1@mymail.nku.edu), and A. Elstun (e-mail: elstuna1 @mymail.nku.edu) are with the Department of Computer Science, Northern Kentucky University, Nunn Drive, Highland Heights, KY 41099, USA.

A. Gasmi is with the Department of Computer Science, ENSICAEN, 14000 Caen, France (e-mail: ashley.gasmi@ ecole.ensicaen.fr).

the structure of the initial social network from its masked version [5]. In the graph modification approaches, utility preservation was discussed in the context of preserving the same average degree distribution and the same shortest paths length [7], [29].

To our knowledge, no previous work has addressed how graph centrality measures are changed between original social networks and anonymized social networks, neither in graph modification approaches, nor in clustering-based approaches. Moreover, the only work that analyses some graph measures (degree, shortest-paths) was performed only for graph modification approaches such as *k*-isomorphism [7] and *k*-symmetry [29].

The remaining of this paper is structured as follows. Section 2 presents a clustering-based social network privacy model, in particular the concepts of edge generalization and *k*-anonymous masked social network. Section 3 briefly describes various graph measures that we comparatively analyze in our experiments, for original and anonymized social networks. Section 4 describes our experiments, and presents our preliminary findings. The paper ends with future work directions and conclusions.

II. Social Networks Anonymization Model

In this paper we use the social network anonymization model introduced in [5]. We briefly summarize it next.

We consider the 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. Each node represents an individual entity. Each edge represents a relationship between two entities. Usually, the set of nodes, $\mathcal{N}$, is described by a set of attributes that are classified into three categories: *identifier* (such as *Name* and *SSN*), *quasi-identifier* (such as *ZipCode* and *Sex*), and *sensitive* (such as *Primary Diagnosis* and *Income*). In this paper, we focus only on social network structure and therefore we will ignore the node attribute values during the anonymization process. For details about how the node attribute values are used during the anonymization process refer to [5].

We allow only binary relationships in our model. Moreover, we consider all relationships as being of the same type and, as a result, we represent them via unlabeled undirected edges. We also consider this type of relationship to be of the same nature as all the other "traditional" quasi-identifier attributes. We will refer to this type of relationship as the *quasi-identifier relationship*. In other words, the graph structure may be known to an intruder and used by matching it with known external structural information, therefore serving in attacks that might lead to identity and/or attribute disclosure [19].

Using the graph structure, an intruder is able to identify individuals due to the uniqueness of the neighborhoods of various individuals. As shown in [17], when the structure of a random graph is known, the probability that there are two nodes with identical 3-radius neighborhoods is less than 2^{-cn}, where n represents the number of nodes in the graph, and c is a constant value, $c > 0$; this means that the vast majority of the nodes can be uniquely identified based only on their 3-radius neighborhood structure.

To achieve anonymity for social networks, we have adapted the *k*-anonymity model [25], [28]. For social network data, the *k*-anonymity model has to impose both the quasi-identifier attribute and the quasi-identifier relationship homogeneity, for groups of at least *k* individuals. We have also reused the generalization technique for the generalization of node attributes' values [25] and we extended it for edges. To our knowledge, the only equivalent methods for the generalization of a quasi-identifier relationship that exist in the research literature appear in [17], [30] and consist of collapsing clusters of nodes together with their component nodes' structure. Edge additions or deletions are currently used, in all the other approaches, to ensure nodes' indistinguishability in terms of their surrounding neighborhood; additions and deletions perturb to a large extent the graph structure and therefore they are not faithful to the original data. We have employed a generalization method for the quasi-identifier relationship similar to the one exposed in [17], [30], but enriched with extra information, that will cause less damage to the graph structure, i.e. a smaller structural information loss.

Let n be the number of nodes from the set $\mathcal{N}$. Using a grouping strategy, one can partition the nodes from this set into v pairwise disjoint clusters: cl_1, cl_2, …, cl_v. For simplicity we assume that the nodes are not labeled (i.e., do not have attributes), and they can be distinguished only based on their relationships. Our goal is that any two nodes from any cluster to be indistinguishable based on their relationships. To achieve this goal, we introduced an edge generalization process, with two components: *edge intra-cluster* and *edge inter-cluster generalization*.

Edge intra-cluster generalization. Given a cluster cl, let $\mathcal{G}_{cl} = (cl, \mathcal{E}_{cl})$ be the subgraph of $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ induced by cl. In the masked data, the cluster cl will be generalized to (collapsed into) a node, and the structural information we attach to it is the pair of values $(|cl|, |\mathcal{E}_{cl}|)$, where $|cl|$ represents the cardinality of the set cl. This information permits assessing some structural features about this region of the network that will be helpful in some applications. From the privacy standpoint, an original node within such a cluster is indistinguishable from the other nodes in the cluster. At the same time, if more internal information was offered, such as the full nodes' connectivity inside a cluster, the possibility of disclosure would be too high, as discussed in [5].

Edge inter-cluster generalization. Given two clusters cl_1 and cl_2, let $\mathcal{E}_{cl1,cl2}$ be the set of edges having one end in each of the two clusters ($e \in \mathcal{E}_{cl1,cl2}$ if and only if $e \in \mathcal{E}$ and $e \in cl_1 \times cl_2$). In the masked data, this set of inter-cluster edges will be generalized to (collapsed into) a single edge and the structural information released for it is the value $|\mathcal{E}_{cl1,cl2}|$.

This information permits assessing some structural features about this region of the network that might be helpful in some applications and it reduces the disclosure risk.

Given a partition of nodes for a social network $\mathcal{G}$, we are able to create an anonymized graph by using edge intra-cluster generalization within each cluster and edge inter-cluster generalization between any two clusters.

Definition 1. (***anonymized social network***): Given an initial social network, modeled as a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, and a partition $\mathcal{S} = \{cl_1, cl_2, \dots, cl_v\}$ of the nodes set $\mathcal{N}$, $\bigcup_{j=1}^{v} cl_j = \mathcal{N}$; $cl_i \cap cl_j = \varnothing$; $i, j = 1..v$, $i \neq j$; the corresponding ***anonymized social network*** $\mathcal{AG}$ is defined as $\mathcal{AG} = (\mathcal{AN}, \mathcal{AE})$, where:

- $\mathcal{AN} = \{Cl_1, Cl_2, \dots, Cl_v\}$, Cl_i is a node corresponding to the cluster $cl_j \in \mathrm{S}$ and is described by the intra-cluster generalization pair $(|cl_j|, |\mathrm{E}_{clj}|)$;
- $\mathcal{AE} \subseteq \mathcal{AN} \times \mathcal{AN}$; $(Cl_i, Cl_j) \in \mathcal{AE}$ iif $Cl_i, Cl_j \in \mathcal{AN}$ and $\exists$ $X \in cl_j$, $Y \in cl_j$, such that $(X, Y) \in \mathcal{E}$. Each generalized edge $(Cl_i, Cl_j) \in \mathcal{AE}$ is labeled with the inter-cluster generalization value $|\mathcal{E}_{cli,clj}|$.

By construction, all nodes from a cluster *cl* collapsed into the generalized (masked) node *Cl* are indistinguishable from each other.

To have the *k*-anonymity property for a masked social network, we need to add one extra condition to Definition 1, namely that each cluster from the initial partition is of size at least *k*. The formal definition of a masked social network that is *k*-anonymous is presented below.

Definition 2. (***k-anonymous anonymized social network***): An anonymized social network $\mathcal{AG} = (\mathcal{AN}, \mathcal{AE})$, where $\mathcal{AN} = \{Cl_1, Cl_2, \dots, Cl_v\}$, and $Cl_j = [(|cl_j|, |\mathrm{E}_{clj}|)]$, $j = 1, \dots, v$ is *k*-anonymous iff $|cl_j| \geq k$ for all $j = 1, \dots, v$.

Example 1: Suppose the social network $\mathcal{G}_{ex}$ depicted in Figure 1 is given. Two possible 3-anonymous social networks $\mathcal{AG}_{e1}$ and $\mathcal{AG}_{e2}$ are depicted in Figure 2.

The algorithm used in the anonymization process, called the *SaNGreeA* (Social Network Greedy Anonymization) algorithm, performs a greedy clustering processing to generate a *k*-anonymous masked social network, given an initial social network modeled as a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$.

Specifically, *SaNGreeA* puts together in clusters nodes that are as similar as possible in terms of their neighborhood structure. To do so, it uses a measure that quantifies the extent to which the neighborhoods of two nodes are similar with each other, i.e. the nodes manifest the same connectivity properties, or are connected / disconnected among them and with others in the same way.

To assess the proximity of two nodes' neighborhoods, we proceed as follows. Given $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, assume that nodes in $\mathcal{N}$ have a particular order, $\mathcal{N} = \{X^1, X^2, \dots, X^r\}$. The neighborhood of each node X^i can be represented as an *n*-dimensional boolean vector $B_i = (b_1^i, b_2^i, , \dots, b_r^i)$, where the j^{th} component of this vector, b_j^i, is 1 if there is an edge $(X^i, X^j) \in \mathcal{E}$, and 0 otherwise, $\forall j = 1, r$; $j \neq i$. We consider the value b_i^i to be *undefined*, and therefore not equal to 0 or 1. We use a classical distance measure to assess the similarity of vectors of this type: the *symmetric binary distance* [15].

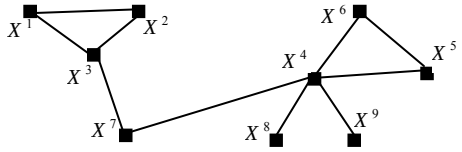
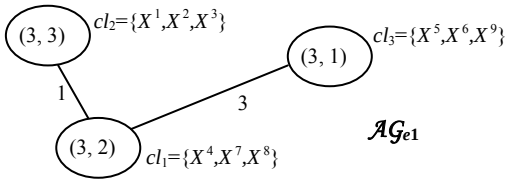


Fig. 1 The Social Network $\mathcal{G}_{ex}$



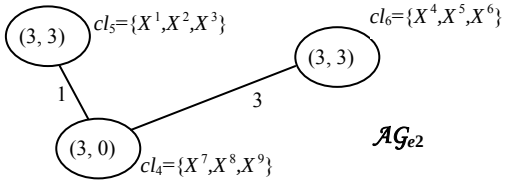


Fig. 2 The 3-anonymous social networks $\mathcal{AG}_{e1}$ and $\mathcal{AG}_{e2}$

Definition 3. (***distance between two nodes***): The ***distance between two nodes*** (X^i and X^j) described by their associated *n*-dimensional boolean vectors B_i and B_j is:

$$dist(X^i, X^j) = \frac{|\{\ell | \ell = 1..r \wedge \ell \neq i, j;\ b_\ell^i \neq b_\ell^j\}|}{r-2}.$$

We exclude from the two vectors' comparison their elements *i* and *j*, which are undefined for X^i and respectively for X^j. As a result, the total number of elements compared is reduced by 2.

In the cluster formation process, our greedy approach will select the closest remaining node to be added to the cluster currently being formed. To assess the structural distance between a node and a cluster we use the following measure.

Definition 4. (***distance between a node and a cluster***): The ***distance between a node*** *X* ***and a cluster*** *cl* is defined as the average distance between *X* and every node from *cl*:

$$dist(X, cl) = \frac{\sum_{Xj \in cl} dist(X, X^j)}{|cl|}.$$

Using the above introduced measures, we explain next how clustering is performed for a given initial social network $\mathcal{G} = (\mathcal{N}, \mathcal{E})$. The clusters are created one at a time. To form a new cluster, a node in $\mathcal{N}$ with the maximum degree and not yet allocated to any cluster is selected as a seed for the new cluster. Then the algorithm gathers nodes to this currently processed cluster until it reaches the desired cardinality *k*. At each step, the current cluster grows with one node. The selected node has to be unallocated yet to any cluster and it will minimize the *dist* measure (see Definition 4).

It is possible, when *n* is not a multiple of *k*, that the last constructed cluster will contain less than *k* nodes. In that case, this cluster needs to be dispersed between the previously constructed groups. Each of its nodes will be added to the cluster that is closest to that node w.r.t. our previously defined distance measure.

A version of the pseudocode of the *SaNGreeA* algorithm that includes node attributes and an additional optimization criterion can be found in [5].

III. Social Network Measures

A variety of social network analyses concentrate on determining how relationships are distributed in a social network between the entities participating in the network. These studies focus on assessing the individual nodes' influence or power in the network. Several graph connectivity and centrality metrics exist that quantify this notion of nodes' influence. Freeman suggested three measures for a node's centrality, as described next [14]. There also are other measures of graph connectivity (radius, diameter) and measures that describe the influence of a node on its network [11]. These social network measures try to capture complex relations between nodes in a network.

In our work, we plan to explore the effect that social network anonymization has on various measures. We investigate if a relationship between such connectivity and centrality measures exists– for the initial social network and for a corresponding anonymized social network. If such measures describing the influence of a node on its network transferred from an original node to its cluster / supernode, then network analysis in various fields (such as viral marketing, communication networks) could be successfully conducted on anonymized networks, while preserving the privacy of individual network nodes. Next, we briefly describe the social network measures analyzed in our experiments.

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

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 **radius of G** is the minimum eccentricity among the nodes of $\mathcal{G}$. Therefore, $radius(G) = \min\{\varepsilon(v) \mid v \in \mathcal{N}\}$.

The **diameter of G** is the maximum eccentricity among the nodes of $\mathcal{G}$. In other words, $diameter(G) = \max\{\varepsilon(v) \mid v \in \mathcal{N}\}$.

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; 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** is the node that has the maximum degree centrality from all nodes from $\mathcal{G}$.

The **betweenness centrality of a node *v*** is the sum of the number of shortest paths between any pair of vertices (except the considered node) going through the node, divided by the number of shortest paths between any pair of vertices. This sum is normalized to [0, 1]. 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 σ_{st} is the number of shortest paths from *s* to *t*, and $\sigma_{st}(v)$ is the number of shortest paths from *s* to *t* that pass through a vertex *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** is the node that has the maximum betweenness centrality from all nodes from $\mathcal{G}$.

The **closeness centrality of a node *v*** is defined as the inverse of the average of shortest paths length between the node *v* and all other nodes from $\mathcal{G}$. This sum is normalized to [0, 1]. In other words, $C_C(v) = \frac{n-1}{\sum_{i=1}^{n} d(v_i,v)}$, where $d(v,w)$ is the length of the shortest path from *v* to *w*. This measure gives the potential for independent communication of a node, or in other words, how much 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 that has the maximum betweenness centrality from all nodes from $\mathcal{G}$.

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

IV. Experiments Design and Results

We designed a series of experiments that allowed us to explore if the proposed graph anonymization algorithm (*SaNGreeA*) preserves some of the graph properties, in particular centrality properties, of social networks. The general framework of our experiments is presented in Figure 3.

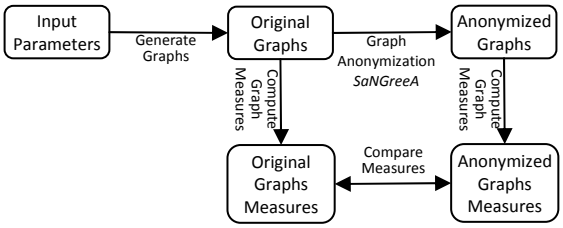


Fig. 3 General framework of the experiments.

We divided our experiment into several phases. In our first phase, labeled Graph Generation in Figure 3, we implemented an R-MAT graph generator [6] and a Random graph generator.

The R-MAT graph generator takes the number of nodes (n), the average node degree (*avg_deg*), and four probabilities as input parameters. The algorithm computes how many edges such a graph has, and for each edge, its location is determined based on the recursive algorithm that divides the adjacency matrix into 4 equal-sized partitions and the location of the edge is probabilistically selected in one of the 4 locations, based on the four probability parameters. Once a partition is found, it is again divided into four sub-partitions until there will be only one location left in the partition. If an edge was already placed on that location, we will repeat this procedure from the beginning (multiple edges between the same pair of nodes is not allowed in our graph model). For all our tests we used the following values for the four probabilities: 0.45, 0.15, 0.15, and 0.25. This choice seems to model better many real-world graphs that follow power-law degree distributions [6]. More details about this algorithm can be found in [6].

The Random graph generator creates a random undirected graph using the Erdos-Renyi model [4]. In this model, each edge is included in the graph with probability p, with the presence or absence of any two distinct edges in the graph being independent. For our generator we use two input parameters: number of nodes (n) and average node degree (*avg_deg*), and we estimate the probability as the *avg_deg* / n. Using this approach, the generated graph will have a slight different average node degree than the input parameter.

We used both graph generator models with various parameter values to create a large number of synthetic graphs on which we performed our experiments. For the number of nodes (n) we used the following values: 10, 25, 50, 75, 100, 250, and 500. For the average node degree (*avg_deg*) we used 2, 3, 4, 5, 8, 10, 25, 50, 75, 100, and 250. Of course, the average node degree is strictly less than the number of nodes (we are not interested in complete graphs in our experiments). Since most of the centrality measures are defined only for connected graphs, for any given combination of input parameters we wanted to generate a connected graph. To achieve this, we generated up to 10,000 graphs and we stopped our graph generator at the first connected graph. In some cases (such as number of nodes = 500, and average node degree = 2) we were not able to generate a connected graph. The list of all generated graphs with the corresponding parameter values is provided in Table I. The total number of generated graphs is 78.

TABLE I
The List of All Generated Graphs

Graph Generator Model	**(*n*, *avg_deg*)**
R-MAT and RANDOM	(10, 2), (10, 3), (10, 4), (10, 5) (25, 2), (25, 3), (25, 4), (25, 5), (25, 8), (25, 10) (50, 3), (50, 4), (50, 5), (50, 8), (50, 10), (50, 25) (75, 4), (75, 5), (75, 8), (75, 10), (75, 25) (100, 4), (100, 5), (100, 8), (100, 10), (100, 25), (100, 50) (250, 5), (250, 8), (250, 10), (250, 25), (250, 50), (250, 100) (500, 8), (500, 10), (500, 25), (500, 50), (500, 100), (500, 250)

In the second phase of this experiment we generated anonymized graphs using the *SaNGreeA* algorithm presented in Section 2. For each generated graph we used various values for k (k as in k-anonymous social network). For n = 10 we used k as 2 and 5; for n = 25, we used k = 2, 5 and 10, and for all other values of n, we used k = 2, 5, 10, 15, and 20. In total 342 anonymized graphs were generated.

In the third phase, we implemented all graph measures described in Section 3. For all 420 graphs (78 generated graphs and 342 anonymized graphs), we computed these graph measures. For an anonymized graph we did not use the weight of an edge between super-nodes, and we considered these graphs as unweighted graphs.

In the last phase of our experiment we compared the original graph measures with the corresponding anonymized graph measures. We are still in the process of analyzing all these results, some preliminary findings are presented next.

Figure 4 shows a sample of the results we obtained for radius and diameter. As expected, both these measures decrease as k increases.

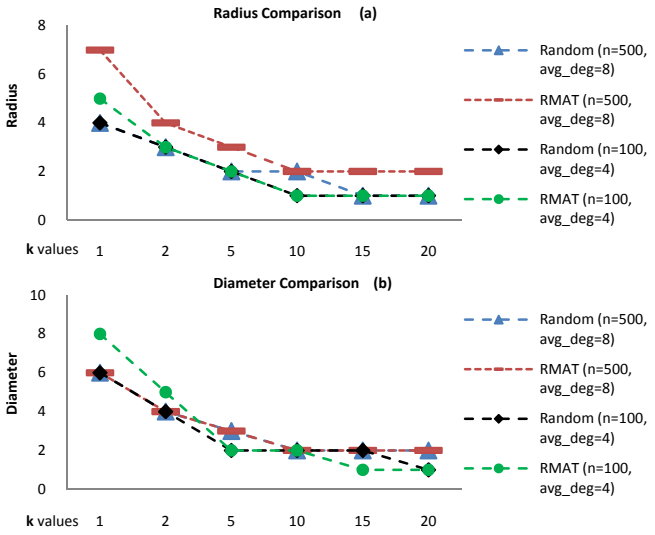


Fig. 4 Radius and diameter values for some of the experiments.

Figure 5 shows partial results with respect to centrality measures. For all measures we report the centrality measure for the anonymized graph divided to the centrality measure for the original graph. The reference value for the original graph is 1 for all three measures. We illustrate these results for four distinct original graphs (2 Random graphs, one with 500 nodes and average node degree 8, and the second one with 100 nodes and average node degree 4, and 2 RMAT graphs with the same number of nodes and average node degrees). For each original graph we created 5 *k*-anonymous graphs, k ∈ {2, 5, 10, 15, 20}.

The degree centrality, illustrated in Figure 5 (a), increases as *k* increases to 5 (for the smaller graphs) or 10 (for the larger graphs) and then decreases. This is due to how *SaNGreeA* algorithm creates clusters. For smaller *k* values, it creates supernodes from nodes highly connected between them and loosely connected to other nodes, which results in lower connectivity between supernodes; this means that the anonymized graph becomes sparser than the original graph. However, when *k* increases, there are not enough similarly connected nodes that could become alone a supernode; as a result, nodes with different connectivity properties are merged into supernodes and the anonymized graph gets closer to the complete graph. We notice the initial increase for degree centrality is steeper for Random graphs than RMAT. This is expected since an original Random graph has a uniform distribution of node degrees.

The betweenness centrality showed in Figure 5 (b) usually decreases for the anonymized graphs. Again, this is because the anonymized graph gets closer to the complete graph as *k* increases, and therefore there are many short paths of length 1. The small increase between $k = 2$ and $k = 5$ is, at the first view, unexpected. This is due to the fact that for small *k* values, the anonymized graph still has variety in supernodes' connectivity, and some of the supernodes gain more control over the shortest paths that exist in the anonymized graph; these nodes have a high betweenness centrality.

The closeness centrality decreases for anonymized graphs when the value of *k* increases as shown in Figure 5 (c). This is again due to the anonymized graph getting closer to the complete graph.

Overall our experiments show a weak correlation between the anonymization level (the *k* value) of a graph and the centrality measures: same changes are observed for graphs of different sizes and with different network properties.

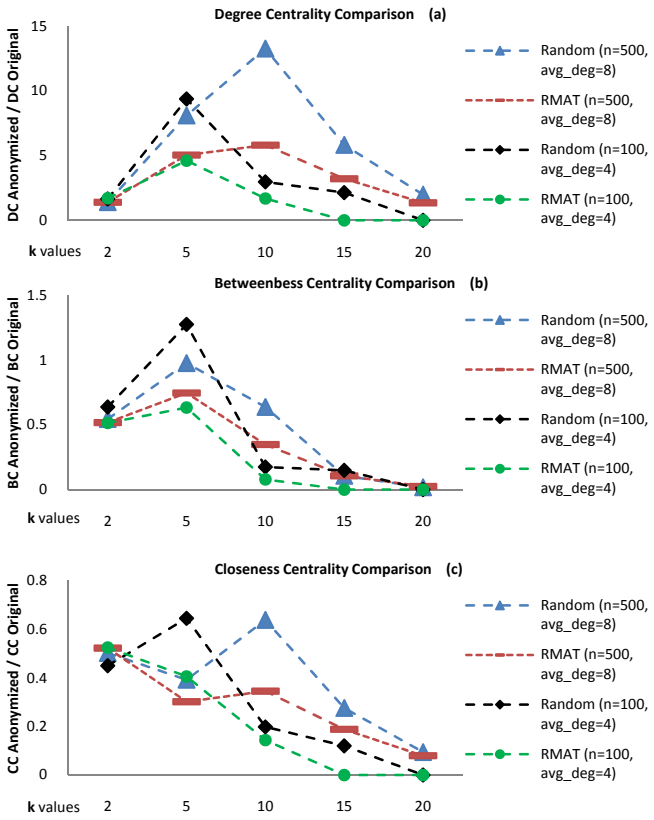


Fig. 5 Centrality measures values for some of the experiments.

V. Conclusions

In this paper we studied a clustering-based anonymization approach with respect to how it preserves the structural content of the initial social network; specifically, we looked at how various graph metrics (centrality measures, radius, diameter etc.) change between the initial and the anonymized social network. Our results showed that there are similarities in how various centrality measures are modified from an original graph to its anonymized versions even if we change the graph size and network properties. We plan to study how other anonymization models behave with respect to centrality measures.

REFERENCES

[1] E.Adar and C.Re, "Managing Uncertainty in Social Networks," *Data Engineering Bulletin*, vol. 30, no. 2, pp. 23-31, 2007.

[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] 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.

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

[7] 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.

[8] G. Cormode, D. Srivastava, T. Yu, and Q. Zhang, "Anonymizing Bipartite Graph Data using Safe Groupings," in *Proc. VLDB'08*, pp. 833-844, 2008.

[9] D. Costa, "Facebook: Privacy Enemy Number One," *PCMag*, Available: http://www.pcmag.com/article2/0,2817,2362967,00.asp, 2010.

[10] 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.

[11] P. Domingos, and M. Richardson, "Mining the network value of customers," in *Proc. KDD'01*, pp. 57-66, 2001.

[12] DS, "Daily Strength," Available: http://www.dailystrength.org, 2006.

[13] C. Dwork, "Differential Privacy: A Survey of Results," *Theory and Applications of Models of Computation*, pp. 1-19, 2008.

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

[15] J. Han and M. Kamber, *Data Mining, Second Edition: Concepts and Techniques*, Morgan Kaufmann, 2006.

[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] J. Kleinberg, "Challenges in Mining Social Network," in *Proc. KDD'07*, pp. 4-5, 2007.

[19] D. Lambert, "Measures of Disclosure Risk and Harm" Journal of Official Statistics, vol. 9, pp. 313-331, 1993.

[20] S. Levine and R. Kurzban, "Explaining Clustering in Social Networks: Towards an Evolutionary Theory of Cascading Benefits," *Managerial and Decision Economics*, vol. 27, pp. 173-187, 2007.

[21] J. Lindamood, R. Heatherly, M. Karantacioglu, and B. Thuraisingham, "Inferring Private Information Using Social Network Data," in *Proc. WWW'09*, pp. 1145-1146, 2009.

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

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

[24] PLM "Patients Like Me," Available: http://www.patientslikeme.com.

[25] V. Rastogi, M. Hay, G. Miklau, and D. Suciu, "Relationship Privacy: Output Perturbation for Queries with Joins," in *Proc. PODS'09*, pp. 107-116, 2009.

[26] RS, "Rareshare," Available: http://www.rareshare.org, 2008.

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

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.