Data and Structural k-Anonymity in Social Networks

Alina Campan and Traian Marius Truta

Department of Computer Science,
Northern Kentucky University,
Highland Heights, KY 41076, U.S.A.
{campana1,trutat1}@nku.edu

Abstract. The advent of social network sites in the last years seems to be a trend that will likely continue. What naive technology users may not realize is that the information they provide online is stored and may be used for various purposes. Researchers have pointed out for some time the privacy implications of massive data gathering, and effort has been made to protect the data from unauthorized disclosure. However, the data privacy research has mostly targeted traditional data models such as microdata. Recently, social network data has begun to be analyzed from a specific privacy perspective, one that considers, besides the attribute values that characterize the individual entities in the networks, their relationships with other entities. Our main contributions in this paper are a greedy algorithm for anonymizing a social network and a measure that quantifies the information loss in the anonymization process due to edge generalization.

Keywords: Privacy, Social Networks, K-Anonymity, Information Loss.

1 Introduction

While the ever increasing computational power, together with the huge amount of individual data collected daily by various agencies are of great value for our society, they also pose a significant threat to individual privacy. Datasets that store individual information have moved from simpler, traditional data models (such as microdata, where data is stored as one relational table, and each row represents an individual entity) to complex ones. The research in data privacy follows the same trend and tries to provide useful solutions for various data models. Although most of the privacy work has been done for healthcare data (usually in microdata form) mainly due to the Health Insurance Portability and Accountability Act regulation [11], privacy concerns have also been raised in other fields, where data usually takes a more complex form, such as location based services [3], genomic data [18], data streams [29], and social networks [9,10,15,31,32].

The advent of social networks in the last few years has accelerated the research in this field. Online social interaction has become very popular around the globe

F. Bonchi et al. (Eds.): PinkDD 2008, LNCS 5456, pp. 33–54, 2009.

and most sociologists agree that this trend will not fade away [27]. More and more social network datasets contain sensitive data. For example, epidemiology researchers are using social network datasets to study the relationship between sexual network structure and epidemic phase in sexually transmitted disease [21,28]. Other social networks datasets in areas such as e-mail communication [23] also benefit from privacy techniques tailored for social networks. Privacy in social networks is still in its infancy, and practical approaches are yet to be developed. A brief overview of proposed privacy techniques in social networks is given in the related work section.

We introduce in this paper a new anonymization approach for social network data that consists of nodes and relationships. A node represents an individual entity and is described by identifier (such as *Name* and *SSN*), quasi-identifier (such as *ZipCode* and *Sex*), and sensitive (such as *Diagnosis* and *Income*) attributes. A relationship is between two nodes and it is unlabeled, in other words, all relationships have the same meaning. To protect the social network data, we mask it according to the k-anonymity model (every node will be indistinguishable with at least other (k-1) nodes) [6,22,24], in terms of both nodes' attributes and nodes' associated structural information (neighborhood). Our anonymization method tries to disturb as little as possible the social network data, both the attribute data associated to the nodes, and the structural information. The method we use for anonymizing attribute data is generalization [22,25]. For structural anonymization we introduce a new method called edge generalization that does not insert into or remove edges from the social network dataset, similar to the one described in [31]. Although it incorporates a few ideas similar to those exposed in the related papers, our approach is new in several aspects. We embrace the k-anonymity model presented by Hay et al. [9,10], but we assume a much richer data model than just the structural information associated to the social network. We define an information loss measure that quantifies the amount of information loss caused by edge generalization (called *structural information loss*). We perform social network data clustering followed by anonymization through cluster collapsing. Our cluster formation process pays special attention to the nodes' attribute data and equally to the nodes' neighborhoods. This process can be user-balanced towards preserving more structural information of the network, as measured by the structural information loss, or the nodes' attribute values, which are quantified by the generalization information loss measure.

The remaining of this paper is structured as follows. Section 2 introduces our social network privacy model, in particular the concepts of edge generalization and k-anonymous masked social network. Section 3 starts by presenting the generalization and structural information loss measures, followed by our greedy social network anonymization algorithm. Section 4 contains comparative results, in terms of both generalization and structural information loss, for our algorithm and one of the existing privacy algorithms. Related work is presented in Section 5. The paper ends with future work directions and conclusions.

2 Social Network Privacy Model

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.

The set of nodes, $\mathcal{N}$, is described by a set of attributes that are classified into the following three categories:

- $I_1, I_2, \ldots, I_m$ are *identifier* attributes such as *Name* and *SSN* that can be used to identify an entity.
- $Q_1, Q_2, \ldots, Q_q$ are *quasi-identifier* attributes such as *Zip_code* and *Sex* that may be known by an intruder.
- $S_1, S_2, \ldots, S_r$ are *confidential* or *sensitive* attributes such as *Diagnosis* and *Income* that are assumed to be unknown to an intruder.

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 privacy attacks that might lead to identity and/or attribute disclosure [12].

While the identifier attributes are removed from the published (masked) social network data, the quasi-identifier and the confidential attributes, as well as the graph structure, are usually released to the researchers/public. A general assumption, as noted, is that the values for the confidential attributes are not available from any external source. This assumption guarantees that an intruder cannot use the confidential attributes values to increase his/her chances of disclosure. Unfortunately, there are multiple techniques that an intruder can use to try to disclose confidential information. As pointed out in the microdata privacy literature, an intruder may use record linkage techniques between quasi-identifier attributes and external available information to glean the identity of individuals. Using the graph structure, an intruder is also able to identify individuals due to the uniqueness of the neighborhoods of various individuals. As shown in [9,10], 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.

A successful model for microdata privacy protection is k-anonymity, which ensures that every individual is indistinguishable with other (k-1) individuals in terms of their quasi-identifier attributes' values [22,24]. For social network data, the k-anonymity model has to impose both the quasi-identifier attributes and the quasi-identifier relationship homogeneity, for groups of at least k individuals.

The generalization of the quasi-identifier attributes is one of the techniques widely used for microdata k-anonymization. It consists of replacing the actual value of an attribute with a less specific, more general value that is faithful to the original. We reuse this technique for the generalization of nodes attributes' values.

To our knowledge, the only method equivalent to our generalization of a quasi-identifier relationship that exists in the research literature appears in [31] and consists of collapsing clusters 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. These methods are equivalent to randomization or perturbation techniques for microdata. We employ a generalization method for the quasi-identifier relationship similar to the one exposed in [31], 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 totally disjoint clusters: cl_1, cl_2, ..., cl_v. For simplicity we assume at this point 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 also indistinguishable based on their relationships. To achieve this goal, we propose an edge generalization process, with two components: edge intra-cluster and edge inter-cluster generalization.

2.1 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 $|X|$ represents the cardinality of the set X. 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. 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 next.

When the cluster size is 2, the intra-cluster generalization doesn't eliminate any internal structural information, in other words the cluster's internal structure is fully recoverable from the masked information (2, 0) or (2, 1). For example, (2, 0) means that the masked node represents two unconnected original nodes. Nevertheless, these two nodes are anyway indistinguishable from one another, inside the cluster, both in the presence and in the absence of an edge connecting them. This means that a required anonymity level 2 is achieved inside the cluster. However, when the number of nodes within a cluster is at least 3, it is possible to differentiate between various nodes if the cluster internal edges, $\mathcal{E}_{cl}$, are provided. Figure 1 shows comparatively several cases when the nodes can be distinguished and when they can be not (i.e., are anonymous) if the full

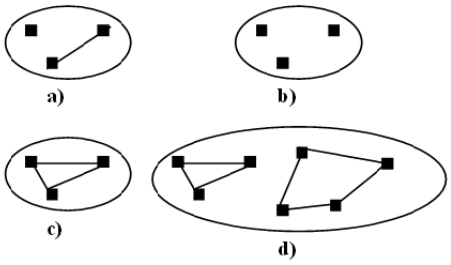

Fig. 1. 3-anonymous (b, c); non 3-anonymous (a); and non 7- anonymous (d)

internal structural information of the cluster was provided. It is easy to notice that a necessary condition that all nodes in a cluster must satisfy in order to be indistinguishable from each other is that all have the same degree. However, this condition is not sufficient, as shown in Figure 1.d, where all the nodes have a degree 2 and they can still be differentiated as belonging to one of the two cycles of the cluster. In this case, the anonymity level is 3, not 7.

2.2 Edge Inter-cluster Generalization

Given two clusters cl_1 and cl_2, let $\mathcal{E}_{cl_1,cl_2}$ be the set of edges having one end in each of the two clusters ($e \in \mathcal{E}_{cl_1,cl_2}$ iff $e \in \mathcal{E}$ 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 $|\mathcal{E}_{cl_1,cl_2}|$. This information permits assessing some structural features about this region of the network that might be helpful in some applications and it does not allow a presumptive intruder to differentiate between nodes within one cluster.

2.3 Masked Social Networks

Let's return to a fully specified social network and how to anonymize it. Given $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, let X^i, $i = 1..n$, be the nodes in $\mathcal{N}$, where $n = |\mathcal{N}|$. We use the term *tuple* to refer only to the corresponding node attributes values (nodes' labels), without considering the relationships (edges) the node participates in. Also, we use the notation $X^i[C]$ to refer to the attribute C's value for the tuple X^i (the projection operation).

Once the nodes from $\mathcal{N}$ have been clustered into totally disjoint clusters $cl_1, cl_2, \ldots, cl_v$, in order to make all nodes in any cluster cl_i indistinguishable from one another in terms of their quasi-identifier attributes values, we generalize each cluster's tuples to the least general tuple that represents all tuples in that group.

There are several types of generalization available. Categorical attributes are usually generalized using generalization hierarchies, predefined by the data owner based on domain attribute characteristics (see Figure 2). For numerical attributes, generalization may be based on a predefined hierarchy or a hierarchy-free model. In our approach, for categorical attributes we use generalization

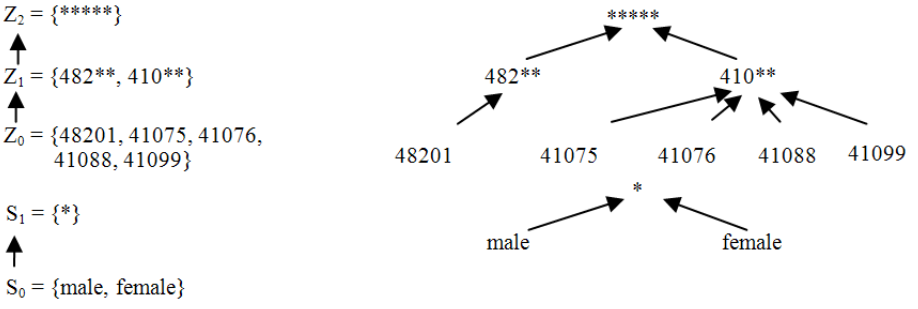

Fig. 2. Domain and value generalization hierarchies for attributes *zip* and *gender*

based on predefined hierarchies at the cell level [16]. For numerical attributes we use the hierarchy-free generalization [13], which consists of replacing the set of values to be generalized with the smallest interval that includes all the initial values. We call generalization information for a cluster the minimal covering tuple for that cluster, and we define it as follows. (Of course, in this paragraph, generalization and coverage refer only to the quasi-identifier part of the tuples).

Definition 1. ***(generalization information of a cluster):*** Let $cl = \{X^1, X^2, \ldots, X^u\}$ be a cluster of tuples corresponding to nodes selected from $\mathcal{N}$, $\mathcal{QN} = \{N_1, N_2, \ldots, N_s\}$ be the set of numerical quasi-identifier attributes and $\mathcal{QC} = \{C_1, C_2, \ldots, C_t\}$ be the set of categorical quasi-identifier attributes. The ***generalization information of cl*** w.r.t. quasi-identifier attribute set $\mathcal{QI} = \mathcal{QN} \cup \mathcal{QC}$ is the "tuple" $gen(cl)$, having the scheme $\mathcal{QI}$, where:

- For each categorical attribute $C_j \in \mathcal{QI}$, $gen(cl)[C_j]$ = the lowest common ancestor in $\mathcal{H}_{C_j}$ of $\{X^1[C_j], \ldots, X^u[C_j]\}$. We denote by $\mathcal{H}_C$ the hierarchies (domain and value) associated to the categorical quasi-identifier attribute C.
- For each numerical attribute $N_j \in \mathcal{QI}$, $gen(cl)[N_j]$ = the interval $[min\{X^1[N_j], \ldots, X^u[N_j]\}, max\{X^1[N_j], \ldots, X^u[N_j]\}]$.

For a cluster cl, its generalization information $gen(cl)$ is the tuple having as value for each quasi-identifier attribute, numerical or categorical, the most specific common generalized value for all that attribute's values from cl tuples. In an anonymized graph, each tuple from cluster cl will have its quasi-identifier attributes values replaced by $gen(cl)$.

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

Definition 2. ***(masked social network):*** Given an initial social network, modeled as a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, and a partition $\mathcal{S} = \{cl_1, cl_2, \ldots, cl_v\}$ of the nodes set $\mathcal{N}$, $\cup_{j=1}^{v} cl_j = \mathcal{N}$; $cl_i \cap cl_j = \emptyset$; $i, j = 1..v, i \neq j$; the corresponding ***masked social network*** $\mathcal{MG}$ is defined as $\mathcal{MG} = (\mathcal{MN}, \mathcal{ME})$, where:

- $\mathcal{MN} = \{Cl_1, Cl_2, \ldots, Cl_v\}$, Cl_j is a node corresponding to the cluster $cl_j \in \mathcal{S}$ and is described by the "tuple" $gen(cl_j)$ (the generalization information of cl_j, w.r.t. quasi-identifier attribute set) and the intra-cluster generalization pair $(|cl_j|, |E_{cl_j}|)$;
- $\mathcal{ME} \subseteq \mathcal{MN} \times \mathcal{MN}$; $(Cl_i, Cl_j) \in \mathcal{ME}$ iif $Cl_i, Cl_j \in \mathcal{MN}$ and $\exists X \in cl_i, Y \in cl_j$, such that $(X, Y) \in \mathcal{E}$. Each generalized edge $(Cl_i, Cl_j) \in \mathcal{ME}$ is labeled with the inter-cluster generalization value $|\mathcal{E}_{cl_i,cl_j}|$.

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 2, 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 3. ***(k-anonymous masked social network):*** A masked social network $\mathcal{MG} = (\mathcal{MN}, \mathcal{ME})$, where $\mathcal{MN} = \{Cl_1, Cl_2, \ldots, Cl_v\}$, and $Cl_j = [gen(cl_j), (|cl_j|, |E_{cl_j}|)], j = 1, \ldots, v$ is k-anonymous iff $|cl_j| \geq k$ for all $j = 1, \ldots, v$.

3 The *SaNGreeA* Algorithm

The algorithm described in this section, 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 $\mathcal{G} = (\mathcal{N}, \mathcal{E})$. Nodes from $\mathcal{N}$ are described by quasi-identifier and sensitive attributes and edges from $\mathcal{E}$ are undirected and unlabeled.

First, the algorithm establishes a "good" partitioning of all nodes from $\mathcal{N}$ into clusters. Next, all nodes within each cluster are made uniform with respect to the quasi-identifier attributes and the quasi-identifier relationship. This homogenization is achieved by using generalization, both for the quasi-identifier attributes and the quasi-identifier relationship, as explained in the previous section.

But how is the clustering process conducted such that a good partitioning is created and what does "good" mean? In order for the requirements of the k-anonymity model to be fulfilled, each cluster has to contain at least k tuples. Consequently, a first criterion to lead the clustering process is to ensure that each cluster has enough elements. As it is well-known, (attribute and relationship) generalization results in information loss. Therefore, a second criterion used during clustering is to minimize the information lost between the initial social network data and its masked version, caused by the subsequent cluster-level quasi-identifier attributes and relationship generalization. In order to obtain good quality masked data, and also to permit the user to control the type and the quantity of information loss he/she can afford, the clustering algorithm uses two information loss measures. One quantifies how much *descriptive* data detail is lost through quasi-identifier attributes generalization - we call this metric the generalization information loss measure. The second measure quantifies how much *structural* detail is lost through the quasi-identifier relationship

generalization and it is called structural information loss. In the remainder of this section, these two information loss measures and the *SaNGreeA* algorithm are introduced.

3.1 Generalization Information Loss

The generalization of quasi-identifier attributes reduces the quality of the data. To measure the amount of information loss, several cost measures were introduced [4,7,13]. In our social network privacy model, we use the generalization information loss measure as introduced and described in [4]:

Definition 4. ***(generalization information loss):*** Let cl be a cluster, $gen(cl)$ its generalization information, and $\mathcal{QI}=\{N_1, N_2, \ldots, N_s, C_1, C_2, \ldots, C_t\}$ the set of quasi-identifier attributes. The ***generalization information loss*** caused by generalizing quasi-identifier attributes of the cl tuples to $gen(cl)$ is:

$$GIL(cl) = |cl| \cdot (\sum_{j=1}^{s} \frac{size(gen(cl)[N_j])}{size(min_{X \in \mathcal{N}}(X[N_j]), max_{X \in \mathcal{N}}(X[N_j]))} + \sum_{j=1}^{t} \frac{height(\Lambda(gen(cl)[C_j]))}{height(\mathcal{H}_{C_j})}),$$

where:

- $|cl|$ denotes the cluster cl's cardinality;
- $size([i_1, i_2])$ is the size of the interval $[i_1, i_2]$, i.e., $(i_2 - i_1)$;
- $\Lambda(w)$, $w \in \mathcal{H}_{C_j}$ is the subhierarchy of $\mathcal{H}_{C_j}$ rooted in w;
- $height(\mathcal{H}_{C_j})$ denotes the height of the tree hierarchy $\mathcal{H}_{C_j}$.

Definition 5. ***(total generalization information loss): Total generalization information loss*** produced when masking the graph $\mathcal{G}$ based on the partition $\mathcal{S} = \{cl_1, cl_2, \ldots, cl_v\}$, denoted by $GIL(\mathcal{G}, \mathcal{S})$, is the sum of the generalization information loss measure for each of the clusters in $\mathcal{S}$:

$$GIL(\mathcal{G}, \mathcal{S}) = \sum_{j=1}^{v} GIL(cl_j).$$

In the above measures, the information loss caused by the generalization of each quasi-identifier attribute value, for any tuple, is a value between 0 and 1. This means that each tuple contributes to the total generalization loss with a value between 0 and $(s+t)$ (the number of quasi-identifier attributes). Since the graph has n tuples, the total generalization information loss is a number between 0 and $n \cdot (s + t)$. To be able to compare this measure with the structural information loss, we chose to normalize both of them to the range [0, 1].

Definition 6. ***(normalized generalization information loss):*** The ***normalized generalization information loss*** obtained when masking the graph $\mathcal{G}$ based on the partition $\mathcal{S} = \{cl_1, cl_2, \ldots, cl_v\}$, denoted by $NGIL(\mathcal{G}, \mathcal{S})$, is the sum of the generalization information loss measure for each of the clusters in $\mathcal{S}$:

$$NGIL(\mathcal{G}, \mathcal{S}) = \frac{GIL(\mathcal{G}, \mathcal{S})}{n \cdot (s+t)}.$$

3.2 Structural Information Loss

We introduce next a measure to quantify the structural information which is lost when anonymizing a graph through collapsing clusters into nodes, together with their neighborhoods.

Information loss in this case quantifies the probability of error when trying to reconstruct the structure of the initial social network from its masked version. There are two components for the structural information loss: the *intra-cluster structural loss* and the *inter-cluster structural loss* components.

Let cl be a cluster of nodes from $\mathcal{N}$, and $\mathcal{G}_{cl} = (cl, \mathcal{E}_{cl})$ be the subgraph induced by cl in $\mathcal{G} = (\mathcal{N}, \mathcal{E})$. When cl is replaced (collapsed) in the masked graph $\mathcal{MG}$ with the node Cl described by the pair $(|cl|, |\mathcal{E}_{cl}|)$, the probability of an edge to exist between any pair of nodes from cl is $|\mathcal{E}_{cl}|/\binom{|cl|}{2}$. Therefore, for each of the real edges from cluster cl, the probability that someone wrongly labels it as a non-edge is $1 - |\mathcal{E}_{cl}|/\binom{|cl|}{2}$. At the same time, for each pair of unconnected edges from cluster cl, the probability that someone wrongly labels it as an edge is $|\mathcal{E}_{cl}|/\binom{|cl|}{2}$.

Definition 7. ***(intra-cluster structural information loss):*** The ***intra-cluster structural information loss*** ($intraSIL$) is the probability of wrongly labeling a pair of nodes in cl as an edge or as an unconnected pair. As there are $|\mathcal{E}_{cl}|$ edges, and $\binom{|cl|}{2} - \mathcal{E}_{cl}$ pairs of unconnected nodes in cl,

$$intraSIL(cl) = \left(\binom{|cl|}{2} - |\mathcal{E}_{cl}|\right) \cdot |\mathcal{E}_{cl}|/\binom{|cl|}{2} + |\mathcal{E}_{cl}| \cdot \left(1 - |\mathcal{E}_{cl}|/\binom{|cl|}{2}\right) = 2 \cdot |\mathcal{E}_{cl}| \cdot \left(1 - |\mathcal{E}_{cl}|/\binom{|cl|}{2}\right).$$

Reasoning in the same manner as above, we introduce the second structural information loss measure.

Definition 8. ***(inter-cluster structural information loss):*** The ***inter-cluster structural information loss*** ($interSIL$)) is the probability of wrongly labeling a pair of nodes (X, Y), where $X \in cl_1$ and $Y \in cl_2$, as an edge or as an unconnected pair. As there are $|\mathcal{E}_{cl_1,cl_2}|$ edges, and $|cl_1| \cdot |cl_2|$ - $|\mathcal{E}_{cl_1,cl_2}|$ pairs of unconnected nodes between cl_1 and cl_2,

$$interSIL(cl_1, cl_2) = (|cl_1| \cdot |cl_2| - |\mathcal{E}_{cl_1,cl_2}|) \cdot \frac{|\mathcal{E}_{cl_1,cl_2}|}{|cl_1| \cdot |cl_2|} + |\mathcal{E}_{cl_1,cl_2}| \cdot \left(1 - \frac{|\mathcal{E}_{cl_1,cl_2}|}{|cl_1| \cdot |cl_2|}\right) = 2 \cdot |\mathcal{E}_{cl_1,cl_2}| \cdot \left(1 - \frac{|\mathcal{E}_{cl_1,cl_2}|}{|cl_1| \cdot |cl_2|}\right).$$

Now, we have all the tools to introduce the total structural information loss measure.

Definition 9. ***(total structural information loss):*** The ***total structural information loss*** obtained when masking the graph $\mathcal{G}$ based on the partition $\mathcal{S} = \{cl_1, cl_2, \ldots, cl_v\}$, denoted by $SIL(\mathcal{G}, \mathcal{S})$, is the sum of all inter-cluster and intra-cluster structural information loss values:

$$SIL(\mathcal{G}, \mathcal{S}) = \sum_{j=1}^{v}(intraSIL(cl_j)) + \sum_{i=1}^{v}\sum_{j=i+1}^{v}(interSIL(cl_i, cl_j)).$$

We analyze the $intraSIL(cl)$ function for a given fixed cluster cl and a variable number of edges in the cluster, $|\mathcal{E}_c l|$, in other words, we consider $intraSIL(cl)$ a function of a variable $|\mathcal{E}_c l|$. Based on Definition 7, this function is (we use f to denote the function and x the variable number of edges):

$$f : \left\{0, 1, \ldots, \binom{|cl|}{2}\right\} \rightarrow \Re,$$
$$f(x) = 2 \cdot x \cdot \left(1 - x / \binom{|cl|}{2}\right).$$

Using the first and second derivative function it can easily be determined that the maximum value the function f takes is for

$$x = \binom{|cl|}{2} / 2 = \frac{|cl| \cdot (|cl| - 1)}{4}.$$

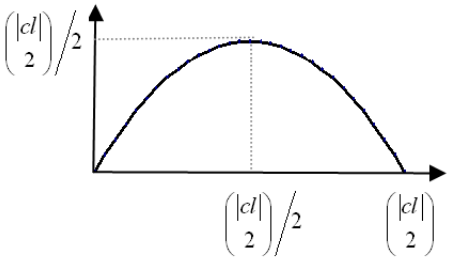

Fig. 3. $intraSIL$ as a function of number of edges for $|cl|$ fixed

Figure 3 shows the graphical representation of the $f(x)$ function. As it can be seen, the smallest values of the function correspond to clusters that are either unconnected graphs (no edges) or completely connected graphs. The maximum function value corresponds to a cluster that has the number of edges equal to half of the number of all the pairs of nodes in the cluster.

A similar analysis, with the same results, can be conducted for the function $interSIL(cl_1, cl_2)$, seen as a function of one variable $|\mathcal{E}_{cl_1,cl_2}|$, when clusters cl_1 and cl_2 are fixed. This function has a similar behavior with $intraSIL(cl)$. Namely, minimum is reached when $|\mathcal{E}_{cl_1,cl_2}|$ is either 0 or the maximum possible value $|cl_1| \cdot |cl_2|$, and the maximum is reached when $|\mathcal{E}_{cl_1,cl_2}|$ is equal to $|cl_1| \cdot |cl_2|/2$.

This analysis suggests that a smaller structural information loss corresponds to clusters in which nodes have similar connectivity properties with one another or, in other words, when cluster's nodes are either all connected (or unconnected) among them and with the nodes in other clusters. We will use this result in our anonymization algorithm.

To normalize the structural information loss, we compute the maximum values for $intraSIL(cl)$ and $interSIL(cl_1, cl_2)$. As illustrated in Figure 3, the maximum value for $intraSIL(cl)$ is $|cl| \cdot (|cl| - 1)/4$. Similarly, the maximum value for $interSIL(cl_1, cl_2)$ is $|cl_1| \cdot |cl_2|/2$. Using Definition 9, we derive the maximum total structural information loss value as:

$$\sum_{j=1}^{v} \frac{|cl_j| \cdot (|cl_j| - 1)}{4} + \sum_{i=1}^{v} \sum_{j=i+1}^{v} \frac{|cl_i| \cdot |cl_j|}{4} =$$
$$\frac{1}{4} \cdot \left(\sum_{j=1}^{v} |cl_j|^2 + 2 \cdot \sum_{i=1}^{v} \sum_{j=i+1}^{v} |cl_i| \cdot |cl_j| \right) - \frac{1}{4} \sum_{j=1}^{v} |cl_j| =$$
$$\frac{1}{4} \left(\sum_{j=1}^{v} |cl_j| \right)^2 - \frac{1}{4} \sum_{j=1}^{v} |cl_j| = \frac{n \cdot (n-1)}{4}.$$

The minimum total structural information loss is 0, and it is obtained for a graph with no edges or for a complete graph.

Definition 10. ***(normalized structural information loss):*** The ***normalized structural information loss*** obtained when masking the graph $\mathcal{G}$ with n nodes, based on the partition $\mathcal{S} = \{cl_1, cl_2, \ldots, cl_v\}$, denoted by $NSIL(\mathcal{G}, \mathcal{S})$, is:

$$NSIL(\mathcal{G}, \mathcal{S}) = \frac{SIL(\mathcal{G}, \mathcal{S})}{(n \cdot (n-1)/4)}.$$

The normalized structural information loss is in the range [0, 1].

3.3 The Anonymization Algorithm

The *SaNGreeA* algorithm puts together in clusters nodes that are as similar as possible, both in terms of their quasi-identifier attribute values, and in terms of their neighborhood structure. This greedy approach tries to minimize the generalization information loss and the structural information loss for the generated k-anonymous masked social network.

To assess the proximity between nodes with respect to quasi-identifier attributes, we use the normalized generalization information loss. However, the structural information loss cannot be computed during the clusters creation process, as long as the entire partitioning is not known. Therefore, we chose to guide the clustering process using a different measure. This measure quantifies the extent in which the neighborhoods of two nodes are similar with each other, i.e., the nodes present 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,

$\mathcal{N} = \{X^1, X^2, \ldots, X^n\}$. The neighborhood of each node X^i can be represented as an n-dimensional boolean vector $B_i = (b_1^i, b_2^i, \ldots, b_n^i)$, 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, $\forall j = 1..n; j \neq i$. We consider the value b_i^i to be *undefined*, and therefore not equal with 0 or 1. We use a classical distance measure for this type of vector, the *symmetric binary distance* [8].

Definition 11. ***(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:

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

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 a node to be added to an existing cluster. To assess the structural distance between a node and a cluster we use the following measure.

Definition 12. ***(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:

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

We note that both distance measures take values between 0 and 1, and they can be used in the cluster formation process in combination with the normalized generalization information loss.

Although this is not formally proved, but shown to be effective in our experiments, by putting together in clusters nodes that are the closest according to the average distance measure, the *SaNGreeA* algorithm will produce a good masked network, with a small structural information loss.

Using the above introduced measures, we explain 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 to minimize the cluster's information loss growth, quantified as a weighted measure that combines $NGIL$ and $dist$. The parameters α and β, with $\alpha + \beta = 1$, control the relative importance given to the total generalization information loss (the parameter α) and the total structural information loss (the parameter β) and are user-defined.

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 whose information loss will minimally increase by that node addition.

The pseudocode for our social network anonymization algorithm is shown next.

```
Algorithm SaNGreeA is

Input G = (N, E) - a social network
       k - as in k-anonymity
       α and β, α + β = 1 - user-defined weight parameters;
       allow controlling the balancing between GIL and SIL.

Output S = {cl_1, cl_2, ..., cl_v}; ∪_{j=1}^{v} cl_j = N; cl_i ∩ cl_j = ∅,
       i, j = 1..v, i ≠ j; |cl_j| ≥ k, j = 1..v - a set of clusters
       that ensures k-anonymity for MG = (MN, ME) so that
       a cost measure is optimized;

S = ∅;
i = 1;
Repeat
       X^seed = a node with maximum degree from N;
       cl_i = {X^seed};
       // N keeps track of nodes not yet distributed to clusters
       N = N - {X^seed};
       Repeat
          X* = argmin_{X∈N}(α · NGIL(G_1, S_1) + β · dist(X, cl_i);
          // X* is the node within N (unselected nodes) that
          // produces the minimal information loss growth when
          // added to cl_i
          // G_1 - the subgraph induced by cl ∪ {X*} in G;
          // S_1 - a partition with one cluster cl ∪ {X*}
          cl_i = cl_i ∪ {X*};
          N = N − {X*};
       Until (cl_i has k elements) or (N == ∅);
       If (|cl_i| < k) then
          DisperseCluster(S, cl_i); // only for the last cluster
       Else
          S = S ∪ {cl_i};
          i++;
       End If;
Until N == ∅;
End SaNGreeA.

Function DisperseCluster(S, cl)
       For every X ∈ cl do
          cl_u = FindBestCluster(X, S);
```

```
            cl_u = cl_u ∪ {X};
        End For;
End DisperseCluster;

Function FindBestCluster(X, S) is
        bestCluster = null;
        infoLoss = ∞;
        For every cl_j ∈ S do
            If α · NGIL(G_1, S_1) + β · dist(X, cl_i) < infoLoss then
                infoLoss = α · NGIL(G_1, S_1) + β · dist(X, cl_i) ;
                bestCluster = cl_j;
            End If;
        End For;
        Return bestCluster;
End FindBestCluster;
```

Because *SaNGreeA* is a greedy algorithm, that selects a solution from the search space (i.e., the set of all partitions of $\mathcal{N}$ consisting of subsets of k or more nodes) based on local optima of the two criterion measures, the algorithm will find a good solution to the anonymization problem, but not the best existing solution. The time complexity of *SaNGreeA* is $O(n^2)$. However, an efficient (sub-exponential) method to find the optimal solution is not known: the k-anonymization for microdata has been proved to be NP-hard [19] and our optimization problem for social network data is similar, with the only difference of having to minimize two measures of the amount of information in the initial data that is not released.

We show next an example that illustrates the concepts of generalization and structural information loss as well as how the obtained solution is dependent of the selection of α and β.

Suppose the social network $\mathcal{G}_{ex}$ depicted in Figure 4 is given. It contains nine nodes, described by the quasi-identifier attributes *age*, *zip* and *gender*. The *age* quasi-identifier is numerical, *zip* and *gender* are categorical - their predefined domain and value generalization hierarchies are presented in Figure 2. The quasi-identifier attributes' values for all nodes are depicted in Table 1.

By running the *SaNGreeA* algorithm for this set of data for ($k = 3$, $\alpha = 1$, and $\beta = 0$) and ($k = 3$, $\alpha = 0$, and $\beta = 1$) respectively, we obtain the 3-anonymous masked social networks $\mathcal{MG}_{e1}$ and $\mathcal{MG}_{e2}$ depicted in Figure 5. We did not show in the figure the generalization information for the clusters, but this can be easily computed; for instance, $gen(cl_2) = \{[25 - 27], 410 * *, male\}$.

In Table 2 we show the information loss measures' values computed based on Definitions 4 - 10. As expected, due to the weights choice, $\mathcal{MG}_{e1}$ is a better solution in terms of total generalization information loss than $\mathcal{MG}_{e2}$ and $\mathcal{MG}_{e2}$ outperforms $\mathcal{MG}_{e1}$ with respect to total structural information loss.

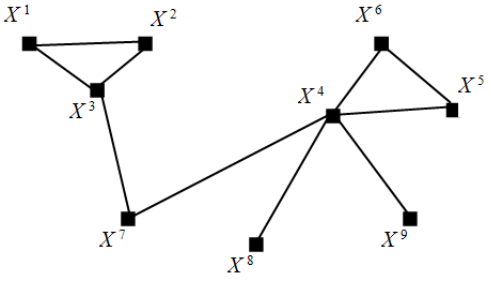

Fig. 4. The Social Network $\mathcal{G}_{ex}$

Table 1. The quasi-identifier attributes' values for $\mathcal{G}_{ex}$ nodes

Node	age	zip	gender
X^1	25	41076	male
X^2	25	41075	male
X^3	27	41076	male
X^4	35	41099	male
X^5	38	48201	female
X^6	36	41075	female
X^7	30	41099	male
X^8	28	41099	male
X^9	33	41075	female

4 Experimental Results

In this section we compare the *SaNGreeA* algorithm and the anonymization algorithm proposed in [31], which is based on collapsing clusters as formed by any classical k-anonymization algorithm for microdata [4,13]. For our experiments, we use the clustering algorithm introduced in [4]. Comparisons of *SaNGreeA* with other existing algorithms for anonymizing social networks [2,9,32] are not feasible, as those algorithms do not take into consideration a full range of quasi-identifier attributes, as we do; usually they consider at most one quasi-identifier attribute and, of course, the quasi-identifier relationship. Another difference that impeded comparison with other algorithms is the incompatibility in how relationships are seen across different anonymization approaches: single type versus multiple types of relationships, relationships with or without attributes etc. Zheleva's algorithm seems to be the only compatible and obviously comparable with ours.

The comparison we present between the *SaNGreeA* algorithm and the Zheleva's algorithm [31] is made with respect to the quality of the results they produce, measured against the normalized generalization information loss and the normalized structural information loss.

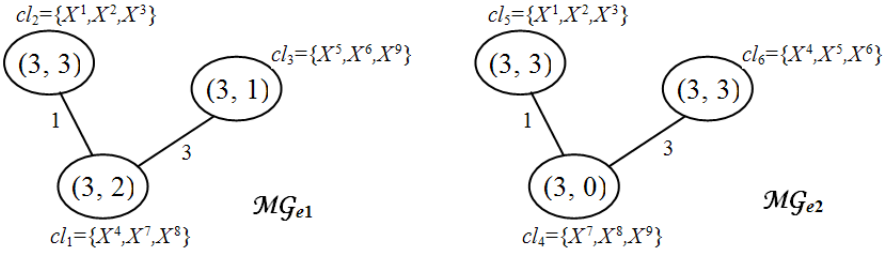

Fig. 5. The k-anonymous masked social networks $\mathcal{MG}_{e1}$ and $\mathcal{MG}_{e2}$

Table 2. Information loss values

$(\mathcal{G}, \mathcal{MG})$	$(\mathcal{G}_{ex}, \mathcal{MG}_{e1})$ with partition $\mathcal{S}_1 = \{\{X^4, X^7, X^8\}, \{X^1, X^2, X^3\}, \{X^5, X^6, X^9\}\}$	$(\mathcal{G}_{ex}, \mathcal{MG}_{e2})$ with partition $\mathcal{S}_2 = \{\{X^4, X^5, X^6\}, \{X^1, X^2, X^3\}, \{X^7, X^8, X^9\}\}$
$GIL, NGIL$	$GIL(\mathcal{G}, \mathcal{S}_1) = 3 \cdot \left(\frac{7}{13} + 0 + 0\right) + 3 \cdot \left(\frac{2}{13} + \frac{1}{2} + 0\right) + 3 \cdot \left(\frac{5}{13} + 1 + 0\right) = 7.730$ $NGIL(\mathcal{G}, \mathcal{S}_1) = \frac{7.730}{9 \cdot 3} = 0.286$	$GIL(\mathcal{G}, \mathcal{S}_2) = 3 \cdot \left(\frac{3}{13} + 1 + 1\right) + 3 \cdot \left(\frac{2}{13} + \frac{1}{2} + 0\right) + 3 \cdot \left(\frac{5}{13} + \frac{1}{2} + 1\right) = 14.307$ $NGIL(\mathcal{G}, \mathcal{S}_2) = \frac{14.307}{9 \cdot 3} = 0.529$
$intraSIL$	$intraSIL(cl_1) = \frac{4}{3}$ $intraSIL(cl_2) = 0$ $intraSIL(cl_3) = \frac{4}{3}$	$intraSIL(cl_4) = 0$ $intraSIL(cl_5) = 0$ $intraSIL(cl_6) = 0$
$interSIL$	$interSIL(cl_1, cl_2) = \frac{16}{9}$ $interSIL(cl_1, cl_3) = 4$ $interSIL(cl_2, cl_3) = 0$	$interSIL(cl_4, cl_5) = \frac{16}{9}$ $interSIL(cl_4, cl_6) = 4$ $interSIL(cl_5, cl_6) = 0$
$SIL, NSIL$	$SIL(\mathcal{G}, \mathcal{S}_1) = 8.444$ $NSIL(\mathcal{G}, \mathcal{S}_1) = 0.469$	$SIL(\mathcal{G}, \mathcal{S}_2) = 5.777$ $NSIL(\mathcal{G}, \mathcal{S}_2) = 0.320$

The two algorithms were implemented in Java; tests were executed on a dual CPU machine with 3.00GHz and 4GB of RAM, running Windows NT Professional. Experiments were performed for a social network with 300 nodes randomly selected from the Adult dataset from the UC Irvine Machine Learning Repository [20]; we refer to this set as $\mathcal{N}$.

In all the experiments, we considered a set of six quasi-identifier attributes: *age*, *workclass*, *marital-status*, *race*, *sex*, and *native-country*. The *age* attribute was the only numerical quasi-identifier, the other five attributes are categorical. Figure 6 depicts the generalization hierarchy for the *native-country* attribute, the categorical attribute with the most developed hierarchy. The remaining four quasi-identifier categorical attributes have the following heights for their corresponding value generalization hierarchies: *workclass* - 1, *marital-status* - 2, *race* - 1, and *sex* - 1. As already explained, for the quasi-identifier numerical attribute we used hierarchy-free generalization [13].

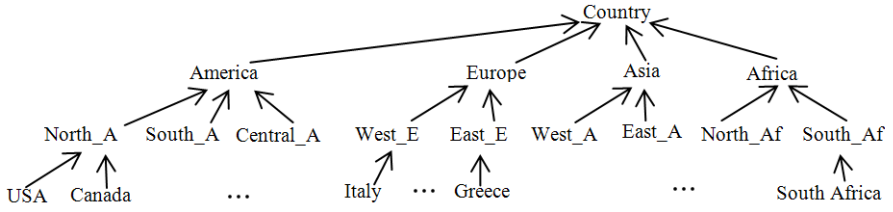

Fig. 6. The value hierarchy for the quasi-identifier attribute *native-country*

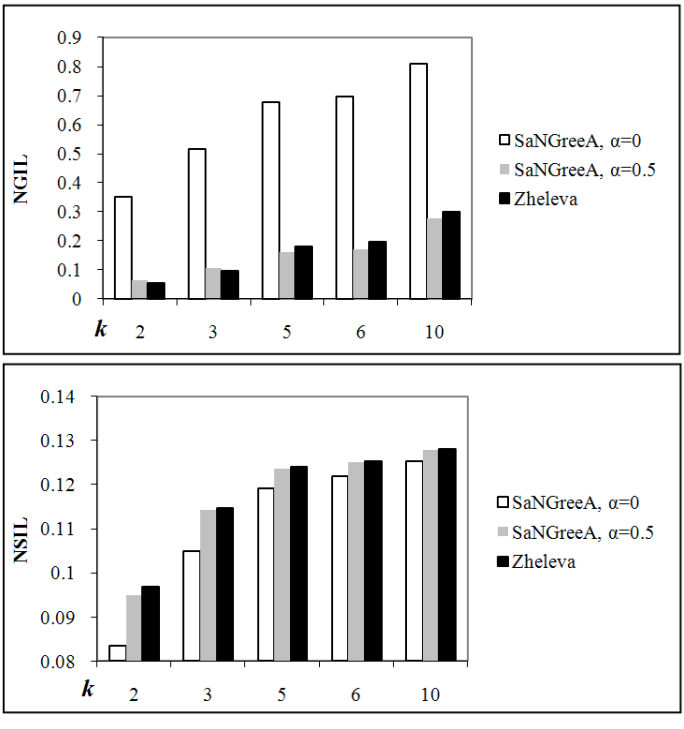

Fig. 7. *NGIL* and *NSIL* for *Random Graph*

Three different synthetic sets of edges were considered, all generated using *GTGraph*, a synthetic graph generator suite [1]. The first edge set corresponds to a random graph with an average vertex degree of 10; we refer to this edge set as $\mathcal{E}_1$. For producing $\mathcal{E}_1$, we used the random graph generator included in the *GTGraph* suite and we replaced with other random edges all but one of the multiple edges between the same pair of vertices. The second edge set we experimented with was generated in agreement with the power law distribution and the small-world characteristic, which are the two most important properties for many real-world social networks [32]; we refer to this edge set as $\mathcal{E}_2$. For

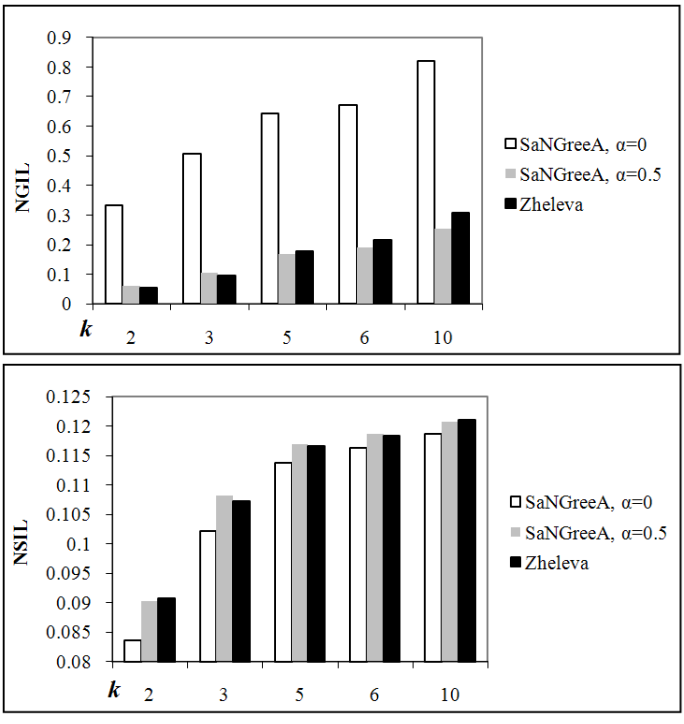

Fig. 8. *NGIL* and *NSIL* for R_MAT Graph, average vertex degree of 9.52

producing $\mathcal{E}_2$, we used the R_MAT graph model [5] and generator included in the *GTGraph* suite. We randomly replaced or removed the multiple edges between the same pair of vertices. The resulting graph $(\mathcal{N}, \mathcal{E}_2)$ had an average vertex degree of 9.52. The third edge set we experimented with was similar with the second one, in the sense that it was generated in agreement with the power law distribution and the small-world characteristic; it differed from the second one on the average vertex degree, which was 5. We refer to this edge set as $\mathcal{E}_3$.

The *SaNGreeA* algorithm and the algorithm introduced in [31] were applied to these three social networks, $\mathcal{G}_1 = (\mathcal{N}, \mathcal{E}_1)$, $\mathcal{G}_2 = (\mathcal{N}, \mathcal{E}_2)$, and $\mathcal{G}_3 = (\mathcal{N}, \mathcal{E}_3)$, for different k values, $k = 2, 3, 5, 6$, and 10. Figures 7, 8, and 9 present comparatively the normalized generalization information loss and the normalized structural information loss values of the results produced by applying the two algorithms, for the graphs $\mathcal{G}_1$, $\mathcal{G}_2$, and respectively $\mathcal{G}_3$, for all considered k values, and for two different value sets for the parameters α and β in the *SaNGreeA* algorithm. The (α, β) occurrences we used are (0.0, 1.0) and respectively (0.5, 0.5). The pair (0.0, 1.0) guides the algorithm towards minimizing the structural information loss, without giving any consideration to the generalization information loss factor. The pair (0.5, 0.5) signifies a request for the algorithm to equally weight both information loss components in the cluster formation process. As expected, while both tested algorithms, with all different parameters selections, produce a

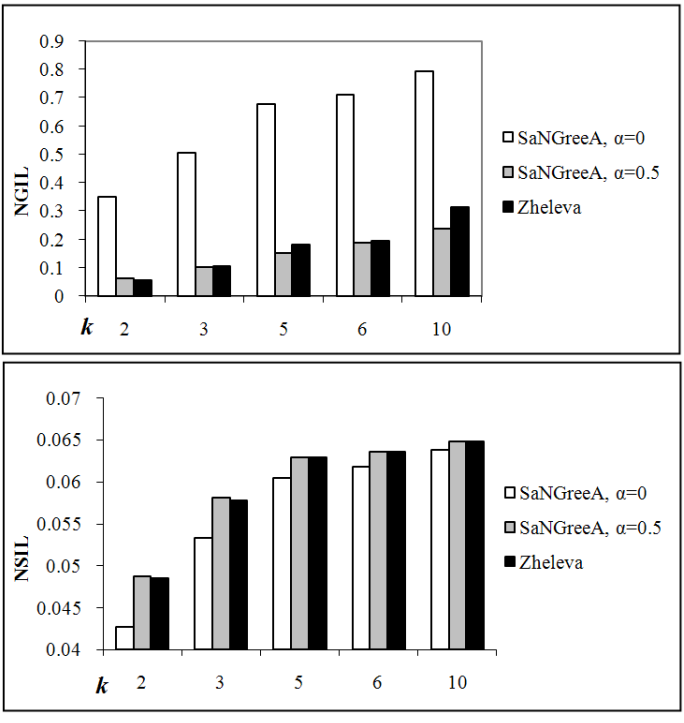

Fig. 9. $NGIL$ and $NSIL$ for R_MAT Graph, average vertex degree of 5

k-anonymized masked social network, the data utility conserved by each solution is different. For the *SaNGreeA* experiments the structural information loss is, in general, smaller than in the Zheleva's algorithm case.

This comes with the cost of greater generalization information loss. Since it is based on defining the weight of generalization/structural information loss, our algorithm is very flexible and allows the user to customize the amount of generalization and/or structural information loss he agrees to in a particular anonymization task. A special note is worth to be made. Our algorithm can be tuned to be equivalent to Zheleva's (when the last one bases its cluster formation on the greedy algorithm explained in [4]), by appropriately setting (α, β) parameters to (1.0, 0.0). The general rule is to set β to a value greater than α's when more structural information needs to be preserved when anonymizing the network; and vice versa, α has to be set to a value greater than β's when more generalization information needs to be preserved.

5 Related Work

The research in social networks privacy is very recent, and many questions are still to be answered. Only a few researchers have explored this integrative field

of privacy in social networks from a computing perspective. We briefly present a short overview of the approaches we are aware of.

Zheleva and Geetor consider the problem where relationships between different individual entities in a network must be protected, and they called this problem link re-identification [31]. Their anonymization approach functions in two steps: first anonymize descriptive data from the graph nodes (the individual entities) to achieve k-anonymity or t-closeness [14], without considering in this step, in any way, the relationships between the network nodes. Their next step is to anonymize the network's structure, by controlled edge removal, in different flavors, each with different success likelihood: edges can all be removed, only a user-specified percentage of them, none of them, or can be generalized at a cluster level. Our work is closest to theirs. However, in our approach we anonymize the social network data at once, i.e., the nodes and edges anonymizations are integrated together in our masking algorithm and occur concurrently.

Other researchers have focused on developing a concept similar to k-anonymity for graph data. Hay et al. defines k-candidate anonymity based on the similarity of neighborhoods, in other words every node has at least k candidate nodes from which it is hard to be distinguished [9,10]. In order to satisfy this property, the graph data suffers a series of random edge additions and deletions. The nodes also do not contain attributes besides an identifier, and the edges are of a single type. Zhou and Pei have a similar social network model, they consider the nodes to be labeled (having one attribute, which can be seen as a quasi-identifier) and that only the near vicinity (1-radius neighborhood) of some target individuals is completely known to an intruder [32]. Their solution generalizes the node labels (attribute values) and adds extra edges to create similar neighborhoods. Their approach guarantees that an adversary with the knowledge of a 1-neigborhood cannot identify any individual with a confidence higher than $1/k$. Liu and Terzi introduced the concept of k-degree anonymous graph if for every node v, there exist at least $k-1$ other nodes in the graph with the same degree as v [15]. They introduce practical anonymization algorithms that are based on principles related to the realizability of degree sequences.

Another approach was introduces by Backstrom, Dwork, and Kleinberg [2]. They consider several possible types of "injection" attacks, in which the intruder is actively involved in the social network before its data will be published in a repository, such that the intruder will be capable to retrieve his own data and to use it as a marker that facilitates the attack. Backstrom's work does not propose a practical method to counter the mentioned attacks.

6 Conclusions and Future Work

In this paper we studied a new anonymization approach for social network data. We introduced a generalization method for edges and a measure to quantify structural information loss. We developed a greedy privacy algorithm that anonymizes a social network. This algorithm can be user-balanced towards preserving more the structural information of the network or the nodes' attribute values.

We envision several research directions that can extend this work:

- Extend the anonymity model to achieve protection against attribute disclosure in social networks. Similar models such as p-sensitive k-anonymity [26], l-diversity [17], (α, k)-anonymity [30], and t-closeness [14] exist for microdata.
- Study the change in utility of an anonymized social network for various application fields.
- Formally analyze how the similarity measure is tied to the total structural information loss measure and improve the greedy selection criteria.

References

1. Bader, D.A., Madduri, K.: GTGraph: A Synthetic Graph Generator Suite (2006), `http://www.cc.gatech.edu/~kamesh/GTgraph/`
2. Backstrom, L., Dwork, C., Kleinberg, J.: Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography. In: International World Wide Web Conference (WWW), pp. 181–190 (2007)
3. Bamba, B., Liu, L., Pesti, P., Wang, T.: Supporting Anonymous Location Queries in Mobile Environments with PrivacyGrid. In: ACM World Wide Web Conference (2008)
4. Byun, J.W., Kamra, A., Bertino, E., Li, N.: Efficient k-Anonymization using Clustering Techniques. In: Kotagiri, R., Radha Krishna, P., Mohania, M., Nantajeewarawat, E. (eds.) DASFAA 2007. LNCS, vol. 4443, pp. 188–200. Springer, Heidelberg (2007)
5. Chakrabarti, D., Zhan, Y., Faloutsos, C.: R-MAT: A Recursive Model for Graph Mining. In: SIAM International Conference on Data Mining (2004)
6. Ciriani, V., Vimercati, S.C., Foresti, S., Samarati, P.: K-Anonymity. In: Secure Data Management In Decentralized Systems, pp. 323–353 (2007)
7. Ghinita, G., Karras, P., Kalinis, P., Mamoulis, N.: Fast Data Anonymization with Low Information Loss. In: Very Large Data Base Conference (VLDB), pp. 758–769 (2007)
8. Han, J., Kamber, M.: Data Mining. In: Concepts and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2006)
9. Hay, M., Miklau, G., Jensen, D., Weiss, P., Srivastava, S.: Anonymizing Social Networks. Technical Report No. 07-19, University of Massachusetts Amherst (2007)
10. Hay, M., Miklau, G., Jensen, D., Towsley, D., Weis, P.: Resisting Structural Reidentification in Anonymized Social Networks. In: Very Large Data Base Conference (VLDB), pp. 102–114 (2008)
11. HIPAA. Health Insurance Portability and Accountability Act (2002), `http://www.hhs.gov/ocr/hipaa`
12. Lambert, D.: Measures of Disclosure Risk and Harm. Journal of Official Statistics 9, 313–331 (1993)
13. LeFevre, K., DeWitt, D., Ramakrishnan, R.: Mondrian Multidimensional K-Anonymity. In: IEEE International Conference of Data Engineering (ICDE), vol. 25 (2006)
14. Li, N., Li, T., Venkatasubramanian, S.: T-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In: IEEE International Conference on Data Engineering (ICDE), pp. 106–115 (2007)

15. Liu, K., Terzi, E.: Towards Identity Anonymization on Graphs. In: ACM SIGMOD International Conference on Management of Data, pp. 93–106 (2008)
16. Lunacek, M., Whitley, D., Ray, I.: A Crossover Operator for the k-Anonymity Problem. In: Genetic and Evolutionary Computation Conference (GECCO), pp. 1713–1720 (2006)
17. Machanavajjhala, A., Gehrke, J., Kifer, D.: L-Diversity: Privacy beyond K-Anonymity. In: IEEE International Conference on Data Engineering (ICDE), vol. 24 (2006)
18. Malin, B.: An Evaluation of the Current State of Genomic Data Privacy Protection Technology and a Roadmap for the Future. Journal of the American Medical Informatics Association 12(1), 28–34 (2005)
19. Meyerson, A., Williams, R.: On the complexity of optimal k-anonymity. In: ACM PODS Symposium on the Principles of Database Systems, pp. 223–228 (2004)
20. Newman, D.J., Hettich, S., Blake, C.L., Merz, C.J.: UCI Repository of Machine Learning Databases (1998), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
21. Potterat, J.J., Philips-Plummer, L., Muth, S.Q., Rothenberg, R.B., Woodhouse, D.E., Maldonado-Long, T.S., Zimmerman, H.P., Muth, J.B.: Risk Network Structure in the Early Epidemic Phase of HIV Transmission in Colorado Springs. Sexually Transmitted Infections 78, 159–163 (2002)
22. Samarati, P.: Protecting Respondents Identities in Microdata Release. IEEE Transactions on Knowledge and Data Engineering 13(6), 1010–1027 (2001)
23. Shetty, J., Adibi, J.: The Enron Email Dataset Database Schema and Brief Statistical Report (2004), `http://www.isi.edu/~adibi/Enron/Enron_Dataset_Report.pdf`
24. Sweeney, L.: K-Anonymity: A Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems 10(5), 557–570 (2002)
25. Sweeney, L.: Achieving k-Anonymity Privacy Protection Using Generalization and Suppression. International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems 10(5), 571–588 (2002)
26. Truta, T.M., Bindu, V.: Privacy Protection: P-Sensitive K-Anonymity Property. In: PDM Workshop, with IEEE International Conference on Data Engineering (ICDE), vol. 94 (2006)
27. Tse, H.: An Ethnography of Social Networks in Cyberspace: The Facebook Phenomenon. The Hong Kong Anthropologist 2, 53–77 (2008)
28. Ward, H.: Prevention Strategies for Sexually Transmitted Infections: Importance of Sexual Network Structure and Epidemic Phase. Sexually Transmitted Infections 83, 43–49 (2007)
29. Wang, T., Liu, L.: Butterfly: Protecting Output Privacy in Stream Mining. In: IEEE International Conference on Data Engineering (ICDE), pp. 1170–1179 (2008)
30. Wong, R.C.W., Li, J., Fu, A.W.C., Wang, K.: (α, k)-Anonymity: An Enhanced k-Anonymity Model for Privacy-Preserving Data Publishing. In: SIGKDD, pp. 754–759 (2006)
31. Zheleva, E., Getoor, L.: Preserving the Privacy of Sensitive Relationships in Graph Data. In: ACM SIGKDD Workshop on Privacy, Security, and Trust in KDD (PinKDD), pp. 153–171 (2007)
32. Zhou, B., Pei, J.: Preserving Privacy in Social Networks against Neighborhood Attacks. In: IEEE International Conference on Data Engineering (ICDE), pp. 506–515 (2008)