P-Sensitive *K*-Anonymity for Social Networks

Roy Ford, Traian Marius Truta, and Alina Campan

***Abstract* — The proliferation of social networks, where individuals share private information, has caused, in the last few years, a growth in the volume of sensitive data being stored in these networks. As users subscribe to more services and connect more with their friends, families, and colleagues, the desire to both protect the privacy of the network users and the temptation to extract, analyze, and use this information from the networks have increased. Previous research has looked at anonymizing social network graphs to ensure their k-anonymity in order to protect their nodes against identity disclosure. In this paper we introduce an extension to this k-anonymity model that adds the ability to protect against attribute disclosure. This new model has similar privacy features with the existing p-sensitive k-anonymity model for microdata. We also present a new algorithm for enforcing p-sensitive k-anonymity on social network data based on a greedy clustering approach. To our knowledge, no previous research has been done to deal with preventing against disclosing attribute information that is associated to social networks nodes.**

Keywords: privacy, social networks, k-anonymity, clustering, greedy algorithm.

I. Introduction

The use of social network sites on the Internet, such as Facebook or MySpace, continues to grow at an exponential rate. The opening of Facebook to non-college membership caused a 500% growth in enrollment in one year [14]. In 2005, before Facebook became a public network, Gross and Acquisti analyzed the profiles of Carnegie Mellon University students and identified privacy implications in the data being stored in this social network [7]. The main privacy concerns reported by Gross and Acquisti were the potential for stalking and re-identification of users based on demographics or images of faces and the possibility of identity theft [7].

Obviously, there is a need to protect the privacy of individuals in social networks. Since social networking has become mainstream only in the last few years, the research in social networks privacy is also 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.

R. Ford (e-mail: fordr1@nku.edu), T. M. Truta (phone: +1-859-572-7551; fax: +1-859-572-5398, e-mail: trutat1@nku.edu), and A. Campan (e-mail: campana1@nku.edu) are with the Department of Computer Science, Northern Kentucky University, Nunn Drive, Highland Heights, KY 41099, USA.

A. Related Work

Most of the existing work had focused on protecting the nodes' identities in a social network [4], [8], [12], [22]. There is a strong similarity between ensuring this type of privacy for social network nodes and preventing against identity disclosure in flat microdata [10]. Therefore, *k*-anonymity, the most popular model that guarantees identity protection in microdata [16], [18], has been extended from its primary form to also work for social network data [4]. With that end in view, the *k*-anonymity model for social networks had to additionally address the anonymization of network's structural information, which itself carries a disclosure "potential". Other researchers have proposed solutions for protecting the confidential links between nodes. Two nodes in a social network may have multiple connections, and some of them represent confidential relationships. Solutions to this link disclosure problem have been analyzed in [9], [21]. Less related to this paper, other contributions in the privacy in social networks field include: active and passive attacks [1], random perturbation [20], and access control / encryption protocols [5], [6]. A good survey of the state of the art in social networks' privacy can be found in [23].

B. Contributions

To our knowledge, this is the first work that extends the existing results on identity protection in social networks [4], [8], [12], [22] to also guard against the disclosure of sensitive information/attributes associated to network's nodes. An equivalent model for flat microdata would be one that guards against attribute disclosure [10].

This paper's contributions are: introducing a new privacy model for social network data entitled *p*-sensitive *k*-anonymity which combines the existing *k*-anonymity model for social networks [4] and the *p*-sensitive *k*-anonymity model for microdata [19], integrating existing algorithms for the *p*-sensitive *k*-anonymity for microdata and the *k*-anonymity for social networks to generate a *p*-sensitive *k*-anonymous social network, and performing experiments that prove the validity of the proposed model and algorithm.

II. *P*-Sensitive *K*-Anonymous Social Networks

We consider a social network to be 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 user of the social network; each edge represents a

relationship or connection between two users in the network.

All nodes in $\mathcal{G}$ are described by a set of attributes. These attributes can be classified as follows

- $I_1, I_2, \ldots, I_m$ are *identifier* attributes such as *Name* and *SSN*.
- $Q_1, Q_2, \ldots, Q_q$ are *quasi-identifier* attributes such as *ZipCode* and *Age*. They may be known from other public datasets and could be potentially used to violate individuals' privacy.
- $S_1, S_2, \ldots, S_r$ are *sensitive* attributes such as *Disease*. These attributes' values must be protected against disclosure.

There are two related aspects in anonymizing a social network modeled as described above. Both the data associated to the social network's nodes, $\mathcal{N}$, (identifier, quasi-identifier and sensitive attributes) and the structural information the network carries about the nodes' relationships, $\mathcal{E}$, have to be properly masked. The resulting masked network data has to protect the nodes against: identity disclosure (i.e. determining who exactly is the individual owning the node) and attribute disclosure (i.e. finding out sensitive data about an individual, but without identity disclosure).

The process of anonymizing the nodes' attributes consists of removing the identifier attributes from the nodes information, ensuring that the quasi-identifier information is at least *k*-anonymous [16], [18], and ensuring that the sensitive attributes are at least *p*-sensitive [19]. *P*-sensitive *k*-anonymity property for nodes' data can be obtained by generalizing the quasi-identifier information, either with hierarchy-free generalization [11] for numeric data, or predefined hierarchies [13] for categorical data. The nodes' data generalization we envision is performed based on a partitioning of the node set $\mathcal{N}$ into distinct *clusters*. The nodes' data generalization is performed at the cluster-level: each cluster will have identical quasi-identifier values for all nodes, a minimum size of *k*, and at least *p* distinct values for each sensitive attribute. The network's structural information (edges) is also masked starting from the established partitioning of $\mathcal{N}$ into clusters. Basically, the detailed connectivity information of the individual nodes in a cluster is replaced with a summary of intra-connectivity and inter-connectivity information of the cluster as a whole. So, the essential task in anonymizing a social network is partitioning the node set $\mathcal{N}$ – how we conduct this step and the reasoning behind it will be explained in the next section.

The goal of the anonymization process is not only to produce a masked *p*-sensitive *k*-anonymous social network (formally defined next), but also to create a good-quality masked social network. The quality of an anonymized social network is given by the amount of information that it preserves from the original unmasked network: lower information loss means higher quality of the anonymous network. As we are dealing with graph data, the measures we use for quantifying information loss need to be sensitive to both the change in the quasi-identifiers attributes and the change in the structural information that occurs due to edge anonymization [4]. To measure the information lost from nodes' quasi-identifier attributes generalization we use the generalized information loss, a measure defined in [2]. For assessing the structural information loss in the edge-anonymization process, we use the measure introduced in [4].

Definition 1 (*generalization information loss*): Let *cl* be a cluster and $Q_I = \{N_1, N_2, .., N_s, C_1, C_2, .., C_t\}$ the set of numerical and categorical quasi-identifier attributes. The ***generalization information loss*** caused by generalizing quasi-identifier attributes of the *cl* nodes is:

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

where:

- *cl*'s generalization information, denoted by *gen*(*cl*), is the "node" having as value for each quasi-identifier attribute, numerical or categorical, the most specific common generalized value for all that attribute values from *cl* nodes (see a formal definition in [3]);
- $|cl|$ denotes the cluster *cl*'s cardinality;
- size($[i_1, i_2]$) is the size of the interval $[i_1, i_2]$, i.e. $(i_2 - i_1)$;
- N_i, $i = 1..s$ are numerical quasi-identifier attributes;
- C_i, $i = 1..t$ are categorical quasi-identifier attributes
- $\Lambda(w)$, $w \in \mathcal{H}_{Cj}$ is the subhierarchy of the C_j's predefined value hierarchy ($\mathcal{H}_{Cj}$) rooted in *w*;
- *height*($\mathcal{H}_{Cj}$) denotes the height of the tree hierarchy $\mathcal{H}_{Cj}$.

To be able to compare this measure with the structural information loss, we normalize it to the range [0, 1]. This is shown in the Definition 2. Detailed justification of Definitions 1 and 2 can be found in [4].

Definition 2 (***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:

$$NGIL(\mathcal{G},\mathcal{S}) = \frac{\sum_{j=1}^{v} GIL(cl_j)}{n \cdot (s+t)}.$$

where:

- *n* is the number of nodes for the graph $\mathcal{G}$;
- $(s + t)$ is the number of quasi-identifier attributes.

Structural information loss 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*. These components occur due to an edge anonymization process [4]. In the anonymized graph, 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 of the cluster. This intra-cluster edge generalization causes an intra-cluster information loss. In a similar way, given any 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 ($e \in \mathcal{E}_{cl1,cl2}$ iff $e \in \mathcal{E}$ and $e \in cl_1 \times cl_2$). In the anonymized graph, 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 process is called inter-cluster edge generalization and induces inter-cluster information loss [4]. Based on this structural generalization method, a structural information loss (*SIL*) measure and a corresponding normalized structural information loss (*NSIL*) measure are derived. Due to the lack of space, for a complete definition we refer the reader to [4].

Given a partition of nodes for a social network $\mathcal{G}$, we are able to create an anonymized graph by the generalization techniques explained above.

Definition 3 (***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 $\mathcal{N}$, $\bigcup_{j=1}^{v} cl_j = \mathcal{N}$; $cl_i \bigcap cl_j = \varnothing$; $i, j = 1..v$, $i \neq j$; 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_i is a node corresponding to the cluster $cl_j \in \mathcal{S}$ and is described by the "tuple" $gen(cl_j)$, the intra-cluster generalization pair ($|cl_j|$, $|\mathcal{E}_{clj}|$), and the projection on all sensitive attributes of the nodes from 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 $\exists\ X \in cl_j$, $Y \in cl_j$, 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 $|\mathcal{E}_{cli,clj}|$.

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 3, 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 4 (***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 $Cl_j = [gen(cl_j), (|cl_j|, |\mathcal{E}_{clj}|)]$, $j = 1, \ldots, v$ is *k*-anonymous iff $|cl_j| \geq k$ for all $j = 1, \ldots, v$.

Now we have all the tools to introduce the *p*-sensitive *k*-anonymous masked social network that combines the above definition with the *p*-sensitive *k*-anonymity property for microdata [19].

Definition 5 (***p-sensitive k-anonymous masked social network***): A masked social network $\mathcal{MG} = (\mathcal{MN}, \mathcal{ME})$, where $\mathcal{MN} = \{Cl_1, Cl_2, \ldots, Cl_v\}$, is *p*-sensitive *k*-anonymous if it is *k*-anonymous and the number of distinct values for each sensitive attribute is at least *p* within each Cl_j, $j = 1..v$.

III. SANGREEA_PK ALGORITHM

The *SaNGreeA_PK* algorithm builds on the work done by Campan et. al. in the areas of *k*-anonymity in social networks [4] and *p*-sensitive *k*-anonymity for microdata [3]. By combining the two algorithms presented in these papers, we developed the *SaNGreeA_PK* algorithm, which is able to perform *p*-sensitive *k*-anonymization for social networks.

The algorithm functions by taking the nodes of the social network and grouping them in a way that ensures *p*-sensitiveness of the formed clusters. The clusters formed will also have cardinality greater than *k*. Of course, some preconditions have to be respected for a social network to be amenable to *p*-sensitive *k*-anonymity, for given *p* and *k* values. For example, the social network must have at least *p* unique values for each of its nodes' sensitive attributes, and at least *k* nodes.

The cluster formation process is performed in a greedy manner. Each cluster is started from an initial seed node and fed with one other node at a time, until it becomes *p*-sensitive and *k*-anonymous. The node to be included in the currently developed cluster is the result of a greedy selection based on the values of the sensitive attributes and the levels of information loss (both through structural and attribute generalization) introduced in the summarization of the quasi-identifiable information. The functions that guide the selection process are: the *diversity* between the cluster being formed and the new node [3], the *NGIL*, and a function called *structural distance* that aims to limit the *SIL* (obviously, (*N*)*SIL* cannot be used as long as a complete partitioning of $\mathcal{N}$ is not known) [4].

The diversity between a cluster and a new node helps achieve the desired level of sensitivity in each cluster. We introduce next this measure. Let X^i, i=1..n, be all the nodes from the social network . We denote an node label information as $X^i = \{k_1^i, k_2^i, \ldots, k_q^i, s_1^i, s_2^i, \ldots, s_r^i\}$, where k^i s are the values for the quasi-identifier attributes and s^i s are the values for the confidential attributes.

Definition 6 (*diversity of two tuples*): The ***diversity of two tuples***, X^i and X^j w.r.t. the sensitive attributes is given by:

$diversity(X^i, X^j) = \sum_{l=1}^{r} w_l \cdot \delta(s_l^i, s_l^j)$, where

$\delta(s_l^i, s_l^j) = \begin{cases} 1, if\ s_l^i <> s_l^j \\ 0, if\ s_l^i = s_l^j \end{cases}$ and $\sum_{l=1}^{r} w_l = 1$ are the weights of the sensitive attributes.

The data owner can choose different criteria to define this weights vector. One good selection of the weight values is to initialize them as inversely proportional to the number of distinct sensitive attribute values in the original dataset. Along this paper we use this choice for the weights in all the experiments.

Definition 7 (*diversity between a tuple and a cluster*): The ***diversity between a tuple*** X^i ***and a cluster*** *cl* is given by $diversity(X^i, cl) = \sum_{l=1}^{r} w_l \cdot \rho(s_l^i, cl)$, where:

$\rho(s_l^i, cl) = \begin{cases} 1, \text{if } s_l^i \text{ does not exist between the } S_l \text{ values in } cl \\ 0, \text{if } s_l^i \text{ exists between the } S_l \text{ values in } cl \end{cases}$ and

$\sum_{l=1}^{r} w_l = 1$ have the same meaning as in **Definition 6**.

The justification that stands behind the selection of the structural distance for guiding cluster formation is presented in [4]. The formal definition of the structural distance measure follows. Assuming that the nodes in $\mathcal{N}$ have an order, $\mathcal{N} = \{X^1, X^2, \ldots, X^n\}$, we represent the neighborhood of each node X^i as an *n*-dimensional boolean vector $B_i = \left(b_1^i, b_2^i, ..., b_n^i\right)$, where $b_j^i = 1$ 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.

Definition 8 (***structural distance between two nodes***): The ***structural distance between two nodes*** (X^i and X^j) described by their associated *n*-dimensional boolean vectors B_i and B_j is:

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

Definition 9 (***structural distance between a node and a cluster***): The ***structural distance between a node*** *X* ***and a cluster*** *cl* is defined as the average distance between *X* and every node from *cl*:

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

The following documents the new *SaNGreeA_PK* algorithm, which combines the *SaNGreeA* algorithm for anonymizing social networks [4] with the *GreedyPKClustering* algorithm that anonymizes microdata to conform to *p*-sensitivity *k*-anonymity [3].

```
Algorithm SaNGreeA_PK is

Input: G = (N, E) - a social network
       k - as in k-anonymity
       p - as in p-sensitivity
       α, β - user-defined weight parameters;
       allow controlling the balancing
       between NGIL and NSIL
Output: S = {cl1, cl2,…, clv}; ⋃_{j=1}^{v} clj = N;

        cli ⋂ clj = ∅, i,j = 1..v, i≠j;

        |clj|≥k, j = 1..v - a set of clusters
        that ensures p-sensitive k-anonymity
        for MG = (MN, ME) so that a cost
        measure is optimized;

S = ∅;
i = 1;
r_seed = a randomly selected node from N;
Repeat
   r_seed = arg max_{r∈N}(diversity(r_seed, r));
  cli = {r_seed};
  N = N - {r_seed};
  Repeat
    // make cli p-sensitive; for that, find
    // the set of most diverse nodes w.r.t. cli
    divcl = argmax_{r∈N}(diversity (r, cli));
    X' = arg min_{X∈divcl}(α * NGIL(G1, S1) + β * sdist(X, cli));
    // G1: subgraph induced by cli∪{X'}in G
    // S1: partition with 1 cluster cli∪{X'}
    cli = cli ∪ {X'};
    N = N - {X'};
  Until (cli is p-sensitive) or (N = ∅);
  If (|cli| < k) and (N ≠ ∅) then
    Repeat
      // add nodes until cli has k nodes
      X' = arg min_{X∈N}(α * NGIL(G1, S1) + β * sdist(X, cli));
      cli = cli ∪ {X'};
      N = N – {X'};
    Until (|cli| ≥ k) or (N = ∅);
  End If;
  If (|cli|≥k and cli is p-sensitive) then
    S = S ∪ {cli};
    i++;
  Else
    // this only happens to last cluster
    DisperseCluster (S, cli);
  End If;
Until N = ∅;
End SaNGreeA_PK;

Function DisperseCluster(S, cl)
  S = S – cl;
  For every r ∈ cl do
    clu = FindBestCluster (r, S);
    clu = clu ∪ {r};
  End For;
End DisperseCluster;
```

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

To illustrate this algorithm, we give an example that shows how *p*-sensitive *k*-anonymity is achieved in a sample social network. Suppose the social network $\mathcal{G}_{ex}$ as shown in Figure 1 is given. The contents of the nodes are given in Table 1. The quasi-identifiers in this example are *age*, *zip* and *gender*, and the sensitive attribute is *illness*. The attribute *age* is numerical and subject to a hierarchy-free generalization. The value generalization hierarchies for the other two quasi-identifiers are given in Figure 2.

By running the *SanGreeA_PK* algorithm for this dataset with $k = 3$, $p = 2$, $\alpha = 0$, and $\beta = 1$, the masked social network $\mathcal{MG}_{ex1}$, shown in Figure 3, is generated. Due to the choice of α and β values, *SanGreeA_PK* guides the cluster formation to optimize the structural information loss and disregards the generalization information loss. When run with $k = 3$, $p = 2$, $\alpha = 1$, and $\beta = 0$, the masked social network $\mathcal{MG}_{ex2}$, shown in Figure 4, is generated. In this case, as $\beta = 0$, the generalization information loss is the only cost metric *SanGreeA_PK* tries to minimize in the cluster formation process.

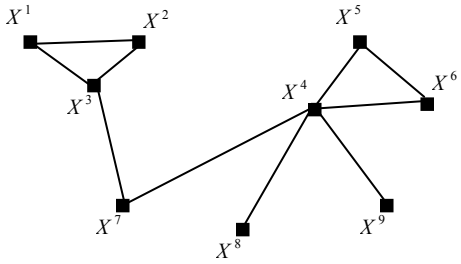

Fig. 1. The social network $\mathcal{G}_{ex}$.

TABLE 1
THE NODES' QUASI-IDENTIFIER AND SENSITIVE ATTRIBUTES IN $\mathcal{G}_{EX}$

Node	Age	Zip	Gender	Illness
X^1	25	41076	Male	Diabetes
X^2	25	41075	Male	Heart Disease
X^3	27	41076	Male	Diabetes
X^4	35	41099	Male	Colon Cancer
X^5	38	48201	Female	Breast Cancer
X^6	36	41075	Female	HIV
X^7	30	41099	Male	Diabetes
X^8	28	41099	Male	HIV
X^9	33	41075	Female	Colon Cancer

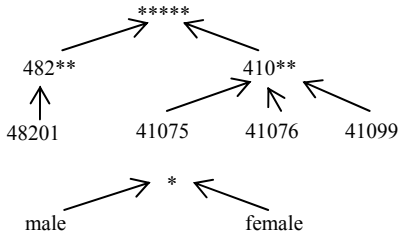

Fig. 2. The value generalization hierarchies for attributes *zip* and *person*

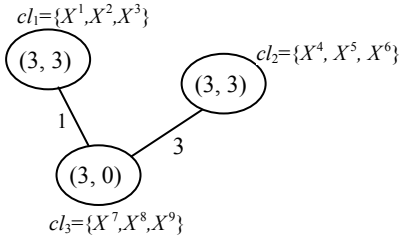

Fig. 3. The social network $\mathcal{MG}_{ex1}$.

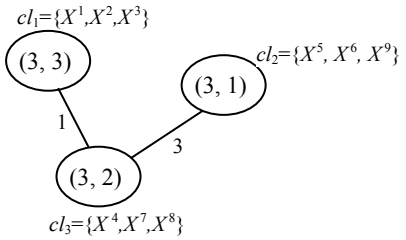

Fig. 4. The social network $\mathcal{MG}_{ex2}$.

When we analyze the resulting graph for both generalization and structural information loss, we produce the results shown in Table 2.

TABLE 2
THE GENERALIZATION AND STRUCTURAL INFORMATION LOSS FOR THE SUMMARIZATIONS OF $\mathcal{G}_{ex}$

$\mathcal{MG}$	GIL	NGIL	SIL	NSIL
$\mathcal{MG}_{ex1}$	14.30	0.53	5.77	0.32
$\mathcal{MG}_{ex2}$	7.73	0.28	8.44	0.46

IV. EXPERIMENTAL RESULTS

In this section we compare the *SaNGreeA_PK* algorithm with the *SaNGreeA* algorithm [4], over various combinations of *k* and *p* values, in terms of the generalization and structural information loss of the masked social networks they generate.

The two algorithms were implemented in Java. The tests were executed on a single CPU machine running at 2.53 GHz with 2GB of RAM and Windows XP Professional.

The algorithms were tested with a social network derived from the Enron e-mail dataset, an ex-employee

status report developed by Shetty et.al. [17], and the Adult dataset from the UC Irvine Machine Learning Repository [15]. The Adult dataset was necessary, as it provided quasi-identifier and sensitive attribute values for the nodes of our test social network. This type of information has been stripped out of the publicly available Enron e-mail dataset, to respect the privacy of the Enron employees. However, the Enron e-mail dataset provided the structural information for our test social network.

The nodes' data we formed contained the quasi-identifier attributes: *role* (coming from the ex-employee status report), *age*, *marital_status*, and *race* (coming from the Adult database). *Age* is the only numeric quasi-identifier; the other three quasi-identifiers are categorical. The heights of their predefined generalization hierarchies are 2 (for *role*), 2 (for *marital_status*), and 1 (for *race*). *Education* and *salary_range* are the sensitive attributes.

The attribute *role* is taken from the Enron ex-employee status report by matching the first and last name on the report with the e-mail database first and last names, resulting in 121 matching records. A Roles table was created that correlated role, education and salary ranges, as shown in Table 3.

TABLE 3
THE ROLES TABLE

Role	Education	Min Salary	Max Salary
Trader	Assoc-voc	40	110
Manager	Bachelors	40	110
Managing Director	Bachelors	70	150
CEO	Doctorate	90	200
President	Doctorate	90	200
N/A	HS-grad	30	60
Director	Masters	70	150
Vice President	Masters	90	200
In House Lawyer	Prof-school	40	110
Director of Trading	Prof-school	70	150
Employee	Some-college	30	60

From this table, the *education* attribute was used to randomly match and select a record from the Adult dataset; that particular record provided the values for the *marital_status*, *age*, and *race* attributes. A salary value was then randomly generated, using a uniform distribution, for each individual, within the range associated to the individual's Enron role. Exact salary values were then transformed into the following reporting ranges: $25-50K, $51-75K, $76-100K, $101-125K, $126-150K, $151-175K, and $176-200K.

The edges of the graph were derived in the same way as in [17], with two users being considered related if they had exchanged 5 e-mails with each other. Once extracted, it was discovered that some nodes had dropped off the graph, as they did not meet the 5 e-mail connection rule with any of the other users. Finally, the edge information was merged with the node information. Any isolated node and any edge for which one end had been already eliminated from the node set were eliminated. The resulting graph consists of 84 nodes and 191 edges.

The created test social network was anonymized with the *SaNGreeA* and *SaNGreeA_PK* algorithms, varying the value of k from 2 to 15 and the values of p from 2 to $min(7, k)$. The experiments were run with two different (α, β) parameter values: (0, 1) and (1, 0). The pair (0, 1) guides the algorithm towards minimizing structural information loss, while (1, 0) reduces information loss due to the generalization of the quasi-identifier attributes. The normalized generalization information loss (*NGIL*) and the normalized structural information loss (*NSIL*) were computed for the resulting masked social networks. Figure 4 presents the resulting *NGIL* and *NSIL* values for $(\alpha, \beta) = (0, 1)$. Figure 5 depicts the same measures for the pair (1, 0).

Looking at the results, we can see that in general there is a correlation between an increase in the values of k and p and an increase in structural and generalization information loss values. The experiments also show that increasing k values (for a fixed p) are reflected in a greater information loss (both *SIL* and *GIL*) than when increasing p values for the same k.

As expected, α and β parameters values controlled the trade-off between *SIL* and *GIL*. For the same k and p values *SIL* is lower when $\alpha = 0$ then when $\alpha = 1$. Similarly, *GIL* is lower when $\beta = 0$ then when $\beta = 1$.

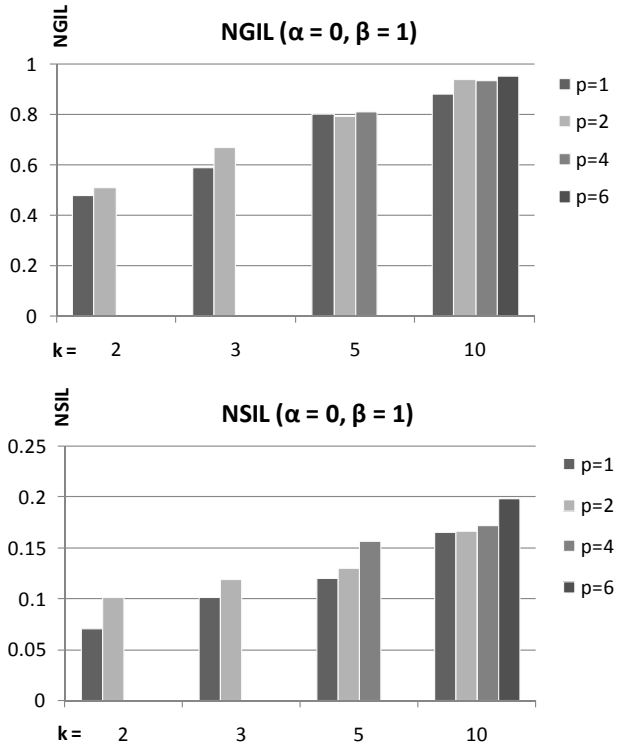

Fig. 4. The *NGIL* and *SGIL* values for the test social network with $\alpha = 0, \beta = 1$.

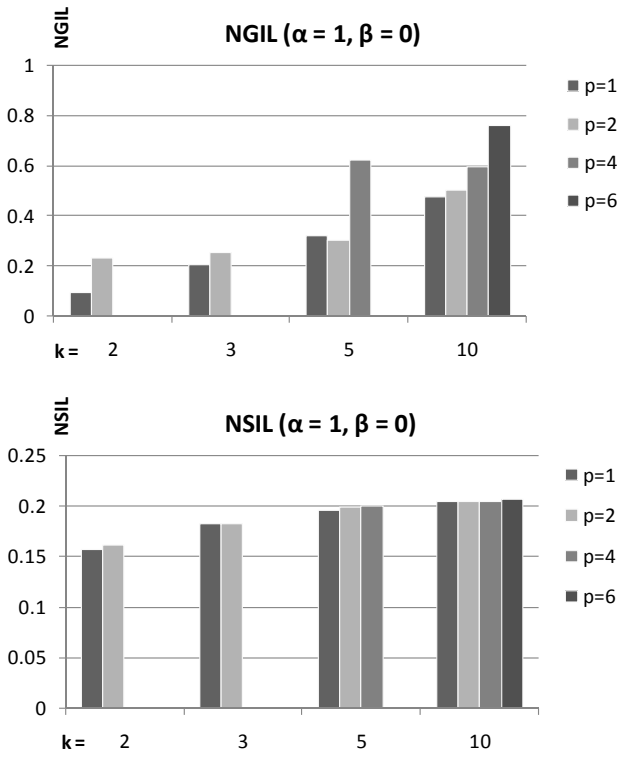

Fig. 5. The *NGIL* and *SGIL* values for the test social network with $\alpha = 1, \beta = 0$.

V. Conclusions and Future Work

In this paper we extended the existing results on identity protection in social networks to also guard against the disclosure of sensitive information/attributes associated to network's nodes. To achieve this extension we introduced a new privacy model for social network data entitled *p*-sensitive *k*-anonymity. We also integrated existing algorithms for the *p*-sensitive *k*-anonymity for microdata and the *k*-anonymity for social networks into a new algorithm entitled *SaNGreeA_PK*. Our experiments showed that the new algorithm generates *p*-sensitive *k*-anonymous social networks with their corresponding information loss is similar to the existing *SaNGreeA k*-anonymity algorithm with only a modest increase in structural and generalization information loss. The new proposed algorithm can also be user-balanced towards preserving more the structural information of the network or the nodes' attribute values.

We consider two possible directions to extend this work:

- Analyze, using social networks from different areas, the utility of the anonymized social network. This may lead to the development of more practical data utility / information loss measures.
- Formally study how the greedy criteria can be improved based on the properties of the social network data and the selected *p* and *k* values.

References

[1] L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Stenography," In *International World Wide Web Conference* (*WWW*), 2007, pp. 181 – 190.

[2] J.W. Byun, A. Kamra, E. Bertino, and N. Li, "Efficient k-Anonymization using Clustering Techniques," In *International Conference on Database Systems for Advanced Applications* (*DASFAA*), 2007, pp. 188 – 200.

[3] A. Campan, T.M. Truta, J. Miller, and R. Sinca, "A Clustering Approach for Achieving Data Privacy," In *International Conference on Data Mining* (*DMIN*), 2007, pp. 321 – 327.

[4] A. Campan and T.M. Truta, "A Clustering Approach for Data and Structural Anonymity in Social Networks," In *Privacy, Security, and Trust in KDD Workshop* (*PinKDD*), 2008.

[5] B. Carminati, E. Ferrari, and A. Perego, "Private Relationships in Social Networks," In *Private Data Management Workshop* (*PDM*), 2007, pp. 163 – 171.

[6] J. Domingo-Ferrer, A. Viejo, F. Sebe, and U. Gonzalez-Nicolas, "Privacy Homomorphism for Social Networks with Private Relationships," In *Computer Networks*, Vol. 52, Issue 15, 2008, pp. 3007 – 3016.

[7] R. Gross and A. Acquisti, "Information Revelation and Privacy in Online Social Networks (The Facebook Case)," In *Workshop on Privacy in the Electronic Society* (*WPES*), 2005, pp. 71 – 80.

[8] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis, "Resisting Structural Re-identification in Anonymized Social Networks," In *Very Large Data Base Conference* (*VLDB*), 2008, pp. 102 – 114.

[9] A. Korolova, R. Motwani, S. U. Nabar, and Y. Xu, "Link Privacy in Social Networks," In *International Conference on Data Engineering* (*ICDE*), 2008, pp. 1355 – 1357.

[10] D. Lambert, "Measures of Disclosure Risk and Harm," In *Journal of Official Statistics*, Vol. 9, 1993, pp. 313 – 331.

[11] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional K-Anonymity," In *IEEE International Conference of Data Engineering* (*ICDE*), 2006, pp. 25.

[12] K. Liu and E. Terzi, "Towards Identity Anonymization on Graphs," In *ACM SIGMOD International Conference on Management of Data*, 2008, pp. 93 – 106.

[13] M. Lunacek, D. Whitley, and I. Ray, "A Crossover Operator for the k-Anonymity Problem," In *Genetic and Evolutionary Computation Conference* (*GECCO*), 2006, pp. 1713 – 1720.

[14] A. McCard and K. Anderson, "Focus on Facebook: Who Are We Anyway," In *Anthropology News*, Vol. 49, No. 3, 2008, pp. 10 – 12.

[15] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, "UCI Repository of Machine Learning Databases", available at: *www.ics.uci.edu/~mlearn/MLRepository.html*, 1998.

[16] P. Samarati, "Protecting Respondents Identities in Microdata Release," In *IEEE Transactions on Knowledge and Data Engineering*, Vol. 13, No. 6, 2001, pp. 1010 – 1027.

[17] J. Shetty, and J. Adibi, "The Enron Email Dataset Database Schema and Brief Statistical Report," available at: *www.isi.edu/~adibi/Enron/Enron.htm*.

[18] L. Sweeney, "K-Anonymity: A Model for Protecting Privacy," In *International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems*, Vol. 10, No. 5, 2002, pp. 557 – 570.

[19] T. M. Truta and V. Bindu, "Privacy Protection: P-Sensitive K-Anonymity Property," In *Privacy Data Management Workshop* (*PDM*), 2006, pp. 94.

[20] X. Ying and X. Wu, "Randomizing Social Networks: A Spectrum Preserving Approach," In *SIAM International Conference on Data Mining* (*SDM*), 2008, pp. 739 – 750.

[21] E. Zheleva and L. Getoor, "Preserving the Privacy of Sensitive Relationships in Graph Data," In *Privacy, Security, and Trust in KDD* (*PinKDD*), *LNCS*, Vol. 4890, 2008, pp. 153 –171.

[22] B. Zhou and J. Pei, "Preserving Privacy in Social Networks against Neighborhood Attacks," In *IEEE International Conference on Data Engineering* (*ICDE*), 2008, pp. 506 – 515.

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