

Comparisons of Randomization and K-degree Anonymization Schemes for Privacy Preserving Social Network Publishing

Xiaowei Ying, Kai Pan, Xintao Wu, Ling Guo
University of North Carolina at Charlotte
{xying, kpan, xwu, lguo2}@uncc.edu

ABSTRACT

Many applications of social networks require identity and/or relationship anonymity due to the sensitive, stigmatizing, or confidential nature of user identities and their behaviors. Recent work showed that the simple technique of anonymizing graphs by replacing the identifying information of the nodes with random ids does not guarantee privacy since the identification of the nodes can be seriously jeopardized by applying background based attacks. In this paper, we investigate how well an edge based graph randomization approach can protect node identities and sensitive links. We quantify both identity disclosure and link disclosure when adversaries have one specific type of background knowledge (i.e., knowing the degrees of target individuals). We also conduct empirical comparisons with the recently proposed K -degree anonymization schemes in terms of both utility and risks of privacy disclosures.

Keywords

Identity Disclosure; Link Disclosure; Social Network Randomization; Privacy Preserving Data Mining.

1. INTRODUCTION

Social networks are of significant importance in various application domains such as marketing, psychology, epidemiology and homeland security. The management and analysis of these networks have attracted increasing interest in the sociology, database, data mining and theory communities. Most previous studies are focused on revealing interesting properties of networks and discovering efficient and effective analysis methods [3–5, 8, 9, 13–15, 17, 20, 22–25].

The nodes in social networks are the individuals and the links among them denote their relationships. Many applications of social networks such as anonymous Web browsing require identity and/or relationship anonymity due to the sensitive, stigmatizing, or confidential nature of user identities and their behaviors. The privacy concerns associated with data analysis over social networks have incurred recent research works [2, 6, 11, 12, 19, 26, 27, 29, 30]. In this paper, we concentrate on two types of privacy breaches: *identity disclosure* and *link disclosure*. The identity disclosure corresponds

to the scenario where the identity of an individual who is associated with a node is revealed while the link disclosure corresponds to the scenario where the sensitive relationship between two individuals is disclosed. We assume all individuals (nodes) and relationships (links) among them are sensitive. Note that sensitive information may be associated with each node as sensitive attributes. We skip *attribute disclosure* in our paper.

To prevent identity and link disclosures, one natural approach is to publish a node-anonymized version of the network that permits useful analysis without disclosing the identity of the individuals represented by the nodes. However, as pointed out in [2, 12], this simple technique of anonymizing graphs by replacing the identifying information of the nodes with random ids does not guarantee identity/link privacy since adversaries may potentially construct a highly distinguishable subgraph with edges to a set of targeted nodes, and then to re-identify the subgraph and consequently the targets in the released anonymized network. Another approach is to further modify edges of the anonymized graphs to prevent identity/link disclosures in the presence of subgraph queries [11, 12, 19, 26, 30]. By introducing false edges in the anonymized graphs, we expect to decrease identity and link disclosures.

In this paper, we focus on one representative edge modification scheme, *Rand Add/Del without replacement*, and study how well it can protect node identities and sensitive links. In *Rand Add/Del*, we randomly add k false edges followed by deleting k true edges from the original graph. This scheme preserves the total number of edges in the original graph. Since the goal of an adversary is to map the nodes/edges in this randomized and anonymized graph to real world entities/relationships, we investigate the relationship between the amount of randomization and the adversary's ability to correctly infer the node identity and the presence of a link. Privacy is jeopardized if adversaries' confidence of prediction is higher than some tolerated threshold or is significantly greater than their a-priori belief (without the exploit of the released randomized graph).

Adversaries usually rely on background knowledge in order to de-anonymize nodes and learn the link relations between de-anonymized individuals from the released perturbed graph. It is challenging to model all types of background knowledge of adversaries in the scenario of publishing social networks with privacy preservation. In [31], the authors listed several types of background knowledge: attributes of vertices, vertex degrees, specific link relationships between some target individuals, neighborhoods of some target individuals, embedded subgraphs, graph metrics (e.g., betweenness, closeness, centrality). In this paper, we focus on one most widely used type of background knowledge, *vertex degree* and quantify both identity disclosure and link disclosure when adversaries know the degrees of target individuals, leaving other other types of back-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

The 3rd SNA-KDD Workshop '09 (SNA-KDD'09), June 28, 2009, Paris, France.

Copyright 2009 ACM 978-1-59593-848-0 ...\$5.00.

ground knowledge for future work.

In our empirical evaluation part, we compare our *Rand Add/Del* with another representative edge modification scheme *K-degree generalization* scheme proposed in [19] in terms of the tradeoff between disclosure risks and utility loss. The *K-degree generalization* scheme modifies a graph via a set of edge addition/deletion operations in order to construct a *K-degree* anonymous graph, in which every node has the same degree with at least $K - 1$ other nodes. Our empirical results show that generalized graph via the *K-degree generalization* scheme generally better preserves structural features than the randomized graph via the *Rand Add/Del*. In our future work, we shall investigate how to reconstruct structural features from the randomized graph rather than calculating structural features from the randomized graph directly. We expect to achieve more accurate structural feature values via reconstruction. It is also worth pointing out that the *K-degree generalization* scheme is designed to only protect the re-identification of individuals while the *Rand Add/Del* can provide both identity and link privacy protection.

1.1 Notations and Disclosure Measures

A network $G(n, m)$ is a set of n nodes connected by a set of m links. The network considered here is binary, symmetric, and without self-loops. Let $A = (a_{ij})_{n \times n}$ be its adjacency matrix, $a_{ij} = 1$ if node i and j are connected and $a_{ij} = 0$ otherwise. Associated with A is the degree distribution $D_{n \times n}$, a diagonal matrix with row-sums of A along the diagonal, and 0's elsewhere. \tilde{G} is the randomized graph obtained by either *Rand Add/Del* or *K-degree generalization*. We denote $\tilde{A} = (\tilde{a}_{ij})_{n \times n}$ be the adjacency matrix of \tilde{G} . Table 1 summarizes the notation used in this paper.

Let Ω denote the set of all individual identifiers in the network: $\Omega = \{Alice, Bob, \dots, Zack\}$, and let $\psi(\cdot)$ be the mapping from the individual identifier to the node random id in the anonymized graph: for any $\alpha \in \Omega$, $\psi(\alpha)$ is the node index of the individual α , and $\psi^{-1}(i)$ is the identity of node i . One natural question for data owners is, compared to not releasing the graph, to what extent releasing an anonymized/randomized graph \tilde{G} jeopardizes the privacy.

To quantify disclosure risk, we define two risk measures: prior risk measure $r(\omega)$ is defined as the adversary's prior confidence on the event ω without the released graph \tilde{G} ; and the posterior risk measure $r(\omega|\tilde{G})$ is defined as the adversary's posterior confidence given the released graph \tilde{G} .

For identity disclosure, we assume the adversary has vertex degree background knowledge, i.e., the target individual's degree is known to adversaries. To make the notation concise, we use d_α to denote the degree of individual α . We use $r(\alpha)$ to denote the adversary's prior confidence on identification of the target individual α . Correspondingly, we use $r(\alpha|d_\alpha, \tilde{G})$ to denote the posterior risk of individual α given the released randomized graph \tilde{G} and the degree of the target individual α (i.e., vertex degree background knowledge). We present our quantification results in Section 2.1.

For link disclosure, adversaries need to first identify target individual nodes (incorporating the vertex degree background knowledge, d_α, d_β , with the released graph \tilde{G}) and then compute the posterior belief of existence of the sensitive link (α, β) . We use $R(a_{\alpha\beta})$ and $R(a_{\alpha\beta}|d_\alpha, d_\beta, \tilde{G})$ to denote the prior risk and posterior risk respectively. We present our results in Section 2.2.

1.2 Organization

The remainder of paper is outlined as follows. In Section 2, we focus on the quantification of identity disclosure and link disclosure

in the released randomized graph when adversaries have vertex degree background knowledge of target individuals. In Section 3, we empirically show how the graph characteristics (including two spectral features and four real features) vary when *Rand Add/Del* and *K-degree Generalization* schemes are applied. In Section 4, we discuss related work including other potential attacks on the randomized graph and other randomization strategies. We conclude and discuss our future work in Section 5.

2. DISCLOSURE ANALYSIS IN RAND ADD/DEL

Throughout this section, we illustrate our theoretical results using empirical evaluations on the US politics book data [16], which contains 105 vertices and 441 edges. As shown in Figure 1(a), nodes represent books about US politics sold by the online bookseller Amazon.com while edges represent frequent co-purchasing of books by the same buyers on Amazon. Nodes are separated into groups according to their political views: "liberal", "neutral", or "conservative". Figure 1(b) shows the histogram of its degree sequence. For example, there are 22 nodes with degree 5 and one node with degree 20. In the remainder of this paper, we use one node (random id 15, identifier label "Breakdown") with degree 5 and the node (random id 30, identifier label "The Price of Loyalty") with degree 20 to illustrate our results.

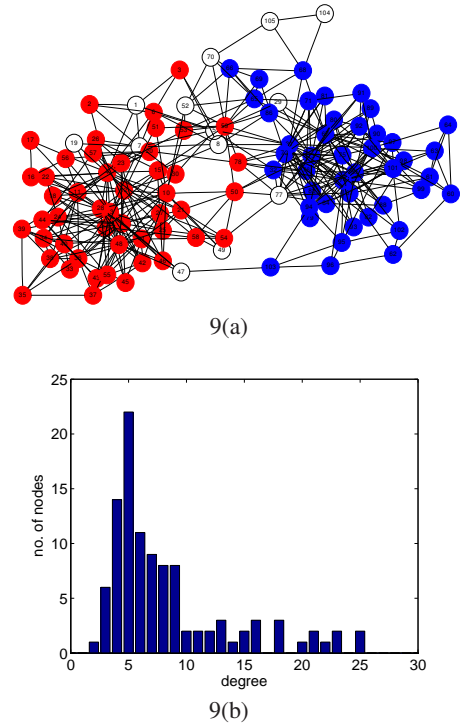


Figure 1: (a) the politics book network; (b) the histogram of its degree sequence.

2.1 Identity Disclosure

In this section, we focus on identity disclosure in the randomized graph. We study the adversary's strategy and then quantify identity disclosure. We assume the adversary has vertex degree background knowledge, i.e., the degree of the target individual is known. The adversary needs to take a guess on the mapping function ψ based on his background knowledge and the released graph \tilde{G} . In other

Table 1: Table of Notations

Symbol	Meaning
G, \tilde{G}	original graph, randomized graph
n, m, N	the graph contains n nodes and m links, $N = \frac{n(n-1)}{2}$
k	randomization parameter: add/delete k links for k times
A, \tilde{A}	adjacency matrices of G and \tilde{G}
a_{ij}, \tilde{a}_{ij}	the (i, j) element in A and \tilde{A}
d_i, \tilde{d}_i	degree of node i in G and \tilde{G}
n_d	number of nodes whose degree is d , $n_d = \{i : d_i = d\} $
Ω	set of individual identifiers in the graph
α, β	two individual identifiers, $\alpha, \beta \in \Omega$
$a_{\alpha\beta}, \tilde{a}_{\alpha\beta}$	link between individual α and β in G and \tilde{G}
$d_\alpha, \tilde{d}_\alpha$	degree of individual α in G and \tilde{G}
$r(\alpha)/r(\alpha \tilde{G})$	prior/posterior risk of individual α , unknown node identities
$R(a_{\alpha\beta})/R(a_{\alpha\beta} \tilde{G})$	prior/posterior risk of link (α, β) , unknown node identities
$\tau_a(\alpha \tilde{G})/\tau_r(\alpha \tilde{G})$	absolute/relative protection measure of α , unknown node ID
$\Gamma_a(a_{\alpha\beta} \tilde{G})/\Gamma_r(a_{\alpha\beta} \tilde{G})$	absolute/relative protection measure of (α, β) , unknown node ID

words, the adversary wants to re-identify which node is corresponding to the target individual α using the background knowledge of degree d_α . To re-identify α in the node set, the adversary can utilize the randomized degree sequence $\tilde{\mathbf{d}} = (\tilde{d}_1, \tilde{d}_2, \dots, \tilde{d}_n)$. Hence, we can write the posterior risk measure $r(\alpha|d_\alpha, \tilde{G})$ as $r(\alpha|d_\alpha, \tilde{\mathbf{d}})$. Let $\hat{\psi}(\cdot)$ denote the adversary's guess of the mapping.

Without the released randomized graph, the background knowledge (such as the true degree of a target individual) cannot be used to enhance the adversary's confidence on the identity mapping. Hence, the prior risk measure $r(\alpha|d_\alpha) = \frac{1}{n}$. Next we deduce the posterior risk measure $r(\alpha|d_\alpha, \tilde{\mathbf{d}})$.

Recall in *Rand Add/Del* scheme, each true edge can remain in the graph with a probability $p_{11} = \frac{m-k}{m}$, and each non-existing link can be added with a probability $p_{10} = \frac{k}{N-m}$, where $N = \binom{n}{2}$. Let d_i and \tilde{d}_i denote the degree of node i in the G and \tilde{G} graph respectively, and \hat{d}_i is the adversary's estimator of d_i .

Lemma 1 shows the calculation of $P(\tilde{d}_i = x|d_i)$, i.e., the probability of a node's degree \tilde{d}_i after randomization given its original degree d_i .

LEMMA 1. *The distribution of \tilde{d}_i is given by*

$$P(\tilde{d}_i = x|d_i) = \sum_{t=0}^x B(t; d_i, p_{11})B(x-t; n-1-d_i, p_{10}), \quad (1)$$

where $B(t; n, p)$ denotes the probability mass function of the binomial distribution with parameter n and p . The expectation and variance of \tilde{d}_i are given by:

$$\mathbf{E}(\tilde{d}_i) = p_{11}d_i + p_{10}(n-1-d_i), \quad (2)$$

$$\mathbf{V}(\tilde{d}_i) = d_i p_{11}(1-p_{11}) + (n-1-d_i)p_{10}(1-p_{10}). \quad (3)$$

Proof. Let d_i^+ denote the remaining true edges after *Add/Del* process, and d_i^- denote the added links by the process. Since each existing or non-existing link are processed independently, d_i^+ and d_i^- follow the binomial distributions $B(d_i, p_{11})$ and $B(n-1-$

$d_i, p_{10})$ respectively:

$$P(d_i^+ = t|d_i) = B(t; d_i, p_{11}) = \binom{d_i}{t} p_{11}^t (1-p_{11})^{d_i-t}. \quad (4)$$

$$P(d_i^- = t|d_i) = B(t; n-1-d_i, p_{10}) \\ = \binom{n-1-d_i}{t} p_{10}^t (1-p_{10})^{n-1-d_i-t}. \quad (5)$$

Since $\tilde{d}_i = d_i^+ + d_i^-$, the distribution of \tilde{d}_i is just the convolution of (4) and (5) and we get (1). Note the d_i^+ and d_i^- are independent, then

$$\mathbf{E}(\tilde{d}_i) = \mathbf{E}(d_i^+) + \mathbf{E}(d_i^-) = p_{11}d_i + p_{10}(n-1-d_i),$$

$$\mathbf{V}(\tilde{d}_i) = \mathbf{V}(d_i^+) + \mathbf{V}(d_i^-) \\ = d_i p_{11}(1-p_{11}) + (n-1-d_i)p_{10}(1-p_{10}).$$

Rearrange (2), we can have the following result:

LEMMA 2. *Given a randomized graph, the moment estimator (ME) of d_i is given by:*

$$\hat{d}_i = \frac{\tilde{d}_i - p_{10}(n-1)}{p_{11} - p_{10}}, \quad (6)$$

and \hat{d}_i is the unbiased estimator of d_i .

The unbiased property is straightforward from (2).

By combining Lemma 1 and Lemma 2, we can calculate the posterior probability $P(d_\alpha|\tilde{d}_i)$ (i.e., the likelihood of the observed node i having the degree d_α in the original graph).

LEMMA 3. *In the randomized graph \tilde{G} , the adversary observes a node i with degree \tilde{d}_i , then the adversary's confidence on $d_i = x$ is given by*

$$P(d_i = x|\tilde{d}_i) = \frac{P(\tilde{d}_i|d_i = x)P(d_i = x)}{\sum_{d=0}^{n-1} P(\tilde{d}_i|d = x)P(d = x)}. \quad (7)$$

When the original degree distribution is unavailable to the adversary, the estimated degree sequence from (6) can be applied instead.

Lemma 3 is a direct result from Bayes' theorem.

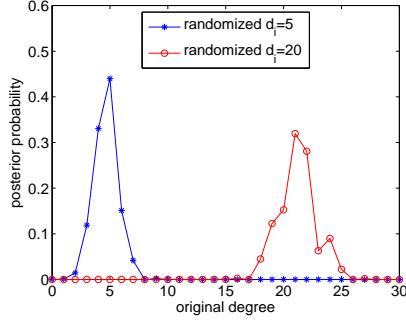


Figure 2: Values of $P(d_i|\tilde{d}_i = 5)$ and $P(d_i|\tilde{d}_i = 20)$ after applying *Rand Add/Del* on polbooks network ($k=10\%m$),

Figure 2 shows values of two posterior probabilities: $P(d_i|\tilde{d}_i = 5)$ and $P(d_i|\tilde{d}_i = 20)$. Generally speaking, the distribution of $P(d_i|\tilde{d}_i)$ is not symmetric, and it skews to the side with larger degree frequency. In Figure 2, for a node with $\tilde{d}_i = 20$, $P(d_i = 21|\tilde{d}_i = 20) > P(d_i = 20|\tilde{d}_i = 20) > P(d_i = 19|\tilde{d}_i = 20)$, this is because the adversary can estimate that, in the original graph $P(d_i = 21) > P(d_i = 20) > P(d_i = 19)$, and Lemma 3 incorporates this information in the calculation. We can also observe that the posterior probability that the original degree value d_i is far away from the observed value \tilde{d}_i tends to be zero. In other words, it is very unlikely that a node's degree has a significant change after perturbation.

Recall our node identification problem is that given the true degree d_α of a target individual α , the adversary aims to discover which node in the randomized graph corresponds to individual α . To the adversary, every node in the randomized graph is possible with probability $P[d_\alpha|\tilde{d}_i]$.

Given a list of posterior probabilities $P(d_\alpha|\tilde{d}_i)$ calculated using Lemma 3, the adversary can make the following probabilistic decision:

$$\hat{\psi}(\alpha) = i, \text{ with probability } \frac{P[d_i = d_\alpha|\tilde{d}_i]}{\sum_{j=1}^n P[d_j = d_\alpha|\tilde{d}_j]}. \quad (8)$$

RESULT 1. Assume the node identities are unknown to the adversary. For any individual $\alpha \in \Omega$, the prior risk measure is

$$r(\alpha|d_\alpha) = \frac{1}{n}. \quad (9)$$

The posterior risk measure, which equals to the accuracy of the probabilistic decision in (8), is then given by:

$$r(\alpha|d_\alpha, \tilde{\mathbf{d}}) = P[\hat{\psi}(\alpha) = \psi(\alpha)] = \frac{P[d_\alpha|\tilde{d}_\alpha]}{\sum_{j=1}^n P[d_j = d_\alpha|\tilde{d}_j]}. \quad (10)$$

In our polbooks example, recall that we select two individuals: α (label "Breakdown") with known degree 5 and β (label "The price of Loyalty") with known degree 20. From Figure 1(b), we can see that there are 22 nodes with degree 5 and only one node with degree 20. Figure 3 shows values of $P(d_i = 5|\tilde{d}_i)$ and $P(d_i = 20|\tilde{d}_i)$. Using Equation 10, we can easily calculate identity disclosure risk, $r(\alpha|d_\alpha = 5) = 0.135$ and $r(\beta|d_\beta = 20) = 0.024$. It is intuitive to learn that identity disclosure risk given the vertex degree background knowledge is dependent on the degree distribution $P(d_i)$ of the original graph.

Another question is how the identity risk disclosure $r(\alpha|d_\alpha)$ varies with the magnitude of randomization. In Figure 4, we show

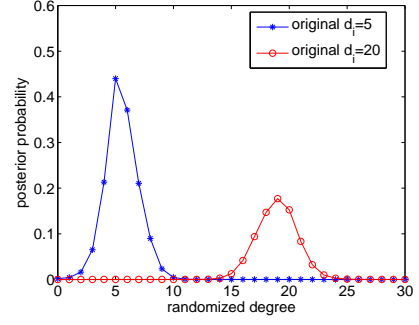


Figure 3: Apply *Rand Add/Del* on polbooks network ($k=10\%m$), values of $P(d_i = 5|\tilde{d}_i)$ and $P(d_i = 20|\tilde{d}_i)$ when \tilde{d}_i varies.

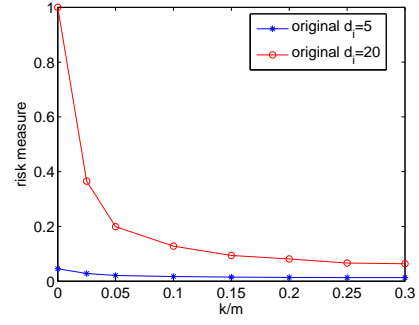


Figure 4: $r(\alpha|d_\alpha)$ vs. k after applying *Rand Add/Del* on polbooks network

how two identity disclosure risks, $r(\alpha|d_\alpha = 5)$ and $r(\beta|d_\beta = 20)$, vary as the perturbation magnitude (k) changes. We can observe that both identity disclosure risks decrease when k increases. The risk value $r(\alpha|d_\alpha = 5)$ is consistently low even if very few or no perturbations are introduced. This is because there are 22 nodes with the degree 5 in the original graph. However, for $r(\beta|d_\beta = 20)$, we can see that randomization can significantly decrease its disclosure risk: the disclosure risk is 100% when we release the anonymized graph without edge randomization while the disclosure risk decreases 0.39 (0.2) when we apply *Rand Add/Del* with $k = 2.5\%m$ (5% m).

2.2 Link Disclosure

The adversary's goal is to predict whether there is a sensitive link between two target individuals $\alpha, \beta \in \Omega$ by exploiting the released graph and individual degrees d_α, d_β . Given the true degrees of α and β and one released graph \tilde{G} , let $R(\alpha, \beta|d_\alpha, d_\beta, \tilde{G})$ denote the posterior risk measure on the link between α and β when the node identities are unknown to the adversary. Similarly, $R(\alpha, \beta)$ is the prior risk measure on link disclosure.

LEMMA 4. For *Rand Add/Del* scheme, the prior and posterior risk measures of the existence of a link between node i and j are given by:

$$P(a_{ij} = 1) = \frac{m}{N}; \quad (11)$$

$$P(a_{ij} = 1|\tilde{a}_{ij}) = \begin{cases} \frac{m-k}{m}, & \text{if } \tilde{a}_{ij} = 1, \\ \frac{k}{N-m}, & \text{if } \tilde{a}_{ij} = 0. \end{cases} \quad (12)$$

where $N = n(n-1)/2$.

Lemma 4 shows the link disclosure risks on the simple scenario where node identities are available to adversaries, i.e., for any target individual $\alpha \in \Omega$, the adversary knows its corresponding index, $\psi(\alpha) = i$, in the released randomized graph.

In general, the adversary does not know individuals' corresponding node indices in the released graph. Instead, the adversary may only have vertex degree background knowledge, i.e., the degrees of target individuals are known.

RESULT 2. *In the scenario where node identities are unknown to the adversary, for any two individuals $\alpha, \beta \in \Omega$, the prior risk measure and the posterior risk measure given \tilde{G} on the link between α and β after applying *Rand Add/Del* scheme are given by:*

$$R(a_{\alpha\beta}) = \frac{m}{n^2 N}, \quad (13)$$

$$R(a_{\alpha\beta}|d_\alpha, d_\beta, \tilde{G}) = \frac{m-k}{m} \left(\frac{P[d_\alpha|\tilde{d}_\alpha]}{\sum_{j=1}^n P[d_j = d_\alpha|\tilde{d}_j]} \right) \left(\frac{P[d_\beta|\tilde{d}_\beta]}{\sum_{j=1}^n P[d_j = d_\beta|\tilde{d}_j]} \right). \quad (14)$$

Proof. Since our risk measures are essentially the accuracy of the adversary's predictions, risk measures can be expressed as:

$$R(a_{\alpha\beta}) = r(\alpha)r(\beta)P(a_{ij} = 1) \quad (15)$$

$$R(a_{\alpha\beta}|d_\alpha, d_\beta, \tilde{G}) = r(\alpha|d_\alpha, \tilde{\mathbf{d}})r(\beta|d_\beta, \tilde{\mathbf{d}})P(a_{ij} = 1|\tilde{a}_{ij}). \quad (16)$$

Combining (11), (12), (9), and (10) into (15) and (16), we have the result on the link risk for *Rand Add/Del* when node identities are unknown.

2.3 Privacy Protection vs. Perturbation k

From the data owner point of view, we are interested in how much perturbation should be introduced to protect privacy. To measure the privacy protection, we thus further define protection measures: the absolute protection measure $\tau_a(\omega)$ and the relative protection measure of $\tau_r(\omega)$. We are interested in relationships between identity (link) privacy protection and the perturbation magnitude k .

Identity Privacy Protection. The absolute and relative identity protection measures are straightforwardly defined as:

$$\tau_a(\alpha|\tilde{\mathbf{d}}) = 1 - r(\alpha|d_\alpha, \tilde{\mathbf{d}}), \quad \tau_r(\alpha|\tilde{\mathbf{d}}) = \frac{1 - r(\alpha|d_\alpha, \tilde{\mathbf{d}})}{1 - 1/n}.$$

Figure 5 shows the histogram distributions of relative protection measures $\tau_r(\alpha|\tilde{G})$ under three different perturbation magnitudes ($k = 5\%, 10\%, 20\%m$). We can easily observe that more nodes are protected when k increases. We can also observe that the distribution generally has skewness, which indicates the majority of nodes are resilient to vertex degree background knowledge attack even under a relatively moderate perturbation. The calculation of $r(\alpha|d_\alpha, \tilde{\mathbf{d}})$ in (10) needs an instance of the randomized graph. In practice, the data owner may expect to determine k before applying *Rand Add/Del* such that the randomized data satisfies some privacy protection threshold. Hence, we should use the expected randomized degree sequence shown in (2) to evaluate the protection measure and choose k such that

$$J(k) = \min_{\alpha \in \Omega} \tau_r[\alpha|\mathbf{E}(\tilde{\mathbf{d}})] \geq 1 - \varepsilon.$$

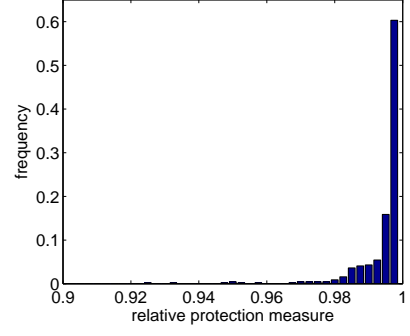


Figure 6: Histogram of $\Gamma_r(a_{\alpha\beta})$ for polbooks network, *Rand Add/Del* ($k = 10\%m$)

Table 2: Perturbation parameter k that meet the protection requirement for polbooks network

$1 - \varepsilon$	k for identity protection	k for link protection
0.5	27	8
0.6	32	9
0.7	59	12
0.8	110	16
0.9	257	37

Link Privacy Protection. Similarly, the link privacy protection measures are shown as:

$$\Gamma_a(a_{\alpha\beta}|\tilde{G}) = 1 - R(a_{\alpha\beta}|d_\alpha, d_\beta, \tilde{G}),$$

$$\Gamma_r(a_{\alpha\beta}|\tilde{G}) = \frac{1 - R(a_{\alpha\beta}|d_\alpha, d_\beta, \tilde{G})}{1 - R(\alpha, \beta)}.$$

Figure 6 shows the histogram of $\Gamma_r(a_{\alpha\beta}|\tilde{G})$ for polbooks network after we apply *Rand Add/Del* scheme ($k = 10\%m$). We can see that all Γ_r values are greater than 90%, and most links have their relative protection measure values close to 1, indicating that the protection of *Rand Add/Del* with $k = 10\%m$ almost achieves the same protection as without a released graph. Formally, we expect to choose a k such that

$$J(k) := \min_{\alpha, \beta} \Gamma_r(a_{\alpha\beta}|\tilde{G}) \geq 1 - \varepsilon. \quad (17)$$

Note that we use $\mathbf{E}(\tilde{\mathbf{d}})$ and $\tilde{a}_{\alpha\beta} = 1$ in calculating (17).

Table 2 shows the minimal k that meets the identity (link) protection requirement in (17) for polbooks network. We can see that *Rand Add/Del* scheme can generally achieve both identity protection and link protection with small or medium perturbations, e.g., $k = 59$ (or $k = 12$) for the relative protection threshold 0.7 of identity privacy (or link privacy). We can also observe that *Rand Add/Del* needs much fewer perturbations to achieve the link protection than the identity protection. This is because the adversary needs to identify the target two individuals before predicting the existence of a link between these two individuals.

3. COMPARISON WITH K -DEGREE GENERALIZATION SCHEME

In this section, we compare the *Rand Add/Del* scheme with the representative generalization based scheme (K -degree) in terms of the tradeoff between privacy protection and utility loss. Since the K -degree scheme is designed to protect the re-identification of in-

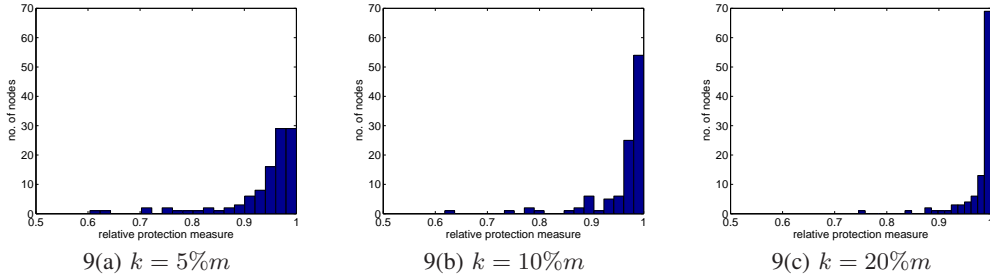


Figure 5: Histogram of $\tau_r(\alpha|\tilde{G})$ for 105 nodes in polbooks network, under *Rand Add/Del* scheme. The skewness of the distribution increases, indicating more nodes are well protected as k increases.

dividuals, we focus on identity privacy protection in empirical evaluations.

Data Sets. In addition to the previous polbook data, we also conducted evaluation on two relatively large data sets, Political blogosphere data (Polblogs) [1] and Enron Email data (Enron). Polblogs compiles the data on the links among US political blogs, containing 1,222 vertices and 16,714 edges. The blogs were labeled as either liberal or conservative, based on incoming and outgoing links and posts around the time of the 2004 presidential election. The original data is a directed graph. Here we simply consider $a_{ij} = 1$ if the two blogs have a link between them. The Enron network was built from email corpus of a real organization over the course covering a 3 years period. We used a pre-processed version of the dataset provided by [21]. This dataset contains 252,759 emails from 151 Enron employees, mainly senior managers. In this paper we focused on emails sent *from and to* these 151 people. An email graph is an undirected and un-weighted graph with edges connecting senders and recipients of emails during the corresponding time periods. The semantics of an edge (u, v) in such a graph is that there has been at least one email communications between u and v . We get the Enron network with 151 vertices and 1,377 edges.

3.1 K -degree Generalization Revisited

In [19], Liu and Terzi investigated how to modify a graph via a set of edge addition (and/or deletion) operations in order to construct a new K -degree anonymous graph, in which every node has the same degree with at least $K - 1$ other nodes. The authors imposed a utility requirement that the minimum number of edge-modifications is made in order to capture the requirement of structural similarity between the original and released randomized graphs. The K -degree anonymity property prevents the re-identification of individuals by the adversaries with a-priori knowledge of the number of social relationships of certain people (i.e., vertex background knowledge). The proposed algorithm is outlined below.

1. Starting from the degree sequence \mathbf{d} of the original graph $G(V, E)$, construct a new degree sequence $\tilde{\mathbf{d}}$ that is K -anonymous and the cost L_1 distance, $L_1(\tilde{\mathbf{d}} - \mathbf{d})$ is minimized.
2. Construct a new graph $\tilde{G}(\tilde{V}, \tilde{E})$ such that $\mathbf{d}_{\tilde{G}} = \tilde{\mathbf{d}}$, $\tilde{V} = V$, and $\tilde{E} = E$ (or $\tilde{E} \cap E \approx E$ in the relaxed version).

The first step is solved by a linear-time dynamic programming algorithm while the second step is based on a set of graph-construction algorithms given a degree sequence. In this paper, we will compare our *Rand Add/Del* with the K -degree algorithm (with simultaneous edge additions and deletions) proposed in [19].

3.2 Identity Privacy Protection vs. Utility Loss

Graph Characteristics vs. Utility. To achieve utility, we expect the released randomized graph should also keep structural properties not much changed or those properties can be reconstructed from the randomized graph. To understand and utilize the information in a network, researches have developed various measures to indicate the structure and characteristics of the network from different perspectives [7]. In this paper, we use the following four representative real space features.

- h , the harmonic mean of the shortest distance. The inverse of the harmonic mean of the shortest distance is also known as the global efficiency.
- Q , the modularity measure. It indicates the goodness of the community structure. It is defined as the fraction of all edges that lie within communities minus the expected value of the same quantity in a graph in which the vertices have the same degrees but edges are placed at random without regard for the communities.
- C , the transitivity measure. The transitivity measure is one type of clustering coefficient, which measures and characterizes the presence of local loops near a vertex.
- SC , the subgraph centrality. It is used to quantify the centrality of vertex i based the subgraphs.

$$SC = \frac{1}{n} \sum_{i=1}^n SC_i = \frac{1}{n} \sum_{i=1}^n \sum_{k=0}^{\infty} \frac{P_i^k}{k!} \quad (18)$$

where P_i^k is the number of paths that start with i and end in i with length of k .

Since it has been shown that the graph spectrum has close relations with the many graph characteristics and can provide global measures for some network properties [20], we also consider the following two spectral features.

- λ_1 , the eigenvalues of the adjacency matrix A . The maximum degree, chromatic number, clique number, and extend of branching in a connected graph are all related to λ_1 .
- μ_2 , the second eigenvalue of the Laplacian matrix defined as $L = D - A$. It can be used to show how good the communities separate, with smaller values corresponding to better community structures.

Table 3 shows our empirical evaluations on three networks: Polbooks, Polblogs, and Enron. For each network, we vary K from 2

Table 3: Identity protection K vs. Feature Changes between *Rand Add/Del* scheme (denoted as *Rand*) and *K-degree Generalization* scheme (denoted as *K-deg*); Lines with $K = 1$ show the feature values of the original networks

	Rand	K-deg	Rand	K-deg	Rand	K-deg	Rand	K-deg	Rand	K-deg	Rand	K-deg
Polbooks	λ_1	λ_1	μ_2	μ_2	h	h	Q	Q	C	C	$SC(\times 10^3)$	$SC(\times 10^3)$
$K = 1$	11.93	11.93	0.32	0.32	2.45	2.45	0.40	0.40	0.34	0.34	2.52	2.52
2	11.64	12.00	0.61	0.43	2.31	2.35	0.37	0.39	0.30	0.33	1.68	2.48
3	11.51	12.05	0.79	0.45	2.28	2.32	0.36	0.39	0.29	0.33	1.43	2.56
4	11.04	12.11	1.16	0.60	2.20	2.28	0.31	0.38	0.22	0.32	0.75	2.53
5	10.50	12.22	1.43	0.60	2.16	2.28	0.26	0.38	0.17	0.33	0.40	2.76
6	10.33	12.30	1.16	0.79	2.16	2.23	0.24	0.36	0.15	0.30	0.34	2.44
7	10.15	12.31	1.41	0.63	2.14	2.27	0.21	0.37	0.13	0.31	0.27	2.68
8	9.83	12.64	1.53	0.65	2.13	2.26	0.15	0.37	0.10	0.32	0.19	3.60
9	9.72	12.72	1.43	0.97	2.13	2.20	0.14	0.34	0.10	0.29	0.17	3.58
10	9.75	12.85	1.61	0.88	2.13	2.19	0.14	0.35	0.1	0.30	0.18	4.12
Polblogs	λ_1	λ_1	μ_2	μ_2	h	h	Q	Q	C	C	$SC(\times 10^9)$	$SC(\times 10^{29})$
$K = 1$	74.08	74.08	0.168	0.168	2.506	2.506	0.405	0.405	0.226	0.226	$1.21(\times 10^{29})$	1.21
2	30.19	74.89	9.30	0.168	2.35	2.500	0.067	0.402	0.027	0.225	10.6	2.73
3	28.55	74.50	10.58	0.168	2.35	2.484	0.024	0.401	0.022	0.223	2.07	1.87
4	28.50	75.16	10.72	0.168	2.35	2.494	0.020	0.401	0.022	0.224	1.96	3.61
5	28.49	75.10	11.11	0.168	2.35	2.475	0.018	0.396	0.022	0.221	1.94	3.40
6	28.47	76.32	10.86	0.168	2.35	2.469	0.019	0.394	0.022	0.222	1.90	1.45
7	28.46	75.82	11.09	0.168	2.35	2.461	0.018	0.395	0.022	0.22	1.88	6.94
8	28.46	76.67	11.14	0.168	2.35	2.462	0.016	0.389	0.022	0.219	1.89	6.25
9	28.46	77.42	10.68	0.168	2.35	2.486	0.019	0.387	0.022	0.221	1.88	4.46
10	28.46	78.42	10.72	0.168	2.35	2.458	0.015	0.385	0.022	0.221	1.88	4.04
Enron	λ_1	λ_1	μ_2	μ_2	h	h	Q	Q	C	C	$SC(\times 10^4)$	$SC(\times 10^6)$
$K = 1$	17.83	17.83	0.80	0.80	2.278	2.278	0.0074	0.0074	0.344	0.344	37.02	0.37
2	13.90	18.16	1.60	0.84	2.096	2.25	0.0046	0.0072	0.127	0.33	0.757	0.51
3	12.69	18.29	3.20	0.86	2.079	2.24	0.0037	0.0072	0.081	0.33	0.220	0.58
4	12.65	18.45	2.99	1.00	2.079	2.17	0.0037	0.0069	0.078	0.31	0.211	0.68
5	12.66	19.31	3.04	0.85	2.078	2.17	0.0037	0.0066	0.080	0.31	0.215	1.62
6	12.63	19.41	2.89	0.84	2.078	2.19	0.0037	0.0065	0.078	0.31	0.206	1.78
7	12.60	20.04	3.04	0.82	2.078	2.19	0.0037	0.0069	0.078	0.31	0.201	3.37
8	12.60	19.92	3.11	0.82	2.079	2.12	0.0037	0.0063	0.079	0.29	0.202	2.98
9	12.61	20.42	2.84	1.45	2.079	2.13	0.0037	0.0061	0.079	0.30	0.204	4.93
10	12.62	21.39	2.96	0.98	2.077	2.05	0.0037	0.0058	0.078	0.29	0.206	12.9

to 10 and apply both *Rand Add/Del* and *K-degree Generalization* schemes. For *Rand Add/Del*, we use the absolute identity protection measure, $\tau_a(\alpha|\mathbf{d}) \geq 1 - 1/K$, to determine the perturbation magnitude k and then generate a randomized network using k .

We can observe that both *Rand Add/Del* and *K-degree* schemes generally decrease structural properties. For example, both Q (indicating the goodness of the community structure) and μ_2 (showing how good the communities separate, with smaller values corresponding to better community structures) increase along K , which indicates the goodness of the community structure is affected due to edge modification. We can also observe from Table 3 that *K-degree* scheme generally better preserves structural features than *Rand Add/Del*. This is because that *K-degree* scheme examines the degree sequence of nodes and chooses a subset of nodes (that violates the *K-degree* anonymity property) for edge modification while *Rand Add/Del* scheme treats all nodes (edges) equally during randomization. We expect that reconstruction methods can be designed for the purely randomized graph so features derived from the reconstructed graph (rather than directly from the released randomized graph) can be more accurate. It is our belief that it is very hard, if not impossible, to figure out reconstruction methods on the released data randomized using *K-degree* scheme. We will investigate reconstruction methods in our future work.

3.3 Further Improvement

Since *Rand Add/Del* randomly adds and deletes edges, a large number of perturbations are applied to those nodes in low risks. As a result, we sacrifice graph utility without further improving identity protection. One natural idea is that we can divide the graph into several blocks according to the degree sequence and apply *Rand Add/Del* separately to each block using different randomization parameters k .

In many real-world networks, we have fewer nodes with high degrees while more nodes with low degrees. By simply partitioning the graph into blocks according to the degree sequence, we expect to introduce fewer perturbations (with better utility preservation) to achieve the same privacy protection. For each block b , we say an existing (or non-existing) link (i, j) is in block b if node i or j is in the block. Let n_b be the number of nodes and m_b be the number of links in block b . We randomly add and delete k_b links, then each existing link remains in the randomized graph with probability $p_{11}^{(b)} = 1 - \frac{k_b}{m_b}$, and each non-existing link is added with probability $p_{10}^{(b)} = \frac{k_b}{N_b - m_b}$ where $N_b = \binom{n_b}{2} - n_b(n - n_b)$. We can use the same methodologies in calculating the identity/link risks except for replacing the overall p_{11} and p_{10} with $p_{11}^{(b)}$ and $p_{10}^{(b)}$. We call this method blockwise random add/delete strategy, or simply *Rand Add/Del-B* for short.

Figure 7 shows preliminary results of *Rand Add/Del-B* on Enron network. In this experiment, we simply divide the graph into two blocks: nodes with degree greater than 30 are in the first block while the rest nodes with high degree frequency values are in the second block. We can observe from Figure 7 that this simple strategy can better preserve graph features than *Rand Add/Del*. We expect to achieve even better utility preservation when we have better block partitions (e.g., using histogram partition algorithms). As we discussed previously, we will also investigate reconstruction methods on the released data using *Rand Add/Del-B* scheme.

4. RELATED WORK

4.1 Advanced Attacks on Randomized Networks

When it comes to an anonymized or randomized graph, the adversary may exploit various a-priori knowledge of the graph such as some topological features or a subgraph. In this section, we briefly discuss whether randomization strategies are resilient to those complex background knowledge based attacks.

4.1.1 Subgraph Attacks

In [2], the authors described a family of subgraph attacks such that an adversary can learn whether edges exist or not between specific targeted pairs of nodes from node-anonymized social networks. The adversary can construct a highly distinguishable subgraph with edges to a set of targeted nodes, and then to re-identify the subgraph and consequently the targets in the released anonymized network. Similarly in [12], Hay and et al. further observed that the structure of the graph itself (e.g., the degree of the nodes or the degree of the node's neighbors) determines the extent to which an individual in the network can be distinguished. Their empirical evaluations showed that edge based randomization (the same as *Rand Add/Del*) can well protect the identification of the vertices since the adversary cannot simply exclude from the candidate set nodes that do not match the structural properties of the target. While it is hard to conduct formal disclosure analysis for *Rand Add/Del* under known subgraph based attacks, we present our following informal discussions. For *Rand Add/Del* strategy, since each link is re-allocated independently, knowing the subgraph cannot enhance the adversary's confidence about the link outside the subgraph. Herein we assume that at least a medium perturbation is applied to the graph, i.e., k is not too small, otherwise the randomized perturbation is not much different from the original one.

4.1.2 Link Prediction based Attacks

In this paper, the effect on privacy due to randomization was quantified by considering only the magnitude information of randomization. It has been well known that graph topological features have close relations with the existence of links and various proximity measures have been exploited to predict the existence of a future link [18] in the classic link prediction task. Although the classic link prediction focuses on network evolution models and the change due to randomization is different with that due to network evolutions, nevertheless, various graph proximity measures used in the classic link prediction could be exploited by adversaries.

The problem of how adversaries may exploit the topological features of the released graph to breach link privacy was recently studied in [27, 28]. The attacking model [27] was based on the distribution of the probability of existence of a link across all possible graphs in the graph space. In [28], the attacking model was to exploit the relationship between the probability of existence of a link and the similarity measure values of node pairs in the released randomized graph. Specifically, the authors investigate how adversaries may exploit proximity measure values (such as common neighbors, Katz measure, Adamic/Adar measure, and Commute time, derived from the released randomized graph after applying *Rand Add/Del*) to breach link privacy. They quantify how much the posterior belief on the existence of a link can be enhanced by exploiting those similarity measures from the *Add/Del* randomized graph.

4.2 Advanced Randomization Strategies

Edge randomization may significantly affect the utility of the released randomized graph. To preserve utility, we expect certain aggregate characteristics (a.k.a., feature) of the original graph should remain basically unchanged or at least some properties can be reconstructed from the randomized graph. However, as shown in Ta-

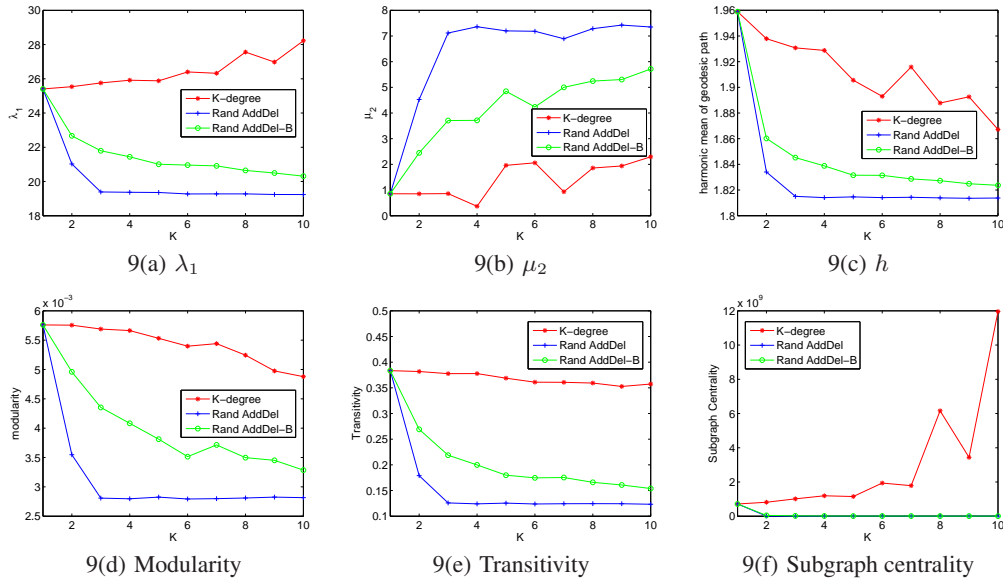


Figure 7: Identity Protection K vs. Feature Change on Enron data

ble 3, many topological features are lost due to randomization. In this section, we present advanced randomization strategies that can preserve structural properties. We would emphasize that it is very challenging to quantify disclosures since the process of feature preserving strategies or generalization strategies are more complicated than that of random strategies.

4.2.1 Random Switch

For *Rand Switch* strategy, we randomly switch a pair of existing edges (t, w) and (u, v) (satisfying that edge (t, v) and (u, w) does not exist in G) to (t, v) and (u, w) and repeat it for k times. This strategy preserves the degree of each vertex, i.e., $\tilde{d}_i = d_i$. To the adversary, every node with the degree d_α in the randomized graph may correspond to the target individual α with the equal probability $\frac{1}{n_{d_\alpha}}$, where n_{d_α} denotes the number on nodes whose degrees equal d_α . Hence, the prior and posterior risk measures are given by

$$r(\alpha|d_\alpha) = r(\alpha|d_\alpha, \tilde{\mathbf{d}}) = \frac{1}{n_{d_\alpha}}. \quad (19)$$

The risk measures are independent with perturbation parameter k , and therefore more switches can not improve the protection on node identities. Since $r(\alpha|d_\alpha) = r(\alpha|d_\alpha, \tilde{\mathbf{d}})$, we naturally have $\tau_a(\alpha|\tilde{\mathbf{d}}) = 1 - \frac{1}{n_{d_\alpha}}$, and $\tau_r(\alpha|\tilde{\mathbf{d}}) \equiv 1$.

4.2.2 Spectrum Preserving Randomization

In [26], Ying and Wu presented a randomization strategy that can preserve the spectral properties of the graph. They focused on the spectrum of networks since the spectrum has close relations with the many graph characteristics and can provide global measures for some network properties. The spectrum of a network is connected to important topological properties such as diameter, presence of cohesive clusters, long paths and bottlenecks, and randomness of the graph. They presented two spectrum preserving randomization methods, *Sptr Add/Del* and *Sptr Switch*, which keep graph spectral characteristics (i.e., the largest eigenvalue of the adjacency matrix and the second smallest eigenvalue of the Laplacian matrix) not much changed during randomization by examining eigenvector val-

ues of nodes to choose where edges are added/deleted or switched. Although they empirically show that the spectrum preserving approach can achieve similar privacy protection as the random perturbation approach, however, they did not derive the formula of the protection measure for either *Sptr Add/Del* or *Sptr Switch* since the number of false edges in the randomization cannot be explicitly expressed.

4.2.3 Markov Chain based Feature Preserving Randomization

In [10, 27], the authors studied the problem of how to generate a synthetic graph matching various properties of a real social network in addition to a given degree sequence. They investigated a switching based algorithm for generating synthetic graphs whose feature values are within a precise range of those of the original graph. In [27], the authors also studied how adversaries exploit the released graph as well as feature constraints to breach link privacy. The adversary can calculate the posterior probability of existence of a link by exploiting the ensemble of graphs with the given degree sequence and the prescribed feature constraints.

4.2.4 Generalization via Edge Modification or Clustering

In addition to the K -degree anonymous graph introduced by Liu and Terzi [19], in [30], Zhou and Pei anonymized the graph by generalizing node labels and inserting edges until each neighborhood is indistinguishable to at least $k - 1$ others. In [6, 11, 29], authors applied a structural anonymization approach called *edge generalization* that consists of collapsing clusters together with their component nodes' structure, rather than add or delete edges from the social network dataset. Although the above proposed approaches would preserve privacy, however, it is not clear how useful the anonymized graph is since many topological features may be lost due to graph reduction (i.e., reducing a subset of vertices and edges into a super-vertex). Although the above proposed approaches would preserve privacy, however, it is not clear how well the anonymized or generalized graph can preserve utility since many topological features may be lost.

5. CONCLUSION AND FUTURE WORK

Edge randomization has been shown a necessity in addition to node anonymization to preserve privacy in the released graph. We have investigated how well the edge randomization scheme *Rand Add/Del* can protect privacy of node identities and sensitive links. We have conducted theoretical analysis and empirical comparisons with the *K-degree Generalization* scheme.

There are some other aspects of this work that merit further research. Among them, we will study how well randomization strategies protect identity and link privacy when adversaries exploit various complex background knowledge in their attacks. Since how to preserve utility (in terms of various structural features) and privacy in the released graph is an important issue in privacy preserving social network analysis, we will continue the study of the trade-off between privacy and utility for various complex randomization strategies. We will also study the scalability issue and conduct empirical evaluations on large social networks. Finally, we will investigate reconstruction methods on the released graph perturbed by various edge modification schemes.

Acknowledgments

This work was supported in part by U.S. National Science Foundation IIS-0546027 and CNS-0831204.

6. REFERENCES

- [1] L. Adamic and N. Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem*, 2005.
- [2] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 181–190, New York, NY, USA, 2007. ACM Press.
- [3] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54, New York, NY, USA, 2006. ACM.
- [4] J. Baumes, M. K. Goldberg, M. Magdon-Ismael, and W. A. Wallace. Discovering hidden groups in communication networks. In *ISI*, pages 378–389, 2004.
- [5] T. Y. Berger-Wolf and J. Saia. A framework for analysis of dynamic social networks. In *KDD*, pages 523–528, 2006.
- [6] A. Campan and T. M. Truta. A clustering approach for data and structural anonymity in social networks. In *PinKDD*, 2008.
- [7] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas. Characterization of complex networks: A survey of measurements. *Advances In Physics*, 56:167, 2007.
- [8] A. Fast, D. Jensen, and B. N. Levine. Creating social networks to improve peer-to-peer networking. In *KDD*, pages 568–573, 2005.
- [9] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99(12):7821–7826, June 2002.
- [10] S. Hanhijarvi, G. C. Garriga, and K. Puolamaki. Randomization techniques for graphs. In *Proc. of the 9th SIAM Conference on Data Mining*, 2009.
- [11] M. Hay, G. Miklau, D. Jensen, D. Towsely, and P. Weis. Resisting structural re-identification in anonymized social networks. In *VLDB*, 2008.
- [12] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava. Anonymizing social networks. *University of Massachusetts Technical Report*, 07-19, 2007.
- [13] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.
- [14] J. M. Kleinberg. Challenges in mining social network data: processes, privacy, and paradoxes. In *KDD*, pages 4–5, 2007.
- [15] Y. Koren, S. C. North, and C. Volinsky. Measuring and extracting proximity in networks. In *KDD*, pages 245–255, 2006.
- [16] V. Krebs. <http://www.orgnet.com/>. 2006.
- [17] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *KDD*, pages 611–617, 2006.
- [18] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559, New York, NY, USA, 2003. ACM.
- [19] K. Liu and E. Terzi. Towards identity anonymization on graphs. In *Proceedings of the ACM SIGMOD Conference*, Vancouver, Canada, 2008. ACM Press.
- [20] A. Seary and W. Richards. Spectral methods for analyzing and visualizing networks: an introduction. *National Research Council, Dynamic Social Network Modelling and Analysis: Workshop Summary and Papers*, pages 209–228, 2003.
- [21] J. Shetty and J. Adibi. The Enron email dataset database schema and brief statistical report. *Information Sciences Institute Technical Report*, University of Southern California, 2004.
- [22] M. Shiga, I. Takigawa, and H. Mamitsuka. A spectral clustering approach to optimally combining numerical vectors with a modular network. In *KDD*, pages 647–656, 2007.
- [23] E. Spertus, M. Sahami, and O. Buyukkocuten. Evaluating similarity measures: a large-scale study in the orkut social network. In *KDD*, pages 678–684, 2005.
- [24] C. Tantipathananandh, T. Y. Berger-Wolf, and D. Kempe. A framework for community identification in dynamic social networks. In *KDD*, pages 717–726, 2007.
- [25] S. White and P. Smyth. Algorithms for estimating relative importance in networks. In *KDD*, pages 266–275, 2003.
- [26] X. Ying and X. Wu. Randomizing social networks: a spectrum preserving approach. In *Proc. of the 8th SIAM Conference on Data Mining*, April 2008.
- [27] X. Ying and X. Wu. Graph generation with prescribed feature constraints. In *Proc. of the 9th SIAM Conference on Data Mining*, 2009.
- [28] X. Ying and X. Wu. On link privacy in randomizing social networks. In *PAKDD*, 2009.
- [29] E. Zheleva and L. Getoor. Preserving the privacy of sensitive relationships in graph data. In *PinKDD*, pages 153–171, 2007.
- [30] B. Zhou and J. Pei. Preserving Privacy in Social Networks Against Neighborhood Attacks. *IEEE 24th International Conference on Data Engineering*, pages 506–515, 2008.
- [31] B. Zhou, J. Pei, and W.-S. Luk. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *SIGKDD Explorations*, 10(2), 2009.