Randomizing Social Networks: a Spectrum Preserving Approach *

Xiaowei Ying, Xintao Wu
Department of Software and Information Systems
Univ. of North Carolina at Charlotte
{xying,xwu}@uncc.edu

Abstract

Understanding the general properties of real social networks has gained much attention due to the proliferation of networked data. The nodes in the network are the individuals and the links among them denote their relationships. Many applications of networks such as anonymous Web browsing require relationship anonymity due to the sensitive, stigmatizing, or confidential nature of the relationship. One general approach for this problem is to randomize the edges in true networks, and only disclose the randomized networks. In this paper, we investigate how various properties of networks may be affected due to randomization. Specifically, we focus on the spectrum since the eigenvalues of a network are intimately connected to many important topological features. We also conduct theoretical analysis on the extent to which edge anonymity can be achieved. A spectrum preserving graph randomization method, which can better preserve network properties while protecting edge anonymity, is then presented and empirically evaluated.

1 Introduction

Many natural and social systems develop complex networks, e.g., the Internet, the World-Wide Web, networks of collaborating movie actors and those of collaborating authors, etc. The nodes in the social network are the individuals and the links among them denote their relationships. Many applications of networks such as anonymous Web browsing require relationship anonymity due to the sensitive, stigmatizing, or confidential nature of relationship. For example, most people prefer to conceal the truth regarding their illegal or unethical behaviors which are customarily disapproved of by society.

Naturally, graph randomization techniques can be applied here. For example, we can remove some true edges and/or add some false edges. Two natural edge-based graph perturbation strategies are shown below.

- *Rand Add/Del*, we randomly add one edge followed by deleting another edge and repeat this process for k times. This strategy preserves the total number of edges in the original graph.
- *Rand Switch*, 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.

*This work was supported in part by the U.S. National Science Foundation NSF IIS-0546027.

After the randomization, the randomized graph is expected to be different from the original one. As a result, the true sensitive or confidential relationship will not be disclosed. We need to know how well the randomization can protect those sensitive links.

On the other hand, the released randomized graph should also keep some properties not much changed or at least some properties can be reconstructed from the randomized graph. Understanding the general properties of real networks has gained much attention due to the proliferation of networked data. Most analysis [11] has been confined to real-space characteristics, e.g., degree sequences, shortest connecting paths, and clustering coefficients.

Since there are numerous characteristics related to networks, it is tedious to evaluate how those characteristics are affected by the randomization process. In this paper, we investigate this problem by focusing on the spectrum of networks since it has been shown the spectrum has close relation with the many graph characteristics and can provide global measures for some network properties. The spectrum of a graph is usually defined as the set of eigenvalues of the graph's adjacency matrix or other derived matrices. The eigenvalues of a network are connected to important topological properties such as diameter, presence of cohesive clusters, long paths and bottlenecks, and randomness of the graph. The associated eigenvectors can also guide to discover clusters. In Section 2, we summarize the properties of the spectrum and associated eigenvectors of graph matrices and their relation to structures of network.

1.1 Contribution

Our contributions are as follows.

- We show theoretically and empirically how the real characteristics of graphes are related with spectral characteristics and how the two edge based pure randomization strategies affect both real and spectral characteristics.
- We develop spectrum preserving randomization methods, *Spctr Add/Del* and *Spctr Switch*, which can better preserve graph characteristics without sacrificing much privacy protection during randomization.

- Our proposed spectrum preserving randomization methods consider the change of both the λ_1 and μ_2 due to randomization. In graph perturbation, researchers have only investigated the problem of comparing the largest eigenvalue of the original and perturbed eigenvalues.

- We conduct privacy analysis for edge based pure randomization strategies and show formally how the randomized graph may be exploited by attackers to improve their a-priori belief on sensitive links.

1.2 Organization The remaining of paper is outlined as follows. In Section 2, we revisit the relationship between the real characteristics and the spectral characteristics (e.g., λ_1 and μ_2). We theoretically show how randomization affects those spectral characteristics and especially we give the bounds of those changes due to randomization in Section 3. In Section 4, we present our spectrum preserving edge randomization approach. We focus on *Spctr Switch* and give detailed theoretical analysis and empirical evaluation. In Section 5, we first show why the edge randomization is resilient to subgraph based attacks and then theoretically show how the randomized graph may be exploited by attackers to improve their a-priori belief on sensitive links. We conclude and discuss our future work in Section 6.

2 Graph Characteristics

2.1 Notation A network or graph $G(V, E)$ is a set of n nodes V connected by a set of m links E. The network considered here is binary, symmetric, connected, and without self-loops. It can be represented as the symmetric adjacency matrix $A_{n\times n}$ with $a_{ij} = 1$ if node i is connected to node j 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. Recall that the degree of a vertex in a network is the number of edges connected to that vertex.

Let λ_i be the eigenvalues of A and $\mathbf{e}_i$ the corresponding eigenvectors, and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. The spectral decomposition of A is $A = \sum_i \lambda_i \mathbf{e}_i \mathbf{e}_i^T$. Since A is irreducible, from the theory of non-negative matrices, the largest eigenvalue λ_1 of A is simple and there exits a unique positive unit eigenvector $\mathbf{e}_1$ such that $A\mathbf{e}_1 = \lambda\mathbf{e}_1$. We call λ_1 the index of G, and call $\mathbf{e}_1 = (x_1, \cdots, x_n)^T$ the principal eigenvector of the graph G where x_i is the ith component of the principal eigenvector.

Another matrix related to A is the Laplacian matrix defined as $L = D - A$. Similarly, let μ_i be the eigenvalues of A and $\mathbf{u}_i$ the corresponding eigenvectors. We have $0 = \mu_1 \leq \mu_2 \leq \cdots \leq \mu_m \leq m$. μ_2 is an important eigenvalue of the Laplacian matrix and can be used to show how good the communities separate, with smaller values corresponding to better community structures. Let $\mathbf{u}_2 = (y_1, \cdots, y_n)^T$ where y_i is the ith component of the eigenvector $\mathbf{u}_2$.

2.2 Spectral vs. Real Characteristics 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 [5]. In this paper, we focus on four real space characteristics of a graph. The first one is the harmonic mean of the shortest distance, h, which is defined in [10] as:

$$h = \{\frac{1}{n(n-1)} \sum_{i\neq j} \frac{1}{d_{ij}}\}^{-1} \tag{2.1}$$

The inverse of the harmonic mean of the shortest distance, also known as the global efficiency, varies between 0 and 1, with $h^{-1} = 0$ when all vertices are isolated and $h^{-1} = 1$ when the graph is complete.

The second one is the modularity measure, Q, which indicates the goodness of the community structure [5]. 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. A value $Q = 0$ indicates that the community structure is no stronger than would be expected by random chance and values other than zero represent deviations from randomness.

The third one is the transitivity measure, C, which is one type of clustering coefficient measure and characterizes the presence of local loops near a vertex. It is formally defined as

$$C = \frac{3N_\Delta}{N_3} \tag{2.2}$$

where N_Δ is the number of triangles and N_3 is the number of connected triples.

The fourth one is the subgraph centrality, SC, which is used to quantify the centrality of vertex i based the subgraphs [6].

$$SC = \frac{1}{n}\sum_{i=1}^{n} SC_i = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{\infty} \frac{P_i^k}{k!} \tag{2.3}$$

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

Throughout this paper, we also focus on two important eigenvalues of the graph spectrum. The first one is the largest eigenvalue (λ_1) of the adjacency matrix A. The eigenvalues of A encode information about the cycles of a network as well as its diameter. Since A contains no self-loops, the sum over all eigenvalues ($\sum_{i=1}^{n} \lambda_i$) is zero. The sum of product pairs ($\sum_{i\neq j} \lambda_i\lambda_j$) is equal to minus the number of edges. And $\sum_{i\neq j\neq k} \lambda_i\lambda_j\lambda_k$ is twice the number of triangles in G. The maximum degree, chromatic number, clique number,

and extend of branching in a connected graph are all related to λ_1. In [12], the authors studied how a virus propagates in a real work and proved that the epidemic threshold for a network is closely related to λ_1. The global subgraph centrality measure can be calculated through eigenvalues of the graph:

$$SC = \frac{1}{n}\sum_{i=1}^{n} e^{\lambda_i}. \tag{2.4}$$

The second one is the second eigenvalue (μ_2) of the Laplacian matrix L, which is also called the algebraic connectivity of the graph. The eigenvalues of L encode information about the tree-structure of G. The spectrum of L contains a 0 for every connected component. The multiplicity of 0 as an eigenvalue is equal to the number of components in G. $\frac{1}{m}\prod_{i=2}^{n}\mu_i$ equals the number of spanning trees of G. The diameter of a general graph is related to μ_m and μ_2 and bounded by

$$Diam(G) \leq \left\lceil \frac{cosh^{-1}(m-1)}{cosh^{-1}\left(\frac{\mu_m+\mu_2}{\mu_m-\mu_2}\right)} \right\rceil$$

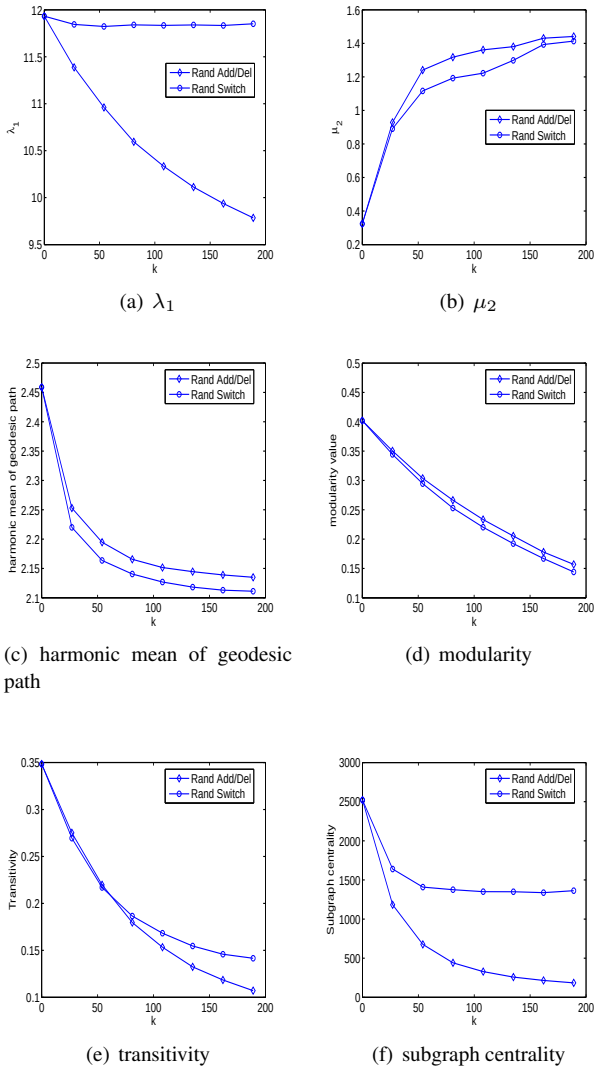
Note that if μ_2 is close to zero, the graph is almost disconnected. Its diameter is small if the eigenvalue gap is large (i.e., $\mu_2 >> \mu_1$). Refer to [11] for more relationships between the spectral and real characteristics of graphs.

3 Spectral Analysis of Graph Perturbation

We are concerned with the connection between the structure of a graph G and the spectrum of a 0-1 adjacency matrix A and Laplacian matrix L of graph G. Intuitively, a local modification of G, such as the addition of an edge between non-adjacent vertices, can be regarded as a perturbation of G.

In Section 3.1, we first empirically show how the spectrum of a graph and some real space characteristics are affected by the random perturbation strategies. In Section 3.2, we conduct the theoretical analysis on how randomization affects the spectrum of a graph and give bounds of the spectrum change.

3.1 Graph Characteristics vs. Perturbation: An Illustrating Example In this section, we empirically show how the graph characteristics (including two spectral, λ_1, μ_2 and four real, harmonic mean of geodesic path, modularity, transitivity, and subgraph centrality) vary when *Rand Add/Del* and *Rand Switch* perturbation strategies are applied. This experiment was conducted on the US politics book data [9], which contains 105 vertices and 441 edges. In this graph, 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



(a) λ_1 (b) μ_2

(c) harmonic mean of geodesic path (d) modularity

(e) transitivity (f) subgraph centrality

Figure 1: Graph characteristic vs. perturbation with varying k for *Rand Add/Del* and *Rand Switch*

are separated into groups according to their political views: "liberal", "neutral", or "conservative".

We can observe from Figure 1 that the changes of spectral measures display similar trends as those of real graph characteristics while applying the two perturbation strategies. Especially, as shown in Figures 1(b), 1(c), 1(d), and 1(e), the μ_2 of the Laplacian matrix displays the very similar pattern as the harmonic mean of geodesic path, modularity, and transitivity. Similarly, as shown in Figures 1(a) and 1(f), the λ_1 of the adjacency matrix displays the similar pattern as the subgraph centrality measure for both *Rand Add/Del* and *Rand Switch* strategies. Networks with community structures is not resilient to random perturbation strategy. This is intuitively reasonable as shown in Figure

1(d). Average vertex-vertex distance may change sharply when edges across communities are switched with edges within communities.

We can also observe neither *Rand Add/Del* nor *Rand Switch* can well preserve the graph characteristics when we increase k to more than 100. Since we have 441 edges in this graph, even the medium randomization ($k = 100$) significantly decreases the utility of the released graph. Generally more perturbation can lead to stronger privacy protection, but it also greatly changes many features of the network, decreasing the information utility. For example, network resilience and community structure are of particular importance in epidemiology where removal of vertices or edges in a contact network may correspond to vaccination of individuals against a disease. Then the epidemiological solution developed from the randomly perturbed graph may not be applicable to the real graph. In Section 4 we shall investigate how to perturb graphs without changing much network structural features, such as resilience and community structure.

3.2 Theoretical Analysis on Spectral Perturbation The theory of graph perturbations is concerned primarily with changes in eigenvalues which result from local modifications of a graph such as adding or deleting an edge. In the following, we let A and $\widetilde{A}$ be the adjacency matrices of the original graph G and the perturbed graph G' with spectra $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and $\widetilde{\lambda}_1 \geq \widetilde{\lambda}_2 \geq \cdots \geq \widetilde{\lambda}_n$ respectively.

LEMMA 3.1. *[4] $\widetilde{\lambda}_1 < \lambda_1$ whenever G' is obtained from G by deleting an edge or vertex. Similarly, $\widetilde{\lambda}_1 > \lambda_1$ whenever G' is obtained from G by adding an edge or a non-isolated vertex.*

Lemma 3.1 shows any proper subgraph of G has smaller index value λ_1 and any supgraph of G has larger index value λ_1. This is also one reason why we only focus on the perturbation strategies which keep the number of edges unchanged. Otherwise, the index of the graph λ_1 may be significantly changed, which will affect many real space graph characteristics.

THEOREM 1. *Weyl's Theorem [8]. Given two $n \times n$ symmetric matrices A and E, assume $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and $\varepsilon_1 \geq \varepsilon_2 \geq \cdots \geq \varepsilon_n$ are their eigenvalues respectively. Let $\widetilde{A} = A + E$, and $\widetilde{\lambda}_1 \geq \widetilde{\lambda}_2 \geq \cdots \geq \widetilde{\lambda}_n$ are its eigenvalues. Then the Weyl's inequalities are*

$$\widetilde{\lambda}_{i+j-1} \leq \lambda_i + \varepsilon_j \leq \widetilde{\lambda}_{i+j-n} \tag{3.5}$$

for $1 \leq i, j, i+j-1, i+j-n \leq n$.

Weyl's theorem states that the eigenvalues of a matrix are perfectly conditioned, i.e., no eigenvalue can move more than the range specified by Equation 3.5.

Some graph features (e.g., the number of vertices n, the number of edges m) remain unchanged after randomization and are assumed to be available to attackers. We also assume the number of perturbations k is available to both data miners and attackers. The reason is that k denotes the magnitude of perturbation which may be needed to analyze the perturbed graph by data miners. In this section, we present to what extent the graph spectrum may change with respect to those graph invariants, specifically, k and n for *Rand Add/Del* and k, n and d_i for *Rand Switch* where d_i is the degree of vertex i.

When $k = 1$, we call the perturbation matrix as the elementary perturbation matrix (EPM). Obviously, the perturbation matrix E when $k > 1$ is the sum of EPMs along the perturbation.

For *Rand Add/Del*, we have two different cases. One is that we add the edge (i, p) and delete an existing edge (i, q). In this case, the EPM has the form as below:

$$E_{(i,p,q)} = \widetilde{A} - A = \begin{pmatrix} 0 & 1 & -1 \\ 1 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} \oplus 0_{n-3}. \tag{3.6}$$

Specifically, $e_{ip} = e_{pi} = 1$, and $e_{iq} = e_{qi} = -1$, where e_{ij} denotes the component of E. The other case is that we add the edge (i, j) and then remove one existing edge (p, q) where i, j, p, q are distinct. Then,

$$E_{(i,j,p,q)} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{pmatrix} \oplus 0_{n-4}. \tag{3.7}$$

Specifically, $e_{ij} = e_{ji} = 1$, and $e_{pq} = e_{qp} = -1$.

For *Rand Switch*, when we switch one pair of edges, (t, w), (u, v) to (t, v) and (u, w), the EPM is:

$$E_{(t,w,u,v)} = \begin{pmatrix} 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \end{pmatrix} \oplus 0_{n-4} \tag{3.8}$$

Specifically, $e_{tw} = e_{wt} = e_{uv} = e_{vu} = -1$, and $e_{tv} = e_{vt} = e_{uw} = e_{wu} = 1$. We can easily derive $\varepsilon_1 = 2$,$\varepsilon_n = -2$, and $\varepsilon_i = 0$ $(2 \leq i \leq n-1)$.

However, when $k > 1$, it is hard to derive directly the eigenvalues of E based on the released k. In the following, we show our result based on the Gershgorin Circle Theorem [8].

THEOREM 2. *Gershgorin Circle Theorem. For an $n \times n$ matrix A, define*

$$R_i = \sum_{j=1, j \neq i}^{n} |a_{ij}| .$$

Then each eigenvalue of A must be in at least one of the disks in the complex plane:

$$C_i(A) = \{z : |z - a_{ii}| \leq R_i\}.$$

RESULT 1. *Let $\varepsilon_1 \geq \varepsilon_2 \geq \cdots \geq \varepsilon_n$ be the eigenvalues of E. For all $i(1 \leq i \leq n)$, we have*

$$\varepsilon_n \leq \left|\lambda_i - \widetilde{\lambda}_i\right| \leq \varepsilon_1 \tag{3.9}$$

or more loosely

$$\left|\lambda_i - \widetilde{\lambda}_i\right| \leq \|E\|_2, \tag{3.10}$$

where for Rand Add/Del*,*

$$\|E\|_2 \leq \min\{2k, n-1\}, \tag{3.11}$$

and for Rand Switch*,*

$$\|E\|_2 \leq 2\min\left\{k, \max_i(\min\{d_i, n-1-d_i\})\right\} \tag{3.12}$$

PROOF. Equation 3.9 and Equation 3.10 can be easily derived from the Weyl's theorem.

Notice that the diagonal elements of E are always 0. Hence,

$$C_i(E) = \{z : |z - e_{ii}| \leq R_i\} = \{z : |z| \leq R_i\}.$$

All these circles are concentric, and all the eigenvalues of A are thus in the circle of the largest radius:

$$\|E\|_2 \leq \max_i\{R_i\}.$$

and $R_i = \sum_{j \neq i} |e_{ij}|$ is actually the totaly number of added and deleted edges of vertex i.

Hence, for *Rand Add/Del*, when $k < n/2$, the worst case is that all the perturbations involve the same vertex; when $k \geq n/2$, the worst case happens when a certain vertex is removed all original edges to its neighbors and adds new edges to all the rest vertices. In this case,

$$\max_i\{R_i\} \leq \min\{2k, n-1\}.$$

and Equation 3.11 follows.

For *Rand Switch*, if one edge is deleted, there must be an edge added to the same vertex. Therefore

$$\frac{1}{2}R_i \leq \min\{d_i, n-1-d_i\},$$

through which we immediately get

$$\max_i R_i \leq 2\min\left\{k, \max_i(\min\{d_i, n-1-d_i\})\right\},$$

and Equation 3.12 follows.

Actually, the bound given in Equation 3.12 is the loose bound in the worst case. It may not accurately reflect the magnitude of spectrum change. In Section 4, we develop our spectrum preserving randomization approach which can control the change of spectrum during the randomization process. Note that all the above results can be easily extended to the Laplacian matrix with some simple adjustment since $\widetilde{L} - L = A - \widetilde{A} = -E$.

4 Spectrum Preserving Randomization

Since many graph structures are shown to have strong association with the spectrum, a very nature idea is whether we can figure out a perturbation strategy such that one or some particular eigenvalues will not significantly change. Hence the new strategy is more probable to better preserve structural characteristics without much scarifying the privacy protection.

Table 1: Conditions on adjusting λ_1 and μ_2 for *Spctr Add/Del*

Condition	Action
$x_ix_j - x_px_q > 0$	$\widetilde{\lambda}_1 > \lambda_1$
$x_ix_j - x_px_q < 0$, and $\lambda_1 - \lambda_2 > \frac{x_i^2+x_j^2+x_p^2+x_q^2}{2(x_px_q - x_ix_j)}$	$\widetilde{\lambda}_1 < \lambda_1$
$y_iy_j - y_py_q > 0$	$\widetilde{\mu}_2 < \mu_2$
$y_iy_j - y_py_q < 0$, and $\mu_3 - \mu_2 > \frac{y_i^2+y_j^2+y_p^2+y_q^2}{2(y_py_q - y_iy_j)}$	$\widetilde{\mu}_2 > \mu_2$

Table 2: Conditions on adjusting λ_1 and μ_2 for *Spctr Switch*

Condition	Action
$(x_t - x_u)(x_v - x_w) > 0$	$\widetilde{\lambda}_1 > \lambda_1$
$(x_t - x_u)(x_v - x_w) < 0$, and $\lambda_1 - \lambda_2 > \frac{x_t - x_u}{x_w - x_v} + \frac{x_w - x_v}{x_t - x_u}$	$\widetilde{\lambda}_1 < \lambda_1$
$(y_t - y_u)(y_v - y_w) > 0$	$\widetilde{\mu}_2 < \mu_2$
$(y_t - y_u)(y_v - y_w) < 0$, and $\mu_3 - \mu_2 > \frac{y_t - y_u}{y_w - y_v} + \frac{y_w - y_v}{y_t - y_u}$	$\widetilde{\mu}_2 > \mu_2$

4.1 Algorithm From matrix perturbation community, researchers have achieved results on the intermediate eigenvalue problem of the second type, i.e., how to determine E such that the eigenvalue λ_1 of $A + E$ can be greater or less than that of A. Specifically, Cvetkovic et al.[4] gave results on how to increase or decrease λ_1 of the adjacency matrix by constructing the noise matrix E based on the principal eigenvector values of the adjacency matrix. We list their re-

sults in the first two rows of Table 1 and Table 2. For example, according to row 1 in Table 1, if we add edge (i, j) and delete edge (p, q) and $x_i x_j - x_p x_q > 0$ stands, λ_1 necessarily increases. Note that x_i denotes the ith component in the principal eigenvector of λ_1.

In this paper, we also need to know whether the eigenvalue μ_2 of the Laplacian matrix L of a particular graph G increases or decreases when an edge is relocated. We derive sufficient conditions on how to adjust μ_2 of the Laplacian matrix for two random strategies *Add/Del* and *Switch*. We summarize our results in the last two rows of Table 1 and 2, leaving the detailed proof in Appendix. Note that μ_2 is the important eigenvalue of the Laplacian matrix L. We use μ_i and $\widetilde{\mu}_i$ to denote the ith smallest eigenvalue of L and $\widetilde{L}$ respectively, and $\mathbf{u}_2$ denotes the eigenvector of μ_2. y_i is the ith component of $\mathbf{u}_2$.

Based on the derived conditions, we propose our spectrum preserving approach which can improve the simple edge randomization by considering the change of spectrum in the randomization process. Here we can determine which edges we should add/remove or switch so that we can control the move of target eigenvalues. As a result, real graph characteristics (or graph utility) are expected to be better preserved. We show our *Spctr Switch* algorithm in Algorithm 4.1.

ALGORITHM 4.1. Spectrum Preserving Graph Randomization through Edge Switch

Input: graph data G, protection threshold ε

1. Derive the adjacency matrix A and the Laplacian matrix L.
2. Calculate the eigenvalues and eigenvectors $(\lambda_1, \lambda_2, \mathbf{e}_1)$ of A and $(\mu_2, \mu_3, \mathbf{u}_2)$ of L respectively.
3. $k = 0$
4. While $J_2(k) \leq 1 - \varepsilon$
5. From graph G, randomly pick one edge (t, w);
6. If $k/2 == 0$
7. Find all the edge combinations such that $\widetilde{\lambda}_1 > \lambda_1$ and $\widetilde{\mu}_2 > \mu_2$;
8. Randomly pick one (u, v), switch (t, w) and (u, v) to (t, v) and (u, w) ;
9. otherwise
10. Find all the edge combinations such that $\widetilde{\lambda}_1 < \lambda_1$ and $\widetilde{\mu}_2 < \mu_2$;
11. Randomly pick one (u, v), switch (t, w) and (u, v) to (t, v) and (u, w) ;
12. k=k+1

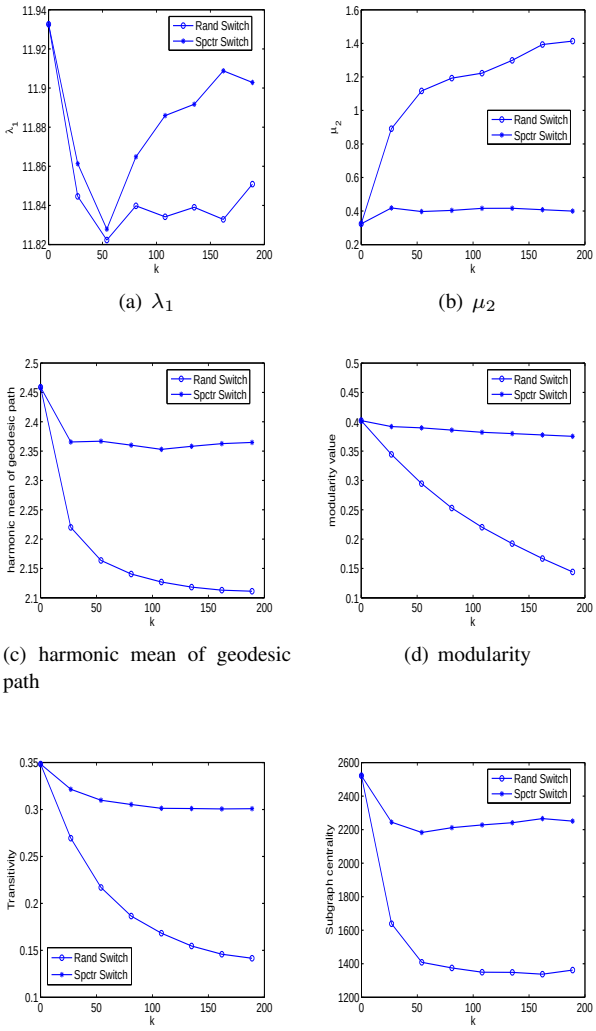
In Row 2 of Algorithm 4.1, we only calculate the first one or two eigenvalues of the corresponding graph matrices. It is not necessary or desirable to calculate the entire eigendecomposition. Note that calculation of the eigenvectors of an $n \times n$ matrix takes in general a number of operations $O(n^3)$. An efficient Lanczos method [3] can be applied to find the second eigenvector of a sparse matrix with $m/(\lambda_3 - \lambda_2)$, where m is the number of edges in the graph. Row 4 gives the loop condition of repeated switch operations (we will discuss details on $J_2(k)$ and the input privacy protection threshold ε in Section 5.2). Rows from 6 to 11 present how to switch based on the sufficient conditions listed in Table 2. Algorithm can be modified to *Spctr Add/Del* with some minor changes: replacing $J_2(k)$ with $J_1(k)$ in Row 4; replacing the switch process with the *Add/Del* process in Row 8 and 11; and finally, in Row 7 and 10 referring to Table 1 for the conditions under which the eigenvalues increase or decrease.

It is ideal to derive the sufficient conditions on how much one or some particular eigenvalues will change. This is the problem of estimating changes in eigenvalues under a wide range of perturbations. The eigenvalues of the perturbed graph can be determined as implicit functions of algebraic and geometric invariants of the original graph. However, this problem has not been solved in the matrix perturbation field.

4.2 Empirical Evaluation Figure 2 shows spectral randomization can significantly better preserve both graph spectrum and real space characteristics of the political book graph data set than the previous random perturbation which does not consider spectrum preserving during the perturbation process. Due to space limitations, we only include comparison between *Spctr Switch* and *Rand Switch*. We can see that *Spctr Switch* can significantly better keep both spectral characteristics and real characteristics close to those computed from the original graph even when we increase the number of switches k to 180. Note that the spectrum preserving approach adjusts both λ_1 and μ_2. The intuition here is that the more eigenvalues we control in perturbation, the more real space characteristics we can preserve in the randomized graph.

We also conducted evaluation on a relatively large data set, political blogosphere data [1]. It compiles the data on the links among US political blogs, containing over 1,000 vertices and 15,000 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.

Table 3 shows the relative change of the spectrum λ_1, μ_2 and the real characteristics (including the harmonic mean of geodesic path h, modularity Q, transitivity C, and subgraph centrality SC) between *Spctr Switch* and *Rand Switch* when we vary k from 300 to 3000. It is easy to observe that *Spctr Switch* preserve both spectrum and real characteristics



(a) λ_1 (b) μ_2

(c) harmonic mean of geodesic path (d) modularity

(e) transitivity (f) subgraph centrality

Figure 2: Graph characteristic vs. varying k between *Spctr Switch* and *Rand Switch*

of the graph much better than *Rand Switch*. In Section 5, we shall show our spectrum preserving randomization can achieve almost the same level of privacy protection as the random perturbation approach. In other words, the proposed approach can better preserve the utility without sacrificing much the privacy protection.

5 Privacy Analysis of Edge Randomization

When it comes to a randomization strategy, we are interested in how well it can preserve the privacy and whether it is resilient to some known attacks. In this section we focus on this issue for *Rand Add/Del* and *Rand Switch* methods, leaving *Spctr Add/Del* and *Spctr Switch* for future work. This is because, first, the process of spectrum preserving strategies are more complicated than that of random strategies; and we believe that the outline of the analysis for random strategies can provide a basis for analyzing the spectrum preserving strategies. Section 5.1 briefly discusses how random strategies are resilient to known subgraph based attacks, followed by our formal analysis on privacy preservation for *Rand Add/Del* and *Rand Switch*, including to what extent the edge randomization can protect the privacy.

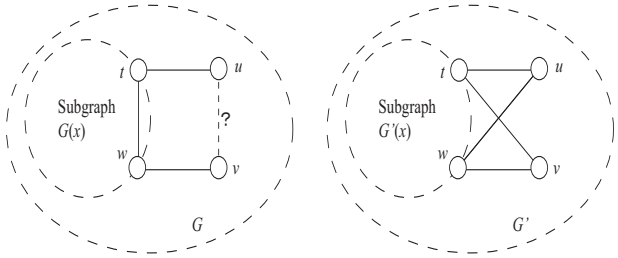
5.1 Resilient to Subgraph Attacks When it comes to anonymized graph, the attackers may have some a-prior knowledge of the graph such as some topological features or a subgraph. In [2] the authors describe a family of attacks such that an adversary can learn whether edges exist or not between specific 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 [7], the authors show by applying subgraph queries the identification of the vertices can be seriously jeopardized. They suggest and empirically evaluate 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.

For *Rand Add/Del* strategy, since each link is reallocated independently, knowing the subgraph cannot enhance the attacker'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.

For *Rand Switch*, will a subgraph prior known to the attackers disclose more beyond the subgraph? In the scenario of [2], the attackers know a subset of vertices $X \subset V$ and all the edges associated to X (denoted by $G(X)$). Although [7] shows that graph randomization can greatly reduce the chance for the attackers to re-identify the subgraph known to them, we here still assume that, in G', the attackers have identified the subgraph corresponding to $G(X)$, denoted by $G'(X)$. Then, from matrix perspective, they know the ith row or column of matrix A, $\widetilde{A}$ and thus E if $i \in X$. Therefore, among the four vertices involved in a switch, if more than three of them are in X, this switch is actually within $G(X)$; and if none of them are in X, the attackers can not utilize the known subgraph. Figure 3 shows a case under which two of the switched vertices are in X. In G', the attackers observe that vertex t, w, u, v form a pattern shown in Figure 3(b). Comparing with the known $G(X)$ shown in Figure 3(a), the attackers know that edge (t, w) is switched to (t, v) and (w, u) must be switched from another edge, probably the unknown (u, v). Then can they be sure

Table 3: Change of the measures for the US political blogs graph where the values in bold font denote the relative change from *Spctr Switch* while those in regular font denote the relative change from *Rand Switch*

k	$\lambda_1(\%)$	$\mu_2(\%)$	$h(\%)$	$Q(\%)$	$C(\%)$	$SC(\%)$
300	0.35, **0.33**	15.24, **15.68**	1.24, **1.13**	4.25, **3.87**	4.83, **4.55**	22.59, **21.67**
600	0.55, **0.51**	25.75, **22.81**	1.94, **1.70**	8.31, **6.91**	9.07, **7.69**	33.05, **30.77**
900	0.68, **0.58**	28.66, **29.83**	2.44, **2.01**	12.16, **9.33**	12.73, **9.88**	39.42, **34.06**
1200	0.77, **0.60**	32.01, **35.18**	2.81, **2.17**	15.82, **11.26**	15.91, **11.49**	43.23, **34.57**
1500	0.83, **0.58**	37.78, **47.38**	3.09, **2.26**	19.31, **12.94**	18.69, **12.65**	45.36, **33.04**
1800	0.85, **0.49**	28.93, **38.11**	3.31, **2.27**	22.61, **14.22**	21.12, **13.35**	46.50, **27.76**
2100	0.82, **0.41**	37.89, **30.05**	3.46, **2.25**	25.78, **15.49**	23.12, **13.89**	45.13, **22.58**
2400	0.79, **0.31**	50.45, **33.37**	3.59, **2.25**	28.82, **16.72**	24.88, **14.35**	43.68, **15.90**
2700	0.75, **0.23**	50.55, **20.22**	3.70, **2.24**	31.77, **17.92**	26.44, **14.78**	42.00, **10.55**
3000	0.69, **0.14**	54.27, **20.35**	3.77, **2.19**	34.53, **19.01**	27.66, **15.07**	39.32, **2.48**



(a) original graph G (b) perturbed graph G'

Figure 3: Resilient to subgraph attacks

of the existence of edge (u, v)? If the total perturbation times k is so small that the attackers are very confident that t, w, u, v are involved in perturbation at most once, definitely they are sure about the existence of edge (u, v). However when k is large, (u, v) can be a false edge switched from elsewhere. Moreover, it is more complicated if there are more than one false edges associated with t or w, for the attackers must guess where (t, w) is switched to. In this case, the attackers have the probability of $1/(c_t c_w)$ to be correct, where c_i denotes the number of false edges associated to vertex i. Similarly, the attackers can not learn beyond the subgraph if only one of the four switched vertices is in X. In summary, switch based randomization is robust to subgraph attack when k is not too small.

5.2 Privacy Analysis When it comes to privacy, we assume it is $a_{ij} = 1$ that people may want to hide, not $a_{ij} = 0$ and attackers are capable of calculating posterior probabilities. We use $P(a_{ij} = 1)$ to denote the users' prior belief about the event of $a_{ij} = 1$ and use $P(a_{ij} = 1 \mid \widetilde{a}_{ij})$ to denote its posterior belief after attackers observe the randomized data $\widetilde{a}_{ij}$. The released data $\widetilde{a}_{ij}$ is regarded as jeopardizing with respect to $a_{ij} = 1$ if $P(a_{ij} = 1 \mid \widetilde{a}_{ij}) > P(a_{ij} = 1)$. To calculate the posteriori probabilities, we need to know how many false edges exist in the perturbed graph. We give in Appendix details on how to compute the expectation value of false edges with *Rand Add/Del* and *Rand Switch* strategy.

We define the absolute measure of protection as

$$\tau_a(i,j) = 1 - \max\{P(a_{ij} = 1 \mid \widetilde{a}_{ij} = 0), P(a_{ij} = 1 \mid \widetilde{a}_{ij} = 1)\} \tag{5.13}$$

Note that the second term in Equation 5.13 can be considered as the maximal suspicion of existing $a_{ij} = 1$. The relative measure of protection is defined as

$$\tau_r(i,j) = \frac{\tau_a(i,j)}{1 - P(a_{ij} = 1)} \tag{5.14}$$

Our following result shows how to calculate the privacy measure.

RESULT 2. *For* Rand Switch*, after k switches, for vertex i, let c_i denote the number of false edges associated to vertex i in graph $\widetilde{G}$, i.e. $c_i = \frac{1}{2}\sum_{j=1}^{n} |\widetilde{a}_{ij} - a_{ij}|$, and $E(c_i)$ is its expectation. Then,*

$$\tau_a(i,j) = (1 - P_i)(1 - P_j), \tag{5.15}$$

$$\tau_r(i,j) = \frac{1 - P_i}{1 - S_i} \cdot \frac{1 - P_j}{1 - S_j}, \tag{5.16}$$

where $P_i = 1 - \frac{E(c_i)}{d_i}$, and $S_i = \frac{d_i}{n-1}$. $E(c_i)$ is shown in Result 5 in the appendix.

Proof. We here assume that the attacker has no other information except each vertex's degree which is kept unchanged in the perturbed data for the *Rand Switch* strategy. Intuitively, $S_i = \frac{d_i}{n-1}$ is the probability that a randomly selected vertex turns out an neighbor of vertex i's. Therefore, the prior probability can be shown as

$$P(a_{ij} = 1) = S_i + S_j - S_i S_j. \tag{5.17}$$

The posterior probability $P(a_{ij} = 1|\widetilde{a}_{ij} = 1)$ is the probability that an edge (i, j) in $\widetilde{G}$ is a true edge in G. $P_i = 1 - \frac{E(c_i)}{d_i}$ is vertex i's proportion of true edges. Hence,

$$P(a_{ij} = 1|\widetilde{a}_{ij} = 1) = P_i + P_j - P_iP_j \tag{5.18}$$

Similarly, $Q_i = \frac{E(c_i)}{n-1-d_i}$ is vertex i's proportion of false edges,

$$P(a_{ij} = 1|\widetilde{a}_{ij} = 0) = Q_i + Q_j - Q_iQ_j \tag{5.19}$$

Notice that P_i is a decreasing function of k and Q_i is an increasing with k, and

$$\lim_{k\to\infty} P_i = \lim_{k\to\infty} Q_i = \frac{d_i}{n-1}.$$

We thus have, $P_i \geq Q_i$. As a result

$$P_i + P_j - P_iP_j \geq Q_i + Q_j - Q_iQ_j.$$

Equation 5.15 and 5.16 is then derived by incorporating Equation 5.18 and 5.17 in Equation 5.14.

For the *Rand Add/Del* strategy, we give the result without proof.

RESULT 3. *For the* Rand Add/Del*, let b be the number of false edges in $\widetilde{G}$, i.e. $b = \frac{1}{4}\sum_{i,j} |\widetilde{a}_{ij} - a_{ij}|$, and $E(b)$ is its expectation. Then,*

$$\tau_a(i, j) = E(b)/m,$$
$$\tau_r(i, j) = \frac{E(b)/m}{1 - m/N},$$

where m is the number of edges and $N = n(n-1)/2$. $E(b)$ is shown in Result 6 in the appendix.

In the following we show how much the spectrum preserving randomization approach can achieve privacy protection. One problem here is that we currently cannot derive the formula of the protection measure $\tau_r(i, j)$ for either *Spctr Add/Del* or *Spctr Switch* since it is hard to calculate the number of false edges in the randomization. Our following explanation show the spectrum preserving approach can achieve similar privacy protection as the random perturbation approach. Consider a simple case in which we only control λ_1. In G, a switch moves λ_1 either up or down. Let ρ be the proportion of up pairs in the graph. Along the perturbation, consider Δk times of perturbation out of total k times. When Δk is small, the graph structure and ρ do not change much. Then, *Rand Switch* produces $\rho\Delta k$ times of up switch and $(1 - \rho)\Delta k$ down switch, while *Spctr Switch* produces exactly $\frac{\Delta k}{2}$ up switch and down switch. Hence only $\left|\frac{1}{2} - \rho\right| \Delta k$ switch in *Spctr Switch* may not produce the same privacy protection as *Rand Switch* does. When k is large, G' under *Rand Switch* tends to be a random graph whose λ_1 does not change much along the perturbation and the switching pair tends to play an equivalent role to the global structure. Hence, ρ tends to be $\frac{1}{2}$ when k increases, which means that *Rand Switch* and *Spctr Switch* do not differ much in protecting the privacy.

5.3 Privacy vs. k The measures of protection (τ_a and τ_r) are defined in terms of one individual edge. In the privacy preserving data mining, one natural question is how many perturbations we need such that we can guarantee the protection for all individual edges are above some threshold. Formally, we expect

- For *Rand Add/Del* strategy,

$$J_1(k) := \min_{i,j} \tau_r(i, j) = \frac{E(b)/m}{1 - m/N} > 1 - \varepsilon.$$

- For *Rand Switch* strategy,

$$J_2(k) := \min_{i,j} \tau_r(i, j) = \min_{i,j} \left\{ \frac{1 - P_i}{1 - S_i} \cdot \frac{1 - P_j}{1 - S_j} \right\} > 1 - \varepsilon.$$

It is easy to check that the protection for all individual edges remains the same with *Rand Add/Del* strategy. The relative measure in *Rand Switch* is a function of k, d_i, and d_j. Our next result shows we only need to consider the protection of the edges that connect the two vertices with the smallest degrees.

RESULT 4. *We re-numerate the vertices by their degree in ascending order: $d_1 \leq d_2 \leq \cdots \leq d_n$,*

$$J_2(k) = \frac{1 - P_1}{1 - S_1} \cdot \frac{1 - P_2}{1 - S_2}, \tag{5.20}$$

PROOF. We first prove that given a fixed k, if two vertices i and j, $d_i \leq d_j$, then

$$\frac{1 - P_i}{1 - S_i} \leq \frac{1 - P_j}{1 - S_j}. \tag{5.21}$$

To a single vertex i, *Rand Switch* strategy actually rearranges the position of 1 and 0 on the ith row of the adjacency matrix. A false edge of vertex i corresponds to a 1 reallocated elsewhere in the ith row of the adjacency matrix. Hence, to produce the same proportion of false edges, the number of 0's in j-th row of adjacency matrix should at least increase to $\frac{d_j}{d_i}(n - 1 - d_i)$:

$$\frac{E(c_i)}{n - 1 - d_i} \leq \frac{E(c_j)}{\frac{d_j}{d_i}(n - 1 - d_i)} \leq \frac{E(c_j)}{\frac{d_j}{d_i}(n - 1 - d_j)},$$

Table 4: τ_r vs. k for two strategies on Political Book data

$1-\varepsilon$	Rand Add/Del	Rand Switch
0.1	48	54
0.2	96	84
0.3	150	114
0.4	210	141
0.5	282	174
0.6	372	210
0.7	492	258
0.8	654	318
0.9	936	420

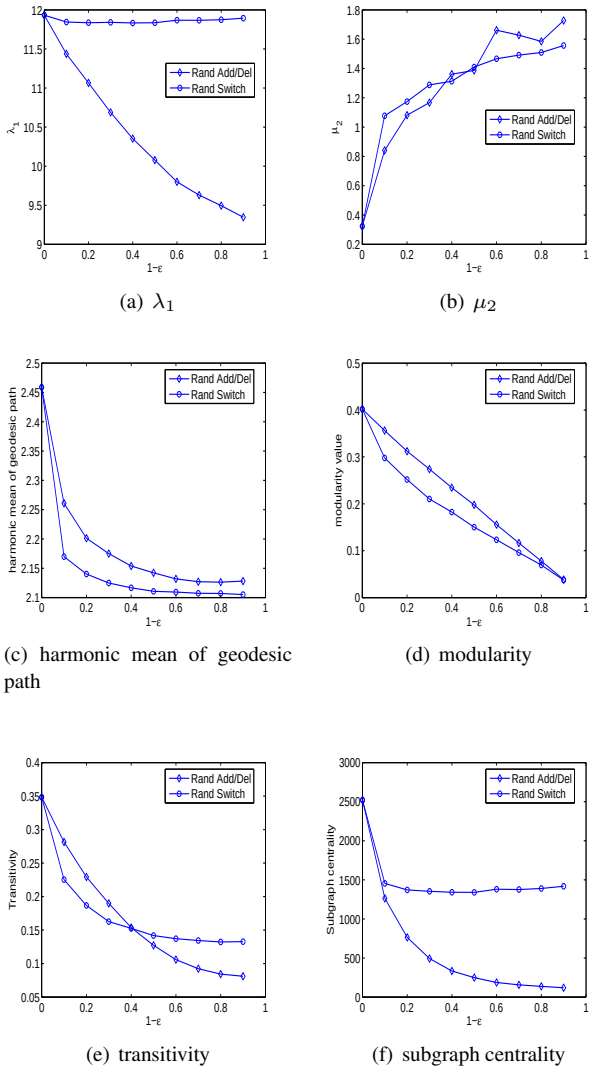
and with some simple deduction Equation 5.21 follows. Since $d_1 \leq d_2 \leq \cdots \leq d_n$, then by the above property, Equation 5.20 stands.

Table 4 shows the number of perturbations we need for *Rand Add/Del* strategy and *Rand Switch* when we aim to achieve different levels of privacy protection $(1-\varepsilon)$. Similary Figure 4 shows how graph characteristics vary with different privacy protection thresholds for both *Rand Add/Del* and *Rand Switch* strategies. We can see the higher the privacy protection we aim, the more perturbation we need, and the less the utility of the graph we can achieve.

6 Conclusion and Future Work

In this paper, we have developed one spectrum preserving randomization approach which can significantly improve the edge based graph randomization methods (*Rand Add/Del* and *Rand Switch*) by increasing the utility of the perturbed graph without sacrificing much the privacy protection. We have also given a bound of graph spectrum changes for pure randomization strategies (i.e., reallocating or switching edges randomly). Since the graph spectrum is closely related to many real graph characteristics, this bound provides a perspective on the extent to which the edge randomization affects the graph structure. Note that the bound derived in this paper can serve as a loose bound for spectrum preserving strategies. In future, we are interested in deriving some (tight) bound of graph spectrum changes for spectrum preserving randomization strategies. We have conducted privacy analysis for pure randomization strategies and will investigate thoroughly how spectrum preserving randomization strategies protect edge privacy.

There are some other aspects of this work that merit further research. Among them, We would conduct, more systematically, empirical evaluation on large social networks from various domains. We would explore potential attacks on randomized social networks especially in the scenario when attackers know additional information. We are trying to figure out the solution for the intermediate eigenvalue problem aiming to derive conditions to adjust any eigenvalue (in addition to λ_1 and μ_2) that may indicate certain structure character of the graph. We are also interested in studying more about the relation between real graph characteristic and graph spectrum. Hence more flexible algorithms can be designed when data owners a-priori know which graph characteristics they would like to preserve.



(a) λ_1 (b) μ_2

(c) harmonic mean of geodesic path (d) modularity

(e) transitivity (f) subgraph centrality

Figure 4: Graph characteristic vs. varying privacy protection on Political Book data

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 Jon 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] G. H. Chaudhuri and C. F. VanLoan. *Matrix Computations*. Johns Hopkins University Press, 1989.

[4] D. Cvetkovic, P. Rowlinson, and S. Simic. *Eigenspaces of Graphs*. Cambridge University Press, 1997.

[5] L. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances In Physics*, 56:167, 2007.

[6] E. Estrada and J. A. Rodrguez-Velzquez. Subgraph centrality in complex networks. *Physical Review E*, 71(056103), 2005.

[7] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava. Anonymizing social networks. *University of Massachusetts Technical Report*, 07-19, 2007.

[8] R. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge, 1985.

[9] V. Krebs. unpublished, http://www.orgnet.com/.

[10] V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Physics Review Letters*, 87, 2001.

[11] A. J. Seary and W. D. 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.

[12] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos. Epidemic spreading in real networks: An eigenvalue viewpoint. *Proceedings of the 22nd International Symposium on Reliable Distributed Systems*, 2003.

A Proof of The Results

A.1 Results on Adjusting μ_2 Let $\mathbf{u}_i$ and $\widetilde{\mathbf{u}}_i$ be the eigenvector corresponding to μ_i and $\widetilde{\mu}_i$. Consider the minimum problem:

$$\min_{x\in S}\left\{\mathbf{x}^T\widetilde{L}\mathbf{x}\right\},$$

where $S=\left\{\mathbf{x}:\mathbf{x}^T\widetilde{\mathbf{u}}_1=0,\text{and }\|\mathbf{x}\|_2=1\right\}$.

Since $\mathbf{u}_1=\widetilde{\mathbf{u}}_1$, $\mathbf{u}_2\in S$. Then

$$\min_{\mathbf{x}\in S}\left\{\mathbf{x}^T\widetilde{L}\mathbf{x}\right\}\le\mathbf{u}_2^T\widetilde{L}\mathbf{u}_2=\mu_2-\mathbf{u}_2^TE\mathbf{u}_2$$

On the other hand, take $\mathbf{x}$ to be $\widetilde{\mathbf{u}}_2$,

$$\widetilde{\mu}_2=\min_{\mathbf{x}\in S}\left\{\mathbf{x}^T\widetilde{L}\mathbf{x}\right\},$$

hence

$$\widetilde{\mu}_2\le\mu_2-\mathbf{u}_2^TE\mathbf{u}_2.$$

When $\mathbf{u}_2^TE\mathbf{u}_2>0$, $\widetilde{\mu}_2<\mu_2$ always holds. With the concrete form of EPM, in *Add/Del* strategy:

$$\mathbf{u}_2^TE\mathbf{u}_2=2(y_iy_j-y_py_q),$$

and in *Switch*:

$$\mathbf{u}_2^TE\mathbf{u}_2=2(y_t-y_u)(y_v-y_w).$$

For the rest part of the table, we focus on the Switch strategy. and *Add/Del* strategy can be proved similarly by using the corresponding perturbation matrix E.

Denote $\lambda_i(M)$ for ith eigenvalues of matrix M sorted in non-decreasing order: $\lambda_1(M)\le\lambda_2(M)\le\cdots\le\lambda_n(M)$. We take $t=1,v=2,u=3,w=4$ without loss of generality. Then, with the second part of the theorem,

$$(y_1-y_3)(y_2-y_4)<0,$$

$$E=\begin{pmatrix}0&1&0&-1&&\\1&0&-1&0&\vdots&\\0&-1&0&1&&\\-1&0&1&0&&\\&\cdots&&&0_{(n-4)\times(n-4)}&\end{pmatrix},$$

Based on Laplacian matrix, we construct our own $\bar{E}$ and $\bar{L}$ needed in the proof: $\bar{E}=(\delta+2)I-E$, and $\bar{L}=L-(\delta+2)I$, where $\delta>0$ is a parameter. Then,

- $\bar{E}$ is positive definite;
- $\lambda_i(\bar{L})=\lambda_i(L)-(\delta+2)$, and $\lambda_i(\bar{L})$ and $\lambda_i(L)$ have the same eigenvector;
- $\bar{L}+\bar{E}=L-E=\widetilde{L}$, and therefore $\mu_2=\lambda_2(\widetilde{L})=\lambda_2(\bar{L}+\bar{E})\ge\lambda_2(\bar{L}+\bar{E}P_2)$ where P_2 is the orthogonal projection onto the subspace spanned by $\{\bar{E}^{-1}\mathbf{u}_1,\bar{E}^{-1}\mathbf{u}_2\}$. (see [4] for more details).

With the similar deduction outlined in [4], we can calculate $\lambda_2(\bar{L}+\bar{E}P_2)$ and thus get a lower bound of $\widetilde{\mu}_2$:

$$\widetilde{\mu}_2\ge\min\{\mu_2-2-\delta+\gamma,\mu_3-2-\delta\},\tag{1.22}$$

where

$$\gamma=\frac{\delta(2+\delta)(4+\delta)}{\delta(\delta+4)-2b\delta+2a}\tag{1.23}$$

and $a=(y_1+y_2-y_3-y_4)^2, b=(y_1-y_3)(y_4-y_2)>0$. γ is an increasing function of δ with range $(0,\infty)$. We thus can always choose $\delta>0$ such that $\gamma=\mu_3-\mu_2$, then we rewrite (Equation 1.22) as

$$\widetilde{\mu}_2\ge\mu_3-2-\delta.$$

Next we deduct the condition under which this lower bound is always greater than μ_2, or equivalently the following inequalities and equation always stands:

$$\begin{cases}\widetilde{\mu}_2\ge\mu_3-2-\delta>\mu_2\\\gamma=\dfrac{\delta(2+\delta)(4+\delta)}{\delta(\delta+4)-2b\delta+2a}\\\gamma=\mu_3-\mu_2\end{cases}\tag{1.24}$$

It is not difficult to show that when

$$\gamma = \mu_3 - \mu_2 > 2 + \frac{a}{b},$$

(Equation 1.24) stands. Since

$$2 + \frac{a}{b} = \frac{(y_1 - y_3)}{(y_4 - y_2)} + \frac{(y_4 - y_2)}{(y_1 - y_3)},$$

when $\mu_3 - \mu_2 > \frac{(y_1-y_3)}{(y_4-y_2)} + \frac{(y_4-y_2)}{(y_1-y_3)}$, $\widetilde{\mu}_2 > \mu_2$ stands. The rest parts of the result are proved.

A.2 The Number of False Edges

RESULT 5. *For* Rand Switch*, denote*

$$c_i = \frac{1}{2} \sum_{j \neq i} |\widetilde{a}_{ij} - a_{ij}|, \tag{1.25}$$

$0 \leq c_i \leq C_i := \min\{d_i, n - 1 - d_i\}$.

Denote q_i as the probability that a switching occurs to vertex i. It can be approximated as

$$q_i \approx \frac{d_i}{m} + \sum_{k \neq i} \frac{d_k}{m} \cdot \frac{d_i - a_{ik}}{m - d_k} \tag{1.26}$$

The expectation of c_i is shown as

$$E(c_i) = (0, 1, 2, \ldots, C_i) \left((1 - q_i)I + q_i P_i\right)^k \mathbf{e}_1.$$

where $\mathbf{e}_1 = (1, 0, 0, \ldots, 0)^T$, $P_i = (p_{st}^{(i)})_{(C_i+1)\times(C_i+1)}$ *and*

$$p_{st}^{(i)} = \begin{cases} \dfrac{t^2}{d_i(n-1-d_i)}, & (s = t-1) \\ \dfrac{t(n-1-2t)}{d_i(n-1-d_i)}, & (s = t) \\ \dfrac{(d_i - t)(n-1-d_i-t)}{d_i(n-1-d_i)}, & (s = t+1) \\ 0, & (otherwise). \end{cases} \tag{1.27}$$

PROOF. The probability that a switching occurs to vertex is a constant. By saying a switch occurs to vertex i, we mean that one of the two switched edges connects to vertex i. Suppose one switch occurs to vertex i. In the ith row of the adjacency matrix $\mathbf{a}_i = (a_{i1}, a_{i2}, \ldots, a_{i,n})$, one component, say a_{ip}, changes from 1 to 0 and another component a_{iq} change from 0 to 1. Equivalently, we replace a 1 in $\mathbf{a}_i$. Since we select the edges uniformly, every 1 (0) has same possibility to become 0 (1). Given r of the k times of switch to vertex i, we first calculate $E(c_i|r)$. The change of c_i follows the Markov chain with the stationary probabilities, and c_i has finite states: $0, 1, \ldots, C_i$. Then, it is easy to establish the transition matrix P_i whose elements $p_{st}^{(i)} = P(c_i^{(n+1)} = s | c_i^{(n)} = t)$ is shown in Equation 1.27. The initial probability distribution vector is $\mathbf{e}_1$. Hence,

$$\begin{aligned} E(c_i|r) &= \sum_{x=0}^{C_i} xP(c_i = x) = (0, 1, 2, \ldots, C_i) P_i^r \mathbf{e}_1. \\ E(c_i) &= \sum_{x=0}^{k} E(c_i|r = x) P(r = x) \\ &= \sum_{x=0}^{k} \binom{k}{x} q_i^x (1 - q_i)^{k-x} E(c_i|r = x) \\ &= (0, 1, 2, \ldots, C_i) \left((1 - q_i)I + q_i P_i\right)^k \mathbf{e}_1. \end{aligned}$$

RESULT 6. *For* Rand Add/Del*, denote*

$$b = \frac{1}{4} \sum_{i,j} |\widetilde{a}_{ij} - a_{ij}|,$$

$0 \leq b \leq B := \min\{m, N - m\}$*, where $N = n(n-1)/2$ is the number of all the possible edges.*

$$E(b) = (0, 1, 2, \ldots, B) P^k \boldsymbol{e}_1, \tag{1.28}$$

where $P = (p_{st})_{(B+1)\times(B+1)}$ *and*

$$p_{st} = \begin{cases} \dfrac{t^2}{m(N-m)}, & (s = t-1) \\ \dfrac{t(N-2t)}{m(N-m)}, & (s = t) \\ \dfrac{(m-t)(N-m-t)}{m(N-m)}, & (s = t+1) \\ 0, & (otherwise). \end{cases}$$