On Randomness Measures for Social Networks

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

Abstract

Social networks tend to contain some amount of randomness and some amount of non-randomness. The amount of randomness versus non-randomness affects the properties of a social network. In this paper, we theoretically analyze graph randomness and present a framework which provides a series of non-randomness measures at levels of edge, node, and the overall graph. We show that graph non-randomness can be obtained mathematically from the spectra of the adjacency matrix of the network. We also derive the upper bound and lower bound of non-randomness value of the overall graph. We conduct both theoretical and empirical studies in spectral geometries of social networks and show our proposed non-randomness measures can better characterize and capture graph randomness than previous measures.

1 Introduction

Social networks have received much attention these days. To understand and utilize the information in a social network, researches have developed various measures to indicate the structure and characteristics of the network from different perspectives [4]. Various properties including its size, density (a measure of the relative number of connections), power-law degree distributions, average distance, small-world phenomenon, clustering co-efficient(the tendency of the network to aggregate in subgroups), community structures etc. have been discovered. For surveys of analysis of large graphs, refer to [4, 11, 17].

Social networks tend to contain some amount of randomness and some amount of non-randomness. Consider an online social network where each node denotes an individual and an edge between two nodes denotes a social interaction between the two individuals. An individual's social network tends to consist of members of the same ethnic group, race, or social class. Intuitively, two friends of a given individual are more likely to be friends with each other than they are with other randomly chosen members. The edge connecting one individual's two friends contains less randomness. However, an individual also tends to have some number of random friends from other groups and those edges between this individual and his random friends contain more randomness.

The amount of randomness versus non-randomness at node/edge levels can clearly affect various properties of a social network. Although randomness plays an important role in understanding the geometry and topology of social networks, very few studies have formally investigated this issue.

In this paper, we analyze graph randomness at all granularity levels, from edge, node to the whole graph. We shall present a framework which provides a series of non-randomness measures at different levels. Non-randomness specified at the edge level can help users quantify how different a given interaction is from random interactions. Similarly, non-randomness specified at the node level can help users quantify how different a given individual is from random nodes (those individuals actually not belonging to this social network). In our framework, we first examine how much non-randomness a given edge (social interaction) has, then measure a node's non-randomness by examining the non-randomness values of edges connecting to this node. Finally, we derive the non-randomness measure of the whole graph by incorporating the non-randomness values of all edges within the whole graph.

Our study analyzes social networks from a spectrum point of view. Graph spectral analysis deals with the analysis of the spectra (eigenvalues and eigenvector components) of the nodes in the graph. It has been shown that there is an intimate relationship between the combinatorial characteristics of a graph and the algebraic properties of its adjacency matrix. Our theoretical results shall show that all our non-randomness measures can be determined by spectral coordinates of nodes in the first k-dimensional spectral space where k corresponds to the number of communities in the graph.

1.1 Our Contributions

Our contributions are summarized as follows.

- We discover spectral geometry properties in social networks that can determine the graph's non-randomness at all granularity levels.

- We present a framework which can quantify graph non-randomness at edge, node, and the overall graph levels. We show that all graph non-randomness measures can be obtained mathematically from the spectra of the adjacency matrix of the network. We present a relative non-randomness measure of the overall graph, which allows quantitative comparisons between various social networks with different sizes and densities or between different snapshots of a dynamic social network. We show that the expectation of the relative non-randomness measure of random graphs generated by the ER model [6] is zero and the relative non-randomness measure of any graph is lower (upper) bounded by the regular (complete) graph.
- We conduct both theoretical and empirical studies in spectral geometries of various real-world social networks and random graphs. Our proposed non-randomness measures can better characterize and capture graph randomness than previous measures [2].
- We also analyze how both edge non-randomness and node non-randomness distribute in real-world social networks and random graphs. Our results show that edge non-randomness and node non-randomness of real-world social networks usually display some high skewed distributions, obeying either a power law or an exponential law. On the contrary, random graphs display approximate normal distributions.

1.2 Paper Outline The rest of this paper is organized as follows. In Section 2 we discuss the notations used in this paper and present preliminaries of spectral graph analysis. In Section 3 we theoretically analyze how coordinates of node points are distributed in the k-dimensional spectral space and why they can be used to measure graph non-randomness. We present our framework and analyze in detail how to derive edge non-randomness, node non-randomness, and the overall graph non-randomness from graph spectrum in Section 4. Section 5 presents our empirical evaluations using various social networks and random graphs. We offer our concluding remarks and discuss future work in Section 6.

2 Preliminaries

Throughout this paper, we use bold lower-case variables, e.g., $\boldsymbol{x}$, to represent vectors; upper-case alphabets, e.g., A, to denote a matrix. $\boldsymbol{\alpha}^T$ refers the transpose of vector $\boldsymbol{\alpha}$. Table 1 summarizes the notation used in this paper.

Table 1: Notations

Symbol	Definition
n	number of nodes in graph G
m	number of edges in graph G
A	the adjacency matrix of graph G
$\Gamma(u)$	the set of the neighbors of node u
$\lambda_i, \boldsymbol{x}_i$	λ_i is the ith largest eigenvalue of A, and $\boldsymbol{x}_i$ is its eigenvector.
x_{ij}	x_{ij} is the jth entry of $\boldsymbol{x}_i$
X	$X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)$
$\boldsymbol{\alpha}_u$	the spectral coordinate of node u: $\boldsymbol{\alpha}_u = (x_{1u}, x_{2u}, \ldots, x_{ku})$
$R(u, v)$	the non-randomness of an edge (u, v).
$R(u)$	the non-randomness of a node (u)
R_G	the non-randomness of the overall graph G

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. 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. Let λ_i be the eigenvalues of A and $\boldsymbol{x}_i$ the corresponding eigenvectors, and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. The spectral decomposition of A is $A = \sum_i \lambda_i \boldsymbol{x}_i \boldsymbol{x}_i^T$. Let $\boldsymbol{x}_i$ be the unit eigenvector of λ_i and let x_{ij} denote the j'th entry of $\boldsymbol{x}_i$.

(2.1)

$$\begin{array}{cccccccc}
 & \boldsymbol{x}_1 & & \boldsymbol{x}_i & & \boldsymbol{x}_k & & \boldsymbol{x}_n \\
 & & & \downarrow & & & & \\
 & x_{11} & \cdots & x_{i1} & \cdots & x_{k1} & \cdots & x_{n1} \\
 & \vdots & & \vdots & & \vdots & & \vdots \\
\boldsymbol{\alpha}_u \rightarrow & x_{1u} & \cdots & x_{iu} & \cdots & x_{ku} & \cdots & x_{nu} \\
 & \vdots & & \vdots & & \vdots & & \vdots \\
 & x_{1n} & \cdots & x_{in} & \cdots & x_{kn} & \cdots & x_{nn}
\end{array}$$

We can see from Formula (2.1) that the eigenvector $\boldsymbol{x}_i$ is represented as a column vector. The row vector $(x_{1u}, x_{2u}, \cdots, x_{nu})$ represents the coordinates of node u in the n-dimensional spectral space. In Section 3, we shall show that only the coordinates of node u in the first k-dimensional spectral space determine the randomness of u where k indicates the number of communities within the graph. Hence we define $\boldsymbol{\alpha}_u = (x_{1u}, x_{2u}, \ldots, x_{ku}) \in \mathbb{R}^{1 \times k}$ as the spectral coordinate of node u in the k-dimensional space.

It has been shown that the eigenvalues of a network are intimately connected to many important topological features. For example, The eigenvalues of A encode information about the cycles of a network as well as its diameter. The maximum degree, chromatic number, clique number, and extend of branching in a connected graph are all related to λ_1. In [18], 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. Refer to [13] for more relationships between the spectral and real characteristics of graphs.

3 Graph Spectral Geometry

In this section, we explore how the spectral coordinate ($\boldsymbol{\alpha}$) of a node point locates in the projected spectral space. Especially we will show that node points locate along k quasi-orthogonal lines when graph G contains k communities [1].

PROPOSITION 1. *For a graph with k communities, the coordinate of node u in k-dimensional space,* $\boldsymbol{\alpha}_u = (x_{1u}, x_{2u}, \ldots, x_{ku}) \in \mathbb{R}^{1\times k}$, *denotes the likelihood of node u's attachment to these k communities. Node points within one community form a line that goes through the origin in the k-dimensional space. Nodes in k communities form k quasi-orthogonal lines in the spectral space.*

PROOF. Consider the division of a graph G into k non-overlapping communities $G_1, G_2, \ldots, G_k$. Let $\boldsymbol{s}_i = (s_{i1}, s_{i2}, \ldots, s_{in})$ be the index vector of community G_i, and s_{ij} equals to 1 if node j belongs to community G_i and 0 otherwise. Note that $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are mutually orthogonal, i.e., $\boldsymbol{s}_i^T \boldsymbol{s}_j = 0$.

For community G_i, we can define its density as

$$D(G_i) := \frac{\#\text{ of edges in } G_i}{\#\text{ of nodes in } G_i}.$$

It can be expressed as

$$D(G_i) = \frac{\boldsymbol{s}_i^T A \boldsymbol{s}_i}{\boldsymbol{s}_i^T \boldsymbol{s}_i}$$

where A is the adjacency matrix of graph G. The density for this division of the graph is

$$\sum_{i=1}^{k} D(G_i) = \sum_{i=1}^{k} \frac{\boldsymbol{s}_i^T A \boldsymbol{s}_i}{\boldsymbol{s}_i^T \boldsymbol{s}_i} \tag{3.2}$$

The task of our graph partition is to maximize Equation (3.2) subject to $s_{ij} \in \{0,1\}$ and $\boldsymbol{s}_i^T \boldsymbol{s}_j = 0$, if $i \neq j$. This optimization problem is NP-complete. However, if we relax $s_{ij} \in \{0,1\}$ to real space, based on the Wielandt's theory [16], we have that the target function reaches the maximum $\sum_{i=1}^{k} \lambda_i$ when taking $\boldsymbol{s}_i$ to be $\boldsymbol{x}_i$.

Hence we can conclude that x_{ij} reflects the degree of node j's attachment to the community G_i.

PROPERTY 1. *A node u belongs to one community G_t if the tth entry of $\boldsymbol{\alpha}_u$, x_{tu}, is much greater than the rest entries and $x_{iu} \approx 0$ for $i \neq t$.*

A node u does not belong to any community if all the entries of $\boldsymbol{\alpha}_u$ are close to 0, or equivalently, $\|\boldsymbol{\alpha}\|_2 \approx 0$. We call such nodes noise nodes.

PROPERTY 2. *If nodes u and v belong to the same community, then*

$$|\cos(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)| \approx 1.$$

If nodes u and v belong to two different communities respectively, then

$$|\cos(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)| \approx 0.$$

Otherwise, if node u belongs to one community G_t and bridging node v locates in the overlap of two communities G_t and G_w, then $|\cos(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)|$ is not close to either 0 or 1.

EXPLANATION. Notice that

$$\cos(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v) = \frac{\boldsymbol{\alpha}_u \boldsymbol{\alpha}_v^T}{\|\boldsymbol{\alpha}_u\|_2 \|\boldsymbol{\alpha}_v\|_2}.$$

When node u and v are in the same community G_t, x_{tu}, we have that x_{tv} is much greater than the rest entries in $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$. Hence

$$\begin{aligned} \frac{\boldsymbol{\alpha}_u \boldsymbol{\alpha}_v^T}{\|\boldsymbol{\alpha}_u\|_2 \|\boldsymbol{\alpha}_v\|_2} &= \frac{\sum_{i=1}^{k} x_{iu} x_{iv}}{\left(\sum_{i=1}^{k} x_{iu}^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{k} x_{iv}^2\right)^{\frac{1}{2}}} \\ &\approx \frac{x_{tu} x_{tv}}{|x_{tu}||x_{tv}|} = \pm 1. \end{aligned}$$

In other words, points $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$ approximately locate along a straight line that goes through the origin.

Similarly, when node u and v are in two different communities G_t and G_w respectively, with $x_{wu} \approx 0$ and $x_{tv} \approx 0$, we have

$$\frac{\boldsymbol{\alpha}_u \boldsymbol{\alpha}_v^T}{\|\boldsymbol{\alpha}_u\|_2 \|\boldsymbol{\alpha}_v\|_2} \approx \frac{x_{tu} x_{tv} + x_{wu} x_{wv}}{|x_{tu}||x_{wv}|} \approx 0,$$

which means that $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$ are approximately orthogonal.

If a bridging node v is in the overlap of two communities S_t and S_w, both tth and wth entries in $\boldsymbol{\alpha}_v$ are not negligible. Hence, $\|\boldsymbol{\alpha}_v\|_2 \approx \left(x_{tv}^2 + x_{wv}^2\right)^{\frac{1}{2}}$. For a node u from G_t, we have

$$\frac{|\boldsymbol{\alpha}_u \boldsymbol{\alpha}_v^T|}{\|\boldsymbol{\alpha}_u\|_2 \|\boldsymbol{\alpha}_v\|_2} \approx \frac{|x_{tu} x_{tv}|}{|x_{tu}| \left(x_{tv}^2 + x_{wv}^2\right)^{\frac{1}{2}}} = \frac{|x_{tv}|}{\left(x_{tv}^2 + x_{wv}^2\right)^{\frac{1}{2}}}.$$

Since neither x_{tv} nor x_{wv} is close to 0, $|\cos(u, v)|$ is not close to either 1 or 0, which indicates that bridging nodes locate between the quasi-orthogonal lines formed by communities, and are also away from the origin.

Figure 2 shows the 2-D spectral geometries of a synthetic network as shown in Figure 1. In Figure 1, there exist two dense subgraphs (denoted by red or blue color respectively), which are separated by one bridging node (node 45, denoted by white color), in addition to some random nodes (denoted by green color). We can observe from Figure 2 that

[1] Communities are loosely defined as collections of individuals who interact unusually frequently.

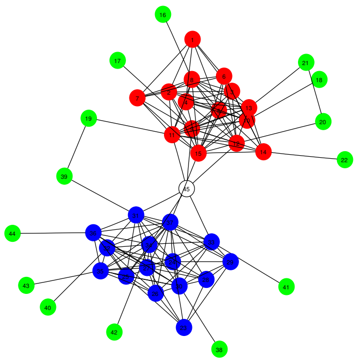


Figure 1: A synthetic network with two communities

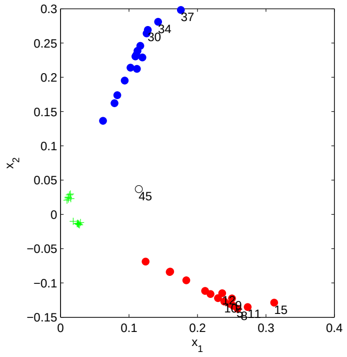


Figure 2: Node coordinates projected in 2-D spectral space

nodes in the two dense subgraphs are projected along two straight and quasi-orthogonal lines in the 2-D spectral space and nodes in green locate around the origin in the projected space. We can also observe that node 45 (white color), which separates two communities, locates away from the origin and between two quasi-orthogonal lines.

4 A framework of measuring graph non-randomness

In this section, we present our framework which can quantify randomness at all granularity levels from edge, node, to the overall graph. We begin with a study of edge non-randomness by spectral coordinates of its two connected nodes in the spectral space. We then define the node non-randomness as the sum of non-randomness values of all edges that connect to it. Similarly, we define the overall graph non-randomness as the sum of non-randomness values of all edges within the the whole graph. The formal definition is given below.

DEFINITION 4.1. *Denote* $\boldsymbol{\alpha}_u = (x_{1u}, x_{2u}, \dots, x_{ku}) \in \mathbb{R}^k$ *as the spectral coordinate of node* u *and* $\boldsymbol{\alpha}_v = (x_{1v}, x_{2v}, \dots, x_{kv}) \in \mathbb{R}^k$ *as the spectral coordinate of node* v.

1. *The edge non-randomness* $R(u, v)$ *is defined as*

$$R(u,v) = \boldsymbol{\alpha}_u \boldsymbol{\alpha}_v^T = \textstyle\sum_{i=1}^k x_{iu} x_{iv}.$$

2. *The node non-randomness* $R(u)$ *is defined as*

$$R(u) = \textstyle\sum_{v \in \Gamma(u)} R(u,v),$$

where $\Gamma(u)$ *denotes the neighbor set of node* u.

3. *The graph non-randomness* R_G *is defined as*

$$R_G = \textstyle\sum_{(u,v) \in E} R(u,v).$$

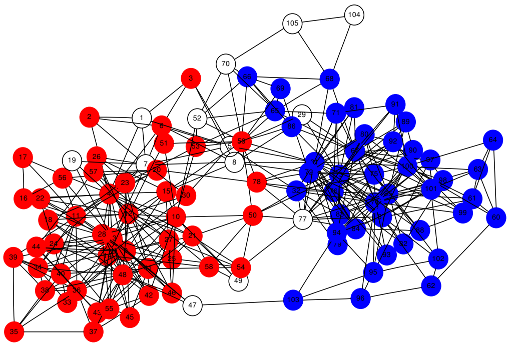


Figure 3: Politics book social network

Throughout this section, we use politics book network [10] as an example to illustrate how we define and calculate graph non-randomness at various levels. The politics book network contains 105 nodes and 441 edges as shown in Figure 3. In this network, 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. Each node is labeled as “liberal”(blue), “neutral”(white), or “conservative”(red). These alignments were assigned separately by Mark Newman based on a reading of the descriptions and reviews of the books posted on Amazon.

Figure 4 shows the 2-D spectral geometries of the politics book network data. We can observe from Figure 4 that the majority of vertices projected in the 2-D spectral space distribute along two straight and quasi-orthogonal lines. It indicates that there exist two communities with sparse edges connecting them. The first up-trend line consists of most nodes in red color while the second down-trend line consists of most nodes in blue color. White nodes distribute either around the origin or between two quasi-orthogonal lines in the projected space.

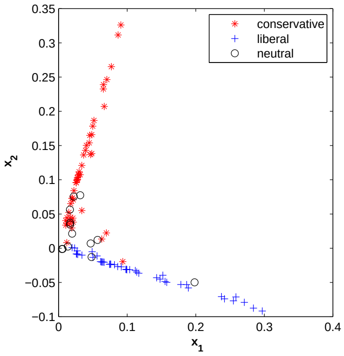


Figure 4: The 2-D spectral geometries of politics book social network

4.1 Edge Non-randomness: $R(u,v)$

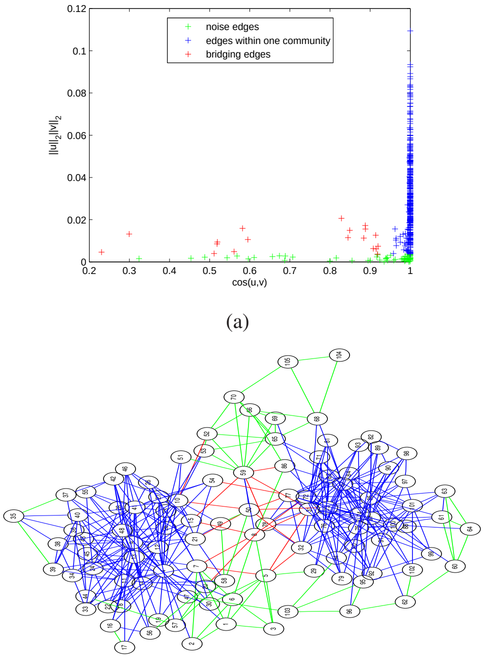
From Section 3, we know that the spectral coordinates of a node reflect its relative attachment to different communities in G. When it comes to the measure of non-randomness of an edge that connects two nodes, intuitively, we need to incorporate the relationship of two nodes' spectral vectors.

The edge non-randomness measure $R(u,v)$ in Definition 1 can be rewritten as

$$R(u,v) = \|\boldsymbol{\alpha}_u\|_2 \|\boldsymbol{\alpha}_v\|_2 \cos(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_u),$$

which is determined by the product of $\|\boldsymbol{\alpha}_u\|_2\|\boldsymbol{\alpha}_v\|_2$ and the cosine of the angle between $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_u$. Generally, $R(u,v)$ tends to be large when u and v are clearly belong to the same community (since $\cos(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_u) \approx 1$). $R(u,v)$ tends to be small when 1) u and v are from two different communities (since $\cos(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_u) \approx 0$); 2) or either node (or both nodes) is noisy (since $\|\boldsymbol{\alpha}_u\|_2\|\boldsymbol{\alpha}_v\|_2 \approx 0$). This intuitively reflects the formation of real world social networks: two individuals within the same community have relatively higher probability to be connected than those in different communities.

Figure 5(a) plots the distribution of edge non-randomness values, where x-axis is the cosine value between $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$ while y-axis denotes the product of the two vector lengths. Figure 5(b) shows a snapshot of different types of 441 edges characterized by edge non-randomness values of politics book network. We can observe that distributions of edge non-randomness values characterized by different regions reflect different types of edges in the original graph: edges with large cosine value (plotted along the vertex line $x = 1$ and denoted by the blue '+') mostly connect two nodes within the same community; edges with small vector length product (green '+' and plotted along the line $y = 0$) mostly connect to non-central nodes; edges plotted in other area forms bridging edges between the two communities. All the above is consistent with our previous explanations in Section 3.



(a)

(b)

Figure 5: Snapshot of different types of edges characterized by edge non-randomness of politics book network

4.2 Node Non-randomness: $R(u)$

A node's non-randomness is characterized by the non-randomness of edges connected to this node. This is well understood since edges in social networks often exhibit patterns that indicate properties of the nodes such as the importance, rank, or category of the corresponding individuals. Result 1 shows how to calculate the node non-randomness using the spectral coordinates as well as the first k eigenvalues of the adjacency matrix.

RESULT 1. *The non-randomness of node u is the length of its spectral vector with eigenvalue weighted on corresponding dimensions:*

$$R(u) = \textstyle\sum_{i=1}^{k} \lambda_i x_{iu}^2 = \boldsymbol{\alpha}_u \Lambda_k \boldsymbol{\alpha}_u^T, \tag{4.3}$$

where $\Lambda_k = diag\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$.

PROOF. Let $\boldsymbol{a}_u$ denote the u'th row of the adjacency matrix A. Since $\boldsymbol{x}_i$ satisfies $A\boldsymbol{x}_i = \lambda_i \boldsymbol{x}_i$ and A is symmetric,

$$\begin{pmatrix} \boldsymbol{a}_1 \\ \vdots \\ \boldsymbol{a}_n \end{pmatrix} \boldsymbol{x}_i = A\boldsymbol{x}_i = \lambda_i \begin{pmatrix} x_{i1} \\ \vdots \\ x_{in} \end{pmatrix}.$$

Hence, $\boldsymbol{a}_u \boldsymbol{x}_i = \lambda_i x_{iu}$, and we have

$$\begin{aligned} R(u) &= \sum_{v \in \Gamma(u)} R(u,v) = \sum_{v=1}^{n} \sum_{i=1}^{k} a_{uv} x_{iu} x_{iv} \\ &= \sum_{i=1}^{k} \left(x_{iu} \sum_{v=1}^{n} a_{uv} x_{iv} \right) \\ &= \sum_{i=1}^{k} x_{iu} \boldsymbol{a}_u \boldsymbol{x}_i = \sum_{i=1}^{k} \lambda_i x_{iu}^2 = \boldsymbol{\alpha}_u \Lambda_k \boldsymbol{\alpha}_u^T. \end{aligned}$$

We can see that the result is elegant since the node non-randomness is actually determined by its vector length weighted by eigenvalues of the adjacency matrix.

Using node non-randomness measure, we can easily separate singleton nodes [2] and noise nodes (with small $R(u)$ values) from those nodes strongly attached to some community (with large $R(u)$ values). We can also identify those nodes bridging across several groups by examining its relative positions to orthogonal lines corresponding to different communties.

4.2.1 Comparison with HITS

Our node non-randomness $R(u)$ can be used to identify those non-random individuals. However, it is different from those traditional link based object ranking methods based on centrality measures. For example, HITS algorithm [9] uses the principle eigenvector to assign *authority/hub* scores to each node. For undirected social networks, since A is now symmetric, *authority* and *hub* scores are the same, which are the principle eigenvector of A^2. Denote $A = X\Lambda X^T$ as the eigen-decomposition of A. Since X is orthogonal, $A^2 = X\Lambda^2 X$, the *authority/hub* scores from HITS algorithm in undirected networks are equivalent to the entries of $\boldsymbol{x}_1$. Therefore, if we are sure that the graph has only one community, our measure is reduced to the HITS score. However, many real-world graphs contain more then one community.

Table 2 compares the difference between the top 10 non-random nodes identified by our measure and the those identified by HITS for polbooks network. We can observe from the Table 2 that top 10 nodes identified by our measures include important nodes from two communities while HITS only identifies nodes from one community. This is because HITS uses $\boldsymbol{x}_1$ only, the scores only reflect relative positions of points along the x_1-axis in Figure 4. Hence they can only discover central nodes in one community (labeled as *liberal*) with the highest density. On the contrary, our node non-randomness measure, which uses the weighted vector length in the k-dimensional spectral space, can successfully discover non-random nodes from all k communities. This empirical evaluation indicates our node non-random measure is different from the traditional centrality measures used to rank nodes.

Table 2: Comparison of top 10 non-random nodes identified by $R(u)$ and HITS.

HITS	label	$R(u)$	label
85	liberal	9	conservative
74	liberal	13	conservative
73	liberal	85	liberal
31	liberal	74	liberal
67	liberal	73	liberal
75	liberal	4	conservative
76	liberal	31	liberal
77	neutral	67	liberal
87	liberal	12	conservative
72	liberal	75	liberal

4.3 Graph Non-randomness R_G and Relative Non-randomness R_G^*

In our framework, the graph non-randomness R_G is defined as the sum of non-randomness values of all edges within the graph. Result 2 shows R_G can be directly calculated using the first k eigenvalues.

RESULT 2. *The graph non-randomness of the overall graph G can be calculated as*

$$R_G = \sum_{(u,v) \in E} R(u,v) = \frac{1}{2} \sum_{u \in G} R(u) = \sum_{i=1}^{k} \lambda_i \tag{4.4}$$

PROOF. The second equation is straightforward since every edge is counted twice in the sum of node non-randomness. For the third equation, denote X as $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)$ where each column is an eigenvector of A: $A\boldsymbol{x}_i = \lambda_i \boldsymbol{x}_i$, hence we have

$$\sum_{(u,v) \in E} R(u,v) = \sum_{u,v} a_{uv} \boldsymbol{\alpha}_u \boldsymbol{\alpha}_v^T = trace(X^T A X) = \sum_{i=1}^{k} \lambda_i.$$

□

The above result is elegant since we can use the sum of the first k eigenvalues to determine the non-randomness of the overall graph. Recall that k indicates the number of communities in the graph. In this paper, we assume the value of k is either specified by domain users or discovered by those graph partition methods. There are tons of work on how to partition graph into k communities (refer to a survey paper [1]).

Chung and Graham indicated the use of the largest eigenvalue λ_1 as an index of the non-randomness of the overall graph since the first eigenvalue of random graphs characterizes the frequency of subgraphs [2]. Our analysis shows

[2] The singletons are degree-zero nodes who joined the network but have never made an interaction with another user in the social network.

that λ_1 may not be an appropriate measure to quantify the graph non-randomness for real-world social networks since they usually contain more than one communities. Actually, we can see that the index of graph non-randomness using λ_1 is a special case of our proposed measure R_G with $k = 1$.

All real networks lie somewhere between the extremes of complete order and complete randomness. While the absolute non-randomness measure R_G can indicate how random a graph G is, it is more desirable to give a relative measure so that graphs with different size and density can be compared. One intuitive approach is comparing the graph's non-randomness value with the expectation of non-randomness value of all random graphs generated by ER model. We can use the standardized measure defined as

$$R_G^* = \frac{R_G - E(R_G)}{\sigma(R_G)}$$

where $E(R_G)$ and $\sigma(R_G)$ denote the expectation and standard deviation of the graph non-randomness under ER model. Our Theorem 1 shows the distribution of R_G.

THEOREM 1. *For a graph G with $k(\ll n)$ communities where each community is generated by ER model with parameter $\frac{n}{k}$ and p, then R_G has an asymptotically normal distribution with mean $(n-2k)p + k$ and variance $2kp(1-p)$ where $p = \frac{2km}{n(n-k)}$.*

PROOF: In G each community has n/k nodes, and hence

$$p = \frac{2m}{k\frac{n}{k}(\frac{n}{k}-1)} = \frac{2km}{n(n-k)}.$$

Let λ_i be the largest eigenvalue of the ith community ($i = 1, 2, \ldots, k$), then $R_G = \sum_{i=1}^{k} \lambda_i$. Since λ_i has the asymptotical normal distribution with mean $(\frac{n}{k} - 2)p + 1$ and variance $2p(1-p)$ [8], then R_G also has the asymptotical normal distribution with mean and variance as in the theorem □.

With Theorem 1, we directly have the following result.

RESULT 3. *The relative non-randomness of the overall graph $G(n, m)$ can be calculated as*

$$R_G^* = \frac{R_G - [(n-2k)p + k]}{\sqrt{2kp(1-p)}}, \tag{4.5}$$

where $p = \frac{2km}{n(n-k)}$.

For any two graphs, G_1 and G_2, if $\mid R_{G_1}^* \mid < \mid R_{G_2}^* \mid$, we can conclude that G_1 is more random than G_2. Since the relative non-randomness measure R_G^* of ER graph approximately follows the standard normal distribution with mean 0 and standard variance 1, we can use $1 - \Phi(R_G^*)$ to indicate the similarity between this graph and a random graph, where $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution. Strictly speaking, $1 - \Phi(R_G^*)$ is the probability that how less likely graph G is actually generated by ER model.

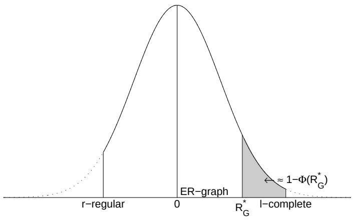


Figure 6: Relative non-randomness measure and its distribution

The relative measure indicates to what extent one real world graph is different from random graphs in terms of probability. As illustrated in Figure 6, when R_G^* is close to 0, the graph G tends to be more likely generated by ER model. From the statistical hypothesis testing point of view, we cannot reject the null hypothesis that G is generated by ER model. On the contrary, when R_G^* is far away from 0, it indicates the graph G is towards extreme ordered graph. We can safely reject the null hypothesis since $1 - \Phi(R_G^*)$ (denoted as the gray region in Figure 6) is significantly small.

Another interesting property illustrated in Figure 6 is that R_G^* of any graph is lower (upper) bounded by that of r-regular (l-complete) graph respectively. For graphs $G(n, m)$ with k communities, we define the r-regular graph as a graph with each node having r neighbors and the l-complete graph here as a graph where each community is a clique of l nodes.

THEOREM 2. *For any graph $G(n, m)$ with k communities, we have*

$$R_{G_{r-regular}}^* \leq R_G^* \leq R_{G_{l-complete}}^*$$

where $R_{G_{r-regular}}^*$ and $R_{G_{l-complete}}^*$ denote the relative non-randomness value of r-regular graph and l-complete graph respectively. Similarly, we have

$$R_{G_{r-regular}} \leq R_G \leq R_{G_{l-complete}}$$

Their expressions are shown in Table 3.

PROOF: See appendix.

Discussion. When it comes to a graph with one community, our graph non-randomness measure R_G is reduced as λ_1 as shown in Equation 4.4. It has been shown in [8] that the largest eigenvalue has asymptotically the normal distribution with mean $(n-2)p + 1$ and variance $2p(1-p)$ when graph G follows ER model with parameter n and p. This

Table 3: Non-randomness measure for different graphs with the same (n, m) and k communities

Graph, $p = \frac{2km}{n(n-k)}$	R_G	R_G^*
ER model	$(n-2k)p+k$	0
r-regular $(m = \frac{krn}{2})$	kr	$-\frac{k}{\sqrt{2kp(1-p)}}$
l-complete $(m = \frac{kl(l-1)}{2})$	$k(l-1)$	$\frac{kl-(n-2k)p-2k}{\sqrt{2kp(1-p)}}$

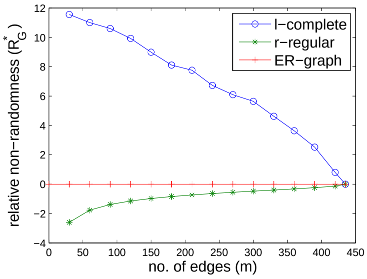


Figure 7: Upper and lower bounds of R_G^* for graphs with $n = 30$, $k = 1$, and varying m

can be considered as a special case of our results shown in Theorem 1.

Theorem 2 shows that r-regular graph and l-complete graph are most non-random graphs among all graphs $G(n, m)$. The relative non-randomness value of r-regular graph reaches the largest negative value while that of l-complete graph reaches the largest positive value. Recall that the expectation of the relative non-randomness value of ER graphs is 0. Figure 7 illustrates how the relative non-randomness values of r-regular graph and l-complete graph vary when the density of graph increases. Note that the number of nodes across all graphs is fixed ($n = 30$). When we increase the number of edges, the range determined by the bounds decreases. In the extreme case of $m = 435$, both relative non-randomness values are zero since the graph is a fully complete graph.

4.4 Comparison with other graph spectra The graph spectrum has been well investigated in the graph analysis field. It has been shown that the eigenvectors of the Laplacian matrix and the normal matrix are good indicators of community clusters [5, 12, 15, 19]. The difference between our non-randomness framework and those traditional spectral clustering methods has two-folds. First, spectral clustering methods aim to minimize the cut between communities while our randomness framework is based on maximizing the densities of communities. Second, in traditional spectral clustering methods, communities are represented by dense clusters in the spectral space of Laplacian or normal matrix while communities in our framework are represented by quasi-orthogonal lines in the spectral space of the adjacency matrix. Our proposed framework can quantify randomness at all edge, node, and overall graph levels using spectra of the adjacency matrix. It is interesting to explore whether similar frameworks can also be derived using spectra of the Laplacian or normal matrix. We will study this issue in our future work.

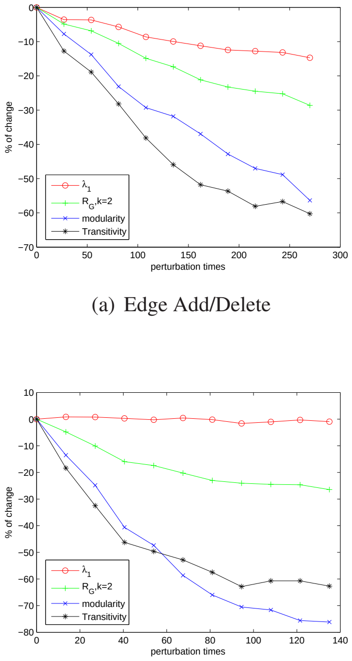
Table 4: Graph non-randomness and characteristics of various social networks

Network	n	m	R_G	R_G^*
synthetic	1000	99820	200	0.02
karate	34	78	11.7	1.22
dolphins	62	159	13.1	1.61
polbooks	105	441	23.5	6.87
Enron	151	869	41.2	4.18
netsci	1589	2742	38.5	128
polblogs	1222	16714	134	187

5 Empirical Evaluations

5.1 Data Sets We used several network data sets in our evaluation. All data sets (except synthetic and Enron data) together with descriptions can be found at `http://www-personal.umich.edu/~mejn/netdata/`. The Enron network was bulit 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 [14] 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 a graph is that there has been at least five email communications between u and v. The synthetic data was generated using ER model with parameters $n = 1000$ and $p = 0.2$.

5.2 Graph non-randomness of various social networks Table 4 shows graph statistics, and graph non-randomness values (calculated using R_G and R_G^*) of various social networks. We can observe that the relative non-randomness measures (R_G^*)) of real world social networks are significantly greater than zero while that of the synthetic random graph is very close to zero. Using R_G^*, we can relatively compare the randomness of graphs with different sizes and densities. For example, we can observe that the network of the dolphins contains less randomness than the karate data since R_G^* of the dolphins (1.61) is greater than that of the karate data (1.22). Furthermore, R_G^* also indicates to what



(a) Edge Add/Delete

(b) Edge Switch

Figure 8: Graph characteristic vs. non-randomness measure for politics book network with various perturbations

extent the graph is different from random graphs. For karate graph, we have $R^*_G = 1.22$ and $1 - \Phi(R^*_G) = 0.11$, which indicates how less likely the karate graph is generated by ER model. Similarly, for dolphins data, we have $R^*_G = 1.61$ and $1 - \Phi(R^*_G) = 0.054$.

We are also concerned with the connection between various real graph characteristics and our graph non-randomness measure (which is derived from graph spectrum). We conducted two types of perturbations on politics book: addition/deletion of randomly chosen edges, and switches of edges. For each perturbed graph, we calculated λ_1, R_G with $k = 2$, and two real graph characteristics (Transitivity and Modularity). The transitivity measure, C, is one type of clustering coefficient measure and characterizes the presence of local loops near a vertex. It is formally defined as $C = \frac{3N_\Delta}{N_3}$ where N_Δ is the number of triangles and N_3 is the number of connected triples. The modularity measure, Q, which indicates the goodness of the community structure [4], 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.

Intuitively, when the magnitude of perturbation increases, we expect the graph tends to lose its structural properties. Figure 8(a) and 8(b) show our empirical evaluations on how Transitivity, Modularity, λ_1, and R_G are changed along perturbations on politics book. One interesting phenomenon is that the relative non-randomness measure always decreases when the magnitude of perturbations increases. We show the strict proof for the perturbation of adding edges in Theorem 3 in Appendix. Furthermore, our graph non-randomness measure R_G can better reflect the change trend indicated by Transitivity and Modularity than λ_1. For example, in Figure 8(b), λ_1 remains almost unchanged even when the graph is significantly perturbed by random switches.

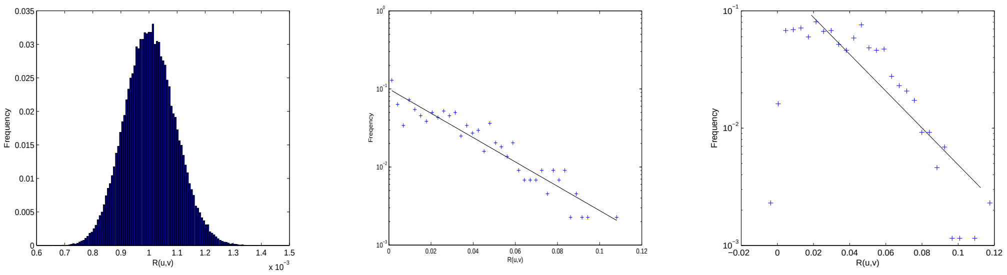
5.3 Distributions of node non-randomness and edge non-randomness

It is well known that the degree distributions in many real-world networks, such as the power-law distribution observed for the Internet and the the Web graph, differ significantly from the Poisson distribution of random graphs [1]. We are interested in the edge non-randomness distribution as well as the node non-randomness distribution in real-world networks and how they are different from synthetic random networks generated by the ER model.

We conducted experiments using three networks: synthetic data generated by the ER model with $n = 1000$ and $p = 0.2$, politics books, and Enron network. Figure 9 (Figure 10) shows distributions of edge (node) non-randomness of these three networks. We can observe from Figure 9(a) and Figure 10(a) that the distributions of both edge non-randomness and node non-randomness follow approximately normal distributions.

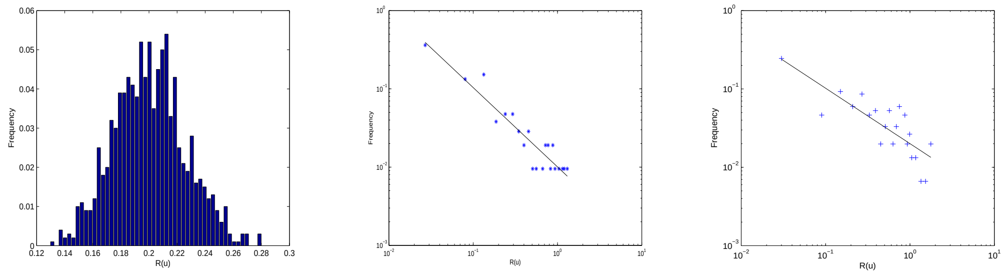
The linear-log plot in Figure 9(b) indicates that edge non-randomness $R(u, v)$ of 441 edges in politics book has a highly skewed form, approximately obeying an exponential law. However, Figure 9(c) shows that edge non-randomness $R(u, v)$ of the majority edges in Enron email network only approximately obeys an exponential law. The log-log plot in Figure 10(b) indicates that node non-randomness $R(u)$ of 105 nodes in politics book clearly follows a power law distribution. However, there is no evidence to display the power law pattern for node non-randomness distribution of 151 nodes in Enron data as shown in Figure 10(c).

The distributions of both edge non-randomness and node non-randomness for real networks are quite different from those for random graphs. We also conducted evaluations on other real-world social networks. Although we cannot reach the conclusion that they definitely follow power law (or exponential law) distributions, our empirical evaluations did show that they are usually highly skewed, with a small number of edges (nodes) having an unusually large non-randomness values and a large number of edges (nodes) having small non-randomness values.



(a) ER synthetic (b) Politics books (c) Enron

Figure 9: Edge non-randomness distributions



(a) ER synthetic (b) Politics books (c) Enron

Figure 10: Node non-randomness distributions

Table 6: Comparison of top k non-random nodes across monthly networks of Enron data

$\mid S_t \mid$	J_t										
	2	3	4	5	6	7	8	9	10	11	12
10	0.67	0.67	0.54	0.67	0.54	0.43	0.43	0.82	0.54	0.54	0.67
20	0.60	0.60	0.54	0.74	0.60	0.33	0.38	0.60	0.54	0.60	0.67
30	0.71	0.58	0.58	0.82	0.62	0.54	0.58	0.71	0.62	0.54	0.76
40	0.63	0.63	0.63	0.82	0.74	0.57	0.57	0.78	0.57	0.57	0.74
50	0.69	0.64	0.67	0.85	0.75	0.56	0.69	0.85	0.69	0.64	0.85

5.4 Evolution of graph non-randomness We are interested in how the graph non-randomness may change for dynamic social networks. We performed the randomness analysis on the monthly email graphs from Enron data. In Table 5, we list graph relative non-randomness values for 12 graphs constructed from Enron dataset from June 2001 to May 2002. Each graph G_t is formed by the total email data in months from 1 to t. We regard there's an edge between node u and v in G_t when there is at least three communications between u and v during this period. We use m_t to denote the number of edges of G_t. We can easily observe that $G_{t-1} \subseteq G_t$ and $m_{t-1} < m_t$. We use $R^*_{G_t}$ ($k = 3$) to denote the relative non-randomness of G_t. We can observe that for most real Enron data sets, $R^*_{G_t} < R^*_{G_{t-1}}$ (except G_3), showing that the relative non-randomness of the graph decreases along the time.

One interesting question here is how those newly added edges in each month are different from randomly added edges. To answer this question, we construct synthetic data sets H_t by randomly adding $m_t - m_{t-1}$ edges to G_{t-1}. We can see in Table 5 that $R^*_{H_t}$ is always less than $R^*_{G_{t-1}}$ since the randomly added edges increase the graph non-randomness. Specifically, the 39 newly added edges in the real graph G_7 decreases the relative non-randomness by 0.10. However, when we randomly add 39 edges to G_6, the relative non-randomness of H_7 decreases by 1.40. This difference indicates those 39 newly added edges in G_7 are significantly different from randomly added edges. On the contrary, 40 newly added edges in G_8 are not significantly different from randomly added edges.

Since the number of nodes are unchanged across all monthly graphs, we are also interested in how the subset of

Table 5: Enron dynamic relative non-randomness ($k = 3$)

	m_t	$R^*_{G_t}$	$R^*_{H_t}$	$R^*_{G_t} - R^*_{G_{t-1}}$	$R^*_{H_t} - R^*_{G_{t-1}}$
G_1	87	12.99	–	–	–
G_2	187	12.41	4.84	-0.58	-8.16
G_3	327	12.45	4.07	0.04	-8.35
G_4	429	9.81	7.14	-2.63	-5.31
G_5	627	7.89	2.01	-1.92	-7.81
G_6	726	7.22	4.29	-0.67	-3.60
G_7	765	7.12	5.82	-0.10	-1.40
G_8	805	5.91	5.74	-1.20	-1.38
G_9	826	5.69	5.22	-0.22	-0.70
G_{10}	851	5.19	4.85	-0.51	-0.84
G_{11}	879	4.88	4.27	-0.31	-0.92
G_{12}	922	4.36	3.56	-0.52	-1.32

individuals identified as top non-random nodes (e.g., top 30) is varied dynamically. We used Jaccard's index measuring the similarity between two subsets. Formally, we define

$$J_t = \frac{\mid S_{t-1} \cap S_t \mid}{\mid S_{t-1} \cup S_t \mid}$$

where S_t denotes the subset of non-randomness nodes from data G_t. We can observe from Table 6 that those non-random nodes do change along the time. Hence, our node non-randomness measure $R(u)$ can be applied in practice to monitor the change of individual's roles in terms of its randomness in the social network.

6 Conclusion and Future Work

The focus of our paper was to present a formal framework characterizing graph non-randomness at various levels. We first proposed a novel measure to characterize the edge non-randomness using spectral coordinates of two connected nodes projected in the k-dimensional spectral space. We then characterized the node non-randomness based on the non-randomness of edges connected to this node and the graph non-randomness based on the non-randomness of all edges within the graph. All non-randomness measures are simple numerical indices which can be derived elegantly from graph spectrum.

We studied several real-world social networks for which our non-randomness measures display useful and desirable properties. We also show that non-randomness measures have different distributions between real-world networks and random graphs. Empirical evaluations on dynamic social networks demonstrated the utility of these measures on monitoring the evolution of graph non-randomness change as well as those non-random (important) nodes.

The major goal of this article has been the development of graph non-randomness measures. We have shown that the graph non-randomness measure is determined by the sum of the first k eigenvalues, where k indicates the number of communities in the graph. We are interested in how different choices of k affect the graph non-randomness. Second, we only validated our measures on small social networks which contain only thousands of nodes. We will consider the computational complexity issues when extremely large networks are available. Finally, we have not investigated in full the relationship between our proposed non-randomness measures (especially the graph non-randomness) with traditional measures. Traditional measures such as modularity can also be used to measure the community structural property. It is interesting to compare our measures with those traditional ones and explore how to apply the proposed non-randomness framework to solve practical problems such as graph partition.

Acknowledgments

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

References

[1] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1):2, 2006.

[2] F. Chung and R. Graham. Sparse quasi-random graphs. *Combinatorica*, 22 (2):217–244, 2002.

[3] D. Cvetkovic and P. Rowlinson. The largest eigenvalue of a graph: A survey. *Linear and multilinear algebra*, 28:3–33, 1990.

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

[5] C. Ding, X. He, H. Zha, M. Gu, and H. Simon. A min-max cut algorithm for graph partitioning and data clustering. In *ICDM*, pages 107–114, 2001.

[6] P. Erdos and A. Renyi. On random graphs i. *Publicationes Mathematicae*, 6:290–297, 1959.

[7] I. Farkas, I. Derenyi, A. Barabasi, and T. Vicsek. Spectra of "real-world" graphs: Beyond the semi-circle law. *Physical Review E*, 64:1, 2001.

[8] Z. Furedi and J. Komlos. The eigenvalues of random symmetric matrices. *Combinatorica*, 1 (3):233?41, 1981.

[9] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[10] V. Krebs. http://www.orgnet.com/. 2006.

[11] M. Newman. The structure and function of complex networks. *SIAM Review*, 45:167, 2003.

[12] M. Newman. Detecting community structure in networks. *The European Physical Journal B - Condensed Matter*, 38(2):321–330, March 2004.

[13] 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.

[14] 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.

[15] J. Shi and J. Malik. Normalized cuts and image segmentation. In *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, page 731, Washington, DC, USA, 1997. IEEE Computer Society.

[16] G. Stewart and J. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

[17] S. Strogatz. Exploring complex networks. *Nature*, 410:268–276, 2001.

[18] 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.

[19] Y. Weiss. Segmentation using eigenvectors: a unifying view. In *IEEE International Conference on Computer Vision*, pages 975–982, 1999.

A Proofs

A.1 Proof of Theorem 2 We first prove the case $k = 1$. When $k = 1$, $R_G = \lambda_1$. Let $d_{\min}$, $d_{\max}$ and $\bar{d}$ be the minimum, maximum, and the average degree. We have the following two inequalities [3]:

$$d_{\min} \leq \bar{d} = \frac{2m}{n} \leq \lambda_1 \leq d_{\max} \tag{1.6}$$

$$\lambda_1 \leq \sqrt{2m - n - 1} \tag{1.7}$$

Assume that $m = rn/2$ for some integer r, then we can construct a r-regular graph with m edges. In r-regular graph $\bar{d} = \frac{2m}{n} = d_{\max} = r$, with inequality (1.6), we have $R_G = \lambda_1 = r$. Since for any graph with the same parameters, we have $\lambda_1 \geq \frac{2m}{n}$. Hence the r-regular graph has the smallest non-randomness value.

The relative non-randomness measure is

$$R^*_{G_{r-regular}} = \frac{r - (n-2)p - 1}{\sqrt{2p(1-p)}}, \tag{1.8}$$

where $p = \frac{r}{n-1}$ for r-regular graph. When n is large, we can further simplify Eq.(1.8) as:

$$R^*_{G_{r-regular}} = -\frac{1}{\sqrt{2p(1-p)}}.$$

Assume that $m = \frac{l(l-1)}{2}$ for some integer l, then we can construct complete graph with node $1, 2, \ldots, l$, leaving the rest nodes isolated. Then, $R_{G_{l-complete}} = l - 1$. Since any graph with the same parameters must involve no less than l non-isolated nodes, and with inequality (1.7), we have

$$\lambda_1 \leq \sqrt{2m - l - 1} = l - 1.$$

Hence the l-complete graph reaches the upper bound. Its relative non-randomness is straightforwardly derived from the definition.

When $k > 1$, it is easy to verify that the minimum and maximum are reached when the graph has k equal-sized r-regular graphs or l-complete graphs. We have the theorem proved. □

A.2 Theorem 3

THEOREM 3. *For graph $G(n, m)$ with k communities and $p = \frac{2m}{n(n-1)} < \frac{1}{2}$ and graph G' obtained by randomly adding edges to G based on ER model with parameter n and Δp satisfying $\Delta p < p$ and $p + \Delta p < \frac{1}{2}$. Assume k communities will not merge. If $R_G - [(n-2k)p + k] \in O(pn)$, then*

$$E(R^*_{G'}) < R^*_G.$$

PROOF: Let A and $\widetilde{A}$ be the adjacency matrix of G and G' respectively, $E = \widetilde{A} - A$. Let λ_i, $\tilde{\lambda}_i$ and ϵ_i be the i-th largest eigenvalue of A, $\widetilde{A}$ and E respectively. With Theorem IV-4.8 in [16], we have

$$R_{G'} = \sum_{i=1}^{k} \tilde{\lambda}_i \leq \sum_{i=1}^{k} \lambda_i + \sum_{i=1}^{k} \epsilon_i = R_G + \sum_{i=1}^{k} \epsilon_i,$$

and hence

$$E(R_{G'}) \leq R_G + E\left(\sum_{i=1}^{k} \epsilon_i\right). \tag{1.9}$$

We know that $E(\epsilon_1) = (n-2)\Delta p + 1$ and with the Semicircle Law [7], we have

$$E(\epsilon_i) \leq 2\sqrt{n\Delta p(1-\Delta p)}, i = 2, 3, \ldots, k. \tag{1.10}$$

Combining Eq.(4.5), (1.9) and (1.10), we have

$$\begin{aligned} E(R^*_{G'}) &= \frac{E(R^*_{G'}) - [(n-2k)(p+\Delta p) + k]}{\sqrt{2k(p+\Delta p)(1-p-\Delta p)}} \\ &\leq \frac{R_G + (n-2)\Delta p + 1 + 2(k-1)\sqrt{n\Delta p(1-\Delta p)}}{\sqrt{2k(p+\Delta p)(1-p-\Delta p)}} \\ &\quad - \frac{(n-2k)(p+\Delta p) + k}{\sqrt{2k(p+\Delta p)(1-p-\Delta p)}} \\ &\leq \frac{R_G - [(n+2k)p + k] + M}{\sqrt{2k(p+\Delta p)(1-p-\Delta p)}} \\ &\quad (\text{let } M = 2(k-1)\left[\Delta p + \sqrt{n\Delta p(1-\Delta p)}\right] + 1) \end{aligned}$$

Hence, to prove $E(R^*_{G'}) < R^*_G$, we need only to show

$$\frac{R_G - [(n+2k)p + k] + M}{R_G - [(n-2k)p + k]} < \frac{\sqrt{(p+\Delta p)(1-p-\Delta p)}}{\sqrt{p(1-p)}}. \tag{1.11}$$

Since $R_G - [(n+2k)p+k] \in O(pn)$ while $M \in O(\sqrt{\Delta p n})$, when n is large, the left-hand side of inequality (1.11) is close to 1. Notice that $p < p + \Delta p < \frac{1}{2}$, the right-hand side of inequality (1.11) is greater than 1 regardless of n, and we have the theorem proved. □