DETECTING HIDDEN CLIQUES FROM NOISY OBSERVATIONS

Yang Liu, Mingyan Liu

Emails: {youngliu, mingyan}@umich.edu.
Electrical Engineering and Computer Science, University of Michigan, Ann Arbor.

ABSTRACT

In this paper we present a methodology to uncover hidden cliques/communities among a set of nodes when observations of their relationships or connectivities are noisy. Existing literature in community detection typically starts with the assumption that the statistical properties of community structure is known a priori, as well as the number of communities, so the task at hand is solely to partition the set into the given number of groups. In practice neither assumption is necessarily true. Motivated by this, we set out to determine a detectability condition (from spectral analysis) prior to performing the partitioning task, and further illustrate how to combine this detectability condition with clustering algorithms to arrive at desirable partitions without a priori information on the clique structure. We validate our results via simulation and make comparison with existing heuristics to demonstrate its advantages.

Index Terms— Community detection, spectral analysis, random matrix

1. INTRODUCTION

In this paper we consider the problem of uncovering clique structure (communities) among a set of nodes or agents when no prior knowledge on the clique structure or the number of cliques is available and observations on the relationship between nodes are noisy. This problem is motivated by clique detection, which has been attracting attentions recently under various contexts. For example, consider malicious campaigns of online crime (e.g., spamming and phishing), which are believed to be highly coordinated and follow certain patterns. To be able to identify such structures would be extremely helpful in designing defensive mechanisms and building more resilient networks.

There is a rich literature in clique detection on graphs, see e.g., [3, 4, 10]. However, most of the current solutions (e.g., clustering algorithms) start with the assumption that the community structure is known a priori, including the number of communities, so the task at hand is solely to develop methodologies to extract the communities from the graph, see e.g., different heuristics developed to target optimal or near-optimal solutions using different measures [1, 3], and to improve the performance of clustering or partitioning algorithms [9]. A prime example is the K-means algorithm, which assumes a priori the knowledge of the right choice of K, the number of clusters to be identified.

On the other hand, in practice we do not always know the clique structures, including the average strength of connection among members of a clique vs. the strength of connection between them and members outside the clique. Nor do we necessarily know how many cliques there are. Moreover our observation or data on the above structural properties are often noisy. For instance, a friend may choose not to communicate with another friend on a given day in a social network. In the case of online malicious campaigns, the external observations are largely incomplete with the clique structure mostly hidden. Motivated by the above, in this paper we seek to address these challenges in the following two questions: 1) How to determine that cliques exist on a graph and what is the detectability condition prior to performing the partition to extract them? 2) Given the detectability condition how to partition a system with no prior information on the clique structure?

For the first step, through spectral analysis we derive a hard constraint on the detectability of clique structures under noises. We show our results are a generalization of a previous work under much simpler settings. Then based on the detectability conditions, we propose a metric with which the cliques can be accurately identified. We validate our results via simulation and make comparison with existing heuristics to demonstrate its advantages. Note that even though we have used the term *detectability*, our focus in this paper is entirely on partitioning. In this sense detectability refers to our ability to partition the set of nodes into distinct cliques.

The remainder of the paper is organized as follows. We formulate the problem in Section 2. Spectral analysis and main results are provided in Section 3. We verify our results with simulation in Section 4. Section 5 concludes our paper.

2. PROBLEM FORMULATION AND PRELIMINARIES

The notion of clique (or group, cluster) generally reflects the following phenomenon: 1) Members within a clique share similar behavior, 2) A member shares higher similarity with others in the same clique than with whose outside the clique (or

The work is partially supported by the NSF under grant CNS-1422211, and by the Department of Homeland Security (DHS) Science and Technology Directorate, Cyber Security Division (DHS S&T/HSARPA/CSD), BAA 11-02 via contract number HSHQDC-13-C-B0015.

those in a different clique). More precisely, consider a system of N nodes (or users, individual, members, to be used interchangeably) interconnected through a graph $\mathcal{G}$. Each node corresponds to a vertex $v_i, i = 1, ..., N$ on this graph. Graph $\mathcal{G}$ is characterized by an adjacency matrix $W = [W_{i,j}]_{i,j=1,2,...,N}$, with edge weights $W_{i,j} \in [0, 1]$ between any two vertices modeling the strength of the connectivity (or similarities) which further captures the probability of generating edges among the users. There are k cliques in the system, denoted by $c_i, i = 1, 2, ..., k$, that arose from the following type of adjacency matrices:

$$W_{i,j} = \begin{cases} w_k^{in}, & i, j \in c_k. \\ \frac{w_i^{out}+w_j^{out}}{2}, & i \in c_i, j \in c_j,\ \forall c_i \neq c_j. \end{cases} \quad (1)$$

Here w_i^{in} and w_i^{out}, $i = 1, 2, \cdots, k$ are constants denoting the in and out-edge weights of clique i, and it is assumed that $w_i^{in} > w_i^{out}, \forall i = 1, 2, ..., k$, reflecting the property 2) of a clique we defined earlier. Denote by $p_i^{in} = \frac{|c_i|}{N}, p_i^{out} = \frac{|\bar{c}_i|}{N}$, i.e., p_i^{in}, p_i^{out} are the fractions of vertices being in and out of community c_i. Naturally we have $p_i^{in} + p_i^{out} = 1, i = 1, ..., k$. First we establish the following proposition (algebraic details omitted).

Proposition 1. *Adjacency matrix W of k-community cliques defined above can be re-written into the following format,*

$$W = \underbrace{\sum_{i=1}^{k} (p_i^{in} w_i^{in} + p_i^{out} w_i^{out}) \cdot (\alpha_i^* \cdot \mathbf{e_i} \cdot \mathbf{e_i^T} + \beta_i^* \cdot \mathbf{1} \cdot \mathbf{1^T})}_{A}$$

$$+ \sum_{i=1}^{k} (p_i^{out} w_i^{in} - p_i^{out} w_i^{out}) \cdot (\alpha_i^+ \cdot \mathbf{u_i} \cdot \mathbf{u_i^T} + \beta_i^+ \cdot \mathbf{1} \cdot \mathbf{1^T}) ,$$

with $\alpha_i^*, \beta_i^*, \alpha_i^+, \beta_i^+$ being constants, $\mathbf{e_i}, \mathbf{u_i}$ being constant vectors and $\mathbf{1}$ the all one vector, and

$$A_{i,j} = \begin{cases} \bar{w}_k, & i, j \in c_k. \\ \frac{\bar{w}_p+\bar{w}_q}{2}, & i \in c_p, j \in c_q. \end{cases}$$

where $\bar{w}_k = p_k^{in} w_k^{in} + p_k^{out} w_k^{out}$.

$\mathbf{e_i}, \mathbf{u_i}$ do not convey any explicit meaning; instead later we will show it is certain equation of them that does matter which does not depend on any entries of them. Often times we do not get to observe directly the interactions between individuals, but may obtain observations when they are on their own; for example when edges on the graph are generated randomly accordingly to the edge weights $W_{i,j}$. We will subsequently model the observed matrix as a combination of W and a random component: $\tilde{W} = W + R$, where $\tilde{W}$ denotes the measured or inferred connectivity matrix and R a random matrix modeling the deviation between the observation process and the average connectivity pattern. Moreover R is a symmetric random matrix with independent elements with zero mean. A typical approach in partitioning networks into cliques is to directly apply spectral analysis to $\tilde{W}$, assuming the number of cliques in the partition is known. A potential issue is that in practice cliques may or may not actually exist which raises the issue of how reliable these spectral methods are and how to evaluate them. This problem was highlighted in a recent work [7] which defined the notion of detectability and showed that there exist conditions under which no spectral method will return meaningful results. In the next section we will extend this result to more general cases and derive the detectability condition for partitioning the set into an arbitrary number of possibly heterogeneous cliques.

3. SPECTRAL ANALYSIS OVER HETEROGENEOUS K-CLIQUES STRUCTURE

3.1. Spectral analysis

We start with spectral analysis of $\tilde{W}$. Based on Proposition 1 we further extract A from $\tilde{W}$ we have the modularity matrix for $\tilde{W}$ defined as follows,

$$\mathcal{M} = \sum_{i=1}^{k} C_i \cdot N_i \cdot \left[\alpha_i^+ \cdot \frac{\mathbf{u_i} \cdot \mathbf{u_i^T}}{N_i} + \beta_i^+ \cdot \frac{\mathbf{1} \cdot \mathbf{1^T}}{N_i} \right] + R ,$$

where $C_i := p_i^{out} w_i^{in} - p_i^{out} w_i^{out}$. Note that the extracted term is a square matrix with entries being the average total degree of our random graph which relates closely to the modularity matrix as introduced in [8] and [7]. Our derivation above generalizes the work in [7] where a simpler case with two homogeneous cliques is analyzed; it is easy to verify that our results reduce to that in [7] with corresponding parameters. Make the following substitution of parameters or normalization

$$\mathbf{u_i} := \frac{\mathbf{u_i}}{\sqrt{N_i}}, \quad \mathbf{v_i} := \frac{\mathbf{1}}{\sqrt{N_i}}, \quad i = 1, ..., k. \quad (2)$$

$$\mathcal{M} = \sum_{i=1}^{k} C_i N_i \cdot \left[\alpha_i^+ \cdot \mathbf{u_i} \cdot \mathbf{u_i^T} + \beta_i^+ \cdot \mathbf{v_i} \cdot \mathbf{v_i^T} \right] + R . \quad (3)$$

Based on above reformulation, we have the following.

Lemma 2. *For above random graph we have,*

$$\frac{\sum_{i=1}^{k} \frac{1}{C^i \cdot N_i}}{\sum_{i=1}^{k} \frac{1}{N_i}} \leq \text{Tr}(z\mathbf{I} - R)^{-1} , \quad (4)$$

where Tr$(\cdot)$ *is the trace function and z is eigenvalue of $\mathcal{M}$.*

Proof. Let $z, \mathbf{b}$ be $\mathcal{M}$'s eigen-value and vector we have

$$\left\{ \sum_{i=1}^{k} C_i N_i \cdot \left[\alpha_i^+ \cdot \mathbf{u_i} \cdot \mathbf{u_i^T} + \beta_i^+ \cdot \mathbf{v_i} \cdot \mathbf{v_i^T} \right] + R \right\} \mathbf{b} = z\mathbf{b} .$$

Make the following change of notations

$$\mathbf{U} = [\sqrt{\alpha_1^+} \mathbf{u_1}, ..., \sqrt{\alpha_k^+} \mathbf{u_k}], \ \mathbf{V} = [\sqrt{\beta_1^+} \mathbf{v_1}, ..., \sqrt{\beta_k^+} \mathbf{v_k}] ,$$

and denote $\mathcal{W}$ as the diagonal matrix with $\mathcal{W}_{i,i} = C_i N_i$, and re-arrange we have,

$$(z\mathbf{I} - R)\mathbf{b} = [\mathbf{U}, \mathbf{V}] \cdot \mathcal{W} \cdot [\mathbf{U}, \mathbf{V}]^{\mathbf{T}} \cdot \mathbf{b} . \quad (5)$$

Multiplying by $[\mathbf{U},\mathbf{V}]^{\mathbf{T}}(z\mathbf{I}-R))^{-1}$ on both sizes we have

$$\mathcal{W}^{-1}=[\mathbf{U},\mathbf{V}]^{\mathbf{T}}\cdot(z\mathbf{I}-R)^{-1}\cdot[\mathbf{U},\mathbf{V}]\,. \tag{6}$$

Matching the terms on the diagonal of the matrices on both sides we have $\forall i=1,...,k$

$$\frac{1}{C^i\cdot N_i}=\sum_{j=1}^{n}\frac{\alpha_i^+\cdot(\mathbf{u_i^T x_j})^{\mathbf{2}}+\beta_i^+\cdot(\mathbf{v_i^T x_j})^2}{z-\lambda_j}\,. \tag{7}$$

Since R is random, so are its eigenvectors x_is. Taking expectation of the above equation, since entries of x_is are mutually independent, the cross terms will cancel out and we have,

$$E\Big[(\sqrt{\alpha_\mathbf{i}^+}\cdot\mathbf{u_i^T x_j})^2+(\sqrt{\beta_\mathbf{i}^+}\cdot\mathbf{v_i^T x_j})^2\Big]=\frac{\sum_{q\in c_i}E|x_j(q)|^2}{N_i}\,,$$

where $x_j(q)$ denote the qth element of $\mathbf{x_j}$. Adding up we have

$$\begin{aligned}&E\Big\{\sum_{i=1}^{k}\Big[(\sqrt{\alpha_\mathbf{i}^+}\cdot\mathbf{u_i^T x_j})^2+(\sqrt{\beta_\mathbf{i}^+}\cdot\mathbf{v_i^T x_j})^2\Big]\Big\}\\&=\sum_{i=1}^{k}\frac{\sum_{q\in c_i}E|x_j(q)|^2}{N_i}\le(\sum_i\sum_{q\in c_i}E|x_j(q)|^2)\cdot(\sum_i\frac{1}{N_i})\\&=E|\mathbf{x_j}|^2\cdot\sum_{i=1}^{k}\frac{1}{N_i}=\sum_{i=1}^{k}\frac{1}{N_i}\,,\end{aligned} \tag{8}$$

here we have used the fact for random matrix R its eigenvectors satisfy $E|\mathbf{x_j}|^2=1$ (see [6] for details), which gives us,

$$\sum_{i=1}^{k}\frac{1}{C^i\cdot N_i}\le\sum_{i=1}^{k}\frac{1}{N_i}\cdot\sum_{j=1}^{n}\frac{1}{z-\lambda_j}=\sum_{i=1}^{k}\frac{1}{N_i}\cdot\mathrm{Tr}(z\mathbf{1}-R)^{-1}\,.$$

Or in another format $\frac{\sum_{i=1}^{k}\frac{1}{C^i\cdot N_i}}{\sum_{i=1}^{k}\frac{1}{N_i}}\le\mathrm{Tr}(z\mathbf{1}-R)^{-1}$. □

Denote LHS of Eqn. (4) as $\bar{C}_{-1}$ and define $\bar{d}$ as the average degree given by

$$\bar{d}=\sum_{i=1}^{k}\frac{|c_i|}{N}\cdot(p_i^{in}w_i^{in}+p_i^{out}w_i^{out})\,. \tag{9}$$

We now present the detectability results.

Theorem 3. *Clique structures can be detected via spectral algorithm if the following condition is met*

$$\bar{C}_{-1}\le\sqrt{N/\bar{d}}\,. \tag{10}$$

Proof. From spectral theory and [6] we know,

$$\mathrm{Tr}(z\mathbf{1}-R)^{-1}=\frac{1}{z}\sum_{q=0}^{\infty}\frac{\mathrm{Tr}R^q}{z^q}\,,\ \mathrm{Tr}R^{2q}=N^{q+1}\cdot\bar{d}^q\cdot\mathrm{Cat}_q\,,$$

where Cat_q is the Catalan number. Therefore we further get

$$\mathrm{Tr}(z\mathbf{1}-R)^{-1}=\frac{1}{2\cdot\bar{d}}\Big[z-\sqrt{z^2-4N\cdot\bar{d}}\Big]\,, \tag{11}$$

from which we get $\frac{z-\sqrt{z^2-4N\cdot\bar{d}}}{2\cdot\bar{d}}\ge\bar{C}_{-1}$. Consider first the LHS we have

$$\frac{z-\sqrt{z^2-4N\cdot\bar{d}}}{2\cdot\bar{d}}=\frac{4N\cdot\bar{d}}{2\cdot\bar{d}\cdot(z+\sqrt{z^2-4N\cdot\bar{d}})}\,, \tag{12}$$

which is a strictly decreasing function of z for $z\ge0$. And moreover we have the leading eigenvalue of $\mathcal{M}$ as $z_1\le\bar{d}\cdot\bar{C}_{-1}+\frac{N}{\bar{C}_{-1}}$. From $z^2-4N\cdot\bar{d}\ge0$ we have a necessary condition for spectral detectability as $z_1\ge\sqrt{4N\cdot\bar{d}}$. We thus have the transition of detectability happening at $\bar{d}\cdot\bar{C}_{-1}+\frac{N}{\bar{C}_{-1}}=\sqrt{4N\cdot\bar{d}}$ which further gives us the detectability constraint $\bar{C}_{-1}\le\sqrt{N/\bar{d}}$. □

Notice $\bar{C}_{-1}$ can be viewed as the inverse of harmonic mean of C_i modulated by the community size N_i. To serve as a partial validation of our results, we consider a special case when there are only two homogeneous cliques as introduced in [6] : $C_1=C_2=C$ and $N_1=N_2=\frac{N}{2}$ and it follows that the inequality in Eqn. (8) holds tightly. Furthermore LHS is $\frac{1}{C}$ which is exactly the results reported in [6].

3.2. Clique identification

We show how to use this condition in the actual partitioning task, by putting it in the context of a given clustering algorithm. We take spectral K-means as the base-line algorithm and show how to combine spectral K-means with the detectability results. The spectral K-means algorithm is introduced in [9] and we refer interested reader to the paper for details.

With an input parameter $\hat{k}$ (number of communities), we execute the spectral K-means clustering and record all results. Upon completion, denote the clustering results as $\hat{c}_i, i=1,2,...,\hat{k}$ with each $\hat{c}_i$ denoting the set of nodes classified to cluster i. We then use these results to estimate the parameters for the graph (e.g., $\mathbf{w^{in}},\mathbf{w^{out}}$) as follows : $\tilde{p}_i^{in}=\frac{|\hat{c}_i|}{N}$, $\tilde{p}_i^{out}=1-\tilde{p}_i^{in}$, $\tilde{w}_i^{in}=\frac{\sum_{j,q\in c_i}\tilde{W}_{j,q}}{|(j,q):j,q\in c_i|}$, $\tilde{w}_i^{out}=\frac{\sum_{j\in c_i,q\notin c_i}\tilde{W}_{j,q}}{|(j,q):j\in c_i,q\notin c_i|}$. We name the set of parameters $\{\tilde{p}_i^{in},\ \tilde{p}_i^{out},\ \tilde{w}_i^{in},\tilde{w}_i^{out},\hat{c}_i\}_{i=1}^{k}$ as the empirically estimated community structure. Denote the tuple by $\mathcal{C}_{\hat{k}}$. We then proceed to calculate parameters $\bar{d}(\mathcal{C}_{\hat{k}})$ and $\bar{C}_{-1}(\mathcal{C}_{\hat{k}})$ for $\mathcal{C}_{\hat{k}}$ and check whether the detectability condition holds: $\bar{C}_{-1}(\mathcal{C}_{\hat{k}})\le\sqrt{N/\bar{d}(\mathcal{C}_{\hat{k}})}$. There are three possibilities as a result of the detectability check: 1) There is no $\hat{k}$ such that $\mathcal{C}_{\hat{k}}$ passes the detectability check. This case essentially implies there is no clique structure detectable from the data. 2) There is a unique $\hat{k}$ that satisfies Eqn.(10), in which case we have identified the only possible hidden clique structure. 3) The trickier case is when there exists multiple $\hat{k}$ such that the detectability check passes[1]. We address this issue in the rest of this section.

[1] In general the K-means clustering algorithm is run repeatedly to identify the best K regardless of the criteria used for selection. This is the case with all algorithms we compare in the next section.

We propose the following optimization problem towards resolving the issue,

$$\max_{\hat{k}} \quad |\bar{C}_{-1} - \sqrt{N/\bar{d}}|, \quad \text{s.t.} \quad \bar{C}_{-1} \leq \sqrt{N/\bar{d}}\,.$$

It is straightforward to see that in the above we have used the detectability gap to capture the clustering performance. The intuition is as follows. $\bar{C}_{-1}$ is the inverse of the harmonic mean of $w_i^{in} - w_i^{out}, i = 1, 2, ..., k$. Therefore the larger the difference between in- and out-weights is , the smaller the $\bar{C}_{-1}$ is and so the easier detecting community structures is. Consider the second term. From spectral theory, we have the following upper bounds on the difference between vector spaces (the norm difference). $d(V_W - V_{\hat{W}}) \leq \frac{||R||_F}{\delta}$, with δ defined as $\delta = \lambda_1(W) - \lambda_2(W)$ (the difference between the largest and second largest eigenvalues). Notice

$$||R||_F = \text{Tr}(R^T R) = \text{Tr}(R^2) = N^2 \cdot \bar{d} \cdot \text{Cat}_1\,, \qquad (13)$$

i.e., with N and Cat_1 being constant and to minimize the distance between two spectral spaces, it is equivalent with minimizing the average degree term $\bar{d}$. Therefore our proposed metric $|\bar{C}_{-1} - \sqrt{N/\bar{d}}|$ is in this sense a combination of above trade-offs.

4. NUMERICAL RESULTS

4.1. Validation

We simulate a network with N nodes, where each node is randomly associated with one of k cliques. In- and Out-statistics of cliques are randomly generated but with the expected In-edge weight being strictly larger than the Out-weights. For each simulation run, an edge is randomly generated for each pair of nodes based on their edge weights (higher weights incurs a higher probability of connection).

$N\backslash k$	2	3	4	5	6
100	56%	57%	46%	55%	85%
200	60%	57%	64%	0.50	58%
500	78%	58%	44%	61%	51%
1000	66%	62%	64%	52%	50%

(a) Detection rate

$N\backslash k$	2	3	4	5	6
100	91%	88%	78%	100%	100%
200	84%	96%	88%	90%	100%
500	100%	94%	86%	96%	94%
1000	92%	92%	88%	90%	92%

(b) Detection rate (within ± 1 neighborhood)

Table 1: Simulation results

We repeat the above experiments and record our algorithm's performance measured by the rate of correctly determining the number of cliques. The results are summarized in Table 1. From the simulation results we see with a reasonable confidence (with majority of the entries in Table 1 being larger than 50%) we can accurately identify the hidden community structure while with very high probability (most being larger than 90% and lower bounded by 80%) our detection results is no further away from the true quantity by 1.

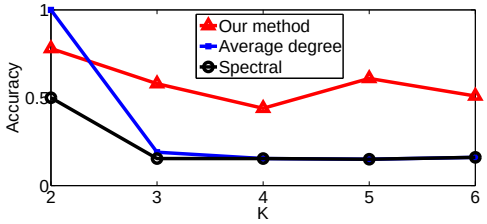
4.2. Performance comparison

We next compare our results with the following two heuristics that have been widely adopted for deciding the number of communities k in algorithms, for example spectral K-means.

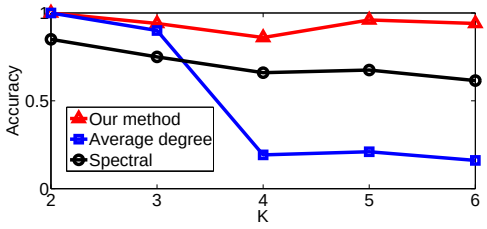
1. Average Degree : The optimal k^* is determined as the number that gives the maximum average difference between In- and Out-degrees [5].

2. Spectral Method : The selection criteria is decided by eigenvalues. Denote all eigenvalues for $\tilde{W}$ as $\lambda_1, ..., \lambda_N$. Then the number of defined as the number k such that sum of eigenvalues $\sum_{i=1}^{k} \lambda_k$ being large enough [11].

Performance comparisons are shown in Figure (1a)-(1b). From Figure (1a) we see after k becomes larger than 2, the detection rates of the above two heuristics drop quickly while our method remains reasonably accurate. In Figure (1b) we show the performance comparison for detection within ± 1 neighbor (i.e., we count an identification as correct when it falls within ± 1 neighborhood of the correct k). Similarly the performance of Average Degree drops quickly while the other two stays relatively steady. However, our method clearly outperforms the Eigenvalue based method.



(a) Perfect detection



(b) Detection within ± 1 neighborhood

Fig. 1: Performance comparison

5. CONCLUSION

In this work we consider a problem of uncovering clique(s) from noisy observations, where hidden connectivities remain unknown. We derived the spectral detectability conditions for a random (dues to noisy observations) adjacency matrix, based on which we propose a metric/methodology to construct the procedure of identifying hidden cliques with no prior information on its structures' statistics. Simulation results validate our method and show clearly its advantages over existing heuristics.

6. REFERENCES

[1] Danail Bonchev and Gregory A Buck. Quantitative measures of network complexity. In *Complexity in Chemistry, Biology, and Ecology*, pages 191–235. Springer, 2005.

[2] Pin-Yu Chen and A.O. Hero. Node removal vulnerability of the largest component of a network. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 587–590, Dec 2013.

[3] Nan Du, Bin Wu, Xin Pei, Bai Wang, and Liutong Xu. Community detection in large-scale social networks. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007*, WebKDD/SNA-KDD '07, pages 16–25, New York, NY, USA, 2007. ACM.

[4] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[5] Petter Holme, Beom Jun Kim, Chang No Yoon, and Seung Kee Han. Attack vulnerability of complex networks. *Physical Review E*, 65(5):056109, 2002.

[6] Raj Rao Nadakuditi and M. E. J. Newman. Graph spectra and the detectability of community structure in networks. *CoRR*, 2012.

[7] R.R. Nadakuditi. On hard limits of eigen-analysis based planted clique detection. In *SSP, 2012 IEEE*, pages 129–132, 2012.

[8] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[9] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On Spectral Clustering: Analysis and an algorithm. In *NIPS*, pages 849–856, 2001.

[10] Keven S Xu, Mark Kliger, and Alfred Hero. Tracking communities of spammers by evolutionary clustering. *ICML*, 2010.

[11] Kuai Xu, Zhi-Li Zhang, and Supratik Bhattacharyya. Profiling internet backbone traffic: behavior models and applications. In *ACM SIGCOMM Computer Communication Review*, volume 35, pages 169–180. ACM, 2005.