ON HARD LIMITS OF EIGEN-ANALYSIS BASED PLANTED CLIQUE DETECTION

Raj Rao Nadakuditi

Department of EECS, University of Michigan, Ann Arbor, MI 48104.

ABSTRACT

We study the problem of detecting or discovering a planted clique embedded in a random graph. Using recent results from random matrix theory, we demonstrate the presence of a phase transition in eigen-analysis based methods for planted clique detection. The transition separates a regime in which eigen-analysis based methods will successfully detect the planted clique and the associated vertices from one in which the planted clique is present but is undetectable. We validate the prediction with numerical simulations.

Index Terms— planted clique, network, random matrix theory

1. INTRODUCTION

The problem of community detection in networks is an active area of research [1, 2]. Communities are groups of vertices within a network that have a high density of within-group connections but a lower density of between-group connections. The challenge is to find such groups accurately and efficiently in a given network and to identify the fundamental limits of such algorithms. The planted clique problem [3, 4, 5, 6], described next, is a variant of the community detection problem.

A clique is a set of vertices such that every two vertices are connected by an edge. The planted clique problem refers to the problem of discovering the vertices associated with a clique that is 'planted' or embedded in a random graph where every two vertices in the random graph are connected by an edge with probability p. In this paper we focus on eigen-analysis based methods for planted clique discovery, which exploit the properties of matrix representations of networks such as the adjacency matrix or its close relative, the modularity matrix [2].

In recent years, significant effort has been devoted to the development of practical algorithms using these and related methods [3, 7, 5, 6]. There has been less work on formal examination of their fundamental limits and the implications for algorithm performance. Here we give an analysis of eigen-analysis based methods for planted clique detection and in the process uncover a fundamental limit that is of practical importance.

Specifically, we uncover the presence of a sharp transition between a regime in which the eigen-spectrum contains information on the planted clique and a regime in which it does not. In the former regime, reliable planted clique detection is possible and current eigen-analysis based algorithms will perform well; in the latter, any method relying on the spectrum to perform planted clique detection must fail in the limit of large networks.

The paper is organized as follows. Section 2 formulates the planted clique problem and motivates the use of eigen-analysis based techniques. Section 3 presents our main results on the phase transition alluded to while Section 4 outlines a proof based on recent results in the mathematical literature. Simulations for validating our theory are presented in Section 5, followed by some concluding remarks in Section 6.

2. PROBLEM FORMULATION

We are given a graph $G = (V, E)$ with a vertex set V comprising of n vertices and edge set E; a k-clique is a subset of k vertices, $V^* \subset V$, $|V^*| = k$ such that all vertices are connected to each other. The vertices outside the planted clique V^* are randomly connected to each other such that any two vertices are connected by an edge with probability p. The k-clique is thus embedded in an (undirected) $G(n, p)$ Erdos-Renyi graph [**?**, 8]. Throughout this paper, we refer to this as a (k, p, n) planted clique problem.

The adjacency matrix A of an undirected network is the $n \times n$ symmetric matrix with elements $A_{ij} = A_{ji} = 1$ if vertices i and j are connected by an edge and 0 otherwise. The adjacency matrix A for the planted clique problem can be modeled as

$$A_{ij} = A_{ji} = \begin{cases} 1 & \text{if } i, j \in V^* \\ 1 & \text{with prob. } p \text{ if } i, j \notin V^* \\ 0 & \text{with prob. } 1-p \text{ if } i, j \notin V^*. \end{cases}$$

The goal is to find the vertices associated with the planted clique given the adjacency matrix A. A close-relative of the adjacency matrix is the so-called modularity matrix [2] B which is defined as

$$B_{ij} = A_{ij} - P_{ij},$$

Work supported by NSF CCF-1116115. Thanks to Mark Newman for feedback, comments and many insightful suggestions and to Luis Rademacher for suggesting this problem.

where P_{ij} is the expected value of the adjacency matrix in a null model that does not contain the planted clique. Here $P_{ij} = p$ so that $B_{ij} = A_{ij} - p$ or equivalently

$$B = A - p\mathbf{1}\mathbf{1}^{\mathbf{T}},$$

where $\mathbf{1}$ denotes the $n \times 1$ vector containing all ones. Let $\overline{B} = \mathbb{E}[B]$ denote the expected value of the modularity matrix B. Then,

$$\mathbb{E}[\overline{B}_{ij}] = \begin{cases} 1-p & \text{if } i,j \in V^* \\ 0 & \text{otherwise.} \end{cases}$$

Thus $\overline{B}$ is a rank-1 matrix whose only non-zero eigenvalue equals $k(1-p)$. The indices of the non-zero entries of the associated (unit-norm) eigenvector u reveals the vertices of the planted clique. This is the motivation for eigen-analysis based planted clique discovery [7]. Of course, we do not have $\overline{B}$. We now describe an algorithm for planted clique detection or discovery using eigen-analysis.

Note that instead of $\overline{B}$, we have

$$B = A - p\,\mathbf{1}\mathbf{1}^T = \overline{B} + X, \tag{1}$$

where X is the deviation between the modularity matrix B and its average value $\overline{B}$. We note that $\mathbb{E}[X_{ij}] = 0$ and that

$$\text{var}[X_{ij}] = \begin{cases} 0 & \text{if } i,j \in V^* \\ p\,(1-p) & \text{otherwise.} \end{cases} \tag{2}$$

Let v denote the eigenvector of B associated with its largest eigenvalue and let v_i for $i = 1, \ldots, n$ denote its i-th element.

Under the null model that does not contain the planted clique, the largest eigenvector of B is asymptotically distributed as $\sqrt{n}v \sim \mathcal{N}(0, I)$ [9]. For a significance level of α corresponding to the false-alarm probability of misidentifying a vertex as part of the clique, we obtain the following procedure for detecting vertices $\widehat{i}$ in the planted clique

$$\boxed{\widehat{i} = \left\{ i : |\sqrt{n}v_i| > \tau_\alpha := F^{-1}_{\mathcal{N}(0,1)}\left(1 - \frac{\alpha}{2}\right) \right\}.} \tag{3}$$

The main result of this paper, stated next is a precise characterization of the fundamental, asymptotic limit of clique discovery using (3).

3. HARD LIMITS OF CLIQUE DETECTION

Theorem 3.1. *Consider a (k,p,n) planted clique problem where the clique vertices are identified using (3) for a significance level α. Then, for fixed $p \in (0,1)$, and $k, n \to \infty$ such that $k/\sqrt{n} \longrightarrow \beta \in (0,\infty)$ we have*

$$\mathbb{P}(\textit{Clique discovered}) \xrightarrow{p} \begin{cases} 1 & \textit{if } \beta > \beta_{\text{crit.}} := \sqrt{\dfrac{p}{1-p}} \\ \alpha & \textit{otherwise,} \end{cases}$$

where $\xrightarrow{p}$ denotes convergence in probability.

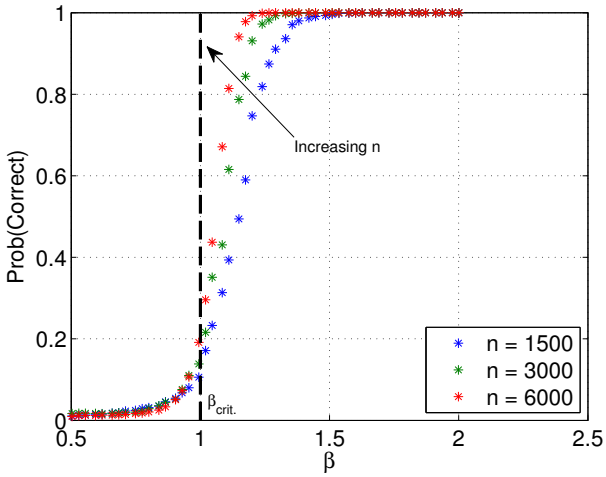


Fig. 1. A plot of the probability that the clique vertices are correctly identified as a function of $\beta = k/\sqrt{n}$ for $n = 1500, 3000, 6000$. Here we set $p = 0.5$ and evaluate the empirical probability over 400 trials. Theorem 3.1 predicts a phase transition in detectability about $\beta_{\text{crit.}} = 1$.

Remark 1. *We conjecture that the above statement holds with probability* 1.

Remark 2. *Note that when $p = 1/2$, $\beta_{\text{crit.}} = 1$. In [4], the authors show that when $k > n^{1/2-\epsilon}$ then reliable clique detection is possible.*

We now characterize the fundamental limit in the setting where the clique is embedded in a sparse random graph with an average degree of c.

Claim 3.2. *For a (k,p,n) planted clique problem, let $p_n = c/n$. Then for sufficiently large $c > 1$, there exists $\beta_{\text{crit.}}$ such that*

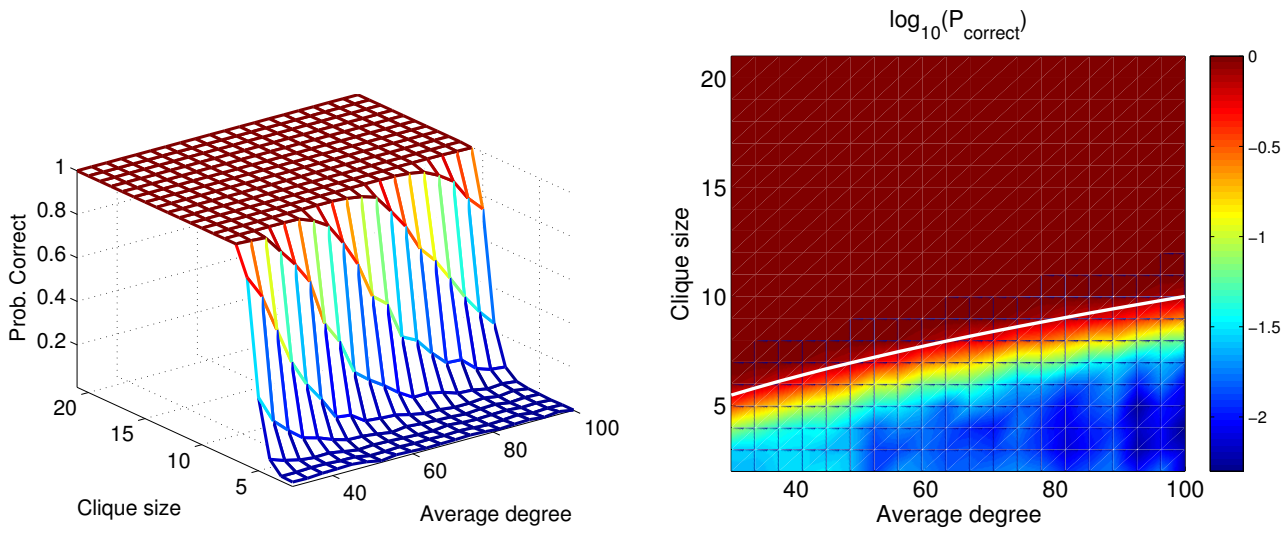
$$\mathbb{P}(\textit{Clique discovered}) \xrightarrow{p} \begin{cases} 1 & \textit{if } k > k_{\text{crit.}} := \sqrt{c} + O\left(\frac{1}{c}\right) \\ \alpha & \textit{otherwise.} \end{cases}$$

Remark 3. *Claim 3.2 follows from plugging in $p_n = c/n$ in Theorem 3.1. The $O(1/c)$ correction term comes from the observations in [10] for sparse Erdos-Renyi graphs.*

We now provide a justification for Theorem 3.1 using recent results from the mathematical literature [11, 12]. We note that similar results can also be derived for multiple disjoint cliques and the planted biclique problem.

4. SKETCH OF THE PROOF

The general form of the matrix B in (1) is that of a rank-1 matrix $\overline{B} = k(1-p)uu^T$ plus a random perturbation matrix X whose entries are independent, zero mean with a 'variance profile' given by (1)).



(a) Phase transition in clique discovery.

(b) Heat map of probability of clique discovery.

Fig. 2. (a) A plot of the probability that the clique vertices are correctly identified as a function of the clique size k and the average degree c of the random graph. This corresponds to the sparse setting described in Claim 3.2; here $n = 100000$, $\alpha = 0.01$ and we used (3) to identify the vertices. (b) A heat map of the logarithm of the probability of correct identification of clique vertices for the same setting as (a). The solid white line corresponds to the phase transition prediction in Claim 3.2 with $k_{\text{crit.}} = \sqrt{c}$.

The matrix X is a so-called Wigner matrix, an ensemble that has been extensively studied in the mathematical literature [13, 14]. In [11, 12], it is shown that for a Wigner matrix with zero mean entries and a variance of $1/n$, there is a phase transition whereby the largest eigenvector v of the rank-1 perturbation of X having form θuu^T exhibits the following behavior in the $n \longrightarrow \infty$ limit

$$|\langle v, u\rangle|^2 \xrightarrow{\text{a.s.}} \begin{cases} \underbrace{1 - \frac{1}{\theta^2}}_{=:\gamma^2} & \text{if } \theta > 1 \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

For finite, but large n we have that

$$\sqrt{n}\, v \sim \begin{cases} \sqrt{n}\,\gamma\, u + \mathcal{N}(0, \sqrt{1-\gamma^2}\, I) & \text{if } \theta > 1 \\ \mathcal{N}(0, I) & \text{otherwise} \end{cases} + o_p(1).$$

Thus for large n and $\theta < 1$, the procedure in (3) will yield

$$\mathbb{P}(\text{Clique vertex identified}) \longrightarrow \alpha.$$

In our setting, the unit-norm eigenvector u has exactly k non-zero entries equal to $1/\sqrt{k} = 1/\sqrt{\beta\sqrt{n}}$, for $k = \beta\sqrt{n}$ as in the hypothesis of Theorem 3.1. Thus for $\theta > 1$ we have

$$\begin{aligned} \mathbb{P}(\text{Clique vertex misidentified}) &= \Phi\left(\frac{\tau_\alpha - \sqrt{n}\gamma\frac{1}{\sqrt{k}}}{\sqrt{1-\gamma^2}}\right) + o_p(1) \\ &= \Phi\left(\frac{\tau_\alpha - n^{1/4}\frac{\gamma}{\sqrt{\beta}}}{\sqrt{1-\gamma^2}}\right) + o_p(1), \end{aligned}$$

where $\Phi(x) = 1/\sqrt{2\pi}\int_{-\infty}^{x}\exp(-x^2/2)dx$ is the CDF of the standard normal distribution. As $n \longrightarrow \infty$, $\tau_\alpha - n^{1/4}\frac{\gamma}{\sqrt{\beta}} \longrightarrow -\infty$ and consequently

$$\mathbb{P}(\text{Clique vertex misidentified}) \longrightarrow 0,$$

and hence

$$\mathbb{P}(\text{Clique vertices identified}) \longrightarrow 1.$$

Tail bounds allow us to establish the stated mode of convergence. Thus, we have shown that so long as $\theta > 1$, the eigen-spectrum will allow the clique to be correctly identified (in the large graph limit). Conversely, for $\theta < 1$, eigen-analysis will yield no evidence of the presence of the clique.

Recall that the argument above revealed a phase transition above a critical θ value of 1 for a rank-1 perturbation of the form θuu^T to a Wigner matrix having zero mean entries with variance $1/n$. To apply this result to our problem, we set $\theta = \frac{k}{\sqrt{n}} \cdot \sqrt{\frac{1-p}{p}}$ and hence obtain Theorem 3.1. The portion of the random matrix corresponding to the clique will be non-random but since that size is $O(\sqrt{n})$, it will not matter in the limit of large network. Thus, what emerges for the analysis is the presence of hard detectability threshold below which the planted clique is present but cannot be detected or discovered using eigen-analysis based techniques.

5. NUMERICAL SIMULATIONS

Figure 1 shows a plot of the probability of correctly identifying the planted clique vertices for the dense setting in Theorem 3.1 in the setting where $p = 0.5$. As n increases we see a phase transition like effect in good agreement with the prediction with a rounding off due to finite size of the graphs. Figure 2 shows a plot of the probability of correctly identifying the planted clique vertices for the sparse setting in Claim 3.2 with $n = 100000$ as a function of the average degree c and size of the clique k. The solid white line corresponds to the curve $\sqrt{c}$. As the figure shows, the agreement between the prediction and the simulations is excellent.

6. CONCLUSIONS

We studied the problem of detecting or discovering a planted clique embedded in a random graph and showed, using recent results from random matrix theory, that there is a hard limit of eigen-analysis based clique discovery or detection.

One might imagine this transition to be a particular property of the eigen-analysis method we have considered. Perhaps a different algorithm, one not based on eigen-analysis techniques, or a different type of clique detection method altogether (such as for example, [5, 6]) would be able to get past this detectability threshold?

Or perhaps the threshold coincides with an algorithmic phase transition of the kind considered in [15]. The answers to these questions remain open and will provide better insight on the ultimate limits of efficient algorithms for planted clique detection and discovery.

7. REFERENCES

[1] A. Clauset, M.E.J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical review E*, vol. 70, no. 6, pp. 066111, 2004.

[2] M.E.J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577, 2006.

[3] N. Alon, M. Krivelevich, and B. Sudakov, "Finding a large hidden clique in a random graph," in *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 1998, pp. 594–598.

[4] U. Feige and R. Krauthgamer, "Finding and certifying a large hidden clique in a semirandom graph," *Random Structures & Algorithms*, vol. 16, no. 2, pp. 195–208, 2000.

[5] S. Brubaker and S. Vempala, "Random tensors and planted cliques," *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pp. 406–419, 2009.

[6] B.P.W. Ames and S.A. Vavasis, "Nuclear norm minimization for the planted clique and biclique problems," *Mathematical Programming*, pp. 1–21, 2011.

[7] F. McSherry, "Spectral partitioning of random graphs," in *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*. IEEE, 2001, pp. 529–537.

[8] M.E.J. Newman, *Networks: An introduction*, Oxford Univ Press, 2010.

[9] F. Benaych-Georges, "Eigenvectors of Wigner matrices: universality of global fluctuations," *Arxiv preprint arXiv:1104.1219*, 2011.

[10] G. Semerjian and L.F. Cugliandolo, "Sparse random matrices: the eigenvalue spectrum revisited," *Journal of Physics A: Mathematical and General*, vol. 35, pp. 4837, 2002.

[11] M. Capitaine, C. Donati-Martin, and D. Féral, "The largest eigenvalues of finite rank deformation of large wigner matrices: convergence and nonuniversality of the fluctuations," *The Annals of Probability*, vol. 37, no. 1, pp. 1–47, 2009.

[12] F. Benaych-Georges and R.R. Nadakuditi, "The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices," *Advances in Mathematics*, 2011.

[13] Z. Bai and J.W. Silverstein, "Wigner matrices and semicircular law," *Spectral Analysis of Large Dimensional Random Matrices*, pp. 15–38, 2010.

[14] T. Tao and V. Vu, "Random matrices: The universality phenomenon for Wigner ensembles," *Arxiv preprint arXiv:1202.0068*, 2012.

[15] D. Achlioptas and A. Coja-Oghlan, "Algorithmic barriers from phase transitions," in *Foundations of Computer Science, 2008. FOCS'08. IEEE 49th Annual IEEE Symposium on*. IEEE, 2008, pp. 793–802.