Community Detection in Random Networks

Ery Arias-Castro

Department of Mathematics,
University of California, San Diego

Joint work with

Nicolas Verzelen
INRA Montpellier, France

Partial support from the *Office of Naval Research*

Community detection

Community detection is an umbrella phrase for the task of understanding how a network (e.g., of friendships) is organized into communities (e.g., groups of friends).

Community detection

Community detection is an umbrella phrase for the task of understanding how a network (e.g., of friendships) is organized into communities (e.g., groups of friends).

Almost all the existing work concerns either

- The modeling of real networks as random graphs.
- The extraction of communities, which amounts to graph partitioning or clustering.

See the excellent review of Fortunato (2010).

Our contribution

Our work is really concerned with the task of *detecting* a community in a network, meaning, addressing the question:

> *Is the network homogeneously connected or is there a salient community?*

Our contribution

Our work is really concerned with the task of *detecting* a community in a network, meaning, addressing the question:

> *Is the network homogeneously connected or is there a salient community?*

Not a whole lot of work directly answering this question (Sun and Nobel, 2008; Butucea and Ingster, 2011; Rukhin and Priebe, 2012).

Some consider a dynamic setting where the goal is to detect a change in the evolution of a network (Heard et al., 2010; Mongiov et al., 2013; Park et al., 2013).

Notation

A network is here modeled as an undirected graph $\mathcal{G} = (\mathcal{E}, \mathcal{V})$ where $\mathcal{V} = [N] := \{1, \ldots, N\}$.

The corresponding adjacency matrix is denoted $\boldsymbol{W} = (W_{i,j}) \in \{0,1\}^{N \times N}$, where $W_{i,j} = 1 \Leftrightarrow (i,j) \in \mathcal{E}$.

For $S \subset \mathcal{V}$, let $\mathcal{G}_S$ denote the subgraph of $\mathcal{G}$ induced by S.

Model

We model a 'homogeneously connected network' as follows.

Erdös-Rényi random graph

$\mathbb{G}(n, p)$ denotes the distribution of a random graph on n vertices where each pair of vertices is connected with probability p, independently of all the other pairs. Equivalently, in terms of adjacency matrix, $(W_{ij} : i < j)$ are IID Bernoulli(p) and $W_{ii} = 0$.

We model the task of detecting a salient community in a network as the following hypothesis testing problem:

We model the task of detecting a salient community in a network as the following hypothesis testing problem:

Under the null hypothesis H_0, $\mathcal{G} \sim \mathbb{G}(n, p_0)$, ie,

$$\mathbb{P}(W_{ij} = 1) = p_0, \quad \forall i \neq j.$$

We model the task of detecting a salient community in a network as the following hypothesis testing problem:

Under the null hypothesis H_0, $\mathcal{G} \sim \mathbb{G}(n, p_0)$, ie,

$$\mathbb{P}(W_{ij} = 1) = p_0, \quad \forall i \neq j.$$

Under the alternative H_S, where $S \subset \mathcal{V}$,

$$\begin{aligned} \mathbb{P}(W_{ij} = 1) &= p_1, \quad \forall i, j \in S, i \neq j, \\ \mathbb{P}(W_{i,j} = 1) &= p_0, \quad \text{otherwise.} \end{aligned}$$

We model the task of detecting a salient community in a network as the following hypothesis testing problem:

Under the null hypothesis H_0, $\mathcal{G} \sim \mathbb{G}(n, p_0)$, ie,

$$\mathbb{P}(W_{ij} = 1) = p_0, \quad \forall i \neq j.$$

Under the alternative H_S, where $S \subset \mathcal{V}$,

$$\begin{aligned} \mathbb{P}(W_{ij} = 1) &= p_1, \quad \forall i, j \in S, i \neq j, \\ \mathbb{P}(W_{i,j} = 1) &= p_0, \quad \text{otherwise.} \end{aligned}$$

We assume the subset S is unknown. For simplicity of exposition, we assume here that its size $n = |S|$ is known. Therefore, we are testing H_0 versus $H_1 = \bigcup_{|S|=n} H_S$.

We model the task of detecting a salient community in a network as the following hypothesis testing problem:

Under the null hypothesis H_0, $\mathcal{G} \sim \mathbb{G}(n, p_0)$, ie,

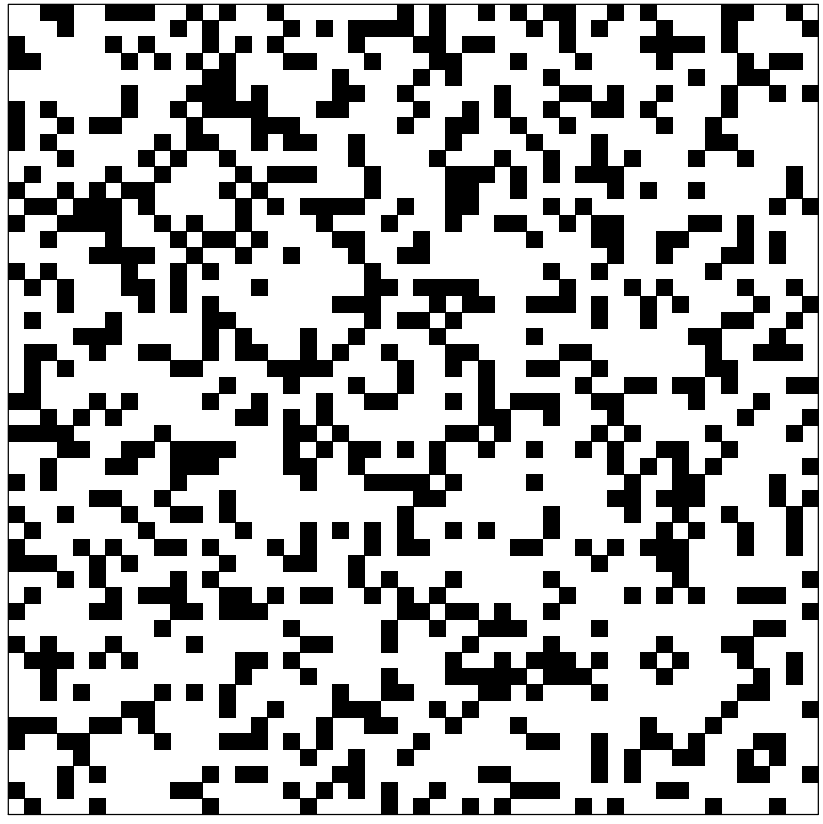
$$\mathbb{P}(W_{ij} = 1) = p_0, \quad \forall i \neq j.$$

Under the alternative H_S, where $S \subset \mathcal{V}$,

$$\begin{aligned} \mathbb{P}(W_{ij} = 1) &= p_1, \quad \forall i, j \in S, i \neq j, \\ \mathbb{P}(W_{i,j} = 1) &= p_0, \quad \text{otherwise.} \end{aligned}$$

We assume the subset S is unknown. For simplicity of exposition, we assume here that its size $n = |S|$ is known. Therefore, we are testing H_0 versus $H_1 = \bigcup_{|S|=n} H_S$.

We assume that $p_1 > p_0$, implying that the connectivity is stronger between nodes in S, so that S is an assortative community.



A realization of the null hypothesis ($N = 50$, $p_0 = 0.3$).
Black = connected; white = not connected.

A realization of the aternative hypothesis, where $S = \{1, \ldots, n\}$ $(N = 50,\ p_0 = 0.3,\ n = 10,\ p_1 = 0.8)$.

Same, but now with S random.
($S = \{2, 10, 13, 33, 38, 15, 8, 32, 40, 3\}$)

We consider an asymptotic setting where

$$N \to \infty, \quad n = n(N) \to \infty, \quad n/N \to 0, \quad n/\log N \to \infty.$$

Also, the probabilities of connection, $p_0 = p_0(N)$ and $p_1 = p_1(N)$, may change with N.

Risk

The (worst-case) *risk* of a test ϕ is defined as:

$$\gamma_N(\phi) = \mathbb{P}_0(\phi = 1) + \max_{S \in \mathcal{S}} \ \mathbb{P}_S(\phi = 0)$$

Risk

The (worst-case) *risk* of a test ϕ is defined as:

$$\gamma_N(\phi) = \mathbb{P}_0(\phi = 1) + \max_{S \in \mathcal{S}} \mathbb{P}_S(\phi = 0)$$

- A sequence of tests $\phi = \phi_N$ is asymptotically powerful if

$$\gamma_N(\phi) \to 0, \quad N \to \infty$$

- A sequence of tests ϕ is asymptotically powerless if

$$\varliminf_{N \to \infty} \gamma_N(\phi) \geq 1$$

(Asymptotically, ϕ is no better than a guess not based on the data.)

Special case: detecting a clique

When $p_1 = 1$, the subset S generates a clique under H_S, meaning that $\mathcal{G}_S$ is a complete graph. When this is the case, we have the following.

Special case: detecting a clique

When $p_1 = 1$, the subset S generates a clique under H_S, meaning that $\mathcal{G}_S$ is a complete graph. When this is the case, we have the following.

Theorem

All tests are asymptotically powerless if

$$\binom{N}{n} p_0^{\frac{n(n-1)}{2}} \to \infty. \tag{1}$$

The clique number of a graph is the size (number of nodes) of one of its largest cliques.

The clique number of a graph is the size (number of nodes) of one of its largest cliques.

Proposition

The test that rejects for large values of the clique number is asymptotically powerful if

$$\binom{N}{n} p_0^{\frac{n(n-1)}{2}} \to 0. \tag{2}$$

Planted clique problem

The test based on the clique number is near-optimal, but computationally intractable.

Planted clique problem

The test based on the clique number is near-optimal, but computationally intractable.

In fact, when $p_0 = 1/2$, we do not know of any polynomial-time algorithm that is asymptotically powerful when $n \ll \sqrt{N}$, while the test based on the clique number is asymptotically powerful when $n \geq C \log N$ with $C > 2/\log 2$ fixed.

Planted clique problem

The test based on the clique number is near-optimal, but computationally intractable.

In fact, when $p_0 = 1/2$, we do not know of any polynomial-time algorithm that is asymptotically powerful when $n \ll \sqrt{N}$, while the test based on the clique number is asymptotically powerful when $n \geq C \log N$ with $C > 2/\log 2$ fixed.

In our setting too, we will see that there is also a substantial discrepancy between what we can achieve with, and what we can achieve without, computational constraints.

The quasi-normal regime

First, assume that

$$\log\left(1 \vee \frac{1}{np_0}\right) = o\left[\log\left(\frac{N}{n}\right)\right] \quad . \tag{3}$$

This includes the case where p_0 is constant, covered by Butucea and Ingster, 2011.

The quasi-normal regime

First, assume that

$$\log\left(1 \vee \frac{1}{np_0}\right) = o\left[\log\left(\frac{N}{n}\right)\right] \quad . \tag{3}$$

This includes the case where p_0 is constant, covered by Butucea and Ingster, 2011.

In this regime, the partial sums

$$W_S = \sum_{i,j \in S, i<j} W_{i,j}$$

are, all together, close to normally distributed.

Introduce the entropy function:

$$H_p(q) = q \log\left(\frac{q}{p}\right) + (1-q)\log\left(\frac{1-q}{1-p}\right), \quad p, q \in (0,1).$$

Introduce the entropy function:

$$H_p(q) = q \log\left(\frac{q}{p}\right) + (1-q)\log\left(\frac{1-q}{1-p}\right), \quad p, q \in (0,1).$$

Theorem

All tests are asymptotically powerless if

$$\zeta := \frac{p_1 - p_0}{\sqrt{p_0}} \frac{n^2}{N} \to 0, \tag{4}$$

and

$$\overline{\lim} \frac{nH_{p_0}(p_1)}{2\log(N/n)} < 1. \tag{5}$$

Recall that $n = n(N)$, $p_0 = p_0(N)$ and $p_1 = p_1(N)$, subject to (3).

The total degree test

The total degree test rejects for large values of

$$W = \sum_{1 \leq i < j \leq N} W_{i,j}.$$

The total degree test

The total degree test rejects for large values of

$$W = \sum_{1 \leq i < j \leq N} W_{i,j}.$$

Proposition

The total degree test is powerful if

$$\zeta = \frac{p_1 - p_0}{\sqrt{p_0}} \frac{n^2}{N} \to \infty. \tag{6}$$

The scan test

Recall

$$W_S = \sum_{i,j \in S, i<j} W_{i,j} \ .$$

The scan test

Recall

$$W_S = \sum_{i,j \in S, i<j} W_{i,j} \ .$$

The scan test rejects for large values of

$$W_n^* = \max_{|S|=n} W_S \ .$$

(Same as the generalized likelihood ratio test when p_0 is known.)

The scan test

Recall

$$W_S = \sum_{i,j \in S, i<j} W_{i,j} \ .$$

The scan test rejects for large values of

$$W_n^* = \max_{|S|=n} W_S \ .$$

(Same as the generalized likelihood ratio test when p_0 is known.)

Proposition

The scan test is powerful if

$$\underline{\lim} \frac{n H_{p_0}(p_1)}{2 \log(N/n)} > 1. \tag{7}$$

When n is unknown

- The total degree test does not require knowledge of n.

When n is unknown

- The total degree test does not require knowledge of n.
- The scan test can be applied at all n's (with a Bonferroni correction) without substantial loss of power.

When p_0 is unknown

When p_0 is unknown, the total degree test is not applicable. In fact, $\frac{p_1 - p_0}{\sqrt{p_0}} \frac{n^2}{N} \to \infty$ is no longer enough for reliable detection, and we suggest another test, called the degree variance test.

When p_0 is unknown

When p_0 is unknown, the total degree test is not applicable. In fact, $\frac{p_1 - p_0}{\sqrt{p_0}} \frac{n^2}{N} \to \infty$ is no longer enough for reliable detection, and we suggest another test, called the degree variance test.

The scan test can be calibrated without loss of power. For example, the scan statistic can be calibrated under $\mathbb{G}(N, \hat{p}_0)$, where $\hat{p}_0 = W/\binom{N}{2}$.

When p_0 is unknown

When p_0 is unknown, the total degree test is not applicable. In fact, $\frac{p_1 - p_0}{\sqrt{p_0}} \frac{n^2}{N} \to \infty$ is no longer enough for reliable detection, and we suggest another test, called the degree variance test.

The scan test can be calibrated without loss of power. For example, the scan statistic can be calibrated under $\mathbb{G}(N, \hat{p}_0)$, where $\hat{p}_0 = W/\binom{N}{2}$.

(See the paper for details.)

Testing in polynomial-time

A question of practical, and theoretical, importance is:

What can be done in polynomial-time?

Testing in polynomial-time

A question of practical, and theoretical, importance is:

What can be done in polynomial-time?

While the total degree (and the degree variance statistic) has complexity $O(N^2)$, computing the scan statistic is intractable.

Maximum degree test

The maximum degree test rejects for large values of

$$\max_i \sum_{j \neq i} W_{ij}.$$

Maximum degree test

The maximum degree test rejects for large values of

$$\max_i \sum_{j \neq i} W_{ij}.$$

Proposition

The maximum degree test is asymptotically powerful if $p_0 \gg \log(N)/N$ and

$$\liminf \frac{n^2}{N \log(N)} \frac{(p_1 - p_0)^2}{p_0} > 2 \ .$$

The maximum degree test is asymptotically powerless if (3) holds and

$$\limsup \frac{\log n}{\log N} < 1, \quad \frac{n^2}{N \log(N)} \frac{(p_1 - p_0)^2}{p_0} \to 0.$$

A first relaxation

Sparse eigenvalue

For a PSD matrix $\boldsymbol{B} \in \mathbb{R}^{N \times N}$ and $1 \leq n \leq N$, define

$$\lambda_n^{\max}(\boldsymbol{B}) = \max_{|S|=n} \lambda^{\max}(\boldsymbol{B}_S) \ ,$$

where $\boldsymbol{B}_S$ denotes the principal submatrix of $\boldsymbol{B}$ indexed by $S \subset [N]$ and $\lambda^{\max}(\boldsymbol{B}_S)$ the largest eigenvalue of $\boldsymbol{B}_S$.

A first relaxation

Sparse eigenvalue

For a PSD matrix $\boldsymbol{B} \in \mathbb{R}^{N \times N}$ and $1 \leq n \leq N$, define

$$\lambda_n^{\max}(\boldsymbol{B}) = \max_{|S|=n} \lambda^{\max}(\boldsymbol{B}_S) \ ,$$

where $\boldsymbol{B}_S$ denotes the principal submatrix of $\boldsymbol{B}$ indexed by $S \subset [N]$ and $\lambda^{\max}(\boldsymbol{B}_S)$ the largest eigenvalue of $\boldsymbol{B}_S$.

Let $\mathbf{1}_S$ be the indicator vector of $S \subset [N]$. We have

$$W_n^* = \frac{1}{2} \max_{|S|=n} \mathbf{1}_S^T \boldsymbol{W} \mathbf{1}_S$$

A first relaxation

Sparse eigenvalue

For a PSD matrix $\boldsymbol{B} \in \mathbb{R}^{N \times N}$ and $1 \leq n \leq N$, define

$$\lambda_n^{\max}(\boldsymbol{B}) = \max_{|S|=n} \lambda^{\max}(\boldsymbol{B}_S) \ ,$$

where $\boldsymbol{B}_S$ denotes the principal submatrix of $\boldsymbol{B}$ indexed by $S \subset [N]$ and $\lambda^{\max}(\boldsymbol{B}_S)$ the largest eigenvalue of $\boldsymbol{B}_S$.

Let $\mathbf{1}_S$ be the indicator vector of $S \subset [N]$. We have

$$W_n^* = \frac{1}{2} \max_{|S|=n} \mathbf{1}_S^T \boldsymbol{W} \mathbf{1}_S \leq \frac{n}{2} \lambda^{\max}(\boldsymbol{W}_S)$$

A first relaxation

Sparse eigenvalue

For a PSD matrix $\boldsymbol{B} \in \mathbb{R}^{N \times N}$ and $1 \leq n \leq N$, define

$$\lambda_n^{\max}(\boldsymbol{B}) = \max_{|S|=n} \lambda^{\max}(\boldsymbol{B}_S) \ ,$$

where $\boldsymbol{B}_S$ denotes the principal submatrix of $\boldsymbol{B}$ indexed by $S \subset [N]$ and $\lambda^{\max}(\boldsymbol{B}_S)$ the largest eigenvalue of $\boldsymbol{B}_S$.

Let $\mathbf{1}_S$ be the indicator vector of $S \subset [N]$. We have

$$W_n^* = \frac{1}{2} \max_{|S|=n} \mathbf{1}_S^T \boldsymbol{W} \mathbf{1}_S \leq \frac{n}{2} \lambda^{\max}(\boldsymbol{W}_S) \leq \frac{n}{2} \lambda_n^{\max}(\boldsymbol{W}).$$

A first relaxation

Sparse eigenvalue

For a PSD matrix $\boldsymbol{B} \in \mathbb{R}^{N \times N}$ and $1 \leq n \leq N$, define

$$\lambda_n^{\max}(\boldsymbol{B}) = \max_{|S|=n} \lambda^{\max}(\boldsymbol{B}_S) \ ,$$

where $\boldsymbol{B}_S$ denotes the principal submatrix of $\boldsymbol{B}$ indexed by $S \subset [N]$ and $\lambda^{\max}(\boldsymbol{B}_S)$ the largest eigenvalue of $\boldsymbol{B}_S$.

Let $\mathbf{1}_S$ be the indicator vector of $S \subset [N]$. We have

$$W_n^* = \frac{1}{2} \max_{|S|=n} \mathbf{1}_S^T \boldsymbol{W} \mathbf{1}_S \leq \frac{n}{2} \lambda^{\max}(\boldsymbol{W}_S) \leq \frac{n}{2} \lambda_n^{\max}(\boldsymbol{W}).$$

We could consider a test based on $\lambda_n^{\max}(\boldsymbol{W})$, but this is still computationally intractable!

A second (convex) relaxation

We consider a convex relaxation of dAspremont et al. (2007):

$$\mathrm{SDP}_n(\boldsymbol{B}) = \max_{\boldsymbol{Z}} \mathrm{Trace}(\boldsymbol{B}\boldsymbol{Z})$$

$$\text{subject to } \boldsymbol{Z} \succeq 0,\ \mathrm{Trace}(\boldsymbol{Z}) = 1,\ |\boldsymbol{Z}|_1 \leq n\ ,$$

where the maximum is over positive semidefinite matrices $\boldsymbol{Z} = (Z_{st}) \in \mathbb{R}^{N \times N}$ and $|\boldsymbol{Z}|_1 = \sum_{s,t} |Z_{st}|$.

A second (convex) relaxation

We consider a convex relaxation of dAspremont et al. (2007):

$$\mathrm{SDP}_n(\boldsymbol{B}) = \max_{\boldsymbol{Z}} \mathrm{Trace}(\boldsymbol{B}\boldsymbol{Z})$$

$$\text{subject to } \boldsymbol{Z} \succeq 0,\ \mathrm{Trace}(\boldsymbol{Z}) = 1,\ |\boldsymbol{Z}|_1 \leq n\ ,$$

where the maximum is over positive semidefinite matrices $\boldsymbol{Z} = (Z_{st}) \in \mathbb{R}^{N \times N}$ and $|\boldsymbol{Z}|_1 = \sum_{s,t} |Z_{st}|$.

We consider the relaxed scan test, which rejects for large values of

$$\mathrm{SDP}_n(\boldsymbol{W}^2)\ . \tag{8}$$

A second (convex) relaxation

We consider a convex relaxation of dAspremont et al. (2007):

$$\mathrm{SDP}_n(\boldsymbol{B}) = \max_{\boldsymbol{Z}} \mathrm{Trace}(\boldsymbol{B}\boldsymbol{Z})$$

$$\text{subject to } \boldsymbol{Z} \succeq 0,\ \mathrm{Trace}(\boldsymbol{Z}) = 1,\ |\boldsymbol{Z}|_1 \leq n\ ,$$

where the maximum is over positive semidefinite matrices $\boldsymbol{Z} = (Z_{st}) \in \mathbb{R}^{N \times N}$ and $|\boldsymbol{Z}|_1 = \sum_{s,t} |Z_{st}|$.

We consider the relaxed scan test, which rejects for large values of

$$\mathrm{SDP}_n(\boldsymbol{W}^2)\ . \tag{8}$$

(This approach is inspired by work of Berthet and Rigollet on detecting a principal component. Here we use $\boldsymbol{W}^2$ instead of $\boldsymbol{W}$ to concentrate the matrix coefficients.)

Proposition

Assume that (3) holds and $n \leq N^{1/2-t}$ for some $t > 0$. Then, the relaxed scan test is powerful if

$$\liminf \frac{n}{\sqrt{N \log(N)}} \frac{(p_1 - p_0)^2}{p_0} > 2 \ . \tag{9}$$

- Assume that $n^2 \ll N$ and $np_0 \gg \log(N/n)$. Then the scan test is powerful when

$$\frac{(p_1 - p_0)^2}{p_0} \succ \frac{\log(N/n)}{n} \ .$$

  Thus we lose a factor of $\sqrt{N/\log(N)}$ when using the relaxed scan test.

- Assume that $n^2 \ll N$ and $np_0 \gg \log(N/n)$. Then the scan test is powerful when

$$\frac{(p_1 - p_0)^2}{p_0} \succ \frac{\log(N/n)}{n} \ .$$

  Thus we lose a factor of $\sqrt{N/\log(N)}$ when using the relaxed scan test.

- But compared to the maximum degree test, which requires

$$\frac{(p_1 - p_0)^2}{p_0} \succ \frac{N \log N}{n^2},$$

  we win a factor of $\frac{1}{n}\sqrt{N \log N}$ with the relaxed scan test.

The non-normal regime

We now assume that

$$np_0 \leq 1, \quad \log\left(\frac{1}{np_0}\right) \succ \log\left(\frac{N}{n}\right) \ , \tag{10}$$

which complements (3).

The non-normal regime

We now assume that

$$np_0 \leq 1, \quad \log\left(\frac{1}{np_0}\right) \succ \log\left(\frac{N}{n}\right) \ , \tag{10}$$

which complements (3).

Let

$$\lambda_0 = Np_0, \quad \lambda_1 = np_1,$$

and define

$$\alpha = \frac{\log \lambda_0}{\log(N/n)} \ . \tag{11}$$

Note that $0 \leq \alpha \leq 1$. We allow $\lambda_0, \lambda_1, \alpha$ to vary with N.

The non-normal regime

We now assume that

$$np_0 \leq 1, \quad \log\left(\frac{1}{np_0}\right) \succ \log\left(\frac{N}{n}\right) \ , \tag{10}$$

which complements (3).

Let

$$\lambda_0 = Np_0, \quad \lambda_1 = np_1,$$

and define

$$\alpha = \frac{\log \lambda_0}{\log(N/n)} \ . \tag{11}$$

Note that $0 \leq \alpha \leq 1$. We allow $\lambda_0, \lambda_1, \alpha$ to vary with N.

The regime (10) includes the *Poisson regime* (λ_0 constant).

- Introduce

$$I_\lambda = \lambda - 1 - \log(\lambda) \ ,$$

  which is intimately related to the size of the largest connected component in $\mathbb{G}(m, \lambda)$.

- Introduce

$$I_\lambda = \lambda - 1 - \log(\lambda) \ ,$$

  which is intimately related to the size of the largest connected component in $\mathbb{G}(m, \lambda)$.

- Recall that

$$\zeta = \frac{p_1 - p_0}{\sqrt{p_0}} \frac{n^2}{N}$$

  controls the performance of the total degree test.

- Introduce

$$I_\lambda = \lambda - 1 - \log(\lambda) \ ,$$

which is intimately related to the size of the largest connected component in $\mathbb{G}(m, \lambda)$.

- Recall that

$$\zeta = \frac{p_1 - p_0}{\sqrt{p_0}} \frac{n^2}{N}$$

controls the performance of the total degree test.

- When $S \subset \mathcal{V}$ is the salient community, we will consider

$$W^*_{k,S} = \max_{T \subset S, |T|=k} W_T \ .$$

(Recall the partial sums $W_T = \sum_{i,j \in T, i<j} W_{i,j}$.)

Recall that

$$p_0 = \frac{\lambda_0}{N},$$

Recall that

$$p_0 = \frac{\lambda_0}{N}, \text{ with } \lambda_0 = \left(\frac{N}{n}\right)^{\alpha},$$

Recall that

$$p_0 = \frac{\lambda_0}{N}, \text{ with } \lambda_0 = \left(\frac{N}{n}\right)^{\alpha}, \text{ and } p_1 = \frac{\lambda_1}{n} \ .$$

Recall that

$$p_0 = \frac{\lambda_0}{N}, \text{ with } \lambda_0 = \left(\frac{N}{n}\right)^{\alpha}, \text{ and } p_1 = \frac{\lambda_1}{n} \ .$$

Theorem

Assume that $\zeta \to 0$. Then all tests are asymptotically powerless in either of the following situations:

Recall that

$$p_0 = \frac{\lambda_0}{N}, \text{ with } \lambda_0 = \left(\frac{N}{n}\right)^{\alpha}, \text{ and } p_1 = \frac{\lambda_1}{n} \ .$$

Theorem

Assume that $\zeta \to 0$. Then all tests are asymptotically powerless in either of the following situations:

$$\lambda_0 \to 0, \quad \lambda_1 \to 0, \quad \overline{\lim} \frac{I_{\lambda_0}}{I_{\lambda_1}} \frac{\log n}{\log N} < 1 \ ; \tag{12}$$

Recall that

$$p_0 = \frac{\lambda_0}{N}, \text{ with } \lambda_0 = \left(\frac{N}{n}\right)^{\alpha}, \text{ and } p_1 = \frac{\lambda_1}{n} \ .$$

Theorem

Assume that $\zeta \to 0$. Then all tests are asymptotically powerless in either of the following situations:

$$\lambda_0 \to 0, \quad \lambda_1 \to 0, \quad \overline{\lim} \frac{I_{\lambda_0}}{I_{\lambda_1}} \frac{\log n}{\log N} < 1 \ ; \tag{12}$$

$$0 < \underline{\lim}\, \lambda_0 \leq \overline{\lim}\, \lambda_0 < \infty, \quad \lambda_1 \to 0 \ ; \tag{13}$$

Recall that

$$p_0 = \frac{\lambda_0}{N}, \text{ with } \lambda_0 = \left(\frac{N}{n}\right)^{\alpha}, \text{ and } p_1 = \frac{\lambda_1}{n} \ .$$

Theorem

Assume that $\zeta \to 0$. Then all tests are asymptotically powerless in either of the following situations:

$$\lambda_0 \to 0, \quad \lambda_1 \to 0, \quad \overline{\lim} \frac{I_{\lambda_0}}{I_{\lambda_1}} \frac{\log n}{\log N} < 1 \ ; \tag{12}$$

$$0 < \underline{\lim} \lambda_0 \leq \overline{\lim} \lambda_0 < \infty, \quad \lambda_1 \to 0 \ ; \tag{13}$$

$$\lambda_0 \to \infty \textit{ with } \alpha \to 0, \quad \overline{\lim} \lambda_1 < 1 \ ; \tag{14}$$

Recall that

$$p_0 = \frac{\lambda_0}{N}, \text{ with } \lambda_0 = \left(\frac{N}{n}\right)^{\alpha}, \text{ and } p_1 = \frac{\lambda_1}{n} \ .$$

Theorem

Assume that $\zeta \to 0$. Then all tests are asymptotically powerless in either of the following situations:

$$\lambda_0 \to 0, \quad \lambda_1 \to 0, \quad \overline{\lim} \, \frac{I_{\lambda_0}}{I_{\lambda_1}} \frac{\log n}{\log N} < 1 \ ; \tag{12}$$

$$0 < \underline{\lim} \, \lambda_0 \leq \overline{\lim} \, \lambda_0 < \infty, \quad \lambda_1 \to 0 \ ; \tag{13}$$

$$\lambda_0 \to \infty \text{ with } \alpha \to 0, \quad \overline{\lim} \, \lambda_1 < 1 \ ; \tag{14}$$

$$0 < \underline{\lim} \, \alpha \leq \overline{\lim} \, \alpha < 1, \quad \overline{\lim} \, (1-\alpha) \sup_{k=n/u_N}^{n} \frac{\mathbb{E}_S[W^*_{k,S}]}{k} < 1 \ . \tag{15}$$

Broad scan test

Let $u_N = \log\log N$. The broad scan test rejects for large values of

$$W_n^{\ddagger} = \sup_{k=n/u_N}^{n} \frac{W_k^*}{k} \ . \tag{16}$$

(Recall the scan statistic $W_k^* = \max_{|T|=k} W_T$.)

Broad scan test

Let $u_N = \log\log N$. The broad scan test rejects for large values of

$$W_n^{\ddagger} = \sup_{k=n/u_N}^{n} \frac{W_k^*}{k} \ . \tag{16}$$

(Recall the scan statistic $W_k^* = \max_{|T|=k} W_T$.)

Theorem (Broad scan test)

The test based on $W_n^{\ddagger}$ is asymptotically powerful if either

$$\limsup \alpha \leq 1, \quad \liminf\ (1-\alpha) \sup_{k=n/u_N}^{n} \frac{\mathbb{E}_S[W_{k,S}^*]}{k} > 1 \ ; \tag{17}$$

or

$$\alpha \to 0 \quad \textit{and} \quad \liminf \lambda_1 > 1 \ . \tag{18}$$

Broad scan test

Let $u_N = \log\log N$. The broad scan test rejects for large values of

$$W_n^{\ddagger} = \sup_{k=n/u_N}^{n} \frac{W_k^*}{k} \ . \tag{16}$$

(Recall the scan statistic $W_k^* = \max_{|T|=k} W_T$.)

Theorem (Broad scan test)

The test based on $W_n^{\ddagger}$ is asymptotically powerful if either

$$\limsup \alpha \leq 1, \quad \liminf\ (1-\alpha) \sup_{k=n/u_N}^{n} \frac{\mathbb{E}_S[W_{k,S}^*]}{k} > 1 \ ; \tag{17}$$

or

$$\alpha \to 0 \quad \textit{and} \quad \liminf \lambda_1 > 1 \ . \tag{18}$$

Hence, the broad scan test achieves the minimax detection boundary delineated by (14) and (15).

The largest connected component test

The largest connected component of an Erdös-Rényi graph undergoes a phase transition. Indeed, the following is well-known.

The largest connected component test

The largest connected component of an Erdös-Rényi graph undergoes a phase transition. Indeed, the following is well-known.

Lemma

Let $\mathcal{C}_m$ denote a largest connected component of $\mathbb{G}(m, \lambda/m)$ and assume that $\lambda \in (0, \infty)$ is fixed. Let η_λ denote the smallest solution of the equation $\eta = \exp(\lambda(\eta - 1))$. Then, in probability,

$$|\mathcal{C}_m| \sim \begin{cases} I_\lambda^{-1} \log m, & \text{if } \lambda < 1 \ ; \\ (1 - \eta_\lambda) m, & \text{if } \lambda > 1 \ . \end{cases}$$

(Recall that $I_\lambda = \lambda - 1 - \log(\lambda)$.)

The largest connected component test rejects for large values of the size of the largest connected component in $\mathcal{G}$, which we denote by $\mathcal{C}_{\max}$.

The largest connected component test rejects for large values of the size of the largest connected component in $\mathcal{G}$, which we denote by $\mathcal{C}_{\max}$.

It is most effective in the subcritical regime where $\overline{\lim}\, \lambda_0 < 1$.

The largest connected component test rejects for large values of the size of the largest connected component in $\mathcal{G}$, which we denote by $\mathcal{C}_{\max}$.

It is most effective in the subcritical regime where $\overline{\lim}\, \lambda_0 < 1$.

A simple analysis goes as follows:

- Under H_0, $|\mathcal{C}_{\max}| \sim I_{\lambda_0}^{-1} \log N$.

The largest connected component test rejects for large values of the size of the largest connected component in $\mathcal{G}$, which we denote by $\mathcal{C}_{\max}$.

It is most effective in the subcritical regime where $\overline{\lim}\, \lambda_0 < 1$.

A simple analysis goes as follows:

- Under H_0, $|\mathcal{C}_{\max}| \sim I_{\lambda_0}^{-1} \log N$.
- Under H_S, $|\mathcal{C}_{\max}| \geq |\mathcal{C}_S|$, where $\mathcal{C}_S$ is the largest connected component in the graph induced by S, with

$$|\mathcal{C}_S| \sim \begin{cases} I_{\lambda_1}^{-1} \log n, & \text{if } \overline{\lim}\, \lambda_1 < 1 \ ; \\ (1 - \eta_{\lambda_1}) n, & \text{if } \lambda_1 > 1 \ . \end{cases}$$

Therefore, in the subcritical regime, the test is asymptotically powerful when

$$\liminf \ \frac{I_{\lambda_0}}{I_{\lambda_1}} \frac{\log n}{\log N} > 1 \ .$$

Therefore, in the subcritical regime, the test is asymptotically powerful when

$$\liminf \ \frac{I_{\lambda_0}}{I_{\lambda_1}} \frac{\log n}{\log N} > 1 \ .$$

Hence, the largest connected component test achieves the minimax detection boundary delineated by (12).

The following result requires a much finer estimate of the size of $\mathcal{C}_{\max}$ under H_S.

The following result requires a much finer estimate of the size of $\mathcal{C}_{\max}$ under H_S.

Theorem (Largest connected component test)

Assume that

$$\overline{\lim}\, \lambda_0 < 1, \quad \log\log(N) = o(\log n), \quad I_{\lambda_0}^{-1} \log(N) \to \infty.$$

The largest connected component test is asymptotically powerful when either $\liminf \lambda_1 > 1$ *or*

$$\lambda_0 \leq \lambda_1 e^{1-\lambda_1} \text{ for } n \text{ large enough},$$

and

$$\liminf \ \frac{I_{\lambda_0}}{\lambda_0 + I_{\lambda_1} - \lambda_0 e^{I_{\lambda_1}}} \frac{\log(n)}{\log(N)} > 1 \ .$$

Theorem (continued)

If we further assume that $n^2 = o(N)$*, then the largest connected component test is asymptotically powerless when* $\lambda_1 < 1$ *for all* n *and*

$$\lambda_0 \geq \lambda_1 e^{1-\lambda_1} \text{ for } n \text{ large enough}$$

or

$$\limsup \frac{I_{\lambda_0}}{\lambda_0 + I_{\lambda_1} - \lambda_0 e^{I_{\lambda_1}}} \frac{\log(n)}{\log(N)} < 1 \ .$$

The limbo detection zone

Our results are loose when

$$0 < \underline{\lim}\, \lambda_0 \leq \overline{\lim}\, \lambda_0 < \infty \ .$$

The limbo detection zone

Our results are loose when

$$0 < \underline{\lim}\, \lambda_0 \leq \overline{\lim}\, \lambda_0 < \infty \ .$$

We saw that

- No test is asymptotically powerful if $\lambda_1 \to 0$.

The limbo detection zone

Our results are loose when

$$0 < \underline{\lim}\, \lambda_0 \leq \overline{\lim}\, \lambda_0 < \infty \ .$$

We saw that

- No test is asymptotically powerful if $\lambda_1 \to 0$.
- The broad scan test is powerful when $\underline{\lim}\, \lambda_1 > 1$.

The limbo detection zone

Our results are loose when

$$0 < \underline{\lim}\, \lambda_0 \leq \overline{\lim}\, \lambda_0 < \infty \ .$$

We saw that

- No test is asymptotically powerful if $\lambda_1 \to 0$.
- The broad scan test is powerful when $\underline{\lim}\, \lambda_1 > 1$.
- The largest connected component test is powerful when $\overline{\lim}\, \lambda_0 < 1$ and $\underline{\lim}\, \lambda_1$ is sufficiently large (less than 1).

The number of triangles test

The number of triangles test rejects for large values of the number of triangles in $\mathcal{G}$, denoted by T.

The number of triangles test

The number of triangles test rejects for large values of the number of triangles in $\mathcal{G}$, denoted by T.

Proposition

When λ_0 and λ_1 are fixed,

$$T \Rightarrow \text{Poisson}(\mu), \quad \mu = \begin{cases} \lambda_0^3/6, & \text{under } H_0; \\ (\lambda_0^3 + \lambda_1^3)/6, & \text{under } H_1. \end{cases}$$

In particular, the test is not *asymptotically powerless if*

$$\limsup \lambda_0 < \infty \quad \text{and} \quad \liminf \lambda_1 > 0 \ .$$

Detection boundary in the Poisson regime

We are lead to the following.

- Assume that $n = N^{\kappa}$ with $0 < \kappa < 1/2$.
- Assume that $\lambda_0 > 0$ is fixed.

Detection boundary in the Poisson regime

We are lead to the following.

- Assume that $n = N^\kappa$ with $0 < \kappa < 1/2$.
- Assume that $\lambda_0 > 0$ is fixed.
- Define $\lambda_1^* = \lambda_1^*(\lambda_0, \kappa)$ as the infimum λ such that, for $\lambda_1 > \lambda_1^*$ fixed, there is an asymptotically powerful test.

Detection boundary in the Poisson regime

We are lead to the following.

- Assume that $n = N^\kappa$ with $0 < \kappa < 1/2$.
- Assume that $\lambda_0 > 0$ is fixed.
- Define $\lambda_1^* = \lambda_1^*(\lambda_0, \kappa)$ as the infimum λ such that, for $\lambda_1 > \lambda_1^*$ fixed, there is an asymptotically powerful test.

Open problem

Find or characterize $\lambda_1^*(\lambda_0, \kappa)$ in a meaningful way.

Detection boundary in the Poisson regime

We are lead to the following.

- Assume that $n = N^\kappa$ with $0 < \kappa < 1/2$.
- Assume that $\lambda_0 > 0$ is fixed.
- Define $\lambda_1^* = \lambda_1^*(\lambda_0, \kappa)$ as the infimum λ such that, for $\lambda_1 > \lambda_1^*$ fixed, there is an asymptotically powerful test.

Open problem

Find or characterize $\lambda_1^*(\lambda_0, \kappa)$ in a meaningful way.

We have some partial results...

Theorem

Write $n = N^\kappa$ with $0 < \kappa < 1/2$, and assume that λ_0 and λ_1 are both fixed. No test is asymptotically powerful in all the following situations:

$$\lambda_1 < 1, \quad \lambda_1^2 e \leq \lambda_0 \ ; \tag{19}$$

$$\lambda_1 < 1, \quad \lambda_1^2 e > \lambda_0, \quad \frac{1-2\kappa}{\kappa} > a_1 := \frac{\log\left(\frac{e\lambda_1^2}{\lambda_0}\right)}{I_{\lambda_1}} \ . \tag{20}$$

Theorem

Write $n = N^\kappa$ with $0 < \kappa < 1/2$, and assume that λ_0 and λ_1 are both fixed. No test is asymptotically powerful in all the following situations:

$$\lambda_1 < 1, \quad \lambda_1^2 e \leq \lambda_0 \ ; \tag{19}$$

$$\lambda_1 < 1, \quad \lambda_1^2 e > \lambda_0, \quad \frac{1-2\kappa}{\kappa} > a_1 := \frac{\log\left(\frac{e\lambda_1^2}{\lambda_0}\right)}{I_{\lambda_1}} \ . \tag{20}$$

(19) is sharp, since the broad scan test is powerful when $\lambda_0 > 0$ and $\lambda_1 > 1$ are fixed. Hence $\lambda_1^* = 1$ when $\lambda_0 \geq e$.

The number of subtrees test

The number of k-trees test rejects for large values of the number of subtrees of size k, where k is chosen of order $\log n$ below.

The number of subtrees test

The number of k-trees test rejects for large values of the number of subtrees of size k, where k is chosen of order $\log n$ below.

Theorem

Assume that λ_1 and λ_0 are both fixed, with $0 < \sqrt{\lambda_0/e} < \lambda_1 < 1$, and that $n = N^\kappa$ with $0 < \kappa < 1/2$, with

$$\limsup \frac{1-2\kappa}{\kappa} < a_2 := \frac{I_{\frac{\lambda_0}{\lambda_1 e}} - I_{\sqrt{\frac{\lambda_0}{e}}}}{\left(1 - \frac{\lambda_0}{\lambda_1 e}\right) I_{\frac{\sqrt{\lambda_1}}{e}}} \,. \tag{21}$$

Then there is a constant $c > 0$ such that the k-tree test with $k = [c \log n]$ is asymptotically powerful.

The number of subtrees test

The number of k-trees test rejects for large values of the number of subtrees of size k, where k is chosen of order $\log n$ below.

Theorem

Assume that λ_1 and λ_0 are both fixed, with $0 < \sqrt{\lambda_0/e} < \lambda_1 < 1$, and that $n = N^\kappa$ with $0 < \kappa < 1/2$, with

$$\limsup \frac{1-2\kappa}{\kappa} < a_2 := \frac{I_{\frac{\lambda_0}{\lambda_1 e}} - I_{\sqrt{\frac{\lambda_0}{e}}}}{\left(1 - \frac{\lambda_0}{\lambda_1 e}\right) I_{\frac{\sqrt{\lambda_1}}{e}}} \,. \tag{21}$$

Then there is a constant $c > 0$ such that the k-tree test with $k = [c \log n]$ is asymptotically powerful.

Compared to (20), there is a discrepancy in constants (ie, $a_1 \neq a_2$).

State of affairs

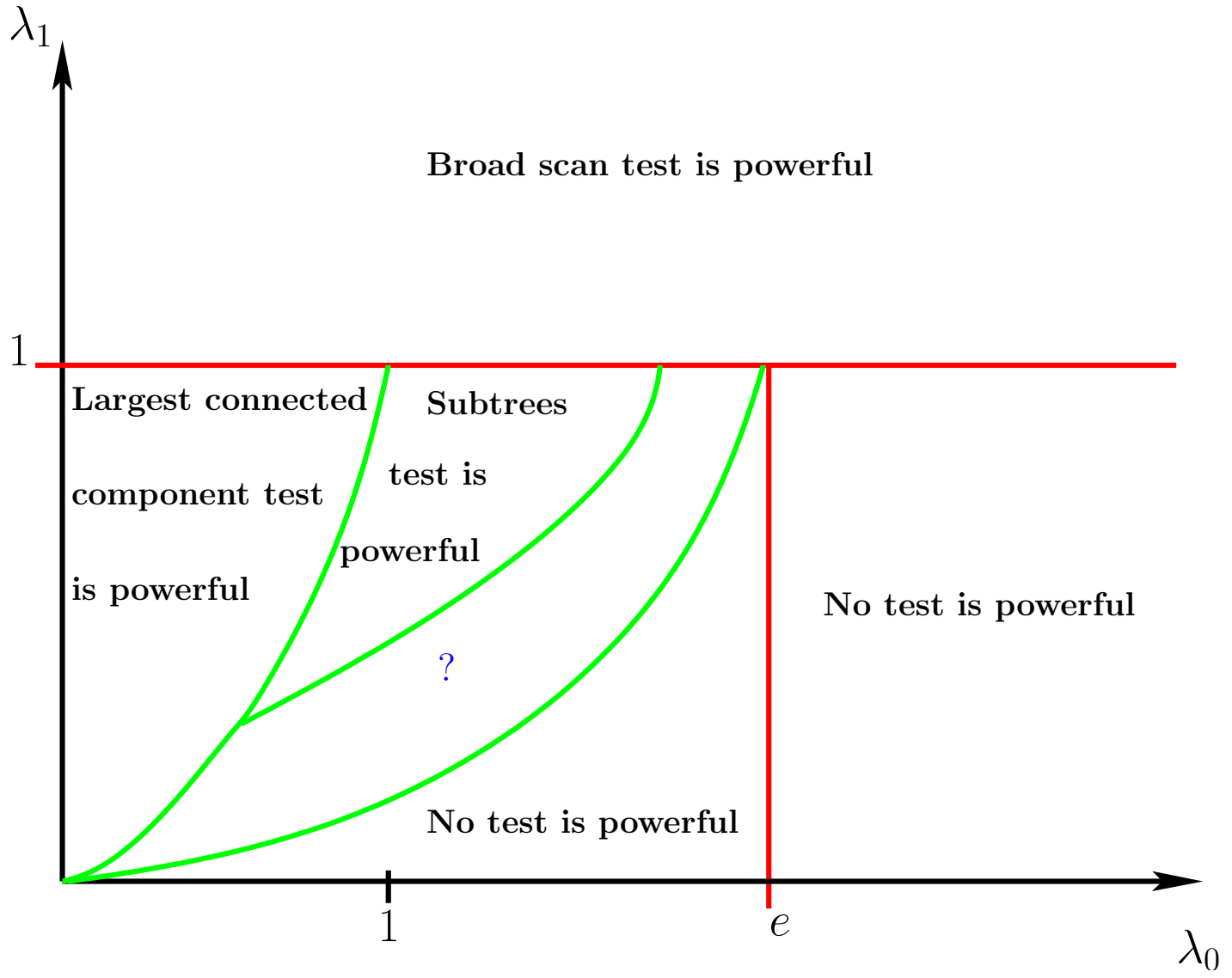


Figure : Detection diagram in the Poissonian asymptotics where λ_0 and λ_1 are fixed and $n = N^\kappa$ with $0 < \kappa < 1/2$.

Testing in polynomial time

The same question remains important:

What can be done in polynomial time?

Testing in polynomial time

The same question remains important:

What can be done in polynomial time?

- The total degree test and the largest connected component test can both be computed in polynomial time.

Testing in polynomial time

The same question remains important:

What can be done in polynomial time?

- The total degree test and the largest connected component test can both be computed in polynomial time.
- The broad scan test and the number of subtrees test seem computationally intractable.

Testing in polynomial time

The same question remains important:

What can be done in polynomial time?

- The total degree test and the largest connected component test can both be computed in polynomial time.
- The broad scan test and the number of subtrees test seem computationally intractable.
- The relaxed scan test and the maximum degree test are still computationally tractable, and their performance is established in the (2nd) paper.

Spectral methods

Let $\beta_1 \geq \cdots \geq \beta_N$ be the eigenvalues of the adjacency matrix $\boldsymbol{W}$.

Spectral methods

Let $\beta_1 \geq \cdots \geq \beta_N$ be the eigenvalues of the adjacency matrix $\boldsymbol{W}$.

Proposition

The test that rejects for large values of β_1 is asymptotically powerful when

$$\lambda_1 \gg \lambda_0 \gg \sqrt{\frac{\log N}{\log \log N}} \ .$$

(Based on asymptotics for β_1 for $\mathbb{G}(m, \lambda)$ by Krivelevich and Sudakov (2003).)

Proposition

The test that rejects for large values of β_2 is asymptotically powerful when $\lambda_0 \succ \log N$ and $\lambda_1 \gg \sqrt{\lambda_0} \vee (\lambda_0 \frac{n}{N})$.

(Based on asymptotics for β_2 for $\mathbb{G}(m, \lambda)$ by Füredi and Komlós (1981) and an extension by Feige and Ofek (2005).)

Note that $\lambda_1 \gg \lambda_0 \frac{n}{N}$ is equivalent to $p_1 \gg p_0$, which is for example true when $\limsup \alpha < 1$.

We did not explore in detail methods based on the Laplacian. (Some quick computations based on perturbation bounds of Chung and Radcliffe (2011) are not promising.)

The number of k-cycles

Consider the test that rejects for large values of C_k, the number of simple k-cycles in $\mathcal{G}$.

The number of k-cycles

Consider the test that rejects for large values of C_k, the number of simple k-cycles in $\mathcal{G}$.

Mossel, Neeman and Sly (2012) use that test to test against a *stochastic block model* alternative.

The number of k-cycles

Consider the test that rejects for large values of C_k, the number of simple k-cycles in $\mathcal{G}$.

Mossel, Neeman and Sly (2012) use that test to test against a *stochastic block model* alternative.

Proposition

The test based on C_k, with $k \to \infty$ and $k = O(\log N)^{1/4}$, is asymptotically powerful when λ_0 and λ_1 are fixed with

$$\lambda_1 > \sqrt{\lambda_0} \vee 1.$$

The number of k-cycles

Consider the test that rejects for large values of C_k, the number of simple k-cycles in $\mathcal{G}$.

Mossel, Neeman and Sly (2012) use that test to test against a *stochastic block model* alternative.

Proposition

The test based on C_k, with $k \to \infty$ and $k = O(\log N)^{1/4}$, is asymptotically powerful when λ_0 and λ_1 are fixed with

$$\lambda_1 > \sqrt{\lambda_0} \vee 1.$$

Although computing C_k seems difficult, Alon and Gutner (2010) provide an a polynomial-time approximation that yields a test with the same property, provided $k = O(\log \log N)$.

No polynomial-time test seems to come close to what the relaxed scan test can achieve when $\lambda_0 \to \infty$.

No polynomial-time test seems to come close to what the relaxed scan test can achieve when $\lambda_0 \to \infty$.

Open problem

Find a polynomial-time test that is asymptotically powerful when $n^2/N = O(1)$, *while* $\lambda_0 \to \infty$ *and* $\lambda_1 = O(1)$.

Papers

E. AC and N. Verzelen (2012). Community Detection in Random Networks. Available on arxiv.org.

N. Verzelen and E. AC (2013). Community Detection in Sparse Random Networks. Available on arxiv.org.

Thank You