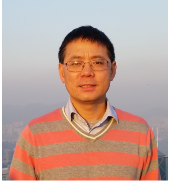
Detecting a small community in a large network

Paxton Turner

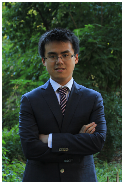
Statistics Department, Harvard University

Collaborators



Jiashun Jin
(CMU)

Tracy Ke
(Harvard)



Anru Zhang
(Duke)

Social networks

Data: $n \times n$ adjacency matrix A (symmetric)

$$A(i,j) = \begin{cases} 1, & \text{an edge between nodes } i \text{ and } j \\ 0, & \text{otherwise} \end{cases}$$

with K perceivable "communities" $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K$

- Community: group of nodes that have more edges within than across
- The upper triangle of A are independent Bernoulli
- Let $W = A - \mathbb{E}[A]$. For a rank-K matrix Ω,

$$A = \underbrace{\Omega}_{\text{main signal}} - \underbrace{\text{diag}(\Omega)}_{\text{secondary signal}} + \underbrace{W}_{\text{noise}}$$

Detecting a small community

$$H_0 : K = 1 \qquad vs. \qquad H_1 : K > 1$$

- **Focus**: Some communities are very small
- E.g., Testing whether there is a small focused group in a large coauthorship network
- Includes **clique detection** as a special case

  (Alon et al, 1998; Arias-Castro–Verzelen, 2014)

Degree-Corrected Block Model (DCBM)

$$\Omega(i,j) = \theta_i\theta_j \cdot \pi_i' P \pi_j, \qquad \Longleftrightarrow \qquad \Omega = \Theta\Pi P \Pi' \Theta$$

- $\Pi = [\pi_1, \ldots, \pi_n]' \in \mathbb{R}^{n,K}$, where $\pi_i \in \mathbb{R}^K$ models community label of node i: when $i \in \mathcal{C}_k$, $\pi_i(\ell) = 1$ if $\ell = k$ and $\pi_i(\ell) = 0$ otherwise.
- $P \in \mathbb{R}^{K,K}$ models community structure: $P(k,\ell)$ is the baseline connecting probability for communities k & ℓ.
- $\Theta = \mathrm{diag}(\theta_1, \ldots, \theta_n)$, where $\theta_i > 0$ models degree heterogeneity of node i
- Reduces to stochastic block model if $\theta_i \equiv 1$.

Degree matching, why χ^2 may lose power

Jin, Ke, Luo (2021)

$$\chi^2\text{-test}: \qquad X = \sum_{i=1}^{n} (d_i - \bar{d})^2, \quad d_i \text{ is degree of node } i$$

- χ^2 is powerful in degree-homogeneous models (SBM) *(Arias-Castro–Verzelen, 2014)*
- Why does it lose power in DCBM?
  - *Degree matching:* Can pair any alternative DCBM with a null model that has the same degree profile (in expectation)

Degree matching, why χ^2 may lose power

Jin, Ke, Luo (2021)

$$\chi^2\text{-test}: \qquad X = \sum_{i=1}^{n} (d_i - \bar{d})^2, \quad d_i \text{ is degree of node } i$$

- χ^2 is powerful in degree-homogeneous models (SBM) *(Arias-Castro–Verzelen, 2014)*
- Why does it lose power in DCBM?
    - *Degree matching:* Can pair any alternative DCBM with a null model that has the same degree profile (in expectation)
    - Proof: Sinkhorn's matrix scaling theorem

The sub-DCBM with $K = 2$

sub-DCBM: **s**everely **u**n-**b**alanced DCBM

As $K = 2$, we only have two communities, $\mathcal{C}_0$ and $\mathcal{C}_1$. For $m \ll n$, suppose

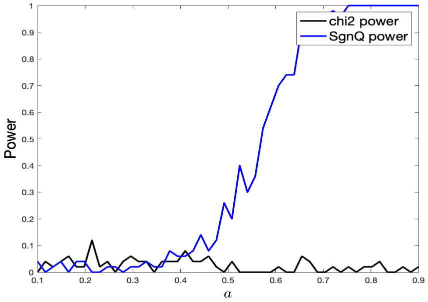
$$P(\text{node } i \text{ belongs to community } \mathcal{C}_k) = \begin{cases} 1 - m/n, & k = 0, \\ m/n, & k = 1 \end{cases}$$

$$\Omega(i,j) = \begin{cases} \theta_i\theta_j \cdot a, & \text{if } i,j \in \mathcal{C}_1, \\ \theta_i\theta_j \cdot c, & \text{if } i,j \in \mathcal{C}_0, \\ \theta_i\theta_j \cdot b, & \text{otherwise.} \end{cases} \qquad \text{where } b = \tfrac{nc-(a+c)m}{n-2m}.$$

Lemma: This parameterizes all sub-DCBMs with $K = 2$.

The sub-DCBM with $K = 2$, II

- We can pair the sub-DCBM with a null model $\Omega(i,j) = \alpha \cdot \theta_i \theta_j$ so that for each node, the expected degrees under the null and alternative are matched
- Naive degree-based χ^2-test is asymptotically powerless
- The SgnQ test (TBA) has much better power



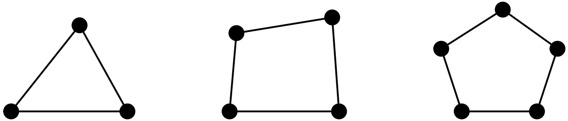
Simulation setting: $(n, m, c) = (100, 10, 0.1)$, $\theta_i \equiv 1$

Next

- The naive χ^2-test may lose power
- **Question**: How to find a *fast*, *broadly implementable*, and *powerful* test?
- **Answer**. The SgnQ test (TBA)
- Proposed by Jin, Ke, Luo (2021) but never studied for severely unbalanced DCBM

Note. Most of our ideas work for general DCBM, but we focus on sub-DCBM for completeness.

The Signed-Quadrilateral (SgnQ) test



- $C_n^{(m)} = \sum_{i_1, i_2, \ldots, i_m (distinct)} A_{i_1 i_2} A_{i_2 i_3} \ldots A_{i_m i_1} = \#\{\text{m-}gons\}$
- Inspired by this, let $m = 4$, $\hat{\eta} = (\mathbf{1}_n A \mathbf{1}_n)^{-1/2} A \mathbf{1}_n$, and apply the cycle count idea above to $\widehat{A} = A - \widehat{\eta}\widehat{\eta}'$:

$$Q_n = \sum_{i,j,k,\ell((distinct)} \widehat{A}_{ij} \widehat{A}_{jk} \widehat{A}_{k\ell} \widehat{A}_{\ell i}$$

- The SgnQ test statistic is

$$\psi_n = \frac{Q_n - 2(\|\hat{\eta}\|^2 - 1)^2}{\sqrt{8(\|\hat{\eta}\|^2 - 1)^4}}$$

Parameter-free limiting null (SgnQ test)

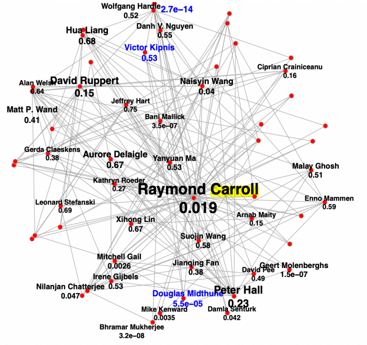
Theorem. Suppose $\Omega = \alpha\theta\theta'$, where $\|\theta\|_1 = n$, $n\alpha \to \infty$, and $\alpha\theta^2_{\max}\log(n^2\alpha) \to 0$. As $n \to \infty$, $\psi_n \to N(0,1)$ in distribution.

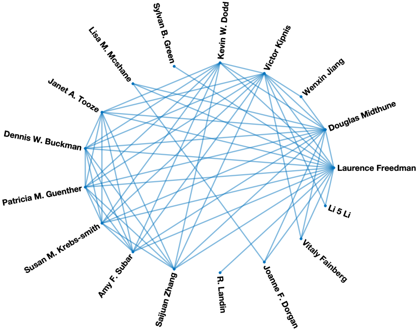
Proof. Mild adaptation of Jin, Ke, Luo (2021).

- **Nontriviality**: DCBM has numerous unknown parameters. It took years' efforts to find a test with an explicit and parameter-free limiting null distribution
- **Applications**: We can obtain an approximate p-value and use it for (a) measuring co-authorship diversity and (b) setting termination rule in a recursive/hierarchical community detection scheme

Carroll's personalized coauthor network

Data: Paper attributes in 36 journals, 1975-2015 *(Ji-Jin-Ke-Li, 2022)*.
The coauthorship network restricted to Carroll and his co-authors.





Left: Carroll's network (only nodes with > 40 degrees are shown). The SgnQ p-value is 0.019. **Right:** An identified small community of 17 authors. Restricted to this sub-network, the SgnQ p-value is 0.682.

Power of the SgnQ test

- $\mathrm{Var}(Q_n) \approx (\|\widehat{\eta}\|^2 - 1)^4 \approx \lambda_1^4$, and by Weyl's theorem, we can't use a rank-1 matrix to well-approx. a rank-K one:

$$Q_n = \sum_{i_1, i_2, i_3, i_4 (distinct)} \widehat{A}_{i_1 i_2} \widehat{A}_{i_2 i_3} \widehat{A}_{i_3 i_4} \widehat{A}_{i_4 i_1}$$

$$\approx \mathrm{trace}([\Omega - \widehat{\eta}\widehat{\eta}']^4) \geq C \sum_{k=2}^{K} \lambda_k^4.$$

- Therefore, the power of the SgnQ test hinges on

$$\frac{\sum_{k=2}^{K} \lambda_k^4}{\lambda_1^2} \asymp \Big(\frac{\lambda_2}{\sqrt{\lambda_1}}\Big)^4$$

Power of SgnQ under the sub-DCBM

In the sub-DCBM with $K = 2$, $|\mathcal{C}_1| = m$ and $|\mathcal{C}_0| = n - m$, and

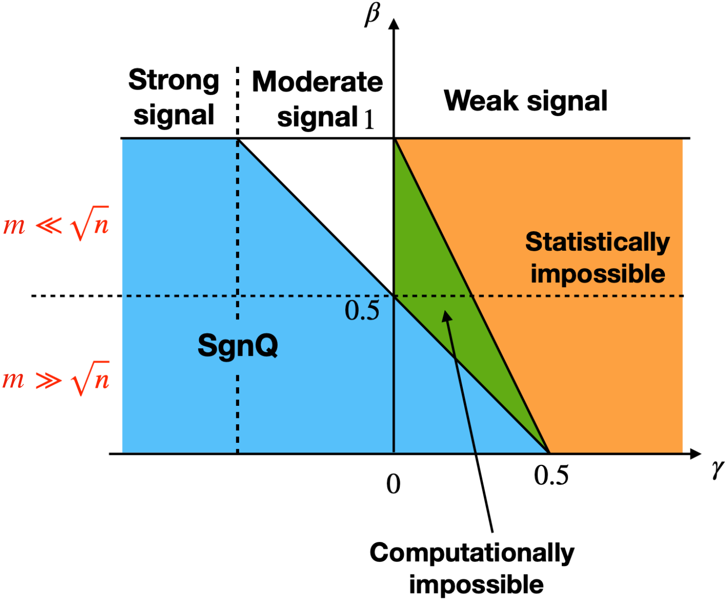
$$\Omega(i,j) = \begin{cases} \theta_i\theta_j \cdot a, & \text{if } i,j \in \mathcal{C}_1, \\ \theta_i\theta_j \cdot c, & \text{if } i,j \in \mathcal{C}_0, \\ \theta_i\theta_j \cdot b, & \text{otherwise.} \end{cases} \qquad \text{where } b = \tfrac{nc-(a+c)m}{n-2m}.$$

Theorem. Consider a sub-DCBM with $K = 2$ where $\theta_{\max} \leq C\theta_{\min}$ and $nc \to \infty$. Suppose as $n \to \infty$,

$$m(a-c)/\sqrt{cn} \to \infty,$$

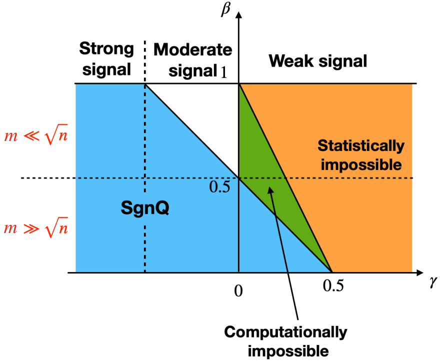
- For any fixed α, the power of level-α SgnQ test $\to 1$.
- Theorem also holds under **severe** degree-heterogeneity (with appropriate regularity conditions)

Next: Phase transition



Phase transition

- **Computationally easy**: There is a polynomial-time test whose sum of type I and type II errors $\to 0$.
- **Statistically possible but computationally hard**: For any poly-time test, sum of type I and type II errors $\to 1$.
- **Statistically impossible**: For any test, the sum of type I and type II errors $\to 1$.



Phase transition (sub-DCBM, $m \gg \sqrt{n}$)

Theorem. In the sub-DCBM with $K = 2$, assume $\theta_{\max} \leq C\theta_{\min}$ and $nc \to \infty$. As $n \to \infty$,

> **Easy:** if $m(a-c)/\sqrt{nc} \to \infty$, then SgnQ test satisfies Type I + Type II error $\to 0$
>
> **Hard:** if $m(a-c)/\sqrt{nc} \to 0$, no poly-time test exists[1] with Type I + Type II error $\to 0$
>
> **Impossible:** if $\sqrt{\frac{n}{m}} \cdot m(a-c)/\sqrt{nc} \to 0$, no test (even non-polytime) has Type I + Type II error $\to 0$

Therefore, there is a gap between **statistically possible** and **computationally possible**, but fortunately, the SgnQ test is optimal among all polynomial time tests.

---

[1] conditionally on the *low-degree conjecture* (Hopkins, 2018)

Phase transition (sub-DCBM, $m \ll \sqrt{n}$)

The case of $m \ll \sqrt{n}$ is more complicated, and how to close the gap between **statistically possible** and **computationally possible** remains an open problem

Theorem. In the sub-DCBM model with $K = 2$, assume $\theta_{\max} \leq C\theta_{\min}$ and $nc \to \infty$. As $n \to \infty$,

> **Easy:** if $m(a-c)/\sqrt{nc} \to \infty$, then SgnQ test satisfies Type I + Type II error $\to 0$
>
> **Hard:** if $\frac{\sqrt{n}}{m} \cdot m(a-c)/\sqrt{nc} \to 0$, no poly-time test exists[2] with Type I + Type II error $\to 0$
>
> **Impossible:** if $\sqrt{\frac{n}{m}} \cdot m(a-c)/\sqrt{nc} \to 0$, no test (even non-polytime) has Type I + Type II error $\to 0$

---

[2]conditionally on the *low-degree conjecture* (Hopkins, 2018)

Phase transition visualization

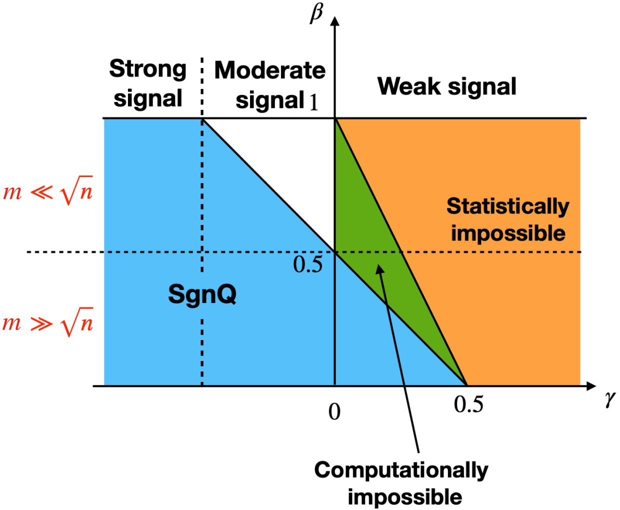
Recall that in the sub-DCBM with $K = 2$, $|\mathcal{C}_1| = m$ and $|\mathcal{C}_0| = n - m$, and

$$\Omega(i,j) = \begin{cases} \theta_i\theta_j \cdot a, & \text{if } i,j \in \mathcal{C}_1, \\ \theta_i\theta_j \cdot c, & \text{if } i,j \in \mathcal{C}_0, \\ \theta_i\theta_j \cdot b, & \text{otherwise.} \end{cases} \qquad \text{where } b = \tfrac{nc-(a+c)m}{n-2m}.$$

For visualization, we fix parameters $\beta, \gamma \in (0,1)$ and let

$$\textbf{Small community size} : m = n^{1-\beta}$$
$$\textbf{Nodewise SNR} : \frac{a-c}{\sqrt{c}} = n^{-\gamma}.$$

Phase transition visualization, II



Orange: $\beta + 2\gamma > 1/2$. Blue: $\beta + \gamma < 1/2$.
Green: $\beta + 2\gamma < 1/2$, $\beta + \gamma > 1/2$, and $\gamma > 0$.

Take home message

- The SgnQ test is fast, has computable p-values, and is powerful against a broad class of alternatives.
- In broad network (or hypergraph) models, degree-based tests may lose power from *degree-matching*.
- If $m \ll \sqrt{n}$, SgnQ test is the optimal polynomial time test. If $m \gg \sqrt{n}$, it is an interesting open problem to close the statistical-computational gap.

Take home message

- The SgnQ test is fast, has computable p-values, and is powerful against a broad class of alternatives.
- In broad network (or hypergraph) models, degree-based tests may lose power from *degree-matching*.
- If $m \ll \sqrt{n}$, SgnQ test is the optimal polynomial time test. If $m \gg \sqrt{n}$, it is an interesting open problem to close the statistical-computational gap.
- Thanks!