# Phase transition for detecting a small community in a large network


## Jiashun Jin[1], Zheng Tracy Ke[2], Paxton Turner[2], and Anru R. Zhang[3]

[1]Department of Statistics, **Carnegie Mellon University**, [2]Department of Statistics, **Harvard University**, and [3]Department of Biostatistics and Bioinformatics, **Duke University**


## Introduction

- **Motivation**
  Detecting a small community in a large network has many applications to statistics and computer science
- **Prior Work**
  Arias-Castro—Verzelen (2014) show that the $\chi^2$ test detects a small community planted in Erdos—Renyi background
  $$\chi^2 = \sum_{i=1}^{n} (d_i - \bar{d})^2, \quad d_i \text{ is degree of node } i$$
- **Limitations of Degree-based $\chi^2$**
  In broad models, the null can be paired to an alternative with the same degree profiles (Jin—Ke—Luo, 2021). Due to this *degree-matching*, degree-based tests lose power.
- **SgnQ: a Higher Moment-based Test**
  The SgnQ test (TBA) counts signed quadrilaterals and achieves better power

## Model and Problem

Let $A$ be the adjacency matrix of a $K$ community degree-corrected block model (DCBM) on $n$ nodes.

**Problem:** Distinguish

$$H_0 : K = 1 \text{ vs } H_1 : K > 1$$

While many of our results apply to larger $K$, we mainly focus on the *severely unbalanced DCBM* (**sub-DCBM**), defined below:

- **Null model ($K = 1$):** Suppose $\Omega_{ij} = \mathbb{P}[A_{ij} = 1] = \alpha\theta_i\theta_j$, where $\|\theta\|_1 = n$
- **Alternative model ($K = 2$):** Let $\mathscr{C}_1$ have size $m \ll n$, and let $\mathscr{C}_0$ be the remaining nodes.

$$\Omega_{ij} = \mathbb{P}[A_{ij} = 1] = \begin{cases} \theta_i\theta_j \cdot a, & \text{if } i, j \in \mathscr{C}_1 \\ \theta_i\theta_j \cdot c, & \text{if } i, j \in \mathscr{C}_0 \\ \theta_i\theta_j \cdot b, & \text{otherwise} \end{cases}$$

where $b = \left[nc - (a + c)m\right]/(n - 2m)$

**Goal**: Give a sharp analysis of SgnQ and assess its optimality via phase transitions

## SgnQ Test

### SgnQ Statistic

1. Form the centered adjacency matrix $\hat{A} = A - \hat{\eta}\hat{\eta}'$, where $\hat{\eta} = (\mathbf{1}_n A \mathbf{1}_n)^{-1/2} A \mathbf{1}_n$
2. Count the *signed quadrilaterals of* $\hat{A}$:

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

### SgnQ Test

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

Let $q_\kappa = \Phi^{-1}(1 - \kappa)$. The level-$\kappa$ SgnQ test rejects if $\psi_n > q_\kappa$ and fails to reject otherwise

## Theorems for SgnQ

- **Asymptotic Normality under Null**

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

- **Power of SgnQ**

**Theorem.** Under the alternative, suppose $\theta_{\max} \le C\theta_{\min}$ and $nc \to \infty$. Suppose as $n \to \infty$, $m(a - c)/\sqrt{cn} \to \infty$, then the power of SgnQ test $\to 1$.

### Short Analysis of SgnQ

Let $\lambda_k$ be the $k^{\text{th}}$ largest singular value of $\Omega$. $\text{Var}(Q_n) \approx (\|\hat{\eta}\|^2 - 1)^4 \approx \lambda_1^4$, and by Weyl's theorem, we can't use a rank-1 matrix to well-approximate a rank-$K$ one:

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

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

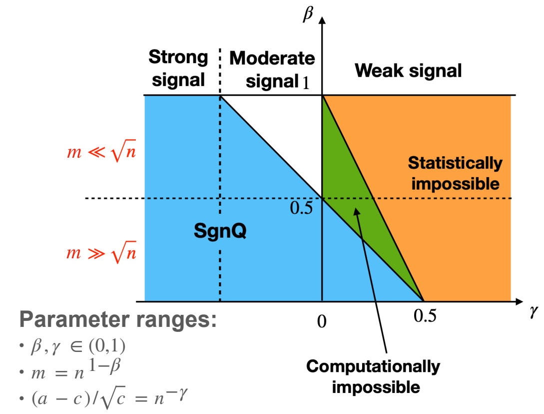
Thus, power of the SgnQ test hinges on

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

## Phase Transitions

### Three phases:

- Computationally **easy**: There is a **poly-time** test (**SgnQ**) whose sum of Type I and Type II errors → 0.
- Statistically possible but computationally **hard**: For any **poly-time** test, sum of Type I and Type II errors → 1.
- Statistically **impossible**: For any test, the sum of Type I and Type II errors → 1.




### Case 1: sub-DCBM, $m \gg \sqrt{n}$

There is a gap between **easy** and **hard**, but fortunately, the SgnQ test is optimal among all polynomial time tests.

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

- **Easy** if $m(a - c)/\sqrt{nc} \to \infty$
- **Hard** if $m(a - c)/\sqrt{nc} \to 0$
- **Impossible** if $\sqrt{n/m} \cdot m(a - c)/\sqrt{nc} \to 0$

### Case 2: sub-DCBM, $m \ll \sqrt{n}$

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

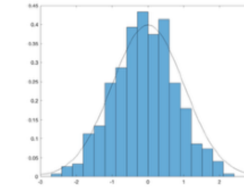
**Theorem**. In the sub-DCBM with $K = 2$, assume $\theta_{\max} \le C\theta_{\min}$ and $nc \to \infty$. As $n \to \infty$,

- **Easy** if $m(a - c)/\sqrt{nc} \to \infty$
- **Hard** if $\sqrt{n/m} \cdot m(a - c)/\sqrt{nc} \to 0$
- **Impossible** if $\sqrt{n/m} \cdot m(a - c)/\sqrt{nc} \to 0$

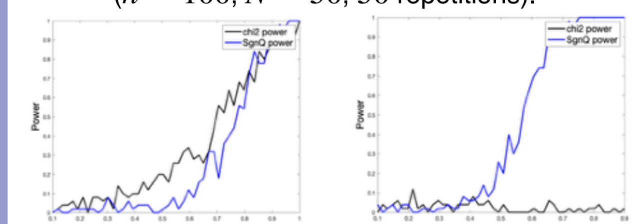
## Numerical Experiments

### Simulated Data

- Null distribution of SgnQ ($n = 500$).

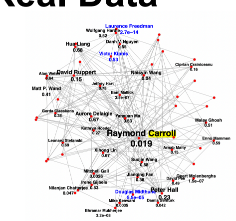


- Power comparison of SgnQ and $\chi^2$ ($n = 100$, $N = 50$, 50 repetitions).




2-community SBM with $P_{11} = a$, $P_{22} = 0.1$, $P_{12} = 0.1$ (left) and $P_{12} = \left[an - (a + 0.1)N\right]/n$ (right).

### Real Data




R. Carroll's personal co-authorship network (Ji et al, 2022)

A small sub-community of 17 authors whose SgnQ p-value is 0.6818.

Raymond Carroll's network has SgnQ p-value 0.019. The right sub-community is well-connected. SgnQ may serve as a *splitting criterion* for hierarchical community detection.

## Bibliography


1. E. Arias-Castro and N. Verzelen. Community detection in dense random networks. *Ann. Stat.*, 42(3):940–969, 2014.
2. K. Bogerd, R. M. Castro, R. van der Hofstad, and N. Verzelen. Detecting a planted community in an inhomogeneous random graph. *Bernoulli*, 27(2):1159–1188, 2021.
3. B. Hajek, Y. Wu, and J. Xu. Computational lower bounds for community detection on random graphs. *COLT*, pp. 899–928. PMLR, 2015.
4. S. B. Hopkins and D. Steurer. Efficient bayesian estimation from few samples: community detection and related problems. *FOCS*, pp. 379–390. IEEE, 2017.
5. P. Ji, J. Jin, Z.T. Ke, and W. Li. Co-citation and co-authorship networks for statisticians (with discussions). *J. Bus. Econ. Statist.*, 40(2), 2022.
6. J. Jin, Z. T. Ke, and S. Luo. Network global testing by counting graphlets. *ICML*, 2018.
7. J. Jin, Z. T. Ke, and S. Luo. Optimal adaptivity of signed-polygon statistics for network testing. *Ann. Stat.*, 49(6):3408–3433, 2021b.