Global community detection using individual-centered partial information networks

Rachel Wang

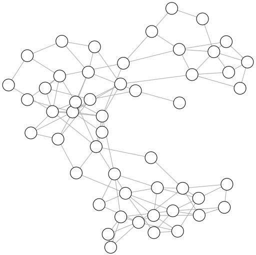
Joint work with Xiao Han and Xin Tong

School of Mathematics and Statistics, University of Sydney

Statistics in the Big Data Era, UC Berkeley, 2022

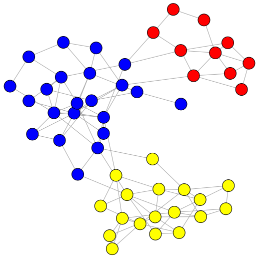
Social network modeling

- $G = (\mathcal{V}, \mathcal{E})$
  - $\mathcal{V}$: set of individuals $[n] := \{1, \ldots, n\}$
  - $\mathcal{E}$: set of edges, assumed to be binary and undirected for simplicity
- Statistical problems: estimating community memberships, subgraph counts, node covariates, ...
- Most of the current literature assumes either a global view of the network or multiple subgraphs (Mukherjee et al. 2021) can be sampled, from which information can be combined.



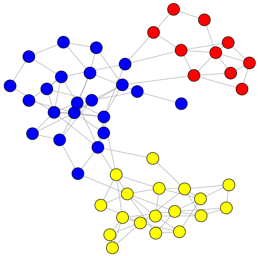
Social network modeling

- $G = (\mathcal{V}, \mathcal{E})$
  - $\mathcal{V}$: set of individuals $[n] := \{1, \ldots, n\}$
  - $\mathcal{E}$: set of edges, assumed to be binary and undirected for simplicity
- Statistical problems: estimating community memberships, subgraph counts, node covariates, ...
- Most of the current literature assumes either a global view of the network or multiple subgraphs (Mukherjee et al. 2021) can be sampled, from which information can be combined.

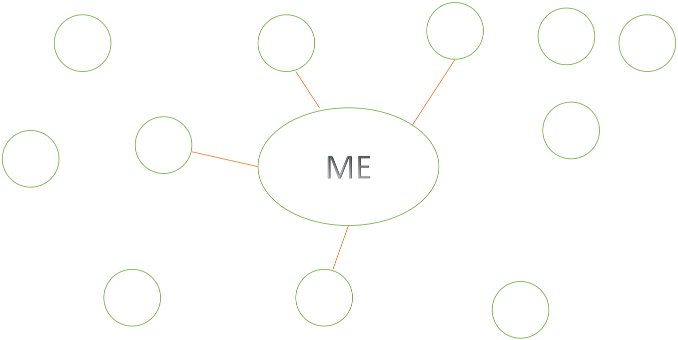


Social network modeling

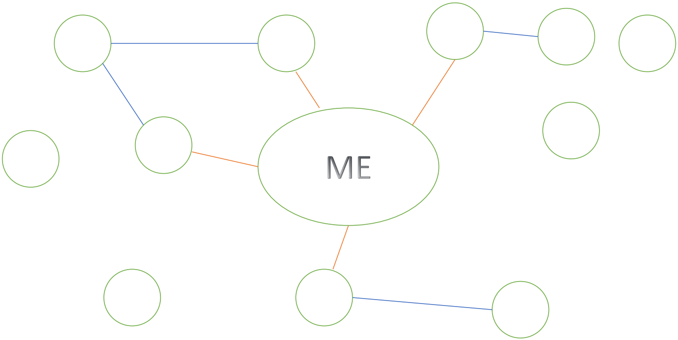
- $G = (\mathcal{V}, \mathcal{E})$
  - $\mathcal{V}$: set of individuals $[n] := \{1, \ldots, n\}$
  - $\mathcal{E}$: set of edges, assumed to be binary and undirected for simplicity
- Statistical problems: estimating community memberships, subgraph counts, node covariates, ...
- Most of the current literature assumes either a global view of the network or multiple subgraphs (Mukherjee et al. 2021) can be sampled, from which information can be combined.



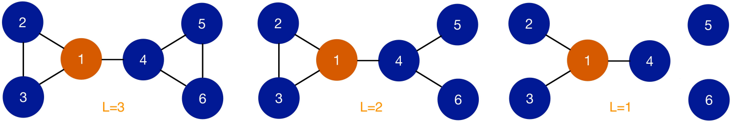
How much does a person understand about the connections in the full network?



Beyond friends?



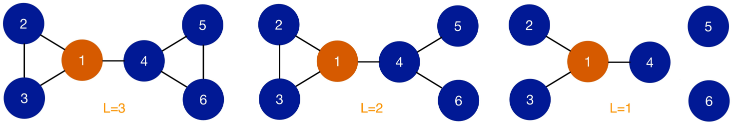
A toy example of individual-centered partial information networks



- An illustration of a network consisting of 6 individuals.
- The left panel is the full network.
- Suppose individual 1 is the person of interest.
- The left, middle and right panel show individual 1's view of the network when their knowledge depth is $L = 3, 2, 1$, respectively.

Key: Characterizing the amount of partial (local) information by path length. We focus on $L = 2$.

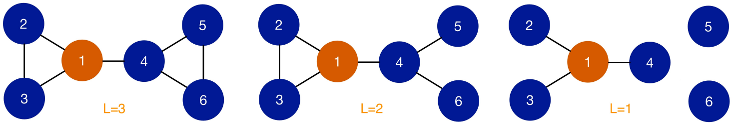
A toy example of individual-centered partial information networks



- An illustration of a network consisting of 6 individuals.
- The left panel is the full network.
- Suppose individual 1 is the person of interest.
- The left, middle and right panel show individual 1's view of the network when their knowledge depth is $L = 3, 2, 1$, respectively.

Key: Characterizing the amount of partial (local) information by path length. We focus on $L = 2$.

A toy example of individual-centered partial information networks



- An illustration of a network consisting of 6 individuals.
- The left panel is the full network.
- Suppose individual 1 is the person of interest.
- The left, middle and right panel show individual 1's view of the network when their knowledge depth is $L = 3, 2, 1$, respectively.

Q: Can one learn the global community memberships based on their partial knowledge of the full network?

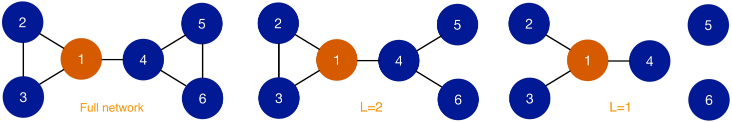
Existing literature

- The structure of our partial network is related to existing network sampling schemes, e.g., egocentric sampling, snowball sampling and respondent driven sampling (RDS).
- Main differences:
  - We are interested in what one instance of partial network can offer, which allows us to compare what network structure is visible to each individual.
  - Most RDS based methods are focused on estimating node covariates, while we are interested in latent community structure.
  - Multiple sampling may not be feasible in networks with restricted access (e.g., a terrorist network)

Preliminaries

- Recall $G = (\mathcal{V}, \mathcal{E})$ is the full network of n individuals.
- G can be represented by a $n \times n$ binary, symmetric adjacency matrix $\mathbf{A} = (a_{ij})$, where $a_{ij} = \begin{cases} 1, & (i,j) \in \mathcal{E}\,, \\ 0, & (i,j) \notin \mathcal{E}\,. \end{cases}$
- Let $\mathbf{B} = (b_{ij})$ be individual 1's perceived adjacency matrix based on knowledge depth $L = 2$.

How is **B** related to **A**?



In this toy example, for individual 1 with knowledge depth $L = 2$:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

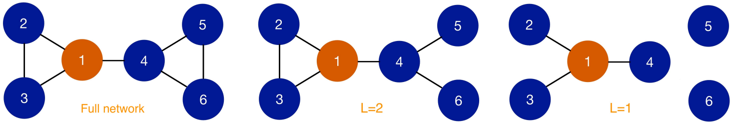
In general,

$$b_{ij} = a_{ij}(1 - \mathbb{I}(a_{1i} = 0)\mathbb{I}(a_{1j} = 0)).$$

It follows that

$$\mathbf{B} = -\mathbf{SAS} + \mathbf{AS} + \mathbf{SA}, \quad \text{where} \quad \mathbf{S} = \operatorname{diag}(a_{11}, \ldots, a_{1n}).$$

How is $\mathbf{B}$ related to $\mathbf{A}$?



In this toy example, for individual 1 with knowledge depth $L = 2$:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

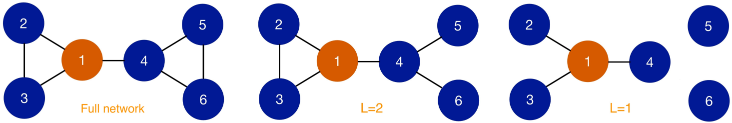
In general,

$$b_{ij} = a_{ij}(1 - \mathbb{1}(a_{1i} = 0)\mathbb{1}(a_{1j} = 0)).$$

It follows that

$$\mathbf{B} = -\mathbf{SAS} + \mathbf{AS} + \mathbf{SA}, \quad \text{where} \quad \mathbf{S} = \text{diag}(a_{11}, \ldots, a_{1n}).$$

How is $\mathbf{B}$ related to $\mathbf{A}$?



In this toy example, for individual 1 with knowledge depth $L = 2$:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

In general,

$$b_{ij} = a_{ij}(1 - \mathbb{I}(a_{1i} = 0)\mathbb{I}(a_{1j} = 0)).$$

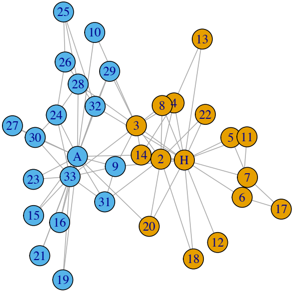
It follows that

$$\mathbf{B} = -\mathbf{SAS} + \mathbf{AS} + \mathbf{SA}, \quad \text{where} \quad \mathbf{S} = \text{diag}(a_{11}, \ldots, a_{1n}).$$

A peak at the results

individual of interest	H	2	3	A	20	32
degrees	16	9	10	17	3	6
fraction of edges	.654	.513	.705	.641	.526	.654
detection accuracy	.559	.706	.941	.706	.941	.794

Zachary's karate club



Preliminaries - low rank assumption

- We assume $\text{rank}(\mathbb{E}\mathbf{A}) = K$ and
- the eigen decomposition (reduced form) $\mathbb{E}\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$.
- $\mathbf{D} = \text{diag}(d_1, \ldots, d_K)$, in which d_i is the i-th largest eigenvalue (by magnitude) of $\mathbb{E}\mathbf{A}$,
- $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_K)$ is the corresponding eigenvector matrix.
- Generate $\mathbf{A}$ as independent Bernoulli from $\mathbb{E}\mathbf{A}$.
- Usually theoretical analysis usually proceeds by noting

$$\mathbf{A} = \underbrace{\mathbb{E}\mathbf{A}}_{\text{signal}} + \underbrace{(\mathbf{A} - \mathbb{E}\mathbf{A})}_{\text{noise}}.$$

Preliminaries - low rank assumption

- We assume $\text{rank}(\mathbb{E}\mathbf{A}) = K$ and
- the eigen decomposition (reduced form) $\mathbb{E}\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$.
- $\mathbf{D} = \text{diag}(d_1, \dots, d_K)$, in which d_i is the i-th largest eigenvalue (by magnitude) of $\mathbb{E}\mathbf{A}$,
- $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_K)$ is the corresponding eigenvector matrix.
- Generate $\mathbf{A}$ as independent Bernoulli from $\mathbb{E}\mathbf{A}$.
- Usually theoretical analysis usually proceeds by noting

$$\mathbf{A} = \underbrace{\mathbb{E}\mathbf{A}}_{\text{signal}} + \underbrace{(\mathbf{A} - \mathbb{E}\mathbf{A})}_{\text{noise}}.$$

Preliminaries - low rank assumption

What about $\mathbf{B} = -\mathbf{SAS} + \mathbf{AS} + \mathbf{SA}$? We can show that

$$\mathbf{B}_E = -\mathbf{S}(\mathbb{E}\mathbf{A})\mathbf{S} + (\mathbb{E}\mathbf{A})\mathbf{S} + \mathbf{S}(\mathbb{E}\mathbf{A})$$

is the "signal" term in the sense that

$$\|\mathbf{B} - \mathbf{B}_E\| \leq \text{smallest singular value of } \mathbf{B}_E$$

.

Theoretical properties of $\mathbf{B}_E$

Theorem 1 (Informal, eigenvalues and eigenvectors)

Suppose that $\mathbf{V}^\top\mathbf{S}\mathbf{V}$ and $\mathbf{I}-\mathbf{V}^\top\mathbf{S}\mathbf{V}$ are invertible. We have $\mathrm{rank}(\mathbf{B}_E)=2K$. Then for $i=-K,\ldots,-1,1,\ldots,K$, $(x_i^{-1},\mathbf{q}_i)$ is an eigenvalue / eigenvector pair of $\mathbf{B}_E$, iff

- x_i is a solution of $\det(\mathbf{H}(x))=0$, where $\mathbf{H}(x)=\mathbf{I}-x\mathbf{D}\mathbf{V}^\top\mathbf{S}\mathbf{V}-x^2\mathbf{D}(\mathbf{I}-\mathbf{V}^\top\mathbf{S}\mathbf{V})\mathbf{D}\mathbf{V}^\top\mathbf{S}\mathbf{V}$.
- $\mathbf{q}_i=\mathbf{S}\mathbf{V}\mathbf{q}_{1i}+(\mathbf{I}-\mathbf{S})\mathbf{V}\mathbf{q}_{2i}$, where $\mathbf{q}_{1i}$ is an eigenvector of $\mathbf{H}(x_i)$ corresponding to the zero eigenvalue, and

$$\mathbf{q}_{2i}=x_i\mathbf{D}\mathbf{V}^\top\mathbf{S}\mathbf{V}\mathbf{q}_{1i}.$$

Let $\mathbf{Q}=(\mathbf{q}_K,\ldots,\mathbf{q}_1,\mathbf{q}_{-1},\ldots,\mathbf{q}_{-K})=\mathbf{S}\mathbf{V}\mathcal{Q}_1+(\mathbf{I}-\mathbf{S})\mathbf{V}\mathcal{Q}_2$.
We can choose $\mathbf{q}_i$'s such that

$$\mathbf{Q}^\top\mathbf{Q}=\mathbf{I}_{2K\times 2K}.$$

Theoretical properties of $\mathbf{B}_E$

Theorem 1 (Informal, eigenvalues and eigenvectors)

Suppose that $\mathbf{V}^\top \mathbf{S} \mathbf{V}$ and $\mathbf{I} - \mathbf{V}^\top \mathbf{S} \mathbf{V}$ are invertible. We have $\text{rank}\,(\mathbf{B}_E) = 2K$. Then for $i = -K, \ldots, -1, 1, \ldots, K$, $(x_i^{-1}, \mathbf{q}_i)$ is an eigenvalue / eigenvector pair of $\mathbf{B}_E$, iff

- x_i is a solution of $\det(\mathbf{H}(x)) = 0$, where $\mathbf{H}(x) = \mathbf{I} - x\mathbf{D}\mathbf{V}^\top \mathbf{S} \mathbf{V} - x^2 \mathbf{D}(\mathbf{I} - \mathbf{V}^\top \mathbf{S} \mathbf{V})\mathbf{D}\mathbf{V}^\top \mathbf{S} \mathbf{V}$.
- $\mathbf{q}_i = \mathbf{S}\mathbf{V}\mathbf{q}_{1i} + (\mathbf{I} - \mathbf{S})\mathbf{V}\mathbf{q}_{2i}$, where $\mathbf{q}_{1i}$ is an eigenvector of $\mathbf{H}(x_i)$ corresponding to the zero eigenvalue, and

$$\mathbf{q}_{2i} = x_i \mathbf{D}\mathbf{V}^\top \mathbf{S} \mathbf{V} \mathbf{q}_{1i} \,.$$

Let $\mathbf{Q} = (\mathbf{q}_K, \ldots, \mathbf{q}_1, \mathbf{q}_{-1}, \ldots, \mathbf{q}_{-K}) = \mathbf{S}\mathbf{V}\mathcal{Q}_1 + (\mathbf{I} - \mathbf{S})\mathbf{V}\mathcal{Q}_2$.
We can choose $\mathbf{q}_i$'s such that

$$\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}_{2K \times 2K}.$$

Theoretical properties of $\mathbf{B}_E$

Let $p_n = \max_{i,j} \mathbb{P}(a_{ij} = 1)$, assume $\min_{j \geq 2} \mathbb{P}(a_{1j} = 1) \sim p_n$, $1 - c > p_n \gg \log n/n$ for some constant $c > 0$.

Theorem 2 (Informal, form of eigenvalues)

With mild conditions on $\mathbf{D}$, $\mathbf{V}$, w.h.p., we have

- $|x_i|^{-1} \sim np_n^{3/2}$ for $i \in [\pm K]$;
- if $p_n \to 0$, with additional conditions, we obtain the high probability expressions of x_i^{-1};
- the expressions of x_i^{-1} suggest $\lambda_{\min} = \lambda_K(\mathbf{V}^\top \mathbf{S} \mathbf{V})$ determines the gap between the smallest eigenvalue (in magnitude) and 0.

Theoretical properties of $\mathbf{B}_E$

Summary so far

- The eigenvectors take the form
  $\mathbf{Q} = (\mathbf{q}_K, \ldots, \mathbf{q}_1, \mathbf{q}_{-1}, \ldots, \mathbf{q}_{-K}) = \underbrace{\mathbf{S}\mathbf{V}\mathcal{Q}_1}_{\text{neighbors of node 1}} + \underbrace{(\mathbf{I} - \mathbf{S})\mathbf{V}\mathcal{Q}_2}_{\text{non-neighbors}}.$
- Order of the eigenvalues, $|x_i|^{-1} \sim np_n^{3/2}$ w.h.p.
- $\lambda_{\min} = \lambda_K(\mathbf{V}^\top \mathbf{S}\mathbf{V})$ detemines the signal strength.
  - $\lambda_{\min}$ influences the performance of spectral clustering $\Rightarrow$ measure of how important individual 1 is $\Rightarrow$ centrality measure
  - $\lambda_{\min}$ lies between 0 and 1.
  - Can be estimated using empirical version of $\mathbf{V}$ from $\mathbf{A}$.
  - When $K = 1$, $\lambda_{\min}$ bears connections to both degree centrality and eigenvector centrality.

Theoretical properties of $\mathbf{B}_E$

Summary so far

- The eigenvectors take the form
  $\mathbf{Q} = (\mathbf{q}_K, \ldots, \mathbf{q}_1, \mathbf{q}_{-1}, \ldots, \mathbf{q}_{-K}) = \underbrace{\mathbf{S}\mathbf{V}\mathcal{Q}_1}_{\text{neighbors of node 1}} + \underbrace{(\mathbf{I} - \mathbf{S})\mathbf{V}\mathcal{Q}_2}_{\text{non-neighbors}}.$
- Order of the eigenvalues, $|x_i|^{-1} \sim np_n^{3/2}$ w.h.p.
- $\lambda_{\min} = \lambda_K(\mathbf{V}^\top\mathbf{S}\mathbf{V})$ detemines the signal strength.
  - $\lambda_{\min}$ influences the performance of spectral clustering $\Rightarrow$ measure of how important individual 1 is $\Rightarrow$ centrality measure
  - $\lambda_{\min}$ lies between 0 and 1.
  - Can be estimated using empirical version of $\mathbf{V}$ from $\mathbf{A}$.
  - When $K = 1$, $\lambda_{\min}$ bears connections to both degree centrality and eigenvector centrality.

Introducing a concrete model

The stochastic block model (SBM, Holland et al. 1983)

- $\mathbb{E}\mathbf{A} = \mathbf{\Pi}\mathbf{P}\mathbf{\Pi}^\top$, where $\mathbf{P} = (p_{kl})$ is a symmetric $K \times K$ matrix, $\mathbf{\Pi} = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n)^\top \in \mathbb{R}^{n\times K}$ and individual i's membership vector $\boldsymbol{\pi}_i \in \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$.
- In this model, when individual i belongs to community k and individual j belongs to community l, we have

$$\mathbb{P}(a_{ij} = 1) = \mathbb{E}a_{ij} = \boldsymbol{\pi}_i^\top \mathbf{P}\boldsymbol{\pi}_j = p_{kl}\,.$$

- Idea: there are $2K$ types of rows in $\mathbf{Q}$ (eigenvectors of $\mathbf{B}_E$); $\mathbf{Q}$ is close to the empirical version $\mathbf{W}$, whose columns are eigenvectors of observed $\mathbf{B}$.

Rationale behind our algorithm

Recall $p_n = \max_{i,j} \mathbb{P}(a_{ij} = 1)$.

Condition

$\min_{k\in[K]} p_{1k} \sim p_n$. $\min_{k\in[K]} \sum_{j\in[n]} \mathbb{I}(\boldsymbol{\pi}_j = \mathbf{e}_k) \geq cn$ and $\sigma_K(\mathbf{P}) \geq cp_n$. Moreover, $1 - c \geq p_n \gg (1/n)^{1/2}$.

Lemma 1 ($2K$ different rows in $\mathbf{Q}$)

For any $2K \times 2K$ orthogonal matrix $\mathbf{O}$, it holds w.h.p. that for $i, j \in [n]$,

$$\boldsymbol{\pi}_i \neq \boldsymbol{\pi}_j \Longrightarrow \left\| \underbrace{\mathbf{Q}(i)}_{1\times 2K} \mathbf{O} - \underbrace{\mathbf{Q}(j)}_{1\times 2K} \mathbf{O} \right\|_2 \geq \sqrt{\frac{2}{cn}},$$

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_j, a_{1i} \neq a_{1j} \Longrightarrow \|\mathbf{Q}(i)\mathbf{O} - \mathbf{Q}(j)\mathbf{O}\|_2 \geq \sqrt{\frac{2}{cn}},$$

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_j, a_{1i} = a_{1j} \Longrightarrow \|\mathbf{Q}(i)\mathbf{O} - \mathbf{Q}(j)\mathbf{O}\|_2 = 0.$$

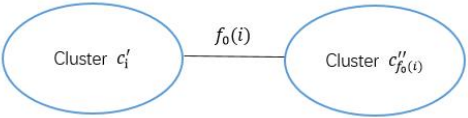
Main algorithm for SBM

1. $\underbrace{\{\mathbf{W}(i) : a_{1i} = 1\}}_{\text{apply } k-\text{means}}$ and $\underbrace{\{\mathbf{W}(i) : a_{1i} = 0\}}_{\text{apply } k-\text{means}}$ $\rightarrow$ return $2K = K + K$ clusters.
2. Merge the $2K$ clusters $\{c'_1, \ldots, c'_K\}$ and $\{c''_1, \ldots, c''_K\}$ into K communities.

Some intuition about the merging step: for $i, j > 1$ and $i \neq j$,

$$\underbrace{\mathbb{P}(b_{ij} = 1 | a_{1i} = a_{1j} = 1)}_{\rightarrow \mathbf{P}^{\mathbf{S},\mathbf{S}}} = \underbrace{\mathbb{P}(b_{ij} = 1 | a_{1i} = 1, a_{1j} = 0)}_{\rightarrow \mathbf{P}^{\mathbf{S},\mathbf{I}-\mathbf{S}}} = \underbrace{\mathbb{P}(a_{ij} = 1)}_{\rightarrow \mathbf{P}} .$$

- Clusters are identifiable only up to label permutation.
- Find the "best" permutation $f_0 : [K] \rightarrow [K]$ to *match* $\widehat{\mathbf{P}}^{\mathbf{S},\mathbf{S}}$ and $\widehat{\mathbf{P}}^{\mathbf{S},\mathbf{I}-\mathbf{S}}$.
- Merge the $2K$ clusters according to f_0.



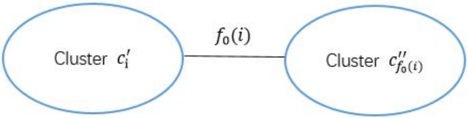
Main algorithm for SBM

1. $\underbrace{\{\mathbf{W}(i) : a_{1i} = 1\}}_{\text{apply } k-\text{means}}$ and $\underbrace{\{\mathbf{W}(i) : a_{1i} = 0\}}_{\text{apply } k-\text{means}} \rightarrow$ return $2K = K + K$ clusters.
2. Merge the $2K$ clusters $\{c'_1, \ldots, c'_K\}$ and $\{c''_1, \ldots, c''_K\}$ into K communities.

Some intuition about the merging step: for $i, j > 1$ and $i \neq j$,

$$\underbrace{\mathbb{P}(b_{ij} = 1 | a_{1i} = a_{1j} = 1)}_{\rightarrow \mathbf{P}^{\mathbf{S},\mathbf{S}}} = \underbrace{\mathbb{P}(b_{ij} = 1 | a_{1i} = 1, a_{1j} = 0)}_{\rightarrow \mathbf{P}^{\mathbf{S},\mathbf{I}-\mathbf{S}}} = \underbrace{\mathbb{P}(a_{ij} = 1)}_{\rightarrow \mathbf{P}} .$$

- Clusters are identifiable only up to label permutation.
- Find the "best" permutation $f_0 : [K] \rightarrow [K]$ to *match* $\widehat{\mathbf{P}}^{\mathbf{S},\mathbf{S}}$ and $\widehat{\mathbf{P}}^{\mathbf{S},\mathbf{I}-\mathbf{S}}$.
- Merge the $2K$ clusters according to f_0.



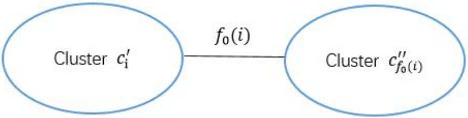
Main algorithm for SBM

1. $\underbrace{\{\mathbf{W}(i) : a_{1i} = 1\}}_{\text{apply } k-\text{means}}$ and $\underbrace{\{\mathbf{W}(i) : a_{1i} = 0\}}_{\text{apply } k-\text{means}} \rightarrow$ return $2K = K + K$ clusters.
2. Merge the $2K$ clusters $\{c'_1, \ldots, c'_K\}$ and $\{c''_1, \ldots, c''_K\}$ into K communities.

Some intuition about the merging step: for $i, j > 1$ and $i \neq j$,

$$\underbrace{\mathbb{P}(b_{ij} = 1 | a_{1i} = a_{1j} = 1)}_{\rightarrow \mathbf{P}^{\mathbf{S},\mathbf{S}}} = \underbrace{\mathbb{P}(b_{ij} = 1 | a_{1i} = 1, a_{1j} = 0)}_{\rightarrow \mathbf{P}^{\mathbf{S},\mathbf{I}-\mathbf{S}}} = \underbrace{\mathbb{P}(a_{ij} = 1)}_{\rightarrow \mathbf{P}} .$$

- Clusters are identifiable only up to label permutation.
- Find the "best" permutation $f_0 : [K] \rightarrow [K]$ to *match* $\widehat{\mathbf{P}}^{\mathbf{S},\mathbf{S}}$ and $\widehat{\mathbf{P}}^{\mathbf{S},\mathbf{I}-\mathbf{S}}$.
- Merge the $2K$ clusters according to f_0.



Consistency of algorithm

Theorem 3 (consistency under SBM)

Under some additional separation condition on $\mathbf{P}$ and $p_n \gg (\log n/n)^{1/4}$, w.h.p. ,

$$\text{Proportion of misclustered nodes} = O(\frac{1}{np_n^2}).$$

That is, the algorithm has the almost exact recovery property.

Extension to the degree-corrected SBM

Adding degree heterogeneity, the degree-corrected stochastic block model (DCSBM, Karrer and Newman 2011)

- $\mathbb{E}(\mathbf{A}|\Theta) = \Theta\mathbf{\Pi}\mathbf{P}\mathbf{\Pi}^\top\Theta$, where $\Theta = \mathrm{diag}(\theta_1, \ldots, \theta_n)$ is the set of degree parameters associated with the nodes.
- When individual i belongs to community k and individual j belongs to community l, we have

$$\mathbb{P}(a_{ij} = 1|\Theta) = \theta_i\theta_j p_{kl}\ .$$

- Assume $\theta_i \in (0, 1]$, $i \in [n]$ are i.i.d. random variables with $\mathbb{E}(\theta_i) = \theta \sim 1$.

Main algorithm for DCSBM

1. $\underbrace{\{\mathbf{W}(i) : a_{1i} = 1\}}_{\text{apply spherical } k-\text{median}}$ and $\underbrace{\{\mathbf{W}(i) : a_{1i} = 0\}}_{\text{apply spherical } k-\text{median}}$ $\rightarrow$ return $2K = K + K$ clusters.
   Spherical k-median (Lei and Rinaldo 2015): e.g., for $\{i \in [n] : a_{1i} = 1\}$,

$$\operatorname{argmin}_{\left\{\substack{\mathbf{x}_i : \mathbf{x}_i \text{ is } 2K\text{-dimensional row vector} \\ \text{with } a_{1i} = r \text{ and } |\{\mathbf{x}_i\}_{a_{1i}=1}| \leq K}\right\}} \sum_{i : a_{1i} = 1, \mathbf{W}(i) \neq 0} \left\| \frac{\mathbf{W}(i)}{\|\mathbf{W}(i)\|_2} - \mathbf{x}_i \right\|_2$$

2. Merge the $2K$ clusters $\{c'_1, \ldots, c'_K\}$ and $\{c''_1, \ldots, c''_K\}$ into K communities, accounting for degree heterogeneity.
3. The consistency result for SBM can be extended to the DCSBM setting.

Main algorithm for DCSBM

1. $\underbrace{\{\mathbf{W}(i) : a_{1i} = 1\}}_{\text{apply spherical } k-\text{median}}$ and $\underbrace{\{\mathbf{W}(i) : a_{1i} = 0\}}_{\text{apply spherical } k-\text{median}}$ $\rightarrow$ return $2K = K + K$ clusters.
   Spherical k-median (Lei and Rinaldo 2015): e.g., for $\{i \in [n] : a_{1i} = 1\}$,
   $$\operatorname{argmin}_{\left\{\substack{\mathbf{x}_i : \mathbf{x}_i \text{ is } 2K\text{-dimensional row vector} \\ \text{with } a_{1i}=r \text{ and } |\{\mathbf{x}_i\}_{a_{1i}=1}| \le K}\right\}} \sum_{i : a_{1i}=1, \mathbf{W}(i) \neq 0} \left\| \frac{\mathbf{W}(i)}{\|\mathbf{W}(i)\|_2} - \mathbf{x}_i \right\|_2$$
2. Merge the $2K$ clusters $\{c'_1, \ldots, c'_K\}$ and $\{c''_1, \ldots, c''_K\}$ into K communities, accounting for degree heterogeneity.
3. The consistency result for SBM can be extended to the DCSBM setting.

Main algorithm for DCSBM

1. $\underbrace{\{\mathbf{W}(i) : a_{1i} = 1\}}_{\text{apply spherical } k-\text{median}}$ and $\underbrace{\{\mathbf{W}(i) : a_{1i} = 0\}}_{\text{apply spherical } k-\text{median}}$ $\rightarrow$ return $2K = K + K$ clusters.
   Spherical k-median (Lei and Rinaldo 2015): e.g., for $\{i \in [n] : a_{1i} = 1\}$,

$$\text{argmin}_{\left\{\substack{\mathbf{x}_i : \mathbf{x}_i \text{ is } 2K\text{-dimensional row vector} \\ \text{with } a_{1i} = r \text{ and } |\{\mathbf{x}_i\}_{a_{1i}=1}| \le K}\right\}} \sum_{i : a_{1i} = 1, \mathbf{W}(i) \neq 0} \left\| \frac{\mathbf{W}(i)}{\|\mathbf{W}(i)\|_2} - \mathbf{x}_i \right\|_2$$

2. Merge the $2K$ clusters $\{c'_1, \ldots, c'_K\}$ and $\{c''_1, \ldots, c''_K\}$ into K communities, accounting for degree heterogeneity.
3. The consistency result for SBM can be extended to the DCSBM setting.

A conditional setting emphasizing individual differences

- Can the clustering error bound reflect neighborhood features of individual 1? Consider conditioning on the neighborhood **S**.
- θ_i generated from a two-component mixture with CDF $yF_1(x) + (1-y)F_2(x)$, $y \in (0,1)$. $\mu_2 \sim 1$ and $\mu_1 \leq \mu_2$. (non-hub vs. hub nodes)
- Let

$$n_{jk} = |\{i : a_{1i} = 1, \theta_i \text{ generated from } F_j, \text{ and } i \in \text{Community } k\}|$$

for $k = 1, \ldots, K$ and $j = 1, 2$.

Consistency of algorithm, conditional setting

Theorem 4 (consistency under DCSBM)

For $\mathbf{S}$ and Θ satisfying some constraints, and

$$\mu_1^{-1} n \sqrt{\frac{1}{p_n \min_{k\in[K]} n_{2k}^2 (n_{2k}\mu_2^2 + n_{1k}\mu_1^2)}} \ll p_n \,,$$

conditioned on $\mathbf{S}$ and Θ w.h.p.,

$$\text{Proportion of misclustered nodes} \\ = O\left(\mu_1^{-1} \sqrt{\frac{1}{p_n \min_{k\in[K]}(n_{2k}\mu_2^2 + n_{1k}\mu_1^2)}}\right).$$

- Knowing more powerful neighbors across all communities helps.
- As a centrality measure for individual, $\lambda_{\min}$ behaves like $\min_{k\in[K]} \frac{n_{1k}\mu_1^2 + n_{2k}\mu_2^2}{n}$.

Consistency of algorithm, conditional setting

Theorem 4 (consistency under DCSBM)

For $\mathbf{S}$ and Θ satisfying some constraints, and

$$\mu_1^{-1} n \sqrt{\frac{1}{p_n \min_{k\in[K]} n_{2k}^2 (n_{2k}\mu_2^2 + n_{1k}\mu_1^2)}} \ll p_n \,,$$

conditioned on $\mathbf{S}$ and Θ w.h.p.,

$$\text{Proportion of misclustered nodes} \\ = O\left(\mu_1^{-1} \sqrt{\frac{1}{p_n \min_{k\in[K]} (n_{2k}\mu_2^2 + n_{1k}\mu_1^2)}}\right).$$

- Knowing more powerful neighbors across all communities helps.
- As a centrality measure for individual, $\lambda_{\min}$ behaves like $\min_{k\in[K]} \frac{n_{1k}\mu_1^2 + n_{2k}\mu_2^2}{n}$.

Consistency of algorithm, conditional setting

Theorem 4 (consistency under DCSBM)

For **S** and Θ satisfying some constraints, and

$$\mu_1^{-1} n \sqrt{\frac{1}{p_n \min_{k\in[K]} n_{2k}^2 (n_{2k}\mu_2^2 + n_{1k}\mu_1^2)}} \ll p_n\,,$$

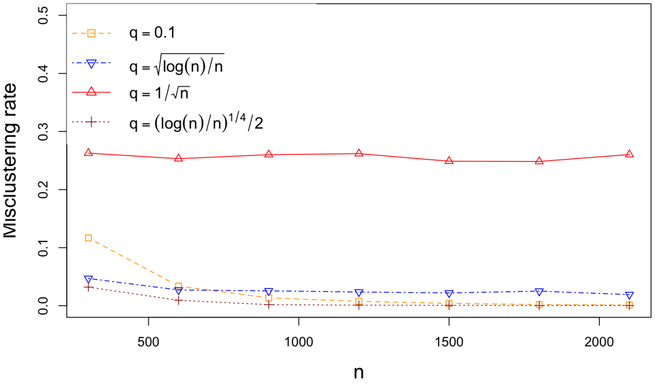
conditioned on **S** and Θ w.h.p.,

$$\text{Proportion of misclustered nodes}$$
$$= O\left(\mu_1^{-1}\sqrt{\frac{1}{p_n \min_{k\in[K]}(n_{2k}\mu_2^2 + n_{1k}\mu_1^2)}}\right).$$

- Knowing more powerful neighbors across all communities helps.
- As a centrality measure for individual, $\lambda_{\min}$ behaves like $\min_{k\in[K]} \frac{n_{1k}\mu_1^2 + n_{2k}\mu_2^2}{n}$.

Simulation with DCSBM

Setting: $\mathbf{P} = \begin{pmatrix} 3q & q \\ q & 3q \end{pmatrix}$. $K = 2$ groups have equal sizes.
$\theta_i \sim$ i.i.d. Unif(0.5,1.5)



Simulation with DCSBM

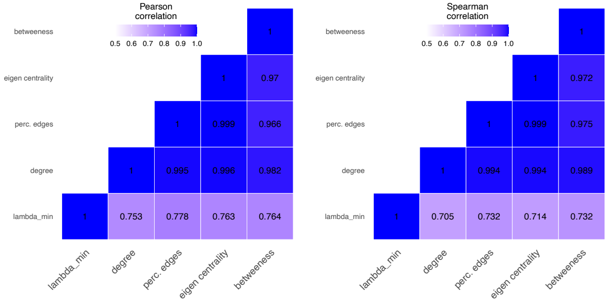
Setting: $\mathbf{P} = \begin{pmatrix} 3q & q \\ q & 3q \end{pmatrix}$. $K = 2$ groups have equal sizes.
$\theta_i \sim$ i.i.d. mixture, $F_1 \sim$Unif$(0.5, 0.75)$, $F_2 \sim$Unif$(0.8, 1.05)$ with proportions $(0.85, 0.15)$.

Table: Correlations between centrality measures and clustering accuracy

	Pearson	Spearman
degree	0.602	0.679
fraction of edges	0.624	0.696
eigen centrality	0.607	0.684
betweenness	0.581	0.700
$\widehat{\lambda}_{\min}$	0.742	0.822

Simulation with DCSBM

Correlations between centrality measures

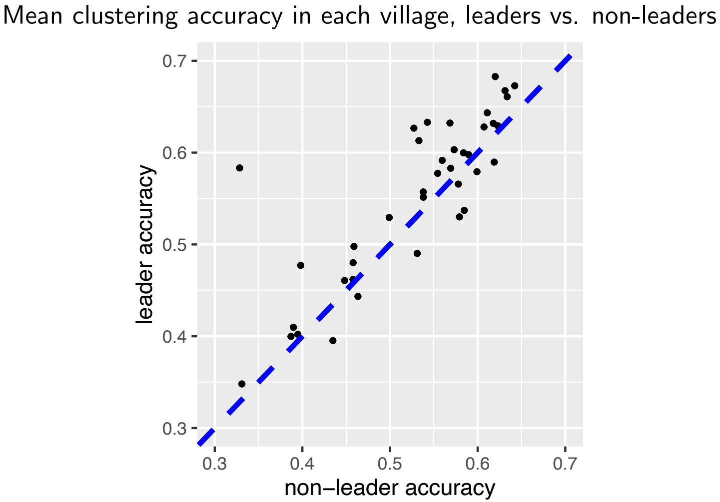


Microfinance in Indian villages

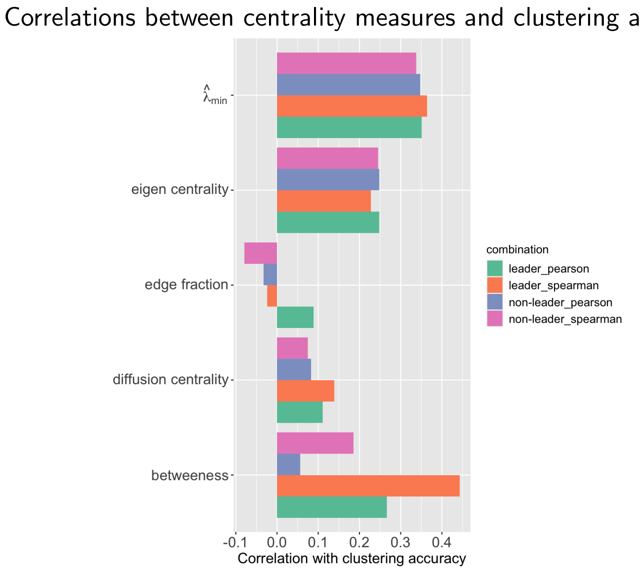
Banerjee et al. *Science* (2013)

- Modeling the spread of information about a microfinance program in Indian villages.
- Social network in each village: households as nodes, each edge is undirected and binary representing any of the 12 relationships collected in the survey (e.g., borrowing / lending money or material goods).
- Caste information as community labels
- Each village has a few predefined leaders serving as "injection" points for information.
- We analyze 39 villages, with the number of households varying between 24-155 and K between 2-4.

Microfinance in Indian villages

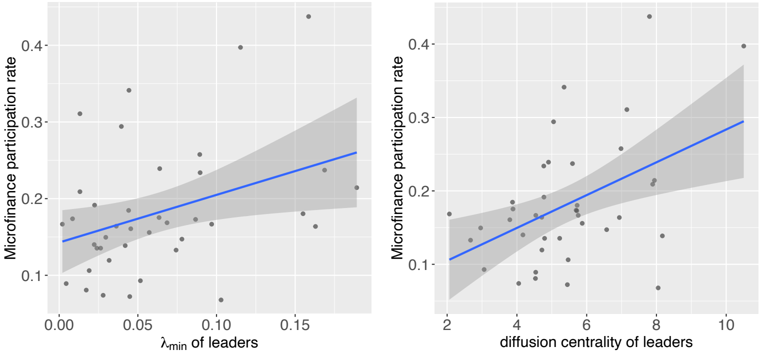


Microfinance in Indian villages



Microfinance in Indian villages

Program participation rate as a function of (left) $\hat{\lambda}_{\min}$ with p-value 0.022; (right) diffusion centrality with p-value 0.003.



Summary and future work

We have introduced an individual-centered partial information framework to study social networks.

- Theoretical properties of the main signal term in the partial adjacency matrix
- Consistent community detection under SBM and DCSBM
- Centrality measure based on eigen gap

Many interesting problems ahead:

- Including only individuals reached by the partial network
- mixed membership block models
- $L = 3$
- Determining K
- Imprecise knowledge about neighbors' neighbors
- Multiple individuals' partial information
- .....

Summary and future work

We have introduced an individual-centered partial information framework to study social networks.

- Theoretical properties of the main signal term in the partial adjacency matrix
- Consistent community detection under SBM and DCSBM
- Centrality measure based on eigen gap

Many interesting problems ahead:

- Including only individuals reached by the partial network
- mixed membership block models
- $L = 3$
- Determining K
- Imprecise knowledge about neighbors' neighbors
- Multiple individuals' partial information
- .....

Acknowledgements

Joint work with

- Dr. Xiao Han, School of Management, University of Science and Technology of China
- Dr. Xin Tong, Marshall School of Business, University of Southern California

Many thanks to

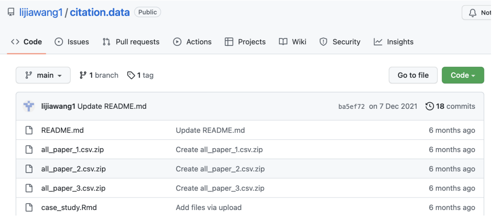
- Prof. Peter Bickel, UC Berkeley
- Dr. Yiqing Xing, Carey School of Business, Johns Hopkins University

External impact of statistical theory and methodology papers

Patterns

CellPress
OPEN ACCESS

Lijia Wang, Xin Tong, and Y. X. Rachel Wang. "Statistics in everyone's backyard: an impact study via citation network analysis."



Happy Birthday Peter! Thank you for your unwavering support and endless inspiration!