Statistical and Computational Phase Transitions in Planted Models

Jiaming Xu

Joint work with Yudong Chen (UC Berkeley)

Acknowledgement: Prof. Bruce Hajek

November 4, 2013
Cluster/Community structure in networks

Network of political webblogs [Adamic-Glance ’05]
Cluster/Community structure in networks

Network of political weblogs [Adamic-Glance ’05]

Social networks: social communities; Metabolic networks: functional communities; Recommendation systems: user and item communities ...
Cluster/Community structure in networks

Network of political webblogs [Adamic-Glance ’05]

Social networks: social communities; Metabolic networks: functional communities; Recommendation systems: user and item communities ...
Q: How to recover hidden cluster structure? → Community Detection
Cluster/Community structure in networks

Network of political webblogs [Adamic-Glance ’05]

Social networks: social communities; Metabolic networks: functional communities; Recommendation systems: user and item communities ...
Q: How to recover hidden cluster structure? → Community Detection
Application: link prediction in social networks, rating prediction in recommendation systems ...
Information theory of community detection

Simple model: Erdős-Rényi type model with "planted" clusters

Information-theoretic view: Converse and achievability for cluster recovery

Computational view: Performance limit of polynomial-time algorithms for cluster recovery
Information theory of community detection

- **Simple model**: Erdős-Rényi type model with “planted” clusters
Information theory of community detection

- **Simple model**: Erdős-Rényi type model with “planted” clusters
- **Information-theoretic view**: Converse and achievability for cluster recovery
Information theory of community detection

- **Simple model**: Erdős-Rényi type model with “planted” clusters
- **Information-theoretic view**: Converse and achievability for cluster recovery
- **Computational view**: Performance limit of polynomial-time algorithms for cluster recovery
Stochastic blockmodel (planted partition model)

A random graph model to generate graph with cluster structure

$n = 5000, r = 10, K = 500, p = 0.999, q = 0.001$. Ref. https://projects.skewed.de/graph-tool.
Stochastic blockmodel (planted partition model)

A random graph model to generate graph with cluster structure

$n = 5000, r = 10, K = 500, p = 0.999, q = 0.001$. Ref. https://projects.skewed.de/graph-tool.

Goal: Exactly recover the hidden clusters given the graph.
Cluster recovery as matrix recovery

Cluster matrix: \( Y_{ij} = 1 \) if \( i \) and \( j \) are in the same cluster; otherwise \( Y_{ij} = 0 \).
Cluster recovery as matrix recovery

Cluster matrix: $Y_{ij} = 1$ if $i$ and $j$ are in the same cluster; otherwise $Y_{ij} = 0$.

True cluster matrix $Y^*$

Observed adjacency matrix $A$
Cluster recovery as matrix recovery

Cluster matrix: $Y_{ij} = 1$ if $i$ and $j$ are in the same cluster; otherwise $Y_{ij} = 0$.

Cluster recovery as a specific matrix recovery problem:

$Y^* \rightarrow A \rightarrow \hat{Y}$
Cluster recovery under stochastic blockmodel

Vast literature on stochastic blockmodel [Holland et al. ’83] and planted partition model [Condon-Karp ’01]:

Two fundamental questions still unclear:

- Information limit: In which regime of $n$, $K$, $p$, $q$, is exact cluster recovery possible (impossible)?
- Computational limit: In which regime of $n$, $K$, $p$, $q$, is exact cluster recovery easy (hard)?
Cluster recovery under stochastic blockmodel

Vast literature on stochastic blockmodel [Holland et al. ’83] and planted partition model [Condon-Karp ’01]:
- [Bickel-Chen ’09] [Rohe et al. ’10] [Mossel et al. ’12] . . .
- [McSherry ’01] [Coja-Oghlan ’10] [Tomozei-Massoulié ’11] [Chaudhuri et al. ’12] [Chen-Sanghavi-Xu ’12]

Two fundamental questions still unclear:
- Information limit: In which regime of $n$, $K$, $p$, $q$, is exact cluster recovery possible (impossible)?
- Computational limit: In which regime of $n$, $K$, $p$, $q$, is exact cluster recovery easy (hard)?
Cluster recovery under stochastic blockmodel

Vast literature on stochastic blockmodel [Holland et al. ’83] and planted partition model [Condon-Karp ’01]:

► [Bickel-Chen ’09] [Rohe et al. ’10] [Mossel et al. ’12] …
► [Karrer-Newman ’11] [Decelle et al. ’11]
Cluster recovery under stochastic blockmodel

Vast literature on stochastic blockmodel [Holland et al. ’83] and planted partition model [Condon-Karp ’01]:

- [Bickel-Chen ’09] [Rohe et al. ’10] [Mossel et al. ’12] …
- [McSherry ’01] [Coja-Oghlan ’10] [Tomozei-Massoulié ’11] [Chaudhuri et al. ’12] [Chen-Sanghavi-Xu ’12] [Heimlicher et al. ’12] [Anandkumar et al. ’13] [Lelarge et al. ’13] …

Two fundamental questions still unclear:

- Information limit: In which regime of $n$, $K$, $p$, $q$, is exact cluster recovery possible (impossible)?
- Computational limit: In which regime of $n$, $K$, $p$, $q$, is exact cluster recovery easy (hard)?
Cluster recovery under stochastic blockmodel

Vast literature on stochastic blockmodel [Holland et al. ’83] and planted partition model [Condon-Karp ’01]:

- [Bickel-Chen ’09] [Rohe et al. ’10] [Mossel et al. ’12] …
- [McSherry ’01] [Coja-Oghlan ’10] [Tomozei-Massoulié ’11] [Chaudhuri et al. ’12] [Chen-Sanghavi-Xu ’12] [Heimlicher et al. ’12] [Anandkumar et al. ’13] [Lelarge et al. ’13] …

Two fundamental questions still unclear:
Cluster recovery under stochastic blockmodel

Vast literature on stochastic blockmodel [Holland et al. ’83] and planted partition model [Condon-Karp ’01]:

- [Bickel-Chen ’09] [Rohe et al. ’10] [Mossel et al. ’12] 
- [McSherry ’01] [Coja-Oghlan ’10] [Tomozei-Massoulié ’11] [Chaudhuri et al. ’12] [Chen-Sanghavi-Xu ’12] [Heimlicher et al. ’12] [Anandkumar et al. ’13] [Lelarge et al. ’13] 

Two fundamental questions still unclear:

- **Information limit**: In which regime of $n, K, p, q$, is exact cluster recovery possible (impossible)?
Cluster recovery under stochastic blockmodel

Vast literature on stochastic blockmodel [Holland et al. ’83] and planted partition model [Condon-Karp ’01]:

- [Bickel-Chen ’09] [Rohe et al. ’10] [Mossel et al. ’12] . . .
- [Karrer-Newman ’11] [Decelle et al. ’11]
- [McSherry ’01] [Coja-Oghlan ’10] [Tomozei-Massoulié ’11]

Two fundamental questions still unclear:

- **Information limit**: In which regime of $n, K, p, q$, is exact cluster recovery possible (impossible)?
- **Computational limit**: In which regime of $n, K, p, q$, is exact cluster recovery easy (hard)?
Cluster recovery under stochastic blockmodel

Our (non-asymptotic) results apply to general setting allowing any $n, K, p, q$.

\[
K = \Theta(n^\beta)
\]

\[
p = 2q = \Theta(n^{-\alpha})
\]

- dense graph separation
- sparse graph small separation

large cluster

small cluster

$K = \Theta(n^\beta)$
Cluster recovery under stochastic blockmodel

Our (non-asymptotic) results apply to general setting allowing any $n, K, p, q$.

\[ K = \Theta(n^\beta) \]

\[ p = 2q = \Theta(n^{-\alpha}) \]
Converse for cluster recovery

\[ \alpha \\
\beta \]

\[ K = \Theta(n^\beta) \]

\[ p = 2q = \Theta(n^{-\alpha}) \]
Converse for cluster recovery

\[ K = \Theta(n^\beta) \]

\[ p = 2q = \Theta(n^{-\alpha}) \]

Proof:

\[ Y^* \rightarrow A \rightarrow \hat{Y}. \] Apply Fano's inequality to lower bound \( P(\hat{Y} \neq Y^*) \) by upper bounding \( I(Y^*;A) \).

Intuition: The observation \( A \) does not carry enough information to distinguish between different possible \( Y^* \).
Converse for cluster recovery

\[ p = 2q = \Theta(n^{-\alpha}) \]

Proof: \( Y^* \rightarrow A \rightarrow \hat{Y} \). Apply Fano’s inequality to lower bound \( \mathbb{P}(\hat{Y} \neq Y^*) \) by upper bounding \( I(Y^*; A) \).
Converse for cluster recovery

$$K = \Theta(n^\beta)$$

$$p = 2q = \Theta(n^{-\alpha})$$

**Proof:** $Y^* \rightarrow A \rightarrow \hat{Y}$. Apply Fano’s inequality to lower bound $P(\hat{Y} \neq Y^*)$ by upper bounding $I(Y^*; A)$.

**Intuition:** The observation $A$ does not carry enough information to distinguish between different possible $Y^*$. 
Converse for cluster recovery

Proof: \( Y^* \rightarrow A \rightarrow \hat{Y} \). Apply Fano’s inequality to lower bound \( \mathbb{P}(\hat{Y} \neq Y^*) \) by upper bounding \( I(Y^*; A) \).

Intuition: The observation \( A \) does not carry enough information to distinguish between different possible \( Y^* \).
Achievability by maximum likelihood estimation

Maximum likelihood estimator: \( \hat{Y} = \arg \max P(A|Y) \)

\[ Y^* \rightarrow A \rightarrow \hat{Y} \]
Achievability by maximum likelihood estimation

Maximum likelihood estimator: \( \hat{Y} = \arg \max \mathbb{P}(A|Y) \)

\[ Y^* \rightarrow A \rightarrow \hat{Y} \]

If \( p > q \), maximum likelihood estimation is equivalent to finding the \( r \) most densely connected subgraphs of size \( K \) in the graph:

\[
\max_{Y} \sum_{i,j} A_{ij} Y_{ij}
\]

s.t. \( Y \) is a cluster matrix.
Achievability by maximum likelihood estimation

Maximum likelihood estimator: \( \hat{Y} = \arg \max \mathbb{P}(A|Y) \)

\[ Y^* \rightarrow A \rightarrow \hat{Y} \]

If \( p > q \), maximum likelihood estimation is equivalent to finding the \( r \) most densely connected subgraphs of size \( K \) in the graph:

\[
\max_{\gamma} \sum_{i,j} A_{ij} Y_{ij}
\]

s.t. \( Y \) is a cluster matrix.

Q: When maximum likelihood estimator equals \( Y^* \)?
Achievability by maximum likelihood estimation

\[ \beta \quad K = \Theta(n^\beta) \]

\[ p = 2q = \Theta(n^{-\alpha}) \]

Proof: Concentration inequality + union bound (needs clever counting argument and peeling technique)

Q: MLE takes an exponential time to solve. Can we achieve information limit via polynomial-time algorithms?
Achievability by maximum likelihood estimation

\[ p = 2q = \Theta(n^{-\alpha}) \]

\[ K = \Theta(n^\beta) \]

Proof: Concentration inequality + union bound (needs clever counting argument and peeling technique)

Q: MLE takes an exponential time to solve. Can we achieve information limit via polynomial-time algorithms?
Achievability by maximum likelihood estimation

\[ p = 2q = \Theta(n^{-\alpha}) \]

\[ K = \Theta(n^\beta) \]

\[ \alpha \]

\[ 1 \]

\[ \beta \]

\[ 1 \]

\[ O \]

Proof: Concentration inequality + union bound (needs clever counting argument and peeling technique)
Achievability by maximum likelihood estimation

\[ K = \Theta(n^\beta) \]

\[ p = 2q = \Theta(n^{-\alpha}) \]

**Proof**: Concentration inequality + union bound (needs clever counting argument and peeling technique)

**Q**: MLE takes an exponential time to solve. Can we achieve information limit via polynomial-time algorithms?
Polynomial-time recovery: convex relaxation of MLE

Cluster matrix $Y$ has low rank:

$$
\begin{bmatrix}
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1
\end{bmatrix}
$$

\[
\text{rank} \begin{bmatrix}
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1
\end{bmatrix} = 2.
\]
Polynomial-time recovery: convex relaxation of MLE

Cluster matrix $Y$ has low rank:

$$
\begin{bmatrix}
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 \\
\end{bmatrix}
$$

$\text{rank} = 2$.

Nuclear norm $\|Y\|_*$ (sum of singular values) is a convex surrogate for rank function.
Polynomial-time recovery: convex relaxation of MLE

Cluster matrix $Y$ has low rank:

$$\text{rank} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} = 2.$$ 

Nuclear norm $\| Y \|_\ast$ (sum of singular values) is a convex surrogate for rank function.

A convex relaxation of MLE [Chen-Sangavi-Xu ’12]:

$$\max_Y \sum_{ij} A_{ij} Y_{ij} \quad \text{s.t.} \quad \| Y \|_\ast \leq n \quad \sum_{ij} Y_{ij} = rK^2, \ Y_{ij} \in [0, 1].$$
Polynomial-time recovery: convex relaxation of MLE

\[ p = 2q = \Theta(n^{-\alpha}) \]

\[ K = \Theta(n^\beta) \]

Proof: Nuclear norm constraint suppresses the random noise and boosts the SNR.

Surprise: Convex relaxation might not be order-optimal when there is a growing number of clusters.
Polynomial-time recovery: convex relaxation of MLE

\[ p = 2q = \Theta(n^{-\alpha}) \]

\[ K = \Theta(n^\beta) \]

Proof: Nuclear norm constraint suppresses the random noise and boosts the SNR.

Surprise: Convex relaxation might not be order-optimal when there is a growing number of clusters.
Proof: Nuclear norm constraint suppresses the random noise and boosts the SNR.
Polynomial-time recovery: convex relaxation of MLE

Proof: Nuclear norm constraint suppresses the random noise and boosts the SNR.

Surprise: Convex relaxation might not be order-optimal when there is a growing number of clusters.
Polynomial-time recovery: convex relaxation of MLE

Proof: Nuclear norm constraint suppresses the random noise and boosts the SNR.
Surprise: Convex relaxation might not be order-optimal when there is a growing number of clusters.
Polynomial-time recovery: counting common neighbor

Similarity between two nodes: The number of common neighbors [Dyer-Frieze ’98].
Polynomial-time recovery: counting common neighbor

Similarity between two nodes: The number of common neighbors [Dyer-Frieze ’98].

**Algorithm**: Each node finds the $K - 1$ most similar nodes.
Polynomial-time recovery: counting common neighbor

Similarity between two nodes: The number of common neighbors [Dyer-Frieze ’98].

**Algorithm**: Each node finds the $K - 1$ most similar nodes.

$$p = 2q = \Theta(n^{-\alpha})$$

$$K = \Theta(n^\beta)$$
Polynomial-time recovery: counting common neighbor

Similarity between two nodes: The number of common neighbors [Dyer-Frieze ’98].

**Algorithm**: Each node finds the $K - 1$ most similar nodes.

\[ K = \Theta(n^\beta) \]

\[ p = 2q = \Theta(n^{-\alpha}) \]
Polynomial-time recovery: counting common neighbor

Similarity between two nodes: The number of common neighbors [Dyer-Frieze ’98].

**Algorithm**: Each node finds the $K - 1$ most similar nodes.

$$p = 2q = \Theta(n^{-\alpha})$$

Proof: Similarity concentrates around its mean.
Polynomial-time recovery: spectral algorithms

Spectral algorithms: based on principal singular vectors (PCA)
Polynomial-time recovery: spectral algorithms

Spectral algorithms: based on principal singular vectors (PCA)
Example: $n = 6^4$, $r = 6$, $K = n^{0.75}$, $p = n^{-0.25}$, $q = p/8$
Polynomial-time recovery: spectral algorithms

Spectral algorithms: based on principal singular vectors (PCA)
Example: $n = 6^4$, $r = 6$, $K = n^{0.75}$, $p = n^{-0.25}$, $q = p/8$

- The $r$ principal singular vectors contain cluster information.
- The bulk of spectrum is caused by the random noise.
Polynomial-time recovery: spectral algorithms

Signal strength ($r$-th largest singular value) is $K(p - q)$; Noise magnitude is $O(\sqrt{np})$.

Signal strength needs to be larger than noise magnitude: $K(p - q) \gg \sqrt{np}$ (Spectral barrier).
Signal strength (r-th largest singular value) is $K(p - q)$; Noise magnitude is $O(\sqrt{np})$.

Signal strength needs to be larger than noise magnitude: $K(p - q) \gtrsim \sqrt{np}$ (Spectral barrier).
Polynomial-time recovery: spectral algorithms

Signal strength ($\beta$-th largest singular value) is $K(p-q)$;
Noise magnitude is $O(\sqrt{np})$.

Signal strength needs to be larger than noise magnitude:
$K(p-q) \gtrsim \sqrt{np}$ (Spectral barrier).

Graph showing signal strength and noise magnitude relationship.
Signal strength ($r$-th largest singular value) is $K(p-q)$; Noise magnitude is $O(\sqrt{np})$.

Signal strength needs to be larger than noise magnitude: $K(p-q) \gg \sqrt{np}$ (Spectral barrier).
Polynomial-time recovery: spectral algorithms

Signal strength ($r$-th largest singular value) is $K(p - q)$; Noise magnitude is $O(\sqrt{np})$. 
Signal strength ({\it r-th} largest singular value) is $K(p - q)$; Noise magnitude is $O(\sqrt{np})$.

Signal strength needs to be larger than noise magnitude: $K(p - q) \gtrsim \sqrt{np}$ (Spectral barrier).
Signal strength ($r$-th largest singular value) is $K(p - q)$; Noise magnitude is $O(\sqrt{np})$.

Signal strength needs to be larger than noise magnitude: $K(p - q) \gtrsim \sqrt{np}$ (Spectral barrier).
Conjecture on computational limit

Conjecture: no polynomial-time algorithm succeeds beyond spectral barrier.

A similar conjecture appears in the planted clique model.
Conjecture on computational limit

\[ O = \alpha, \beta \]

\[ p = 2q = \Theta(n^{-\alpha}) \]

\[ K = \Theta(n^{\beta}) \]

\[ \frac{1}{2} \]

\[ 1 \]

Conjecture: no polynomial-time algorithm succeeds beyond spectral barrier.
A similar conjecture appears in the planted clique model.
Review: Conjecture in planted clique model

\[ A = K \text{ clique} + \text{Ber}(0.5) \]
Review: Conjecture in planted clique model

\[
A = K_{\text{clique}} + \text{Ber}(0.5)
\]

- Feasible if and only if \( K > 2 \log_2 n \)
- Simple algorithm by picking the \( K \) nodes with highest degree works if \( K = \Omega(\sqrt{n \log n}) \)
- Spectral algorithm works if \( K = \Omega(\sqrt{n}) \) [Alon et al. '98]
- Belief: No polynomial-time algorithm works if \( K = o(\sqrt{n}) \)
Review: Conjecture in planted clique model

\[ A = K \text{ Ber}(p) + \text{Ber}(q) \]

- Feasible if and only if \( K > 2 \log_2 n \)
- Simple algorithm by picking the \( K \) nodes with highest degree works if \( K = \Omega(\sqrt{n \log n}) \)
- Spectral algorithm works if \( K = \Omega(\sqrt{n}) \) [Alon et al. '98]
- Belief: No polynomial-time algorithm works if \( K = o(\sqrt{n}) \)

Planted dense subgraph model: \( p, q \in [0, 1] \)
Planted dense subgraph model

\[ p = 2q = \Theta(n^{-\alpha}) \]

\[ K = \Theta(n^\beta) \]

Conjecture: no polynomial-time algorithm succeeds beyond the spectral barrier.
Planted dense subgraph model

\[ p = 2q = \Theta(n^{-\alpha}) \]

\[ K = \Theta(n^\beta) \]

Conjecture: no polynomial-time algorithm succeeds beyond the spectral barrier.
Planted dense subgraph model

Conjecture: no polynomial-time algorithm succeeds beyond the spectral barrier.
Concluding remarks

- Simple model: Stochastic blockmodel (planted partition model).

- If $K = \Theta(n)$, cluster structure can be recovered up to the information limit via polynomial-time algorithms.

- If $K = o(n)$, cluster structure can be recovered up to the information limit via exponential-time algorithms but might not via polynomial-time algorithms due to spectral barrier.

- Conjecture on existence of big gap between information and computational limit also appears in planted dense subgraph model.

- Future work: prove the conjecture by assuming no polynomial-time algorithm detects hidden clique of size $o(\sqrt{n})$ in the planted clique model.
Concluding remarks

- Simple model: Stochastic blockmodel (planted partition model).
- If $K = \Theta(n)$, cluster structure can be recovered up to the information limit via polynomial-time algorithms.
- If $K = o(n)$, cluster structure can be recovered up to the information limit via exponential-time algorithms but might not via polynomial-time algorithms due to spectral barrier.
- Conjecture on existence of big gap between information and computational limit also appears in planted dense subgraph model.
- Future work: prove the conjecture by assuming no polynomial-time algorithm detects hidden clique of size $o(\sqrt{n})$ in the planted clique model.
Concluding remarks

- Simple model: Stochastic blockmodel (planted partition model).
- If $K = \Theta(n)$, cluster structure can be recovered up to the information limit via polynomial-time algorithms.
- If $K = o(n)$, cluster structure can be recovered up to the information limit via exponential-time algorithms but might not via polynomial-time algorithms due to spectral barrier.
Concluding remarks

- Simple model: Stochastic blockmodel (planted partition model).
- If $K = \Theta(n)$, cluster structure can be recovered up to the information limit via polynomial-time algorithms.
- If $K = o(n)$, cluster structure can be recovered up to the information limit via exponential-time algorithms but might not via polynomial-time algorithms due to spectral barrier.
- Conjecture on existence of big gap between information and computational limit also appears in planted dense subgraph model.
Concluding remarks

- Simple model: Stochastic blockmodel (planted partition model).

- If $K = \Theta(n)$, cluster structure can be recovered up to the information limit via polynomial-time algorithms.

- If $K = o(n)$, cluster structure can be recovered up to the information limit via exponential-time algorithms but might not via polynomial-time algorithms due to spectral barrier.

- Conjecture on existence of big gap between information and computational limit also appears in planted dense subgraph model.

- Future work: prove the conjecture by assuming no polynomial-time algorithm detects hidden clique of size $o(\sqrt{n})$ in the planted clique model.
Gap between information and computational limit

Search version

- Planted Clique
- Planted dense Subgraph
- Sparse PCA
- Planted Submatrix
- Planted Partition

Hypothesis testing version

- Planted Clique
- Planted dense Subgraph
- [Ma&Wu '13]
- [Berthet&Rigollet '13]
- Sparse PCA
- Planted Submatrix
- Sparse PCA