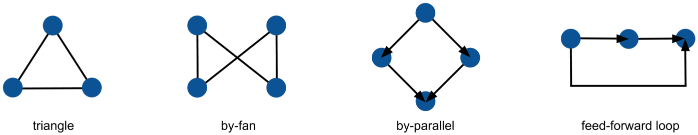
Network Community Detection Using Higher Order Interactions

Xianshi Yu
Joint work with Ji Zhu
Aug 03, 2020

Table of contents

Real Data Observation



Certain subgraphs are abundant. [4,5,6]

triange: social network; world wide webs
by-fan: transcriptional gene regulation network; neural network
by-paralle: neural network; food web

...

Real Data Observation

The pattern of observed subgraphs could reflect the underlying community memberships.

- How to exploit the configuration of $\triangle$'s in community detection?
- How to model the abundance and distribution of $\triangle$?

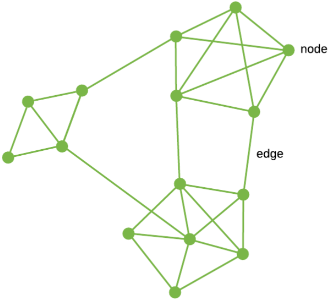
Notations

- $\{1, 2, \ldots, n\}$ index the nodes.
- $A_{n \times n}$ (adjacency matrix) describes edges of an observed network.
- $A_{ij} = 1$ if nodes i and j have an edge, $A_{ij} = 0$ otherwise.
- $\Delta_{n \times n \times n}$ represents triangles: $\Delta_{ijk} = 1$ iff nodes i, j and k share a triangle. $\Delta_{ijk} = 0$ otherwise.

$\star$ $\Delta_{ijk} = A_{ij} A_{jk} A_{jk}$

Goal

∘ $g_i \in \{1, 2, \ldots, K\}$ is the community index of node i.
∘ $Z_{n \times K}$ is a matrix of all zeros, except that $Z_{ig_i} = 1$ for all i.

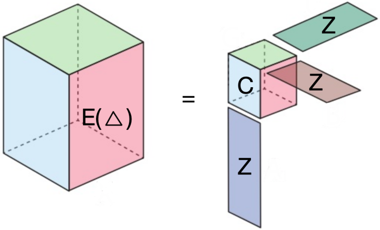


Ideas

Consider the formation of triangles as depending only on community membership

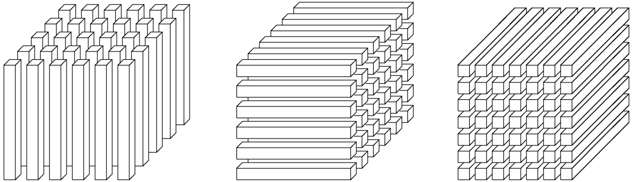
$$P(\Delta_{ijk} = 1) = C_{g_i, g_j, g_k}$$

$$\text{i.e. } \mathbb{E}(\Delta_{n\times n\times n}) = C_{K\times K\times K} \times_1 Z \times_2 Z \times_3 Z$$



Definition of n-mode product ▸ ref

To obtain $C_{K\times K\times K} \times_1 Z_{n\times K}$
Consider mode-1 fibers of C



(a) Mode-1 (column) fibers: $c_{.jk}$

(b) Mode-2 (row) fibers: $c_{i.k}$

(c) Mode-3 (tube) fibers: $c_{ij.}$

Calculate the products of each mode-1 fiber $C_{.jk}$ and Z, namely $ZC_{.jk}^T$.
Arrange the resulted vectors accordingly to form an $n \times K \times K$ tensor.

Ideas

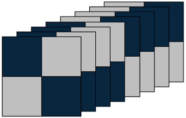
Remarks

$$\mathbb{E}(\Delta) = C \times_1 Z \times_2 Z \times_3 Z \tag{1}$$

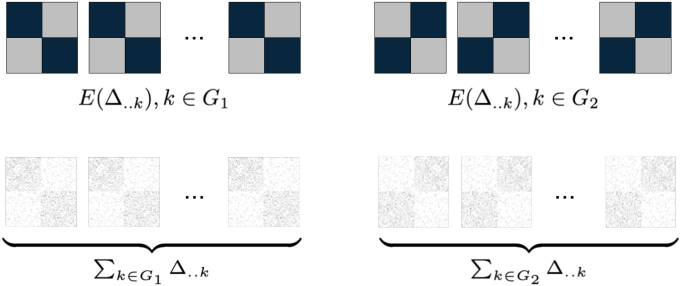
- In Stochastic Block Model (SBM), community structure is on $E(A)$.
- (1) does not define a generative model, but a 'constraint' on network models. SBM satisfies (1).

Ideas

How to infer $g_i, i \in \{1, \ldots, n\}$ from Δ?

$E(\Delta) =$  $=$ 

Ideas



- $\mathbb{E}(\Delta)$ is a block tensor, and slices of it, i.e $\mathbb{E}(\Delta_{..k})$, are block matrices. Spectral decomposition of $\mathbb{E}(\Delta_{..k})$ reveals g_i.
- $\Delta_{..k}$ is likely very sparse.

Method

Algorithm

Input: $A_{n\times n}, K$

Output: $\hat{Z}$

Step 1. Calculate $\Delta = (A_{i,j}A_{j,k}A_{i,k})_{i,j,k}$

Step 2. Obtain an initial estimate $Z^0_{n\times K}$

Step 3. Calculate sums of slices of $\Delta_{n\times n\times n}$ w.r.t groups defined by Z^0.
i.e. $S^l_\Delta := \Delta \times_3 Z^0_{.l}$, for $l \in \{1, \cdots, K\}$.

Step 4. Apply synchronized spectral decomposition to S^l_Δ,
$l \in \{1, \cdots, K\}$ to obtain $\hat{U}_{n\times K}$

Step 5. Perform K-means on $\hat{U}$ to obtain $\hat{Z}$

Method—synchronized spectral decomposition (SynSD)

$$\hat{U} = \underset{U_{n\times K}^T U = I}{\arg\max} \sum_{l=1}^{K} \| U^T S_\Delta^l U \|_F^2$$

- SynSD finds ‘shared singular vectors’.
- SynSD is a Grassmann manifold optimization problem, for which ample literature and packages are available.

Experiment on lawyers' co-work network

71 attorneys (1104 edges)

seniority	status	gender	office	age
range:[1,32] median:7	partner:36 associate:35	man:53 woman:18	Boston:48 Hartford:19 Providence:4	range:[26,67] median:39

practice	law school
litigation:41 corporate:30	harvard, yale:15 ucon:28 other:28

Experiment on lawyers' co-work network

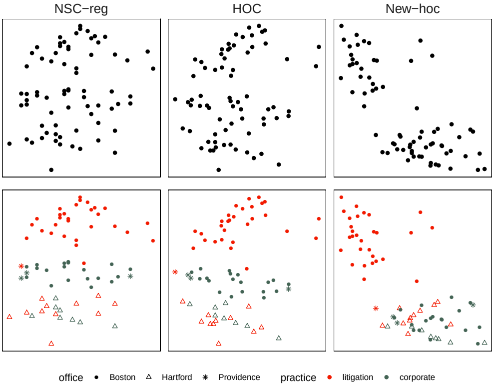
Benchmarks

- Normalized Spectral Clustering with regularization (NSC-reg)
- High-Order Clustering [1, 2] (HOC)
  —perform NSC on $\sum_{k=1}^{n} \Delta_{..k}$

Initial estimate Z^0 of the new method

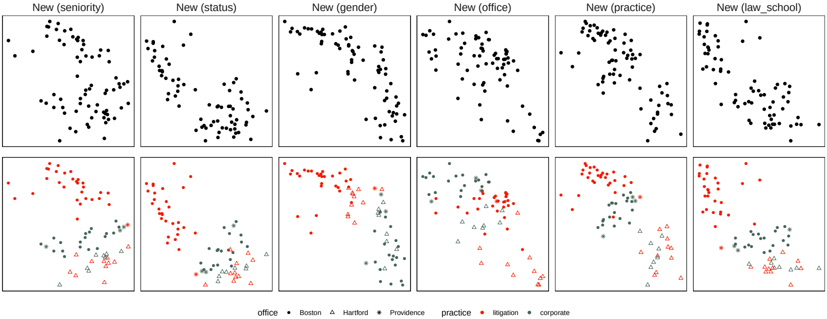
- HOC
- Attribute of lawyers

Experiment on lawyers' co-work network



- Column 1&2: leading two eigenvectors of NSC-reg and HOC
- Column 3: $\hat{U}$ of the new method ($K = 2$)

Experiment on lawyers' co-work network

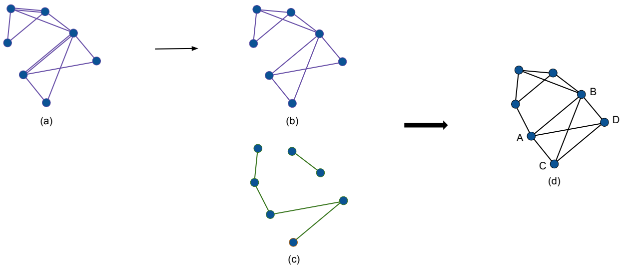


$\hat{U}$ obtained when attributes in bracket are taken as Z^0

Model & Theory

A model that satisfies $\mathbb{E}(\Delta) = C \times_1 Z \times_2 Z \times_3 Z$
— triangle mechanism + edge mechanism

- $T \sim$ independent Bernoulli with $\dddot{P}_{n\times n\times n} = \dddot{B} \times_1 Z \times_2 Z \times_3 Z$
- $\ddot{A} \sim$ independent Bernoulli with $\ddot{P}_{n\times n}$
- $A_{i,j} = \max\{\ddot{A}_{i,j}, \mathbf{1}(\sum_k T_{i,j,k} > 0)\}$



Model & Theory

Conditions

- $C_1 \ln^4 n \leq \dddot{d} \leq C_2 n^{\frac{2}{5}}$, where $\dddot{d} = n^2 \dddot{P}_{\max}$
- $\ddot{d} \leq C_3 n^{\frac{1}{3}} \dddot{d}^{\frac{1}{6}}$, where $\ddot{d} = n\ddot{P}_{\max}$

Notations

- $\sigma^l_{\min} = \sigma_K(\mathbb{E}(S^l_\Delta))$, $\sigma^l_{\max} = \sigma_1(\mathbb{E}(S^l_\Delta))$

Theorem 1

With $n > N$ and S^l_Δ calculated from a fixed Z^0, if the conditions above are satisfied, and $\hat{U}$ is the global optimum of synchronized spectral decomposition, then with probability at least $1 - n^{-r}$,

$$\frac{|\mathcal{M}|}{n} \leq C_4 K \frac{K\dddot{d} + \dddot{d}^{\frac{1}{2}} \sum_{l=1}^K \sigma^l_{\max}}{\sum_{l=1}^K (\sigma^l_{\min})^2}.$$

Here $\mathcal{M}$ is the set of misclustered nodes and N, r, $C_1, \cdots, C_4$ are absolute constants.

Remark

In the simple case of the model where $\ddot{P} = 0$, consider n growing, if Z^0 is independent of A with a confusion matrix proportionally constant and $\dddot{B}/\dddot{P}_{\max}$ constant, then the upper bound in Theorem 1 is $O(\dddot{d}^{-1/2})$.

Simulation results are in the manuscript

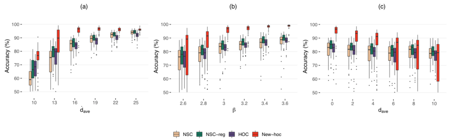


Figure 6: Accuracy with varying parameter settings: In (a), (b) and (c), we let one of $\dddot{d}_{\text{ave}}$, β and $\ddot{d}_{\text{ave}}$ to vary respectively. The fixed parameters for each plot are set to $\dddot{d}_{\text{ave}} = 15$, $\beta = 3$ and $\ddot{d}_{\text{ave}} = 0$, the same to those in Figure 5.

Conclusion

◦ A new method for network community detection that takes as input the observed $\triangle$'s.
◦ The new method tries to explain edges in networks not only by affiliation but also by roles in higher order interaction.
◦ Consistency theory that involves analyzing dependent objects.

Future work:
Other subgraphs, e.g. $\bowtie$, $\diamond$, V.

References

Previous methods that use triangles

1. Benson, A. R., Gleich, D. F. and Leskovec, J. (2016). Higher-order organization of complex networks. Science, 353(6295), 163-166.
2. Paul, S., Milenkovic, O. and Chen, Y. (2018). Higher-Order Spectral Clustering under Superimposed Stochastic Block Model. *arXiv preprint* arXiv:1812.06515.
3. Vandecappelle, M., Boussé, M., Van Eeghem, F. and De Lathauwer, L. (2016). Tensor decompositions for graph clustering. Internal Report, (16-170).

Abundance of subgraphs in real world networks.

4. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002). Network motifs: simple building blocks of complex networks. Science, 298(5594), 824-827.
5. Mangan, S. and Alon, U. (2003). Structure and function of the feed-forward loop network motif. Proceedings of the National Academy of Sciences, 100(21), 11980-11985.
6. Yaveroğlu, Ö. N., Malod-Dognin, N., Davis, D., Levnajic, Z., Janjic, V., Karapandza, R., ... and Pržulj, N. (2014). Revealing the hidden language of complex networks. Scientific reports, 4, 4547.