Triangles & networks

presented by Alex Hayes on 2021-02-17

The impossibility of low-rank representations for triangle-rich complex networks

by C. Seshadhri, Aneesh Sharma, Andrew Stolman & Ashish Goel

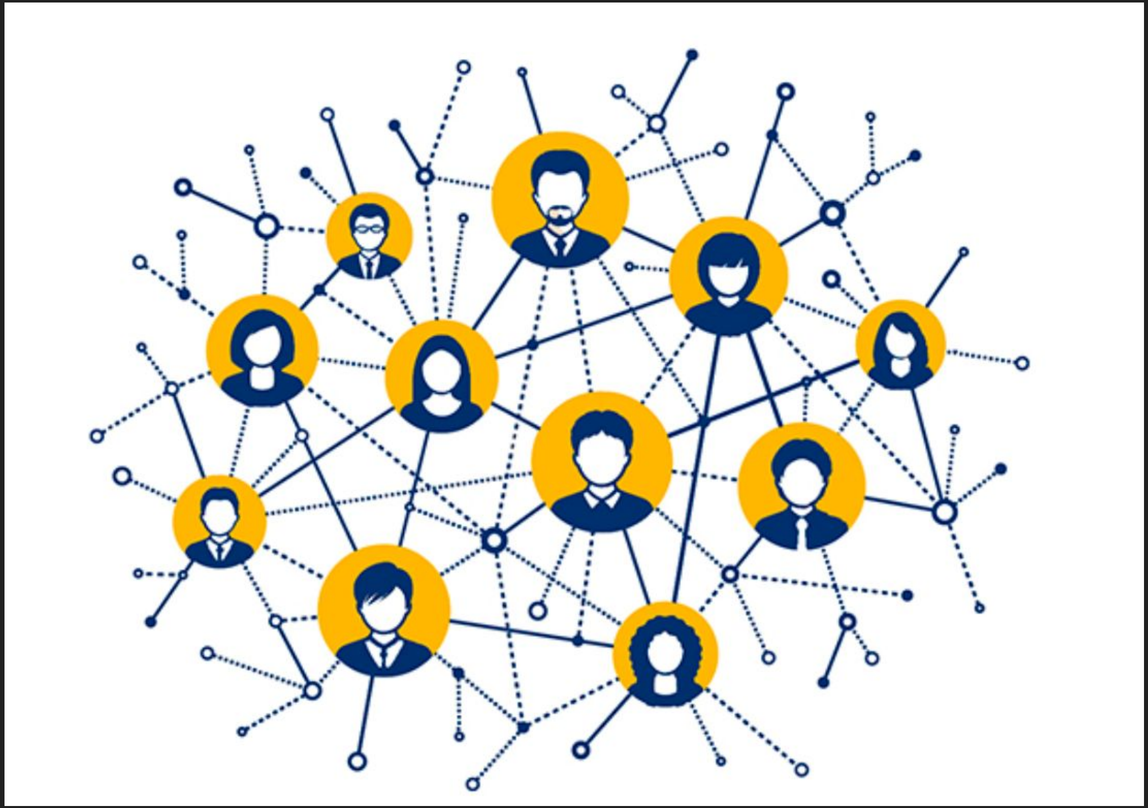
This paper is small component of a large literature on triangles in networks

We’re going to think exclusively about networks

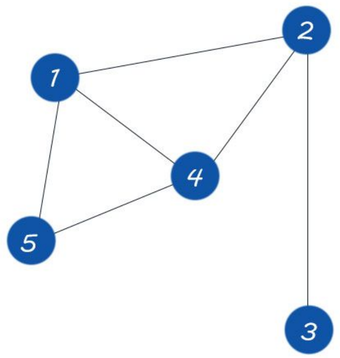
network = nodes + edges

nodes are items under consideration

edges are relationships between those items



Networks can be represented as adjacency matrices



$$A = \begin{array}{cc} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{array}{ccccc} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{array} \end{array}$$

What do real life networks look like?

1. Sparse
2. Transitive
3. Low diameter
4. Skewed degree distribution

So far generative models have had a hard time doing all of these at once. Typically you have to choose two out of (1), (2) and (3)

We need to talk about "reproducing the phenomenon"

Claim: We should use generative models that generates data like real life data*

* Generating data like real world data is neither necessary (think LPM) nor sufficient (curve fitting isn't causal) for good inference!

What do real life networks look like?

1. Sparse
2. Transitive
3. Low diameter
4. Skewed degree distribution

The paper we read this week is about difficulties with transitive closure (equivalently, having enough triangles)

So far generative models have had a hard time doing all of these at once. Typically you have to choose two out of (1), (2) and (3)

What is a triangle?

Number triangles in an undirected graph = trace(A A A) / 6

since $(AAA)_{ij}$ is the number of the paths of length 3 from i to j

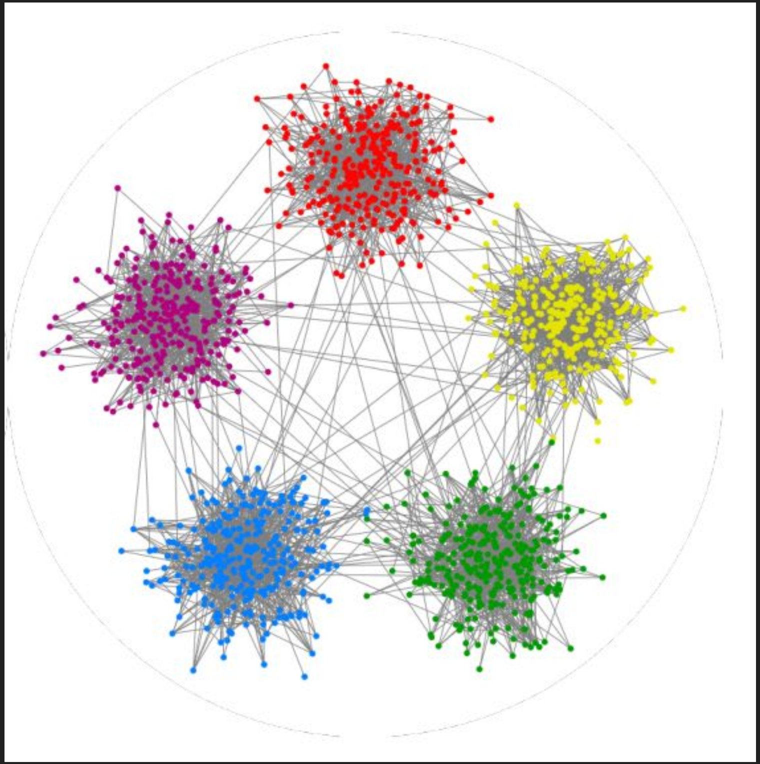
Key idea of the paper: the canonical generative model for networks generates networks with too few triangles

* This is a theoretical result

The baseline generative model for networks

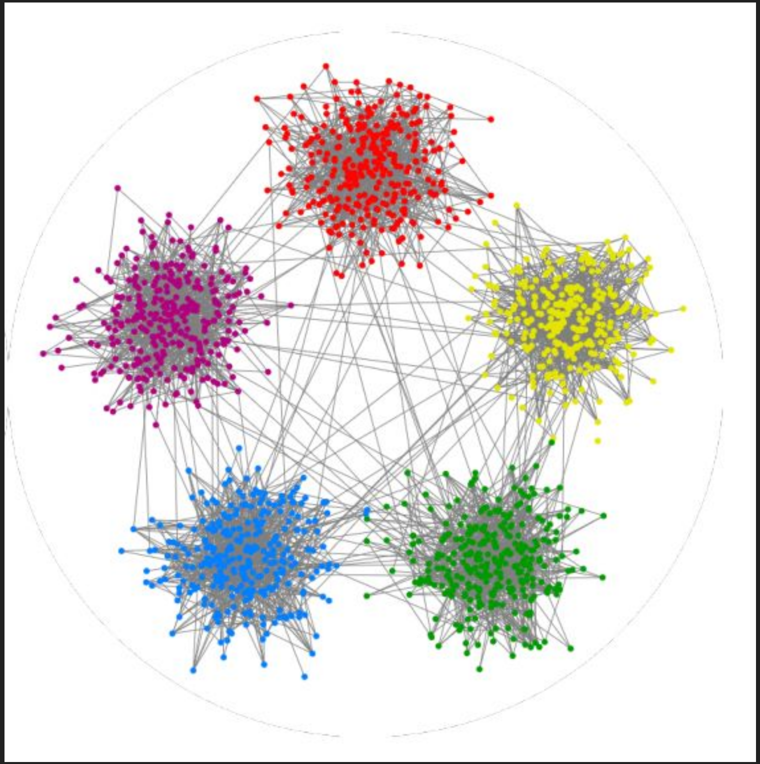
Stochastic blockmodels

- n nodes
- k communities
- each node belongs to one community
- each community has a distinct probability of connecting to a different community



Stochastic blockmodels: inference

- Learn the community memberships of each node
- Learn the relationship strengths within and between communities
- Many fitting methods (spectral, bayesian, likelihood, etc)
- Massive amounts of theory over the last decade

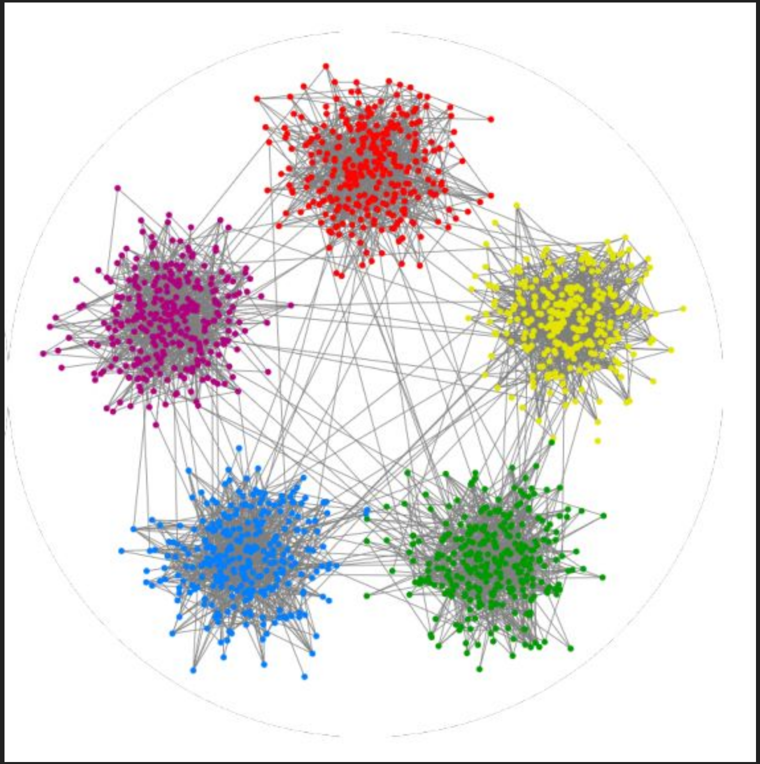


Stochastic blockmodels: extensions

Improve realism of models by extending basic SBM:

- Mixed community membership
- Overlapping community membership
- Degree-correction
- Etc, etc

I start with degree-corrected mixed membership stochastic blockmodels in applied work



Rank of a network

The population adjacency matrix of SBMs has low rank

$$E(A \mid X) = X B X'$$

n by k matrix of dummy variables indicating community membership

k by k matrix of community mixing probabilities

The rank of E(A|X) is the number of communities k

This intuition generalizes super well!

"Generalized Random Dot Product Graph"

$$E(A \mid X) = X\,B\,X'$$

n by k matrix where X_ij measures participation of node i in community j

k by k matrix of community mixing probabilities

The rank of E(A) is still the number of communities k

Stochastic blockmodels are the archetype for network models of the form

$$P(\text{edge between } i \text{ and } j \mid X\_i, X\_j)$$

$$= E(A\_ij \mid X\_i, X\_j)$$

$$= X\_i \, B \, X\_j'$$

X_i is an “embedding” of node i, B is the “mixing matrix”, X_j is an “embedding” of node j

* We assume all edges are independent of each other conditional on X

Main result (informal): Stochastic blockmodels (and their more appealing extensions, random dot product networks) are incapable of generating networks that are both sparse and transitive.

Formal result

Definition 1.1. *For parameters $c > 1$ and $\Delta > 0$, a graph G with n vertices has a (c, Δ)-*triangle foundation *if there are at least Δn triangles contained among vertices of degree at most c. Formally, let S_c be the set of vertices of degree at most c. Then, the number of triangles in the graph induced by S_c is at least Δn.*

Theorem 1.2. *Fix $c > 4, \Delta > 0$. Suppose the expected number of triangles in $G \sim \mathcal{G}_V$ that only involve vertices of expected degree c is at least Δn. Then, the rank of V is at least* $\min(1, \mathrm{poly}(\Delta/c))n/\lg^2 n$.

* I'm not going to talk about the proof at all

Population adjacency matrix of blockmodel conditional on latent X

Since $E(A \mid X)$ has low rank, stochastic blockmodels must have low expected triangle count

Population adjacency matrix of blockmodel conditional on latent X

Since $E(A \mid X)$ has low rank, ~~stochastic blockmodels~~ must have low expected triangle count

The situation is probably much worse than this in practice since graph embeddings in general appear to basically be learning stochastic blockmodels despite all the deep learning “non-linear” bluster

Recap: why the result matters

Everybody is fitting blockmodels

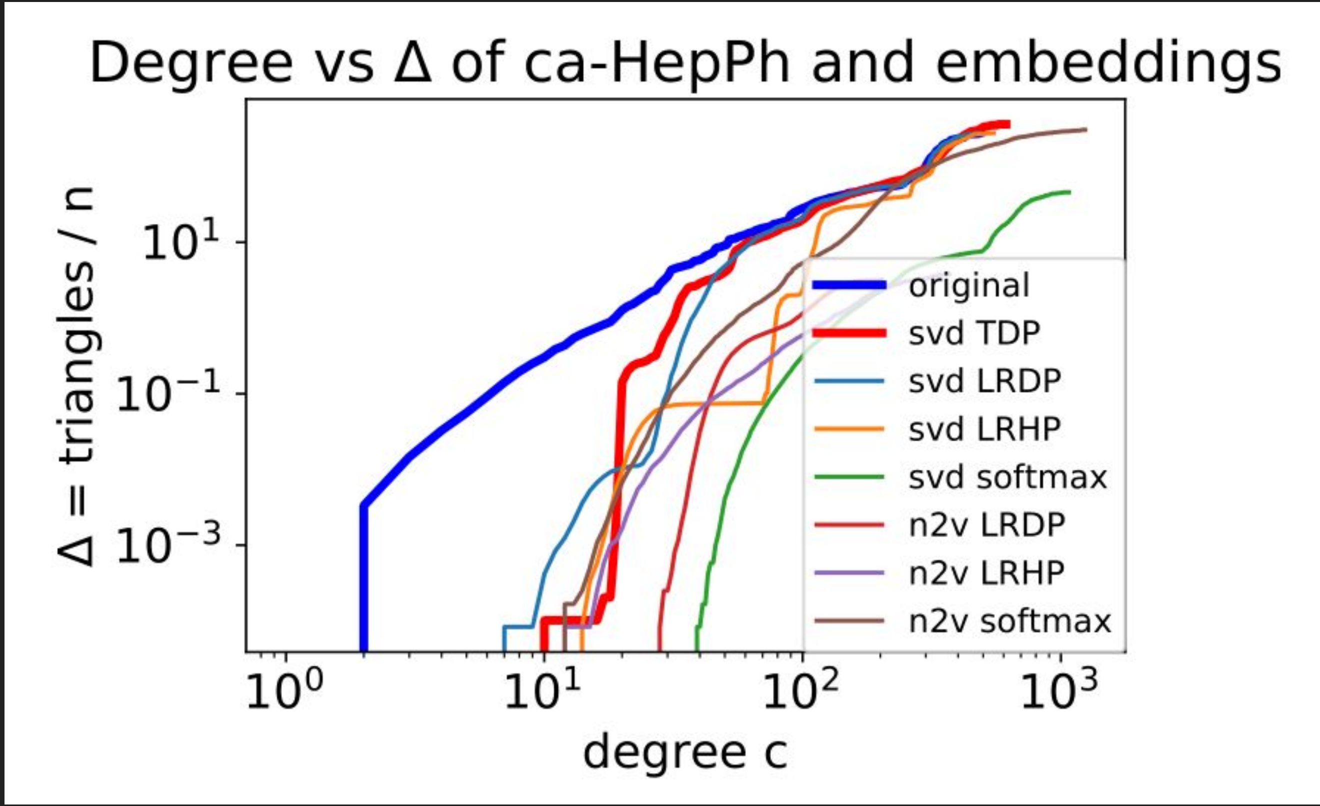
- Statisticians on purpose
- Computer scientists on accident

Blockmodels don’t generate data that looks like real world data

This has been folk wisdom for some time

Now we have a proof

Simulations



Should we care? Maybe

Depends on the context. Often, low-degree nodes are boring and largely uninformative, so who cares if we our generative model doesn’t put them in enough triangles

Sometimes you really care!

- A new user joins Facebook and you want to recommend friends to them!
- Anything involving small, local communities

How to handle misspecification when it matters

1. Don’t throw out stochastic blockmodels!!

There are extremely compelling reasons to fit stochastic blockmodels

Best discrete M-estimators for the graphon

i.e. perform well under misspecification

Graphon the most general vertex exchangeable graph model

The Annals of Statistics
2014, Vol. 42, No. 1, 29–63
DOI: 10.1214/13-AOS1173

CO-CLUSTERING SEPARATELY EXCHANGEABLE NETWORK DATA[1]

By David Choi and Patrick J. Wolfe

Carnegie Mellon University and University College London

This article establishes the performance of stochastic blockmodels in addressing the co-clustering problem of partitioning a binary array into subsets, assuming only that the data are generated by a nonparametric process satisfying the condition of separate exchangeability. We provide oracle inequalities with rate of convergence $\mathcal{O}_P(n^{-1/4})$ corresponding to profile likelihood maximization and mean-square error minimization, and show that the blockmodel can be interpreted in this setting as an optimal piecewise-constant approximation to the generative nonparametric model. We also show for large sample sizes that the detection of co-clusters in such data indicates with high probability the existence of co-clusters of equal size and asymptotically equivalent connectivity in the underlying generative process.

1. Introduction. Blockmodels are popular tools for network modeling that see wide and rapidly growing use in analyzing social, economic and biological systems; see Zhao, Levina and Zhu (2011) and Fienberg (2012) for recent overviews. A blockmodel dictates that the probability of connection between any two network nodes is determined only by their respective block memberships, parameterized by a latent categorical variable at each node.

Fitting a blockmodel to a binary network adjacency matrix yields a clustering of network nodes, based on their shared proclivities for forming connections. More generally, fitting a blockmodel to any binary array involves

Received December 2012; revised September 2013.
[1]Supported in part by the US Army Research Office under PECASE Award W911NF-09-1-0555 and MURI Award W911NF-11-1-0036; by the UK EPSRC under Mathematical Sciences Established Career Fellowship EP/K005413/1 and Institutional Sponsorship Award EP/K503459/1; by the UK Royal Society under a Wolfson Research Merit Award; and by Marie Curie FP7 Integration Grant PCIG12-GA-2012-334622 within the 7th European Union Framework Program.
AMS 2000 subject classifications. Primary 62G05; secondary 05C80, 60B20.
Key words and phrases. Bipartite graph, network clustering, oracle inequality, profile likelihood, statistical network analysis, stochastic blockmodel and co-blockmodel.

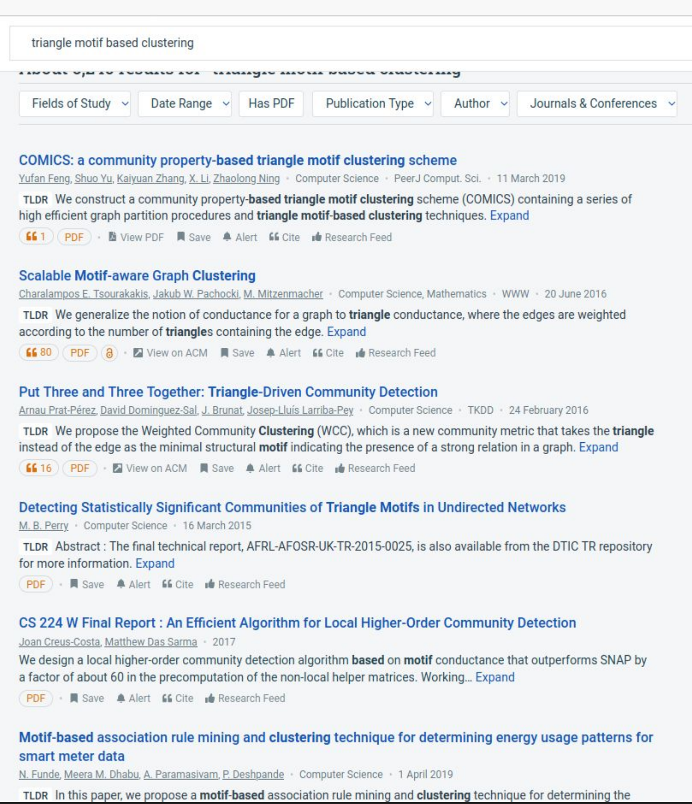
2. Fancier models

- High/infinite dimensional stochastic blockmodels
  - Jing Lei’s graph root distributions
  - Karl is working on something similar
- Hierarchical stochastic blockmodels
- Local models
  - Unpublished but very cool pre-print from Karl Rohe from a couple years ago
- Microclustering

Very cool, very nascent work

3. Better graph embeddings

- High dimensional embeddings
- Embed motif participation rather than edges
- ???
- Something actually non-linear, lol



triangle motif based clustering

Fields of Study | Date Range | Has PDF | Publication Type | Author | Journals & Conferences

COMICS: a community property-based triangle motif clustering scheme
Yufan Feng, Shuo Yu, Kaiyuan Zhang, X. Li, Zhaolong Ning · Computer Science · PeerJ Comput. Sci. · 11 March 2019
TLDR We construct a community property-**based triangle motif clustering** scheme (COMICS) containing a series of high efficient graph partition procedures and **triangle motif-based clustering** techniques. Expand
1 · PDF · View PDF · Save · Alert · Cite · Research Feed

Scalable Motif-aware Graph Clustering
Charalampos E. Tsourakakis, Jakub W. Pachocki, M. Mitzenmacher · Computer Science, Mathematics · WWW · 20 June 2016
TLDR We generalize the notion of conductance for a graph to **triangle** conductance, where the edges are weighted according to the number of **triangles** containing the edge. Expand
80 · PDF · View on ACM · Save · Alert · Cite · Research Feed

Put Three and Three Together: Triangle-Driven Community Detection
Arnau Prat-Pérez, David Dominguez-Sal, J. Brunat, Josep-Lluís Larriba-Pey · Computer Science · TKDD · 24 February 2016
TLDR We propose the Weighted Community **Clustering** (WCC), which is a new community metric that takes the **triangle** instead of the edge as the minimal structural **motif** indicating the presence of a strong relation in a graph. Expand
16 · PDF · View on ACM · Save · Alert · Cite · Research Feed

Detecting Statistically Significant Communities of Triangle Motifs in Undirected Networks
M. B. Perry · Computer Science · 16 March 2015
TLDR Abstract : The final technical report, AFRL-AFOSR-UK-TR-2015-0025, is also available from the DTIC TR repository for more information. Expand
PDF · Save · Alert · Cite · Research Feed

CS 224 W Final Report : An Efficient Algorithm for Local Higher-Order Community Detection
Joan Creus-Costa, Matthew Das Sarma · 2017
We design a local higher-order community detection algorithm **based** on **motif** conductance that outperforms SNAP by a factor of about 60 in the precomputation of the non-local helper matrices. Working… Expand
PDF · Save · Alert · Cite · Research Feed

Motif-based association rule mining and clustering technique for determining energy usage patterns for smart meter data
N. Funde, Meera M. Dhabu, A. Paramasivam, P. Deshpande · Computer Science · 1 April 2019
TLDR In this paper, we propose a **motif-based** association rule mining and **clustering** technique for determining the

Questions?