What is the core distribution of a graph telling us?

Sonja Petrović

Illinois Institute of Technology, Chicago

Joint work with:
Vishesh Karwa (Carnegie Mellon / Harvard),
Michael Pelsmajer (IIT),
Despina Stasi (Univ. of Cyprus / IIT)
Dane Wilburne (IIT)

arXiv:1410.7357 - v2 soon.

- AS2015 -

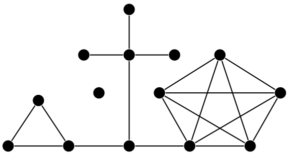
Genova, June 2015

Setting: statistical models for random graphs

How to capture node importance?

In some applications, it matters not just to how many other nodes a particular node in the network is connected, but also to which other nodes it is connected.

→ Is degree-centric analysis suitable? ←



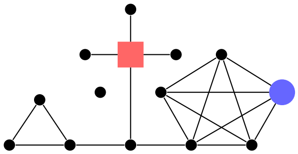
- Examples: information dispersal, the spread of disease or viruses, or robustness to node failure...
- Social network setting: record 'node celebrity status'.

Setting: statistical models for random graphs

How to capture node importance?

In some applications, it matters not just to how many other nodes a particular node in the network is connected, but also to which other nodes it is connected.

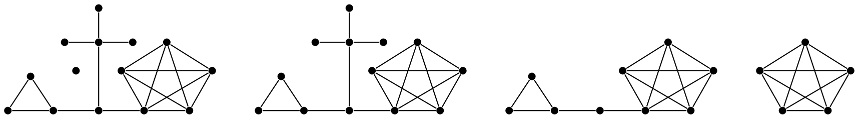
$\rightarrow$ Is degree-centric analysis suitable? $\leftarrow$



- Examples: information dispersal, the spread of disease or viruses, or robustness to node failure...
- Social network setting: record 'node celebrity status'.

Classifying vertices: coreness (a.k.a. shell index)

[Seidman83]: A *k*-core decomposition of a graph captures precisely this:



Any vertex may live in many cores, but only one shell.

Vast literature on:

- Fast computation of shell indices;
- Interesting applications and heuristic studies.

Not surfaced in stats literature so far:

- A rigorous statistical model for networks relying on core structure.

$\rightarrow$ Core structure is summarized by shell distribution. $\leftarrow$

The shell distribution model.

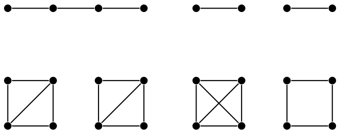
- $G = g$: a random instance of a graph on n nodes
- $n_i(g)$: number of vertices in shell i; p_i: the "shel parameter"

$$P(G = g; p) = \varphi(p) \prod_{i=0}^{n-1} {p_i}^{n_i(g)}$$

Exponential family form

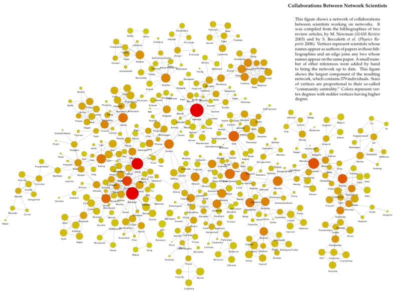
$$P(G = g; p) = \exp\{\sum_{i=0}^{n-2} n_i(g)\theta_i - \psi(\theta)\}.$$

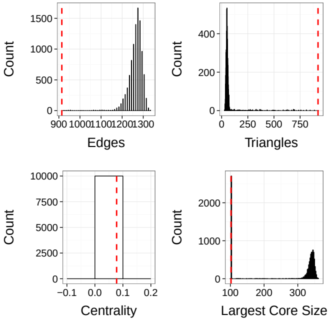
- Shell $\not\leftrightarrow$ degree distribution:
- Erdös-Rényi not a formal submodel
- Log-linear structure only on 'atomic level'.

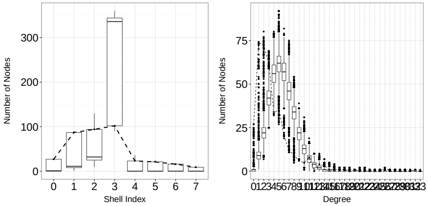


Sampling from the model - Authorship dataset

The largest connected component of the network science co-authorship network (379 nodes)

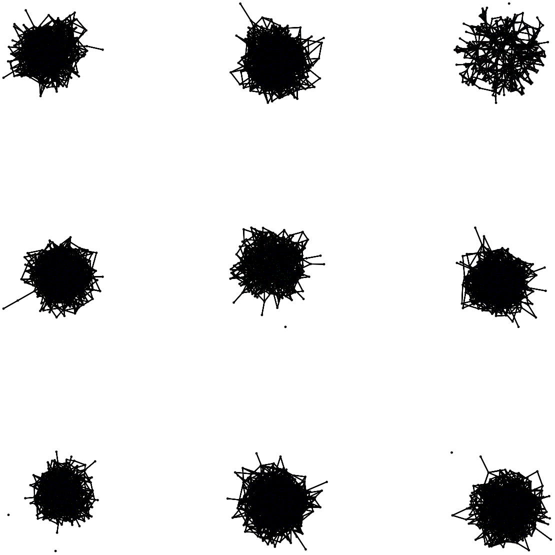






Typical graphs from the model - Authorship dataset

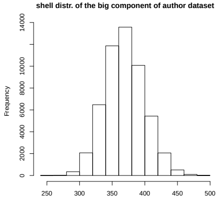
... what to do with this??



Exploring structure of graphs within a fiber

- The author network component core distribution can be realized with graphs that have from about 250 to 500 triangles.
- Simulations: examples from $n = 18$ to $n = 57$ nodes, algorithm never visited the same graph twice, min and max number of triangles differ by a factor of 2 or 3.

A typical histogram of number of triangles:



So what do we have?

- Model that provides necessary formalism for using k-cores in statistical considerations
- Algorithm for constructing all graphs with given shell structure
- MCMC algorithm for simulating from the model

3 problems (... or: the usual ERGM suspects)

- Model fitting questions lead to three important subproblems;

* Solving these is crucial for MLE estimates and goodness of fit tests *

1) Existence of MLE - captured by the model polytope:

Theorem

The polytope of all shell distribution vectors is a dilate of a simplex.
All realizable lattice points lie on the boundary of this polytope.
The MLE never exists for a sample of size 1.

2) Sampling from the fibers (via the Metropolis algorithm):

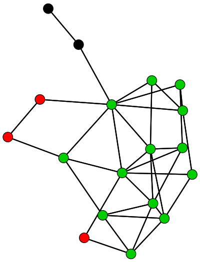
Algorithm

Randomly construct a graph with a given shell distribution.
Constructs all graphs with positive probability.
Experiments: fast graph discovery.

3 problems (continued) (... or: the usual ERGM suspects)

3) Sampling from the model: direct sampling intractable

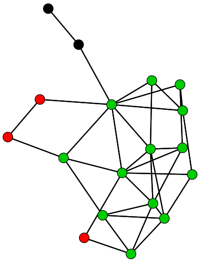
- Sampson data set: 18 monks in a New England Monastery: $n_S(g) = (0, 2, 3, 15, 0, 0, \ldots)$
- MCMC scheme: "tie-no-tie" proposal [Caimo et al] - good mixing
- Probability of accepting: $\pi = \min\left(1, \prod_i p_i^{n_i(g')-n_i(g)}\right)$.

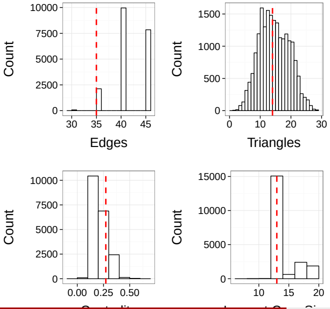


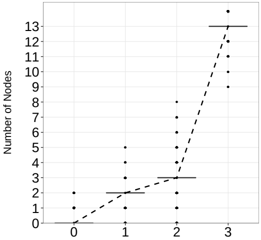
3 problems (continued) (... or: the usual ERGM suspects)

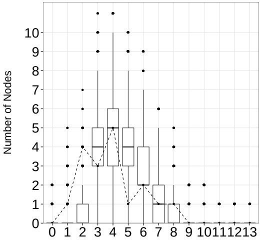
3) Sampling from the model: direct sampling intractable

- Sampson data set: 18 monks in a New England Monastery: $n_S(g) = (0, 2, 3, 15, 0, 0, \dots)$
- MCMC scheme: "tie-no-tie" proposal [Caimo et al] - good mixing
- Probability of accepting: $\pi = \min\left(1, \prod_i p_i^{n_i(g') - n_i(g)}\right)$.

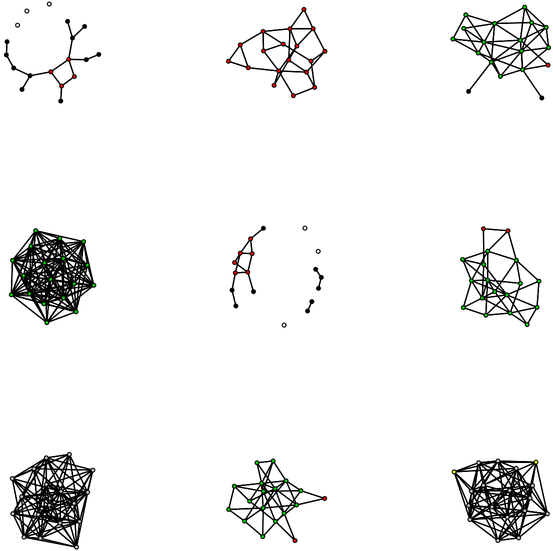








Model degeneracy! - example



Extending the model family

Introduce a parameter for the degeneracy of a graph:

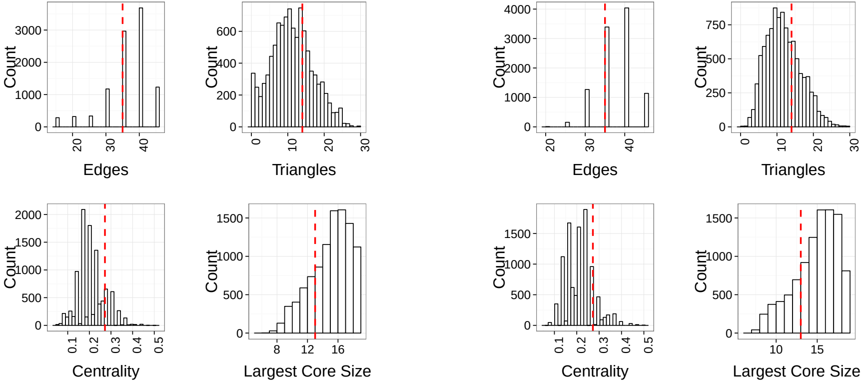
$$P(G = g; p, m) = \varphi(p) \prod_{i=0}^{n-1} {p_i}^{n_i(g)}, \text{if } g \in \mathcal{G}_{n,m} := \{G : dgen(G) \leq m\}.$$

It means that all graphs under this model will have degeneracy at least m.

- Treat m as a parameter (that needs to be estimated)
- analogous to choosing the number of components in a mixture model vs. assuming that it is known.
- We treat m as fixed - select the observed value of m.
- Estimation - open; but at least the new model is not degenerate.

Simulations - Sampson network

Two submodels: support graphs with degeneracy ≤ 3, or $= 3$ (observed).



- Note heavier tails in one model
- Parameter used = good estimate of MLE (moment equations)
  (expected shell distrib. under the MLE very close to observed)

Simulations - Sampson network - typical graphs