# What is the *shell distribution* of a graph telling us?

Vishesh Karwa

Based on joint work with
Michael J. Pelsmajer (IIT)
Sonja Petrović (IIT)
Despina Stasi (Univ of Cyprus/IIT)
Dane Wilburne (IIT)

arXiv:1410.7357 - v2 soon. (Monday?)

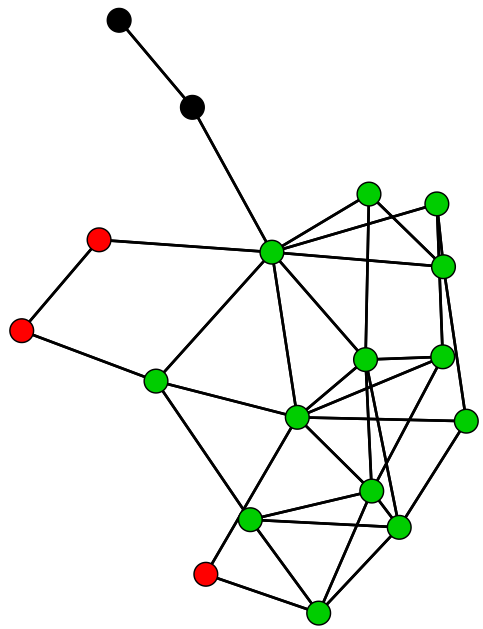
Carnegie Mellon University
Harvard University

AMS Sectional Meeting
Oct 4, 2015

# The $k$-core decomposition of a graph

## Definition (Seidman83)

*The $k$-core of a graph $G$ is the maximal subgraph in which every vertex has degree at least $k$. The shell index of a vertex $i$ is the highest $k$ such that $i$ is contained in the $k$-core of $G$.*



# "Modeling" a graph via its Core decomposition

A core decomposition has been used as a descriptive tool to explain many properties of observed graphs, such as:

1. Core-Periphery or the rich club structure
2. Importance of a node in a network - Robust degree of a node
3. Visualization of network topology by peeling it into layers

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

# Why do we care?

No clear understanding of what the core structure really represents:

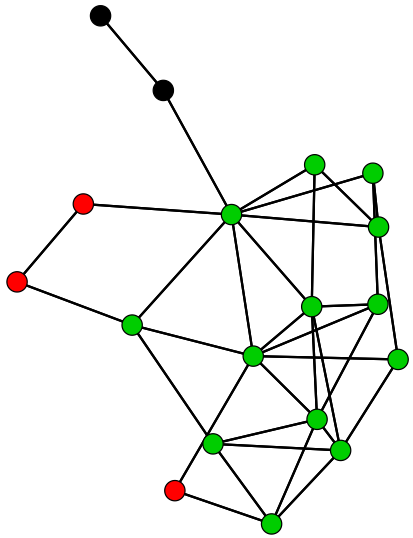
1. 1983,2006: Shell index measures the importance of a node.
2. 2007: Wait, it does not.
3. 2010: But wait, if you take this into account the degrees it does...

How do we make this question precise?
What *properties* of a network does the core structure really capture?

Goal:
How to make the core decomposition a tool for statistical modeling rather than a descriptive analysis?

# Summarizing the $k$-core decomposition

- Recall shell index of a vertex $i$ is the highest $k$ such that $v$ is contained in the $k$-core.
- Shell sequence is the sequence of shell indices of each node.
- Shell distribution is the histogram of shell sequence.



$$n_s(g) = \{0, 2, 3, 13, 0, 0, ..., 0\}$$

# Enter: Exponential random graph models

$$P(G, \theta) = \exp\left\{\sum_{i=1}^{k} \theta_i t_i(g) - \psi(\theta)\right\}$$

- ERGMs are natural statistical tools to model networks through their summary statistics.
- Large growing literature on ERGMs - posses both good and bad (but fixable) properties, see Rinaldo et al. [2009].
- Embed the core structure in the ERGM framework and study it's properties.

# The Family of Shell distribution ERGMs

- $\mathcal{G}_{n,m} := \{g : dgen(g) = m\}^*$
- $m =$ degeneracy parameter
- $\{n_0(g), \ldots, n_i(g), \ldots, n_{m-1}(g)\} =$ shell distribution
- $p_i =$ shell index parameter

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

- For a fixed value of $m$, defines a sub model.

*Can also define the model on $\mathcal{G}_{n,\leq m} = \{g : dgen(g) \leq m\}$

# Exponential family form

- $\theta_0, \ldots, \theta_{m-1}$ = vector of natural parameter where $\theta_i = \log \frac{p_i}{p_m}$

$$P(G = g) = \exp\left\{ \sum_{i=0}^{m-1} n_i(g)\theta_i - \psi(\theta) \right\}.$$

where $\psi(\theta) = \log \sum_{g \in \mathcal{G}_{n,m}} \exp\left\{ \sum_{j=0}^{m-1} n_j(g)\theta_j \right\}$.

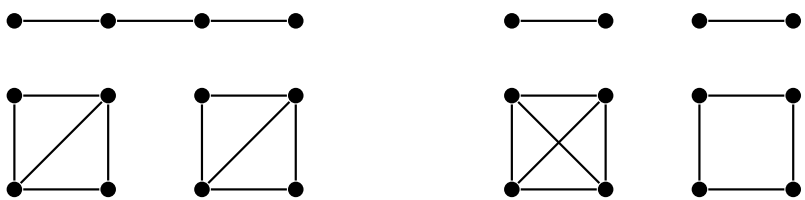
# Exponential family form

- $\theta_0, \ldots, \theta_{m-1}$ = vector of natural parameter where $\theta_i = \log \frac{p_i}{p_m}$

$$P(G = g) = \exp\left\{ \sum_{i=0}^{m-1} n_i(g)\theta_i - \psi(\theta) \right\}.$$

where $\psi(\theta) = \log \sum_{g \in \mathcal{G}_{n,m}} \exp\left\{ \sum_{j=0}^{m-1} n_j(g)\theta_j \right\}$.

- Same degree distribution, different shell distribution.



- Erdös-Rényi *not* a sub model.
- Log-linear model only in "atomic" level.

# Three Inference tasks on ERGMS

1. Characterize the Marginal Polytope - the convex hull of sufficient statistics, conditions for existence of MLE
2. Sampling random graphs from the model - estimation of MLE or Bayesian Inference
3. Sample graphs from the *Fiber* - the set of all graphs with fixed shell distribution - Useful for goodness of fit testing, understanding the space of graphs with fixed shell distribution.

# Marginal Polytope of the model $\mathcal{P}(\mathcal{G}_{n,\leq m})$

- The unrestricted Model $\mathcal{P}(\mathcal{G}_{n,\leq n-1})$

**Theorem**

The marginal polytope of $\mathcal{P}(\mathcal{G}_{n,\leq n-1})$ is a dilate of a simplex.
All realizable lattice points lie on the boundary of this polytope.
The MLE of $\mathcal{P}(\mathcal{G}_{n,n-1})$ never exists for a sample of size $1$.

# Marginal Polytope of the model $\mathcal{P}(\mathcal{G}_{n,\leq m})$

- The unrestricted Model $\mathcal{P}(\mathcal{G}_{n,\leq n-1})$

**Theorem**

The marginal polytope of $\mathcal{P}(\mathcal{G}_{n,\leq n-1})$ is a dilate of a simplex.
All realizable lattice points lie on the boundary of this polytope.
The MLE of $\mathcal{P}(\mathcal{G}_{n,n-1})$ never exists for a sample of size $1$.

- The restricted Model $\mathcal{P}(\mathcal{G}_{n,\leq m})$

**Theorem**

The marginal polytope of $\mathcal{P}(\mathcal{G}_{n,\leq m})$ is a dilate of a simplex.
If $n > 2m$, the polytope has a non-empty interior and the MLE may exist.

# Marginal Polytope of the model $\mathcal{P}(\mathcal{G}_{n,\leq m})$

- The unrestricted Model $\mathcal{P}(\mathcal{G}_{n,\leq n-1})$

**Theorem**

The marginal polytope of $\mathcal{P}(\mathcal{G}_{n,\leq n-1})$ is a dilate of a simplex.
All realizable lattice points lie on the boundary of this polytope.
The MLE of $\mathcal{P}(\mathcal{G}_{n,n-1})$ never exists for a sample of size $1$.

- The restricted Model $\mathcal{P}(\mathcal{G}_{n,\leq m})$

**Theorem**

The marginal polytope of $\mathcal{P}(\mathcal{G}_{n,\leq m})$ is a dilate of a simplex.
If $n > 2m$, the polytope has a non-empty interior and the MLE may exist.

Note - In general, $\mathcal{P}(\mathcal{G}_{n,=m})$ is better behaved than $\mathcal{P}(\mathcal{G}_{n,\leq m})$.

# An MCMC algorithm to Sample from the model

- MCMC scheme: TNT (tie-no-tie) sampler [Hunter et al, Caimo-Friel]
  - instead of selecting a dyad at random whose state it will flip, it first selects a set of either non-edges or edges and swaps one of them: re-weighs the probability of selecting the dyads.
  - better mixing properties.
- Probability of accepting:

$$\pi = \min\left(1, \prod_i p_i^{n_i(g')-n_i(g)} \cdot \frac{P(g' \to g)}{P(g \to g')}\right).$$

- Issue: Computing $n_i(g') - n_i(g)$.

# Understanding the structure of the fiber

## Algorithm to sample graphs with fixed Shell Distribution

1. Constructs an arbitrary graph with a given shell distribution.
2. Does so with positive probability for each graph in the fiber.
3. Fast graph discovery.

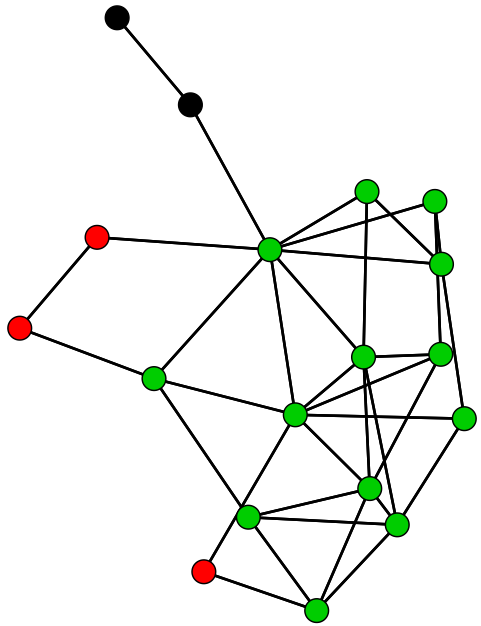
## Bounds on complementary sufficient statistics in the fiber,e.g.,

### Proposition

*The maximum number of triangles for a graph with sorted shell sequence* $s_1 \leq \ldots \leq s_n = m$ *is*

$$\binom{m}{3} + \sum_{i=1}^{n-m} \binom{s_i}{2}.$$

# Application to Sampson Data



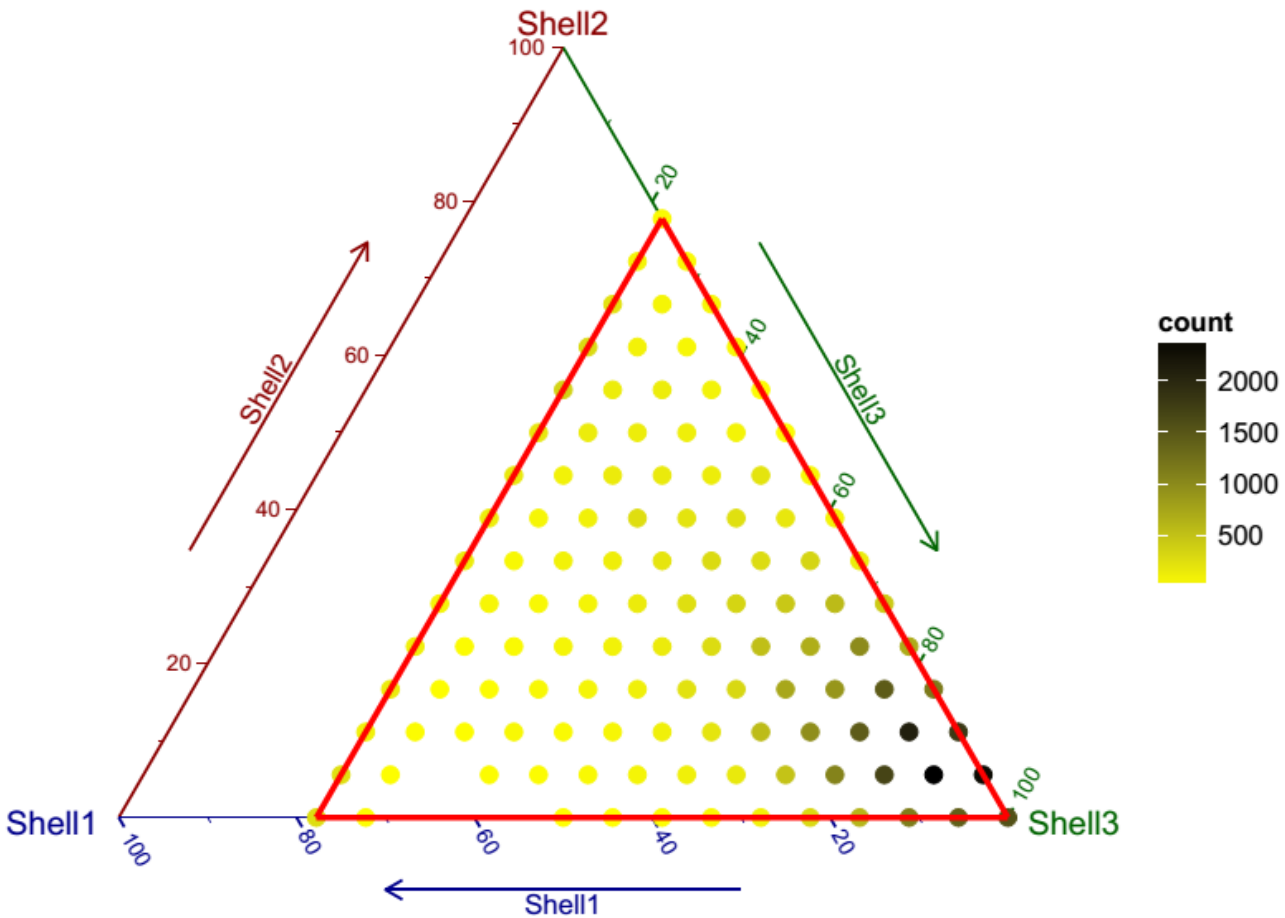
- Sampson data set: $18$ monks in a New England Monastery

$$n_S(g) = (0, 2, 3, 13)$$

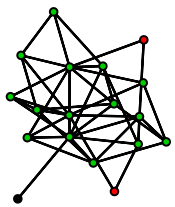
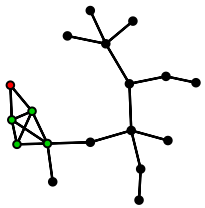
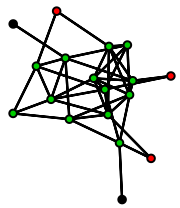
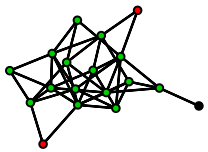
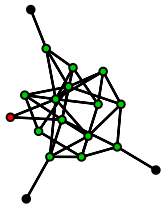
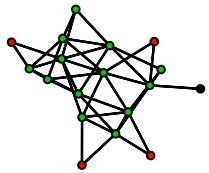
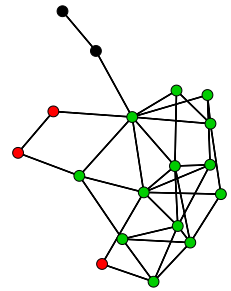
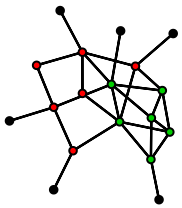
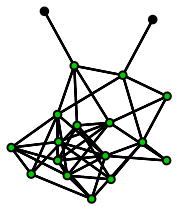
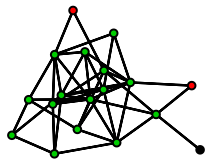
- $\hat{\theta}_{mle} = (-7.95, 2.79, 0.91)$ Estimated using MCMC MLE.
- $\hat{p}_{mle} = (0.00, 0.82, 0.13, 0.05)$.

# The Polytope for the Sampson Data

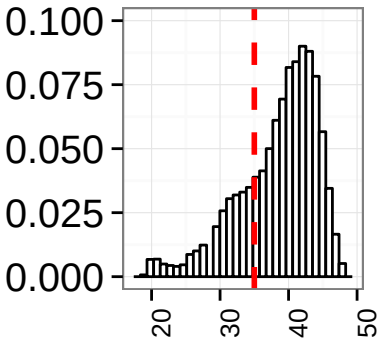
- Samples from $\hat{\theta}_{mle}$ using a 40,000 step MCMC using TNT proposal



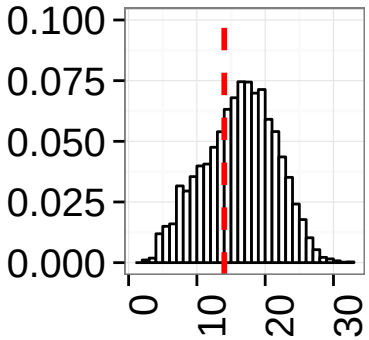
# Typical Graphs from the models

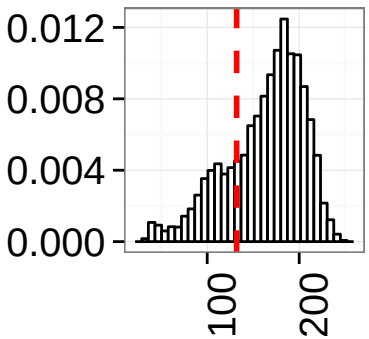
# Histogram of various summary Statistics



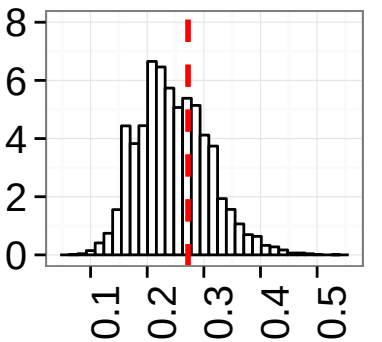

Edges



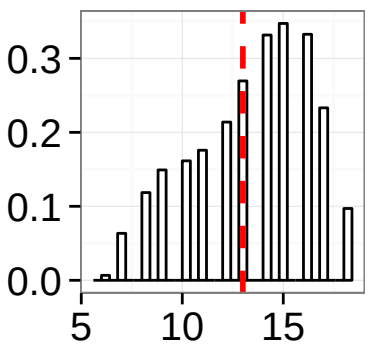

Triangles



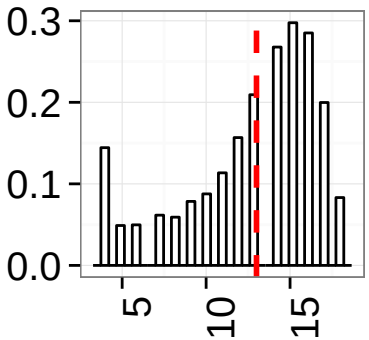

2 stars



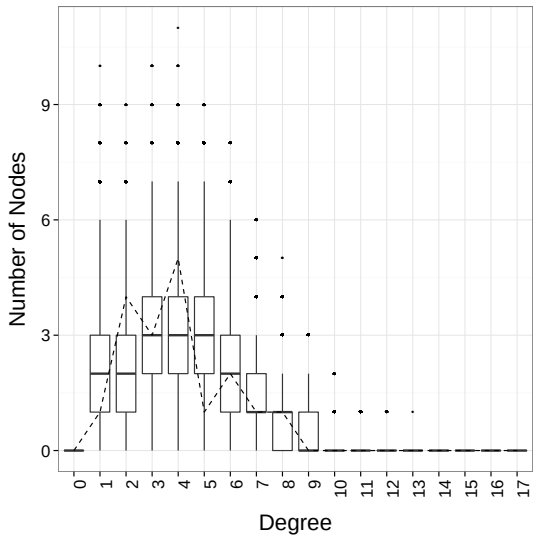

Centrality




Size of largest shell




Size of innermost shell




Degree

# Open Problems

1. Generate uniform random samples from $\mathcal{G}_{n,m}$ and $\mathcal{G}_{n,\leq m}$.
2. Asymptotic formula for the number of graphs in a fiber (e.g. Barvinok and Hartigan for degree sequence)
3. Better Markov chain proposals that move rapidly in the marginal polytope space.
4. Local computation of core distribution to speed up MCMC

Questions?

Thank you for your attention!

arXiv:1410.7357 - v2 soon

# Questions?

Thank you for your attention!