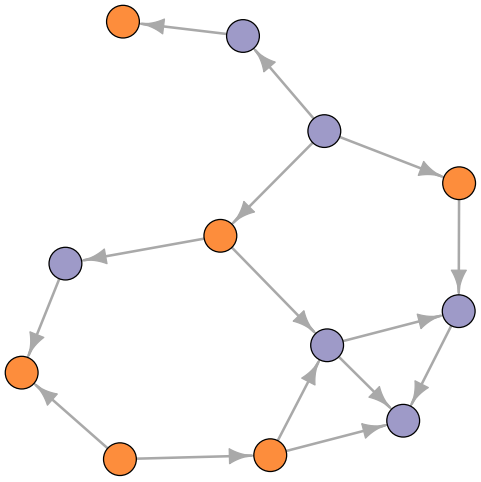
Modeling Network Data

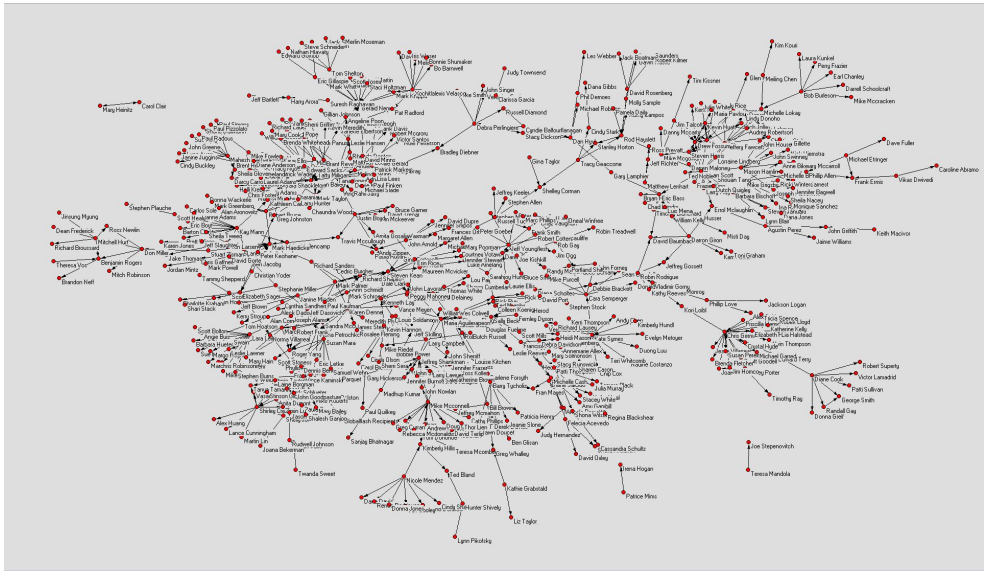
PJ Wolfe

Chair of Statistics
Royal Society Research Fellow
University College London

IISc Workshop on High Dimensional Network Analytics
Bangalore, India, 17 December 2013

Motivation





- Networks capture high-dimensional yet sparse dependency structure
  - Relations, flows, (pairwise) interactions
  - Co-occurrences, correlation, closeness in some metric space
- How do we appropriately elicit models for, and draw inferences about, datasets that take the form of (large) networks?
- One approach is to develop simple null or 'baseline' models for such data, and then check how the data differ from what they predict

Outline

Modeling Simple, Undirected Graphs

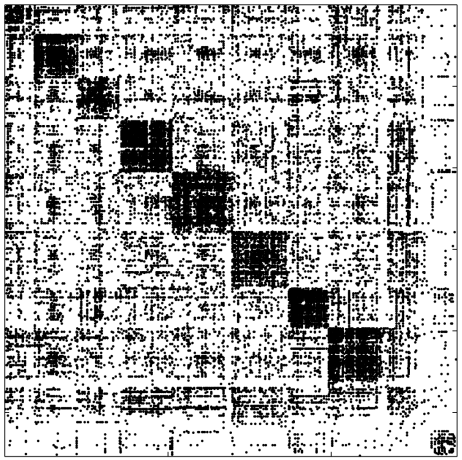
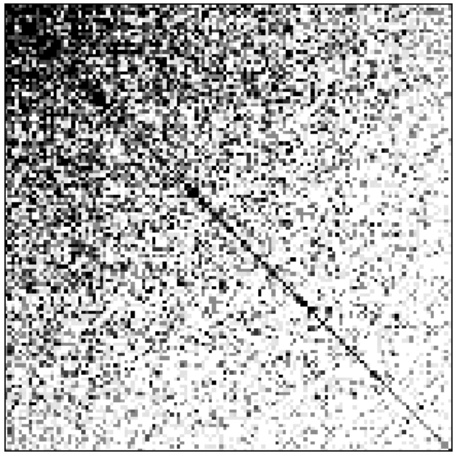
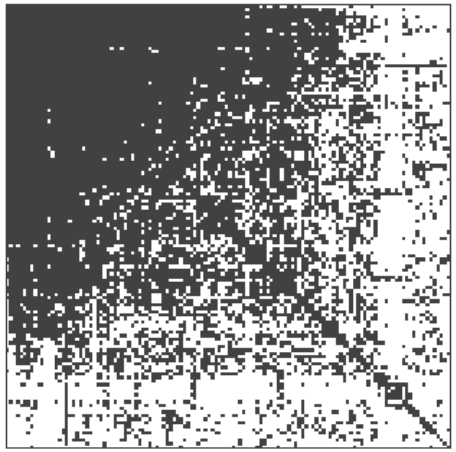




Figure: Adjacency matrices of three network datasets

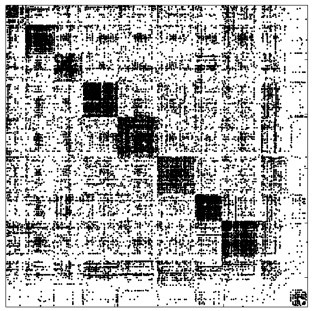
- An undirected, unweighted network on n nodes may be encoded by an $n \times n$ symmetric binary adjacency matrix X
- A simple model specifies $\{X_{ij}, i < j\}$ as independent coin tosses, with edge probability $\mathbb{P}\{X_{ij} = 1\} = p_{ij}$, non-edge probability $1 - p_{ij}$
- If all p_{ij} are fixed to a single value p, we recover the classical random graph model posited by Erdős, Rényi, and others

Null Models

- Why (and how) are null models useful?
  1. Baseline points of comparison for assessing goodness of fit (score tests, analysis of variance: explained vs. unexplained)
  2. Residuals-based analyses (exploratory data analysis, detection of outliers, etc.)
- Traditional approach: Elicit a model and prove what happens when its parameters are estimated from data
  - Estimator consistency (recover 'truth' as data grow large)
  - Leads to classical notion of statistical 'significance'
- Null model approach: Specify a simplistic baseline model intended to capture 'uninteresting' variability in the data
  - Inspect residuals—'observed-minus-expected' values
  - Simpler to specify, easier (hopefully) to fit

Example of a Consistency Result

Consistency results presume the data to follow a particular model. The 'stochastic block' model provides an example:



- Assume all n network nodes form exactly k groups
- Assume coin-toss edges whose probabilities depend only on the group membership of their pair of nodes

Example of a consistency theorem (Choi, Airoldi, & W, 2010):

- Suppose the number of groups k grows with n at a rate no greater than $\sqrt{n}$, and that the expected degree of each node grows with n at a rate greater than the cube of $\log n$.
- Then, as n grows large, it is possible to recover each node's group membership with an overall error rate tending toward zero (convergence in probability under maximum likelihood)

Null Model Approach

- What if we remove the assumption that 'groups' gave rise to the data, and assume only independent coin tosses with potentially arbitrary coins?
- We cannot fit this model to a single network (why not?), but if we have a set of 'candidate' edge probabilities $\{p_{ij}\}_{i<j}$ in mind, we can appeal to a null model result as follows:
- We can check if 'expected' edge summaries under this model deviate from the actual edge summaries we observe
- To do this, consider carving up the data into groups of nodes and averaging the p_{ij}'s as well as the observed edges $\{X_{ij}\}$
- Recall that the 'expected' edge value at location i, j is p_{ij}, so we are going down the path of 'observed-minus-expected'

Example of a Deviations Result

- Arrange observed, expected averages into symmetric matrices $\hat{\theta}, \theta$, whose dimension will be equal to the number of groups
- To check whether the observed averages deviate from expected, we need a 'distance' between $\hat{\theta}$ and θ
- The usual choice would be the sum of squared differences across elements of $\hat{\theta}$ and θ.
- Consider instead the weighted sum $\sum_{a \leq b} n_{ab} D(\hat{\theta}_{ab} || \theta_{ab})$, with n_{ab} the block size induced by groups a and b, and

$$D(p||p') = p \log \frac{p}{p'} + (1-p) \log \frac{1-p}{1-p'}$$

  is a distance-like quantity satisfying $D(p||p') \geq 2(p-p')^2$
- We now obtain a deviations-based theorem as follows...

Example of a Deviations Result

Theorem (Choi, Airoldi, & W, 2011)

Let $\{X_{ij}\}_{i<j}$ *be comprised of* $\binom{n}{2}$ *independent* Bernoulli(p_{ij}) *trials, and let* $\mathcal{G} = \{1, \ldots, k\}^n$. *Then with probability at least* $1 - \delta$,

$$\max_{g \in \mathcal{G}} \sum_{a \leq b} n_{ab} D(\hat{\theta}_{ab}^{(g)} || \theta_{ab}^{(g)}) \leq n \log k + (k^2 + k) \log \left(\frac{n}{k} + 1 \right) + \log \frac{1}{\delta}$$

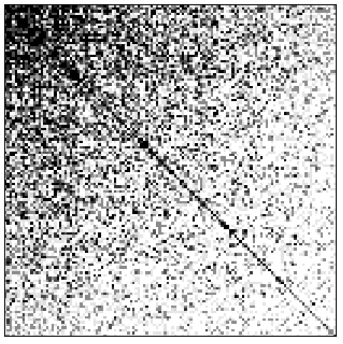
Proof sketch:

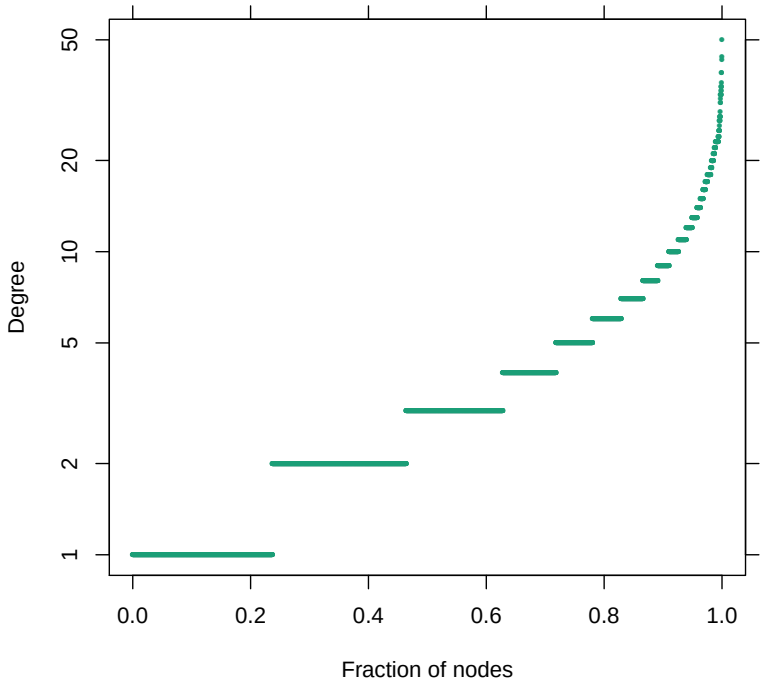
- For fixed g, the probability of any realization of $\hat{\theta}$ is first bounded by $\exp\{-\sum_{a \leq b} n_{ab} D(\hat{\theta}_{ab} \,||\, \theta_{ab})\}$
- A counting argument then yields a deviation result in terms of $(n/k + 1)^{k^2+k}$
- Finally, a union bound is applied so that the result holds uniformly over all k^n possible choices of g

Outline

Degree Models

- The degree of each network node refers to its number of neighbors, which we may write as $X_{i+} = \sum_{j=1}^{n} X_{ij}$. Note that $X_{++} = \sum_{i=1}^{n} X_{i+}$ gives twice the number of total edges
- Stochastic block models do not capture degree heterogeneity often present in network data (co-authorship network, right)
- An alternative approach is to model each node's link-forming propensity via a latent variable. We now consider this approach





A Log-Additive Parameterization

- Consider a simple log-additive latent variable model given by

$$\log p_{ij} = \alpha_i + \alpha_j, \quad 1 \leq i < j \leq n,$$

  with vector α specifying n unknown parameters
- Let $\varepsilon = \{\varepsilon_{ij} : i \neq j\}$ be a family of smooth functions mapping pairs of reals to reals, with $\varepsilon_{ij}(x, y) = \varepsilon_{ji}(y, x)$
- If we then specify p_{ij} as

$$\log p_{ij} = \alpha_i + \alpha_j + \varepsilon_{ij}(\alpha_i, \alpha_j),$$

  we can extend this model to other link functions; e.g., a logit-link model has $\varepsilon_{ij}(x, y) = -\log\{1 + \exp(x + y)\}$

The Connection to Degree Models

- When the observed graph is not too star-like, we show that a maximum likelihood estimate of p_{ij} under this model exists and is close to $\tilde{p}_{ij}$, where

$$\tilde{p}_{ij} = \frac{X_{i+} X_{j+}}{X_{++}},$$

whenever the functions in ε and their derivatives are controlled

- Furthermore $\tilde{\alpha}$ is close to a maximum likelihood estimate of parameter vector α, where

$$\tilde{\alpha}_i = \log X_{i+} - \frac{1}{2} \log X_{++} \qquad \text{for } i = 1, \ldots, n$$

such that

$$\tilde{p}_{ij} = \exp(\tilde{\alpha}_i + \tilde{\alpha}_j)$$

- Thus we obtain nearly a maximum likelihood estimate of α_i by a monotone transformation of the ith nodal degree X_{i+}

Intuition

- Intuition: if p_{ij} is small, then a Bernoulli(p_{ij}) random variable behaves like a Poisson random variable having the same mean
- Under the log-linear parameterization, the corresponding Poisson log-likelihood $\sum_{i<j} X_{ij} \log p_{ij} - p_{ij}$ becomes

$$\sum_{i<j} X_{ij}(\alpha_i + \alpha_j) - \exp(\alpha_i + \alpha_j) = \sum_{i=1}^{n} \alpha_i X_{i+} - \sum_{i \neq j} \exp(\alpha_i + \alpha_j),$$

  thus a stationary point must have $X_{i+} = \sum_{j \neq i} \exp(\alpha_i + \alpha_j)$
- When α is set to $\tilde{\alpha}$, observe

$$\sum_{j \neq i} \exp(\tilde{\alpha}_i + \tilde{\alpha}_j) = \frac{X_{i+}}{X_{++}} \sum_{j \neq i} X_{j+} = X_{i+} \left(1 - \frac{X_{i+}}{X_{++}}\right),$$

  and so we see that each component of the Poisson score evaluated at $\tilde{\alpha}$ is precisely X_{i+}^2 / X_{++}.

Theorem (Perry & W, 2011)

Consider the family of models $\log p_{ij} = \alpha_i + \alpha_j + \varepsilon_{ij}(\alpha_i, \alpha_j)$, *having the property that for all pairs* i, j *and all choices of* k, l, *and* m, *the functions* ε_{ij}, $\partial\varepsilon_{ij}/\partial\alpha_k$, $\partial^2\varepsilon_{ij}/(\partial\alpha_k\partial\alpha_l)$, *and* $\partial^3\varepsilon_{ij}/(\partial\alpha_k\partial\alpha_l\partial\alpha_m)$, *are sub-exponential in* $\alpha_i + \alpha_j$ *with constant* C_0.

Then if all observed degrees X_{i+} *satisfy* $1 \leq X_{i+}^2 \leq \varepsilon_0 X_{++}$, *with* $\varepsilon_0 \leq \{15(C_0 + 1)\}^{-2}$, *there exists a solution* $\hat{\alpha}$ *to the likelihood equation under these models such that*

$$\|\hat{\alpha} - \tilde{\alpha}\|_\infty \leq 10(C_0 + 1)\varepsilon_0.$$

- One can also show that the corresponding estimates of p_{ij} are within a universal constant of their maximizers, as well as the value of the log-likelihood itself

Example: Co-Authorship Network Shown Earlier

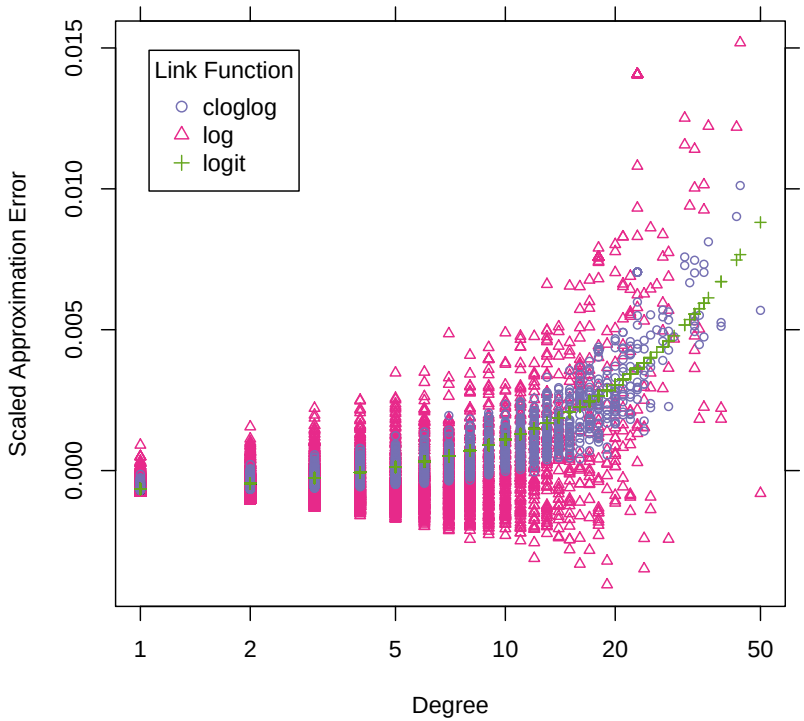


Figure: Scaled approximation error $(\hat{\alpha}_i - \tilde{\alpha}_i)/(C\varepsilon_0)$, with $C = 10\,(C_0+1)$

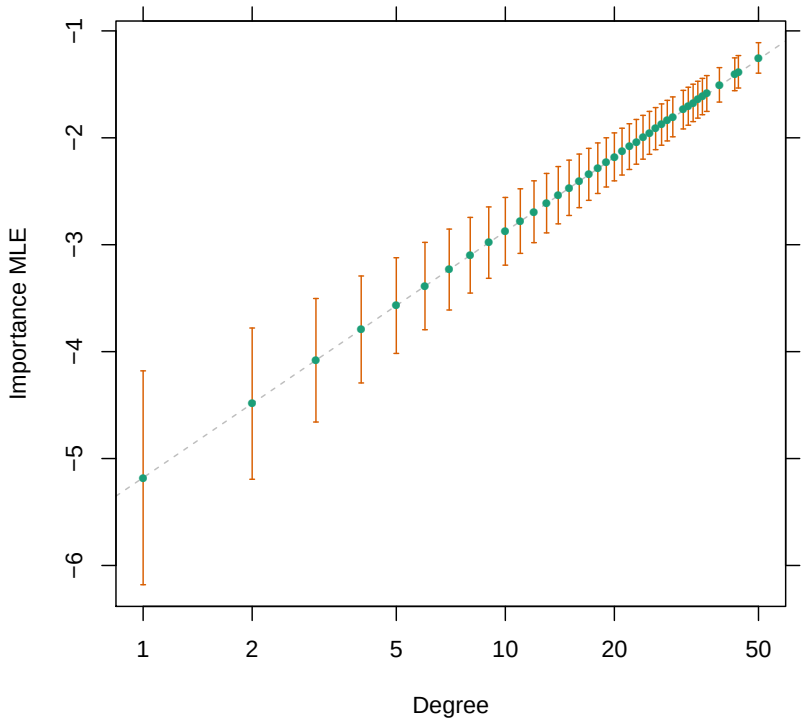


Figure: Standard errors, $\hat{\alpha}_i$, and $\tilde{\alpha}_i$ (dashed), as a function of X_{i+}

- C_0 is equal to 0 for the log link, 1/2 for the complementary log-log link, and 1 for the logistic function link

Approximation Error Across Popular Datasets

Dataset	n	X_{++}	$\max X_{i+}$	Link	Valid %	$\frac{\|\hat{\alpha}-\tilde{\alpha}\|_2}{\sqrt{n}\, C\, \varepsilon_0}$	$\frac{\|\hat{\alpha}-\tilde{\alpha}\|_\infty}{C\, \varepsilon_0}$
KARATE	34	156	17	cloglog	0	0.004	0.01
				log	0	0.006	0.02
				logit	0	0.009	0.03
FOOTBALL	115	1226	12	cloglog	0	0.02	0.02
				log	0	0.005	0.01
				logit	0	0.02	0.03
JAZZ	198	5484	100	cloglog	6	0.004	0.02
				log	7	0.002	0.02
				logit	4	0.005	0.02
CELEGANS	453	4050	237	cloglog	5	5e-04	0.004
				log	36	6e-04	0.009
				logit	5	6e-04	0.005
POLBLOGS	1224	33430	351	cloglog	42	9e-04	0.006
				log	50	0.001	0.02
				logit	38	0.002	0.01
NETSCIENCE	1461	5484	34	cloglog	63	0.002	0.01
				log	75	0.003	0.02
				logit	46	0.001	0.01
POWER	4941	13188	19	cloglog	93	0.001	0.01
				log	97	0.002	0.02
				logit	80	0.001	0.01
HEP-TH	7610	31502	50	cloglog	87	9e-04	0.01
				log	94	0.001	0.02
				logit	78	8e-04	0.009

Outline

Modularity and Community Detection

- Our degree-based model posits that edges are independent and appear with a probability p_{ij} depending exclusively on the corresponding pair of nodal parameters (α_i, α_j)
- What about variability left unexplained by this model...? Recall our stochastic block model notion of a nodal partition g
- Newman (2004) suggested to evaluate community structure for any g based on the notion of network modularity Q:

$$Q(g) = \sum_{i<j} \left(x_{ij} - \frac{x_{i+} x_{j+}}{x_{++}} \right) \delta(g_i, g_j),$$

with $\delta(g_i, g_j) = 1$ iff $g_i = g_j$ (nodes i, j are in the same group)

- This boils down to a graph-based residuals analysis with respect to our earlier family of degree-based models

Generalized Modularity

- The appearance of degree centrality is certainly suggestive of a 'null' model. . . Is there indeed a valid 'alternative' model underlying modularity?
- The key is the following generalization, due to Reichardt & Stefan (2006), which incorporates a scale parameter $\gamma \geq 1$:

$$Q_\gamma(g) = \sum_{i<j} (X_{ij} - \gamma\, p_{ij}^{(0)})\, \delta(g_i, g_j)$$

- Here $p_{ij}^{(0)}$ is the probability of edge $i \sim j$ appearing in the absence of group structure—as per a degree-based model
- The modularity $Q_\gamma(g)$ of node partition g is hence motivated as a sum of residuals between the observed within-community edges and the null expected within-community edges

Modularity Alternate

- The null model featured in modularity $Q_\gamma(g)$ may be defined by an n-vector α of nodal parameters as seen earlier:

$$\text{logit}\, p_{ij}^{(0)} = \alpha_i + \alpha_j$$

(Observed degrees are a sufficient statistic for this model)

- The alternate requires a partition g and strength factor $\lambda > 0$:

$$\text{logit}\, p_{ij} = \text{logit}\, p_{ij}^{(0)} + \lambda\, \delta(g_i, g_j)$$

- We can show that the corresponding log-likelihood $\ell(\lambda, g)$ is essentially a monotone transformation of $Q_\gamma(g)$, with

$$\gamma = \gamma(\lambda) = \frac{\exp(\lambda) - 1}{\lambda} \geq 1$$

Likelihood Formalism of Modularity

Theorem (Perry & W, 2011)

Let the null and alternate models for modularity $Q_\gamma(g)$ be as defined previously, and consider the difference in log-likelihoods under models $\lambda > 0$ vs. $\lambda = 0$, where $\gamma = \gamma(\lambda) = [\exp(\lambda) - 1]/\lambda$.

If eventually $\max_{ij}\{p_{ij}^{(0)}\}$ is small enough, then this difference can be written as

$$\ell(g, \lambda) - \ell(g, 0) = \lambda\, Q_\gamma(g) + \mathcal{O}(\lambda^2)$$

- This result justifies modularity as an (approximate) generalized likelihood ratio test statistic, and opens the door to formal hypothesis testing, uncertainty quantification, and the like

Summary

1. Basic Null Model Concepts
   - Example of a consistency result
   - A related null model deviations result
2. A Family of Degree-Based Null Models
   - Parameterization and fitting
   - Data examples
3. Null Models and Community Detection
   - A null model for residuals-based community detection
   - Likelihood-based evaluation of community structure

References

- P. O. Perry and P. J. Wolfe, "Null models for network data," in review for *Biometrika*, DOI arXiv:1201.5871.
- P. O. Perry and P. J. Wolfe, "Point process modeling for directed interaction networks," in review for *J. Roy. Statist. Soc., B*, DOI arXiv:1011.1703.
- D. S. Choi, P. J. Wolfe, and E. O. Airoldi, "Stochastic blockmodels with growing number of classes," *Biometrika*, DOI arXiv:1011.4644, in press.
- E. O. Airoldi, D. S. Choi, and P. J. Wolfe, "Confidence sets for network structure," *Statistical Analysis and Data Mining*, vol. 4, pp. 559–563, 2011, DOI arXiv:1105.6245. (Preliminary version: NIPS 2011.)
- P. O. Perry and P. J. Wolfe, "Dual approaches to network science," presented at the MIT WIDS network symposium, 2011.

NSF-DMS/MSBS/CISE, DARPA, ARO MURI and PECASE support is gratefully acknowledged