NETWORKS 2021

MANCHESTER 1824
The University of Manchester

Goodness of Fit Tests for Random Multigraph Models

Termeh Shafie
Department of Social Statistics
The Mitchell Centre for Social Network Analysis
The University of Manchester

multigraph representation of network data

graphs where multiple edges and self-edges are permitted

- can appear directly in applications (although scarce)
- can be constructed by different kinds of aggregations in graphs
  - aggregation based on vertex attributes
  - aggregation based on edge attributes

composition

structure

multigraph representation of network data

- multigraphs represented by their edge multiplicity sequence

$$\mathbf{m} = (m_{ij} : (i,j) \in R)$$

where R is the canonical site space for undirected edges

$$R = \{(i,j) : 1 \leq i \leq j \leq n\}$$

$$(1,1) < (1,2) < \ldots < (1,n) < (2,2) < (2,3) < \ldots < (n,n)$$

- the number of vertex pair sites is given by $r = \binom{n+1}{2}$
- edge multiplicities as entries in a matrix

$$\mathbf{m} = \begin{bmatrix} m_{11} & m_{12} & \ldots & m_{1n} \\ 0 & m_{22} & \ldots & m_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & m_{nn} \end{bmatrix} \qquad \mathbf{m}+\mathbf{m}' = \begin{bmatrix} 2m_{11} & m_{12} & \ldots & m_{1n} \\ m_{12} & 2m_{22} & \ldots & m_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ m_{1n} & m_{2n} & \ldots & 2m_{nn} \end{bmatrix}$$

Shafie, T. (2015). A Multigraph Approach to Social Network Analysis. *Journal of Social Structure*, 16.

random multigraph models

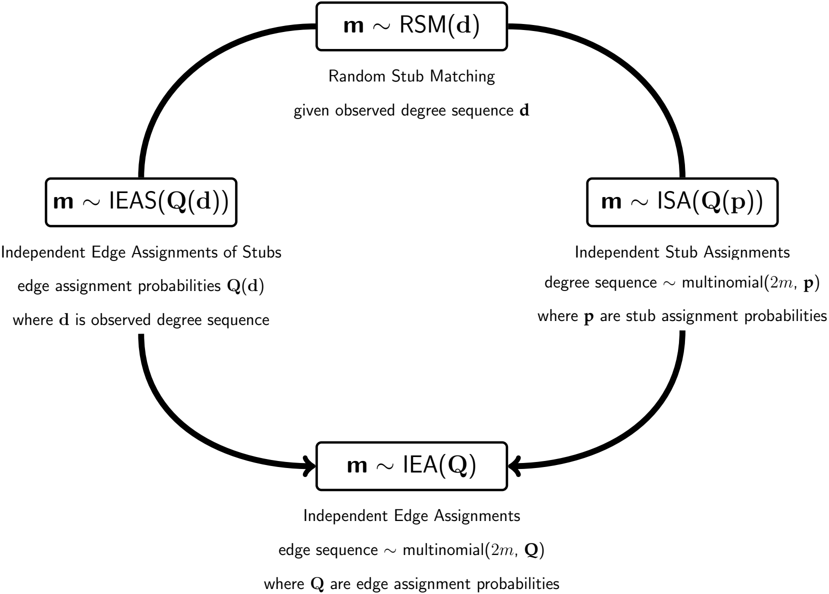
1. **random stub matching (RSM)**
   - edges are assigned to sites given fixed degree sequence $\mathbf{d} = (d_1, \ldots, d_n)$
   - probability that an edge is assigned to site $(i, j) \in R$

$$Q_{ij} = \begin{cases} \binom{d_i}{2} \Big/ \binom{2m}{2} & \text{for } i = j \\ d_i d_j \Big/ \binom{2m}{2} & \text{for } i < j \end{cases}$$

2. **independent edge assignment (IEA)**
   - edges are independently assigned to vertex pairs in site space R
   - edge assignment probabilities $\mathbf{Q} = (Q_{ij} : (i, j) \in R)$
   - $\mathbf{m}$ is multinomial distributed with parameters m and $\mathbf{Q}$
   - statistics for analysing local and global structure are easily derived
   - two variants:
     - **independent edge assignment of stubs (IEAS)**
     - **independent stub assignment (ISA)**

Shafie, T. (2016). Analyzing local and global properties of multigraphs. *The Journal of Mathematical Sociology*, 40(4), 239-264.

random multigraph models



$\mathbf{m} \sim \mathsf{RSM}(\mathbf{d})$

Random Stub Matching

given observed degree sequence $\mathbf{d}$

$\mathbf{m} \sim \mathsf{IEAS}(\mathbf{Q}(\mathbf{d}))$

Independent Edge Assignments of Stubs

edge assignment probabilities $\mathbf{Q}(\mathbf{d})$

where $\mathbf{d}$ is observed degree sequence

$\mathbf{m} \sim \mathsf{ISA}(\mathbf{Q}(\mathbf{p}))$

Independent Stub Assignments

degree sequence $\sim$ multinomial($2m$, $\mathbf{p}$)

where $\mathbf{p}$ are stub assignment probabilities

$\mathbf{m} \sim \mathsf{IEA}(\mathbf{Q})$

Independent Edge Assignments

edge sequence $\sim$ multinomial($2m$, $\mathbf{Q}$)

where $\mathbf{Q}$ are edge assignment probabilities

goodness of fit

gof measures between observed and expected edge multiplicity sequence under simple or composite hypothesis

- test statistics: S of Pearson and A of information divergence type
- expected values of the Pearson statistic are derived
- exact distributions of the test statistics are numerically investigated

answers sought to the following:

- are significance levels of test statistics for small number of edges far from those of the asymptotic distribution?
- is the convergence of the cdf's of test statistics slow or rapid?
- does the convergence speed depend on specific parameters in models?
- can better approximations to the actual distributions be obtained using adjustments of the χ^2-distributions?
- can power approximations be made for small number of edges?
- how does RSM influence the distributions of statistics?
- how can RSM be tested?

tests of a simple multigraph hypothesis

edge multiplicities according to $\mathrm{IEA}(\mathbf{Q})$ and correct model $\mathbf{Q_0} = \mathbf{Q}$ tested:

- the Pearson statistic

$$S_0 = \sum\sum_{i \leq j} \frac{(m_{ij} - mQ_{0ij})^2}{mQ_{0ij}} = \sum\sum_{i \leq j} \frac{m_{ij}^2}{mQ_{0ij}} - m \overset{asymp}{\sim} \chi^2(r-1)$$

- the divergence statistic

$$D_0 = \sum\sum_{i \leq j} \frac{m_{ij}}{m} \log \frac{m_{ij}}{mQ_{0ij}} \quad \text{and} \quad A_0 = \frac{2m}{\log \mathrm{e}} D_0 \overset{asymp}{\sim} \chi^2(r-1)$$

tests of a composite multigraph hypothesis

the composite multigraph hypothesis

- ISA for unknown **p**
- IEAS for unknown **d**

parameters have to be estimated from data **m**

when correct model is tested:

- the Pearson statistic

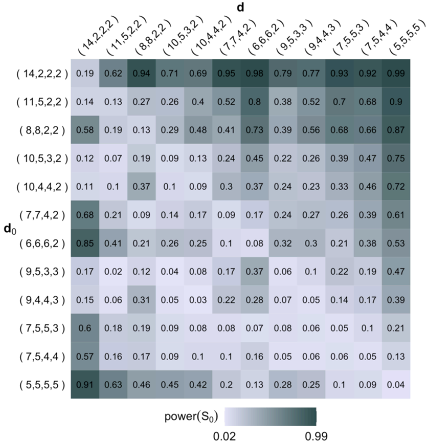
$$\hat{S} = \sum_{i \le j}\sum \frac{(m_{ij} - m\hat{Q}_{ij})^2}{m\hat{Q}_{ij}} = \sum_{i \le j}\sum \frac{m_{ij}^2}{m\hat{Q}_{ij}} - m \overset{asymp}{\sim} \chi^2(r-n)$$

- the divergence statistic

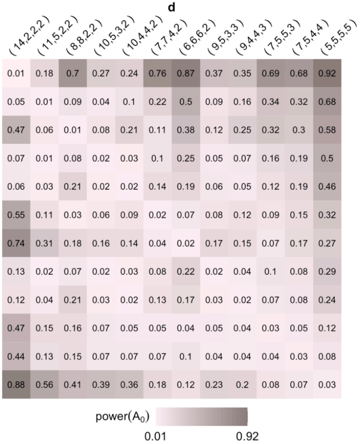
$$\hat{D} = \sum_{i \le j}\sum \frac{m_{ij}}{m} \log \frac{m_{ij}}{m\hat{Q}_{ij}} \qquad \text{and} \quad \hat{A} = \frac{2m}{\log \mathrm{e}}\hat{D} \overset{asymp}{\sim} \chi^2(r-n)$$

goodness of fit: test illustrations

power when simple IEAS($\mathbf{d}_0$) hypotheses are tested against IEAS($\mathbf{d}$) models for multigraphs with $n = 4$, $m = 10$ and $\alpha(\chi^2_9) = 0.04$



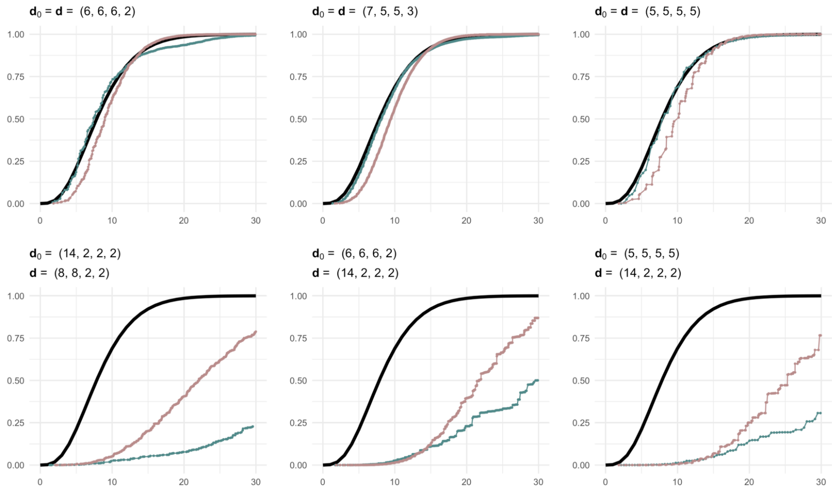
	(14,2,2,2)	(11,5,2,2)	(8,8,2,2)	(10,5,3,2)	(10,4,4,2)	(7,7,4,2)	(6,6,6,2)	(9,5,3,3)	(9,4,4,3)	(7,5,5,3)	(7,5,4,4)	(5,5,5,5)
(14,2,2,2)	0.19	0.62	0.94	0.71	0.69	0.95	0.98	0.79	0.77	0.93	0.92	0.99
(11,5,2,2)	0.14	0.13	0.27	0.26	0.4	0.52	0.8	0.38	0.52	0.7	0.68	0.9
(8,8,2,2)	0.58	0.19	0.13	0.29	0.48	0.41	0.73	0.39	0.56	0.68	0.66	0.87
(10,5,3,2)	0.12	0.07	0.19	0.09	0.13	0.24	0.45	0.22	0.26	0.39	0.47	0.75
(10,4,4,2)	0.11	0.1	0.37	0.1	0.09	0.3	0.37	0.24	0.23	0.33	0.46	0.72
(7,7,4,2)	0.68	0.21	0.09	0.14	0.17	0.09	0.17	0.24	0.27	0.26	0.39	0.61
(6,6,6,2)	0.85	0.41	0.21	0.26	0.25	0.1	0.08	0.32	0.3	0.21	0.38	0.53
(9,5,3,3)	0.17	0.02	0.12	0.04	0.08	0.17	0.37	0.06	0.1	0.22	0.19	0.47
(9,4,4,3)	0.15	0.06	0.31	0.05	0.03	0.22	0.28	0.07	0.05	0.14	0.17	0.39
(7,5,5,3)	0.6	0.18	0.19	0.09	0.08	0.08	0.07	0.08	0.06	0.05	0.1	0.21
(7,5,4,4)	0.57	0.16	0.17	0.09	0.1	0.1	0.16	0.05	0.06	0.06	0.05	0.13
(5,5,5,5)	0.91	0.63	0.46	0.45	0.42	0.2	0.13	0.28	0.25	0.1	0.09	0.04



	(14,2,2,2)	(11,5,2,2)	(8,8,2,2)	(10,5,3,2)	(10,4,4,2)	(7,7,4,2)	(6,6,6,2)	(9,5,3,3)	(9,4,4,3)	(7,5,5,3)	(7,5,4,4)	(5,5,5,5)
(14,2,2,2)	0.01	0.18	0.7	0.27	0.24	0.76	0.87	0.37	0.35	0.69	0.68	0.92
(11,5,2,2)	0.05	0.01	0.09	0.04	0.1	0.22	0.5	0.09	0.16	0.34	0.32	0.68
(8,8,2,2)	0.47	0.06	0.01	0.08	0.21	0.11	0.38	0.12	0.25	0.32	0.3	0.58
(10,5,3,2)	0.07	0.01	0.08	0.02	0.03	0.1	0.25	0.05	0.07	0.16	0.19	0.5
(10,4,4,2)	0.06	0.03	0.21	0.02	0.02	0.14	0.19	0.06	0.05	0.12	0.19	0.46
(7,7,4,2)	0.55	0.11	0.03	0.06	0.09	0.02	0.07	0.08	0.12	0.09	0.15	0.32
(6,6,6,2)	0.74	0.31	0.18	0.16	0.14	0.04	0.02	0.17	0.15	0.07	0.17	0.27
(9,5,3,3)	0.13	0.02	0.07	0.02	0.03	0.08	0.22	0.02	0.04	0.1	0.08	0.29
(9,4,4,3)	0.12	0.04	0.21	0.03	0.02	0.13	0.17	0.03	0.02	0.07	0.08	0.24
(7,5,5,3)	0.47	0.15	0.16	0.07	0.05	0.05	0.04	0.05	0.04	0.03	0.05	0.12
(7,5,4,4)	0.44	0.13	0.15	0.07	0.07	0.07	0.1	0.04	0.04	0.04	0.03	0.08
(5,5,5,5)	0.88	0.56	0.41	0.39	0.36	0.18	0.12	0.23	0.2	0.08	0.07	0.03

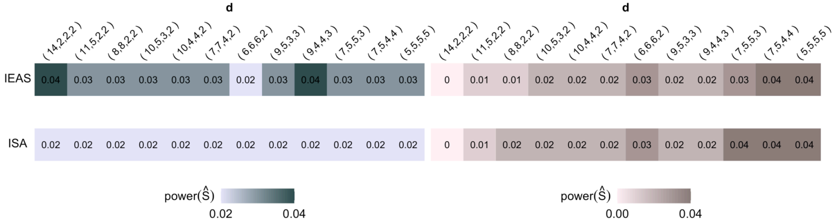
goodness of fit: test illustrations

null and non-null distributions of S_0 and A_0, and the χ^2_9-distribution when simple IEAS($\mathbf{d}_0$) hypotheses are tested against IEAS($\mathbf{d}$)



goodness of fit: test illustrations

probabilities of false rejection (top) and power (bottom) when composite IEAS and ISA hypotheses are tested against IEAS(**d**) models for multigraphs with $n = 4$, $m = 10$ and $\alpha(\chi^2_6) = 0.04$

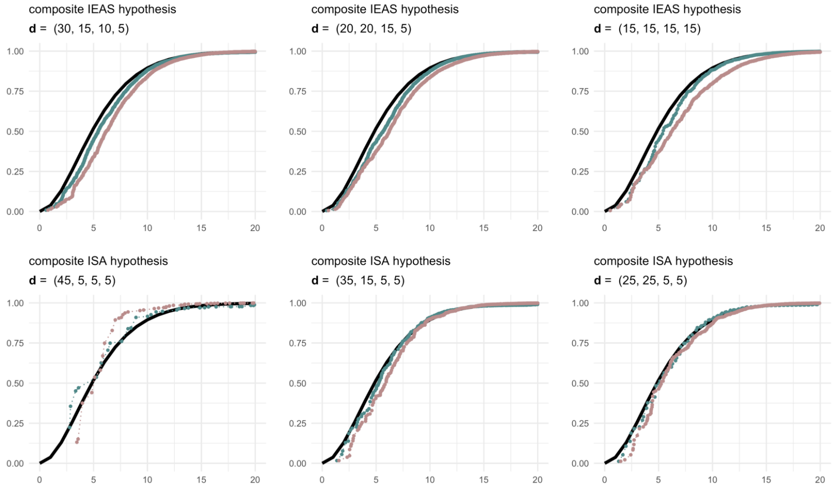


	d											
	(14,2,2,2)	(11,5,2,2)	(8,8,2,2)	(10,5,3,2)	(10,4,4,2)	(7,7,4,2)	(6,6,6,2)	(9,5,3,3)	(9,4,4,3)	(7,5,5,3)	(7,5,4,4)	(5,5,5,5)
IEAS	0.04	0.03	0.03	0.03	0.03	0.03	0.02	0.03	0.04	0.03	0.03	0.03
ISA	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02

	d											
	(14,2,2,2)	(11,5,2,2)	(8,8,2,2)	(10,5,3,2)	(10,4,4,2)	(7,7,4,2)	(6,6,6,2)	(9,5,3,3)	(9,4,4,3)	(7,5,5,3)	(7,5,4,4)	(5,5,5,5)
IEAS	0	0.01	0.01	0.02	0.02	0.02	0.03	0.02	0.02	0.03	0.04	0.04
ISA	0	0.01	0.02	0.02	0.02	0.02	0.03	0.02	0.02	0.04	0.04	0.04

goodness of fit: test illustrations

non-null distributions of $\hat{S}$, $\hat{A}$, and the χ^2_6-distribution when composite IEAS and ISA hypotheses tested against RSM($\mathbf{d}$) models


composite IEAS hypothesis

d = (30, 15, 10, 5)

composite IEAS hypothesis

d = (20, 20, 15, 5)

composite IEAS hypothesis

d = (15, 15, 15, 15)

composite ISA hypothesis

d = (45, 5, 5, 5)

composite ISA hypothesis

d = (35, 15, 5, 5)

composite ISA hypothesis

d = (25, 25, 5, 5)

summary of error probabilities

		simple IEAS($\mathbf{d}_0$) hypothesis			composite hypothesis	
model		$\mathbf{d}_0 = \mathbf{d}$	Flat $\mathbf{d}_0 \neq \mathbf{d}$	Skew $\mathbf{d}_0 \neq \mathbf{d}$	IEAS	ISA
IEAS	Flat $\mathbf{d}$	$\alpha_{S_0} > \alpha_{A_0}$	$\beta_{S_0} < \beta_{A_0}$	$\beta_{S_0} < \beta_{A_0}$	$\alpha_{\hat{S}} \leq \alpha_{\hat{A}}$	$\beta_{\hat{S}} > \beta_{\hat{A}}$
	Skew $\mathbf{d}$	$\alpha_{S_0} > \alpha_{A_0}$	$\beta_{S_0} < \beta_{A_0}$	$\beta_{S_0} < \beta_{A_0}$	$\alpha_{\hat{S}} > \alpha_{\hat{A}}$	$\beta_{\hat{S}} \geq \beta_{\hat{A}}$
		simple ISA($\mathbf{d}_0/2m$) hypothesis			composite hypothesis	
		$\mathbf{d}_0 = \mathbf{d}$	Flat $\mathbf{d}_0 \neq \mathbf{d}$	Skew $\mathbf{d}_0 \neq \mathbf{d}$	IEAS	ISA
ISA	Flat $\mathbf{d}$	$\alpha_{S_0} \geq \alpha_{A_0}$	$\beta_{S_0} \leq \beta_{A_0}$	$\beta_{S_0} < \beta_{A_0}$	inconclusive	$\alpha_{\hat{S}} \leq \alpha_{\hat{A}}$
	Skew $\mathbf{d}$	$\alpha_{S_0} > \alpha_{A_0}$	$\beta_{S_0} \leq \beta_{A_0}$	$\beta_{S_0} < \beta_{A_0}$	$\beta_{\hat{S}} < \beta_{\hat{A}}$	$\alpha_{\hat{S}} > \alpha_{\hat{A}}$
		simple IEAS($\mathbf{d}_0$) or ISA($\mathbf{d}_0/2m$) hypothesis			composite hypothesis	
		$\mathbf{d}_0 = \mathbf{d}$	Flat $\mathbf{d}_0 \neq \mathbf{d}$	Skew $\mathbf{d}_0 \neq \mathbf{d}$	IEAS	ISA
RSM	Flat $\mathbf{d}$	$\beta_{S_0} \geq \beta_{A_0}$	inconclusive	$\beta_{S_0} = \beta_{A_0}$	$\beta_{\hat{S}} > \beta_{\hat{A}}$	$\beta_{\hat{S}} > \beta_{\hat{A}}$
	Skew $\mathbf{d}$	$\beta_{S_0} \leq \beta_{A_0}$	$\beta_{S_0} = \beta_{A_0}$	$\beta_{S_0} < \beta_{A_0}$	$\beta_{\hat{S}} \geq \beta_{\hat{A}}$	$\beta_{\hat{S}} > \beta_{\hat{A}}$

conclusions

- even for very small m, the null distributions of the test statistics under IEA are fairly well approximated by their asymptotic distributions
- the convergence of the cdf's of test statistics are rapid and depend on the parameters in models
- approximations to the actual distributions can be obtained using adjustments of the χ^2-distributions yielding better power
- the influence of RSM on both test statistics is substantial for small m, implying a shift of their distributions towards smaller values compared to what holds true for the null distributions under IEA

the tests in R

https://github.com/termehs/multigraphr

```
# install.packages("devtools")
devtools::install_github("termehs/multigraphr")
```