Structural Identification and Estimation of Peer Effects Using Network Data

Presentation for Big Data Seminar

Fei (Michael) WANG

University of Konstanz, April 11th, 2019

Roadmap

1. Basics of Graph Theory and Network
2. Peer Effects and Reflection Issue
3. The Game and Model of Network
4. Dataset, Descriptive Statistics
5. Estimation and Simulation

Networks and Graphs

A network can be seen as any collection of objects (nodes or vertices) in which some pairs of these objects are connected by *links or edges*. There are two basic network parameters:

- *Number of nodes*, or N, represents the number of components in the system. We will often call N the size of network
- *Number of edges (links)*, or L, represents total number of interactions.
- To express asymmetric information, we also introduce the *directed* edges, for example, that A points to B but not vice versa.

Networks and Graphs

Figure 1 gives two examples:

- Undirected network: $N = 4, L = 4$
- Directed network: $N = 4, A \to B, B \to C$, etc.

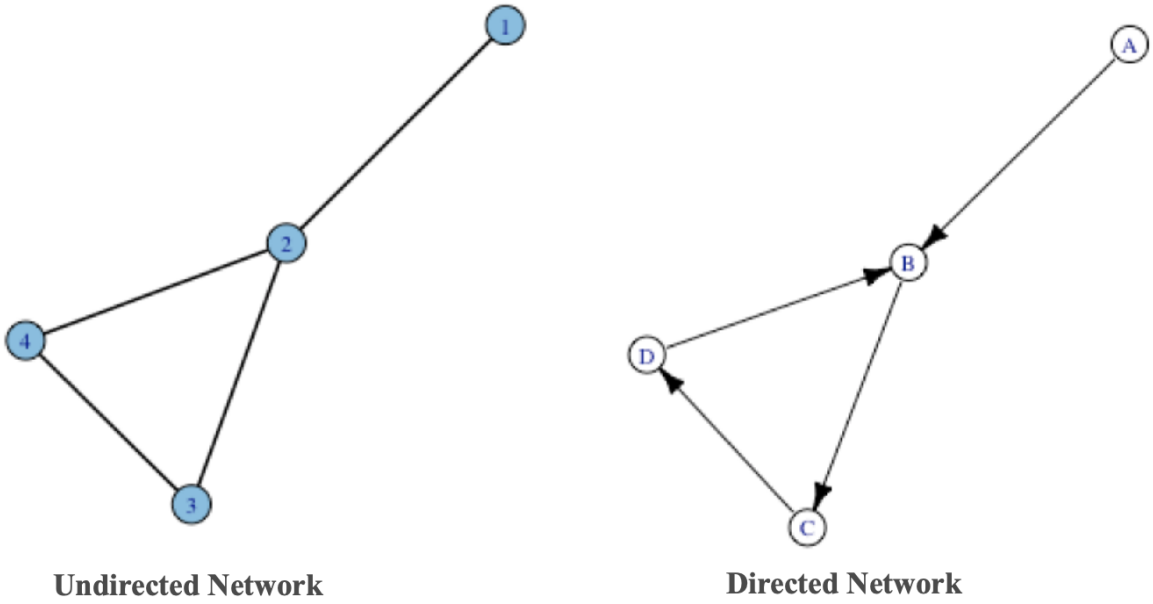
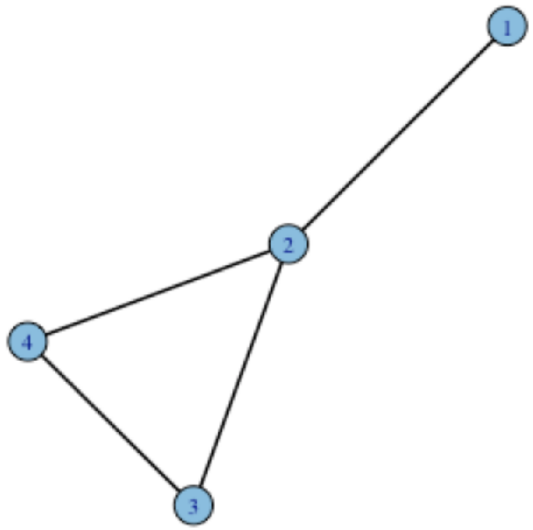


Figure 1: Network Examples

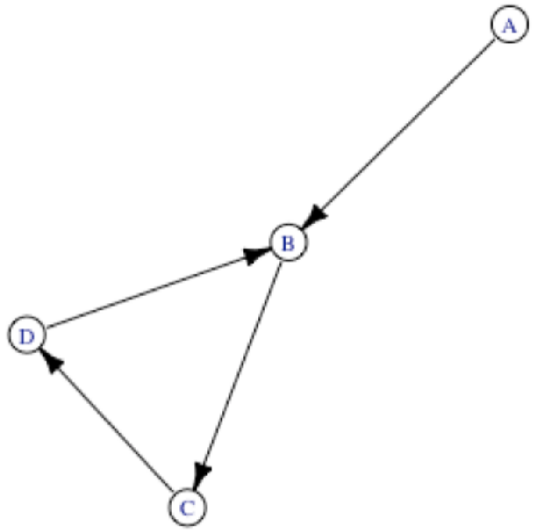
Networks and Graphs

We can use matrix to represent *undirected* or *directed* network:

$$G1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \end{array} \qquad G2 = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$



Undirected Network



Directed Network

Network: representation and measurement

We can use matrix to represent *undirected* or *directed* network:

- The set $N = \{1, \cdots, n\}$ is the set of *nodes* that are involved in a network of relationships.
- A *graph* (N, G) consists of a set of nodes $N = \{1, \cdots, n\}$ and a real-valued $n \times n$ matrix G, where g_{ij} represents the relation between i and j.
- This matrix is often referred to as the **adjacency matrix**. It is standard to use the values of either 0 or 1 to represent the unweighted network.
- In the case in which the entries of G take on more than two values and can track the intensity level of relationships, the graph is referred to as a *weighted* graph.

Network: examples

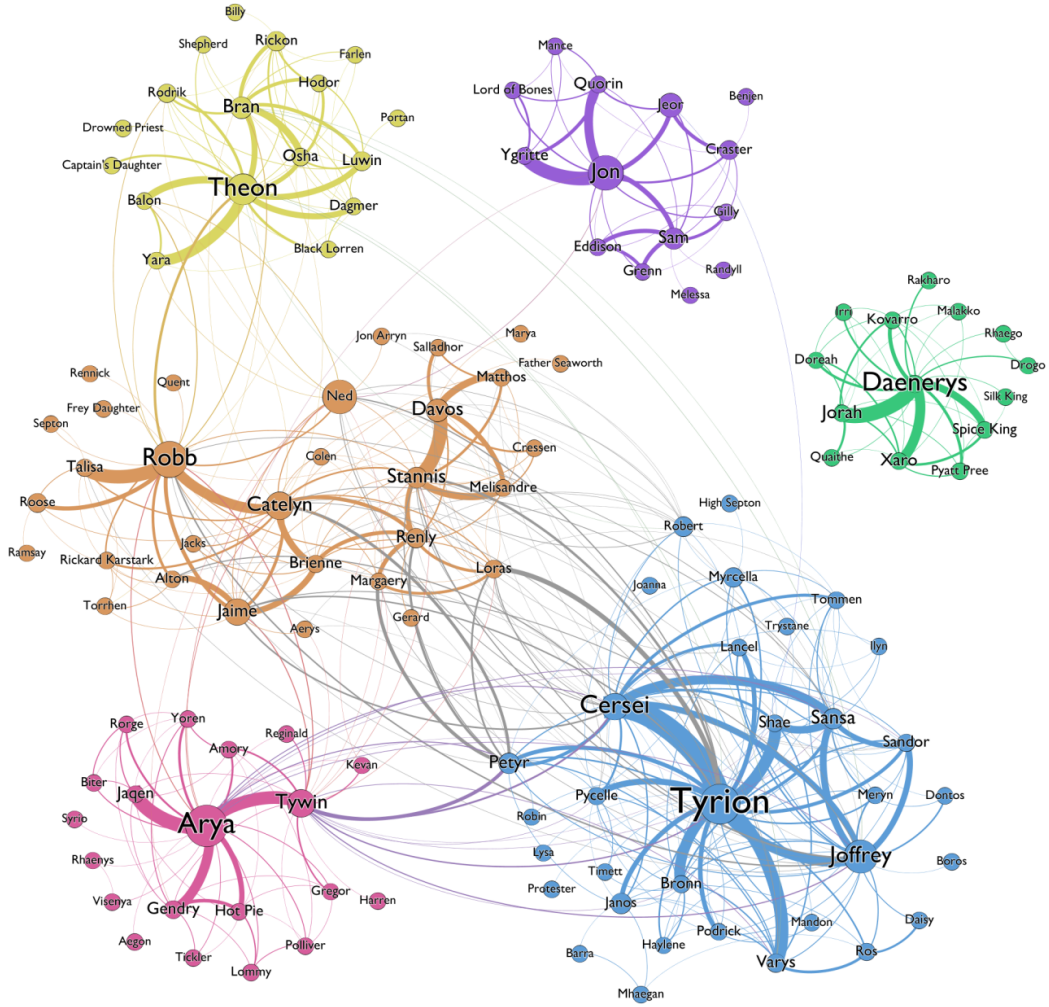


Figure 2: Network of Thrones: Season 2(Andrew, 2018)

Network: examples

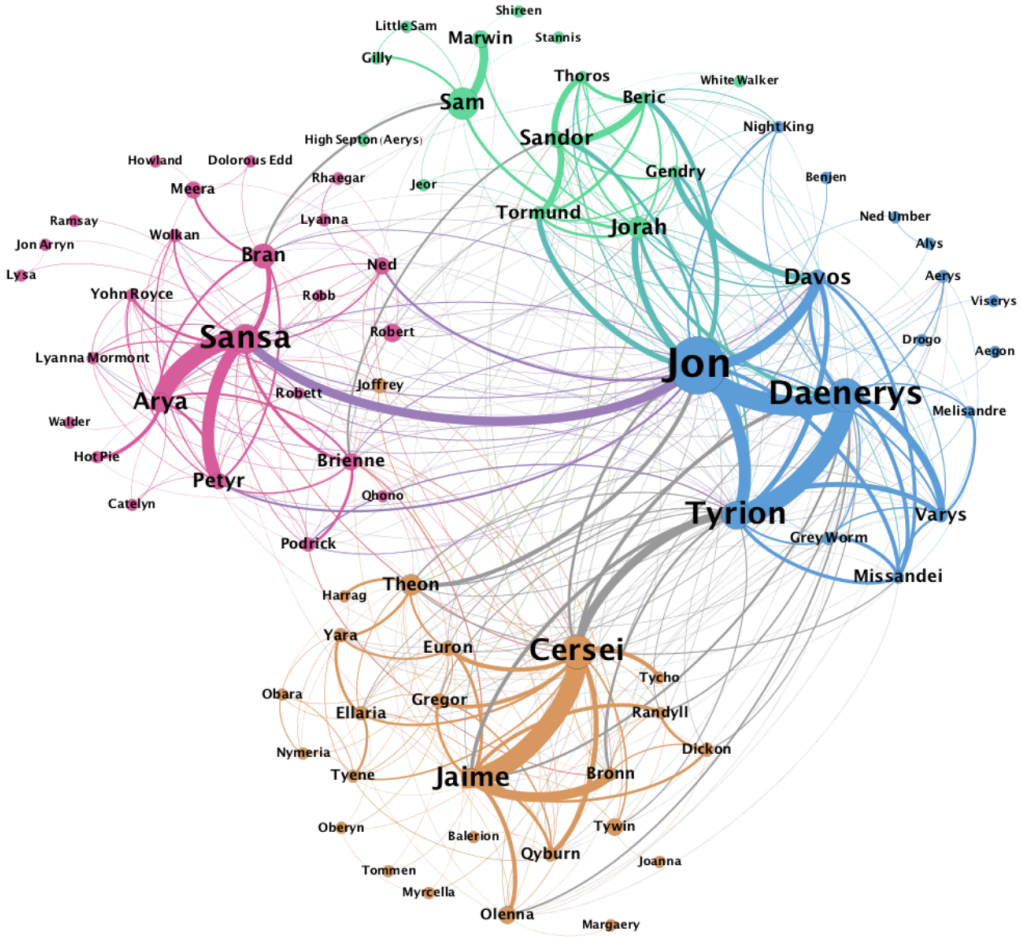
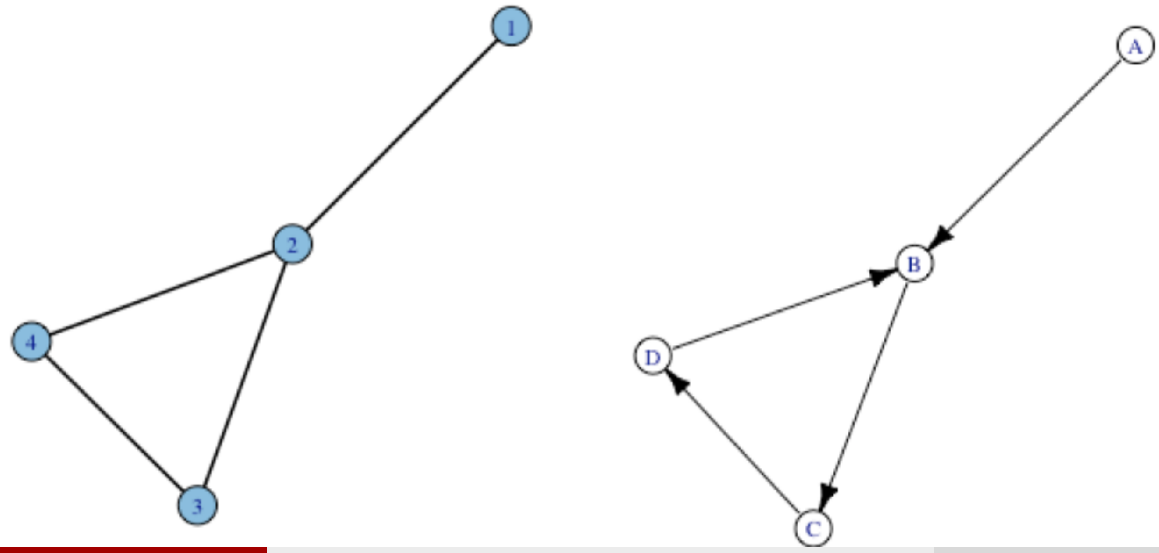


Figure 3: Network of Thrones: Season 7(Andrew, 2018)

Network: representation and measurement

To study networks properly, we need certain measurements linked to properties of networks:

- *degree* of node represents the number of links it has to other nodes. In the undirected network in Figure 1: $k_1 = 1, k_2 = 3, k_3 = 2, k_4 = 2$. In directed networks, we distinguish between *incoming* degree and *outgoing* degree.
- A *geodesic* between nodes i and j is a shortest path between these nodes; that is, a path with no more links than any other path between these nodes.



Network: representation and measurement

Degree distribution examples:

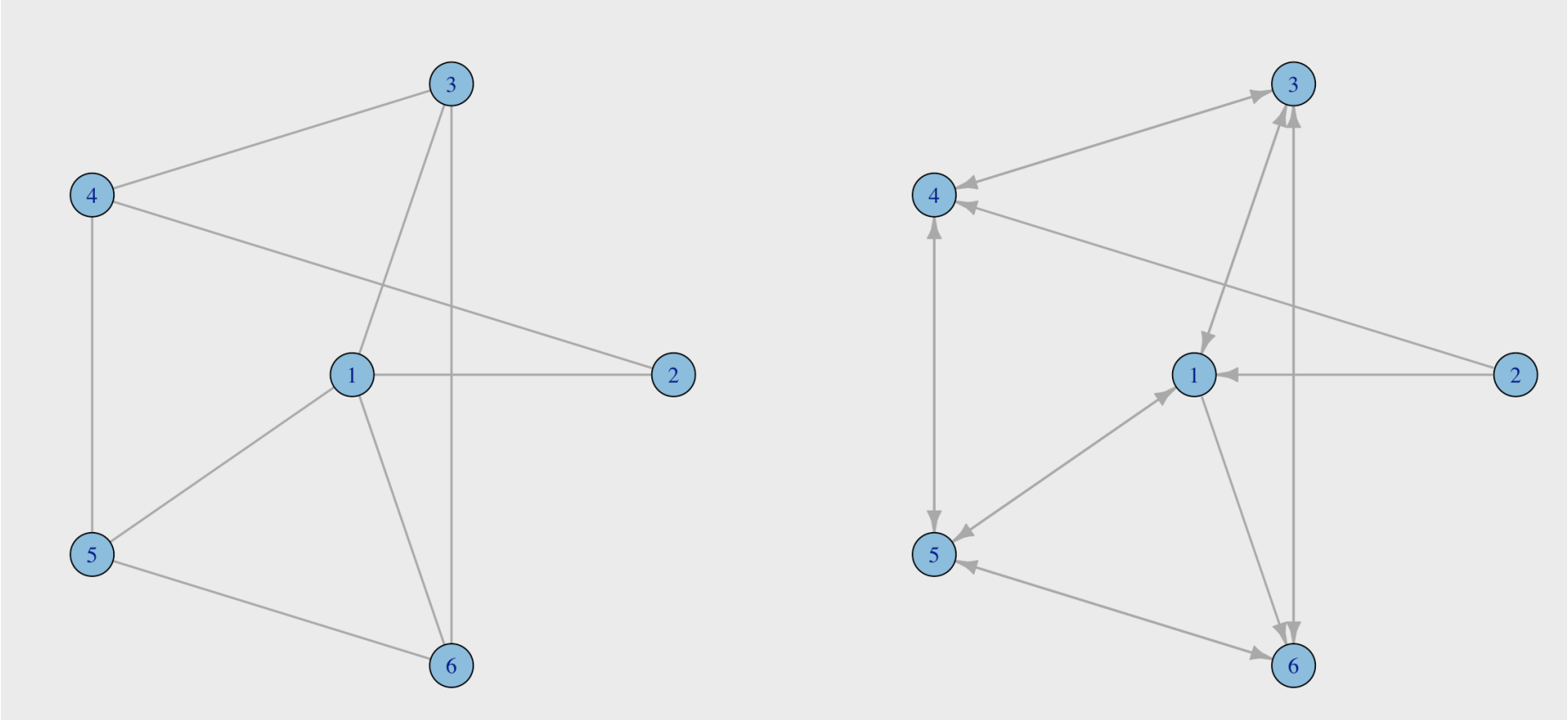
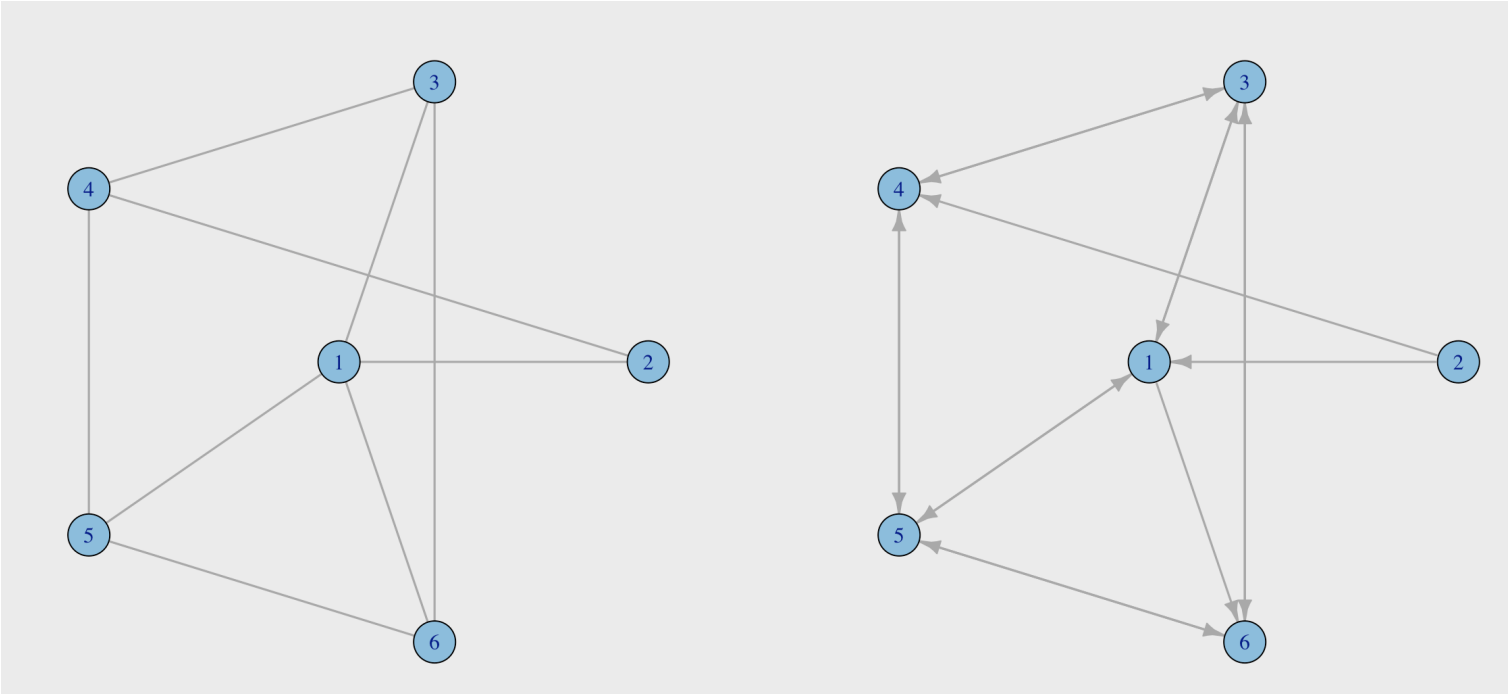


Figure 4: G1: Undirected network; G2: Directed network

Network: representation and measurement

Degree distribution examples:

G1	all	4	2	3	3	3	3
G2	in	3	0	3	3	3	3
	out	3	2	3	2	3	2



Degree distribution example:

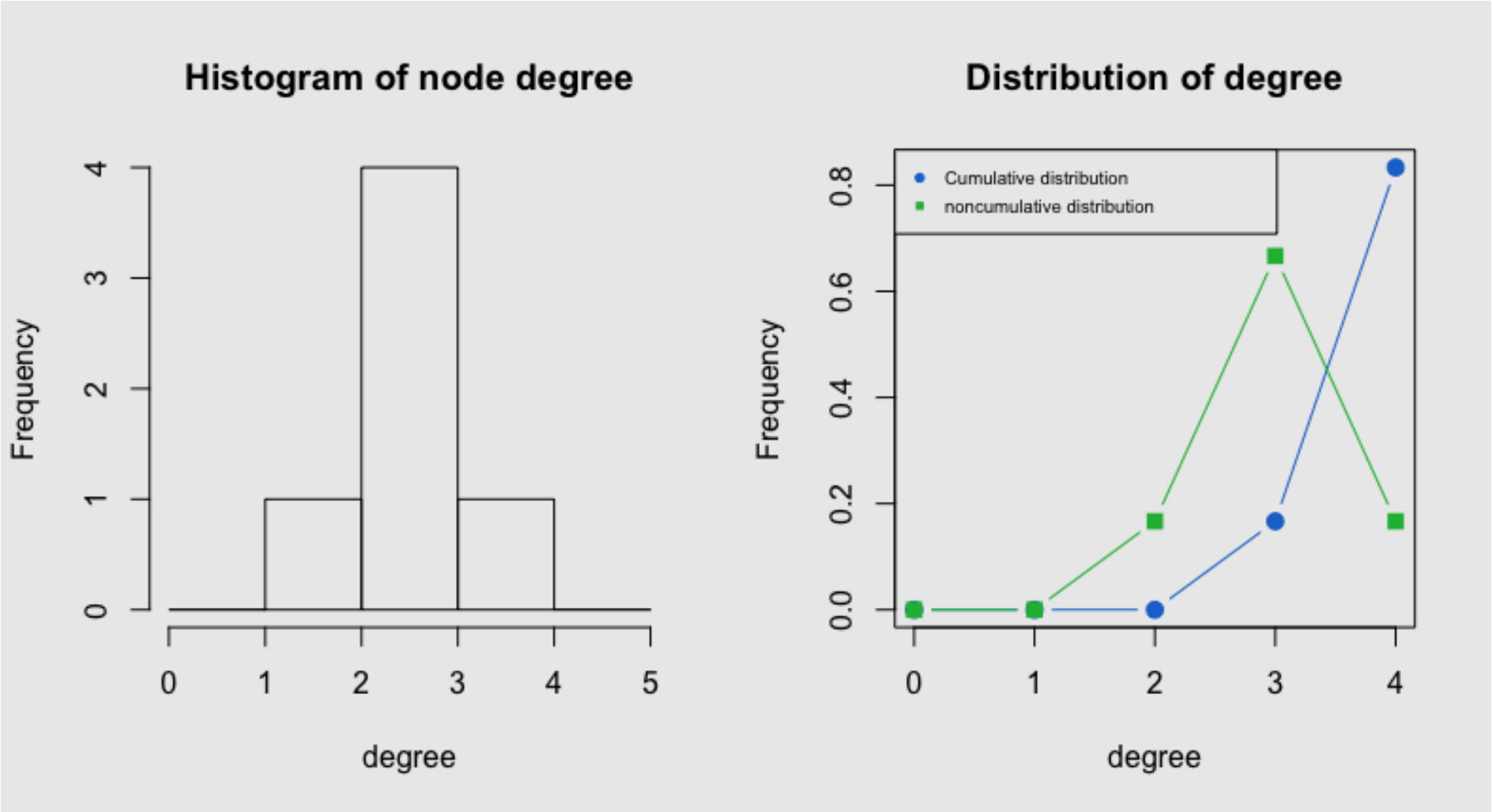
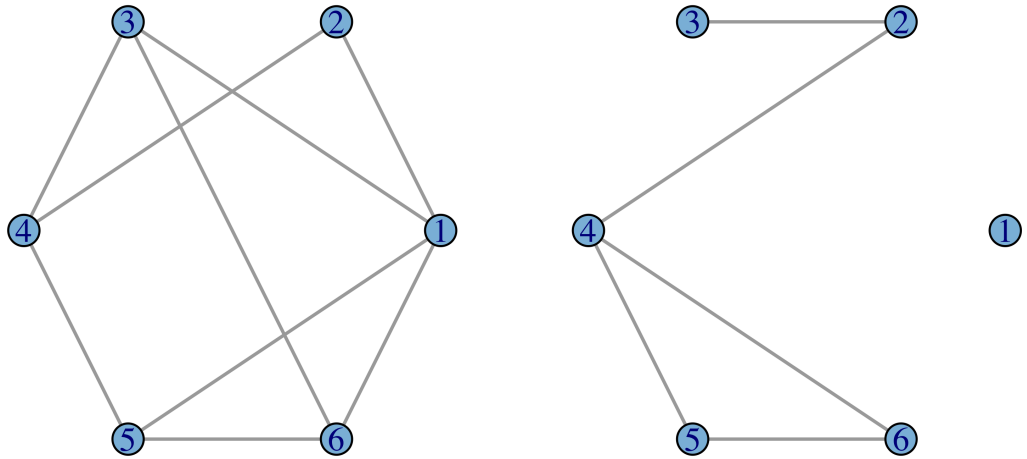


Figure 5: Degree distribution of G1

Network: representation and measurement

Geodesic examples:

	Node1	Node2	Node3	Node4	Node5	Node 5
G1node2	1	0	2	1	2	2
G2node2	Na	0	1	1	2	2
G1	Average degree		3	Average geodesic		1.4
G2	Average degree		1.67	Average geodesic		1.7



Random Networks Vs Scale-Free Networks

Random Networks:

- Consider a set of nodes $N = \{1, \cdots, n\}$, and let a link between any two nodes, i and j, be formed with probability p, where $0 < p < 1$. The **formation of links is independent**. For n nodes, the all possible combination (or links) are:

$$\binom{n}{2} = \frac{n!}{2!(n-2)!} = \frac{n(n-1)}{2} \tag{1}$$

- Then, any given network that has m links on n nodes has a probability of

$$p^m(1-p)^{\frac{n(n-1)}{2}-m} \tag{2}$$

Random Networks Vs Scale-Free Networks

Random Networks:

- Define the average degree of a random network as λ

$$\lambda = E[k] = p(N-1) \tag{3}$$

- When $N \to \infty$, we can derive the *degree distribution of a random network* as the Poisson distribution

$$p_k = e^{-\lambda}\frac{\lambda^k}{k!} \tag{4}$$

Random Networks Vs Scale-Free Networks

Scale-Free Networks or Albert and Barabási (2002) Model:

- Start with m_0 nodes, the links between which are chosen arbitrarily, as long as each node has at least one link. The network develops following two steps:
  1. Growth: at each timestep we add a new node with m ($\leq m_0$) links that connect the new node to m nodes already in the network.
  2. The probability $\Pi(k)$ that a link of the new node connects to node i depends on the degree k_i as

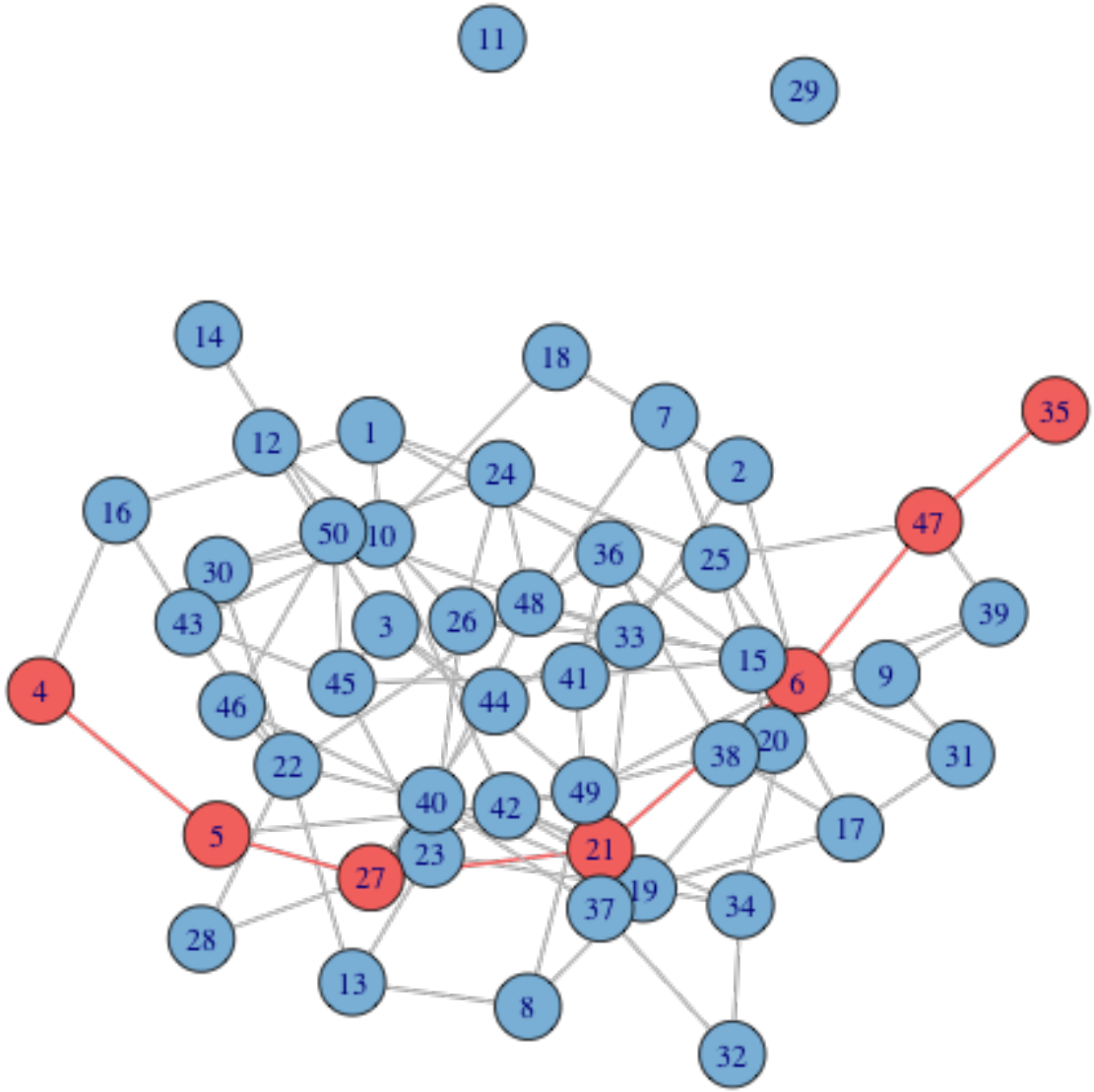
$$\Pi(k_i) = \frac{k_i}{\sum_j k_j} \tag{5}$$

- *Preferential attachment* is a probabilistic mechanism: Equation (5) implies that if a new nodes has a choice between a degree-two and a degree-four node, it is twice as likely that it connects to the degree-four node, which can provide more resources. Degree distribution follows power law:

$$p_k \sim k^{-\gamma} \tag{6}$$

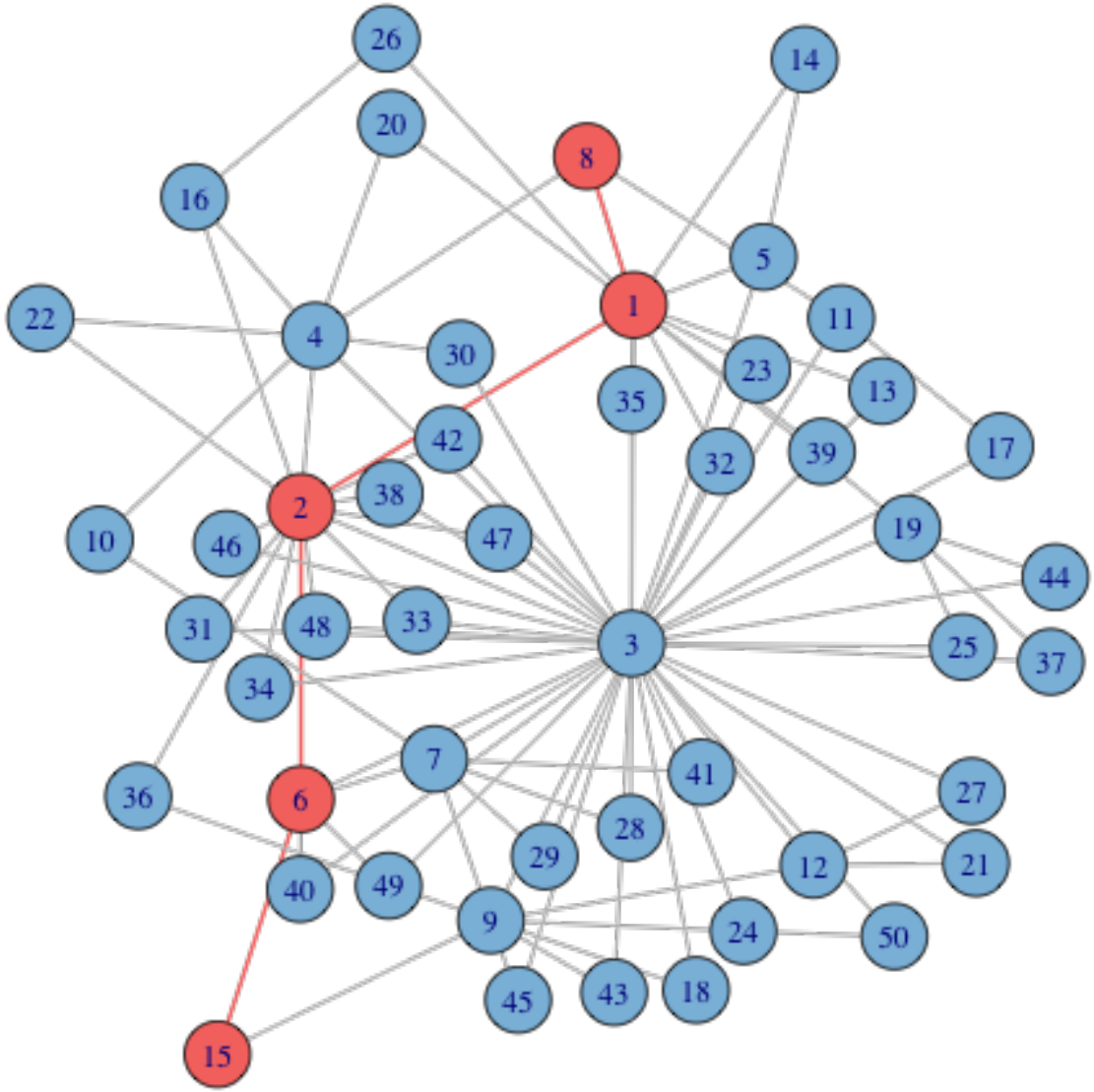
Random Networks: example

- $N = 50$
- $p = 0.079$
- $E[k] = 3.88$
- $E(g) = 2.83$
- $d = 6$

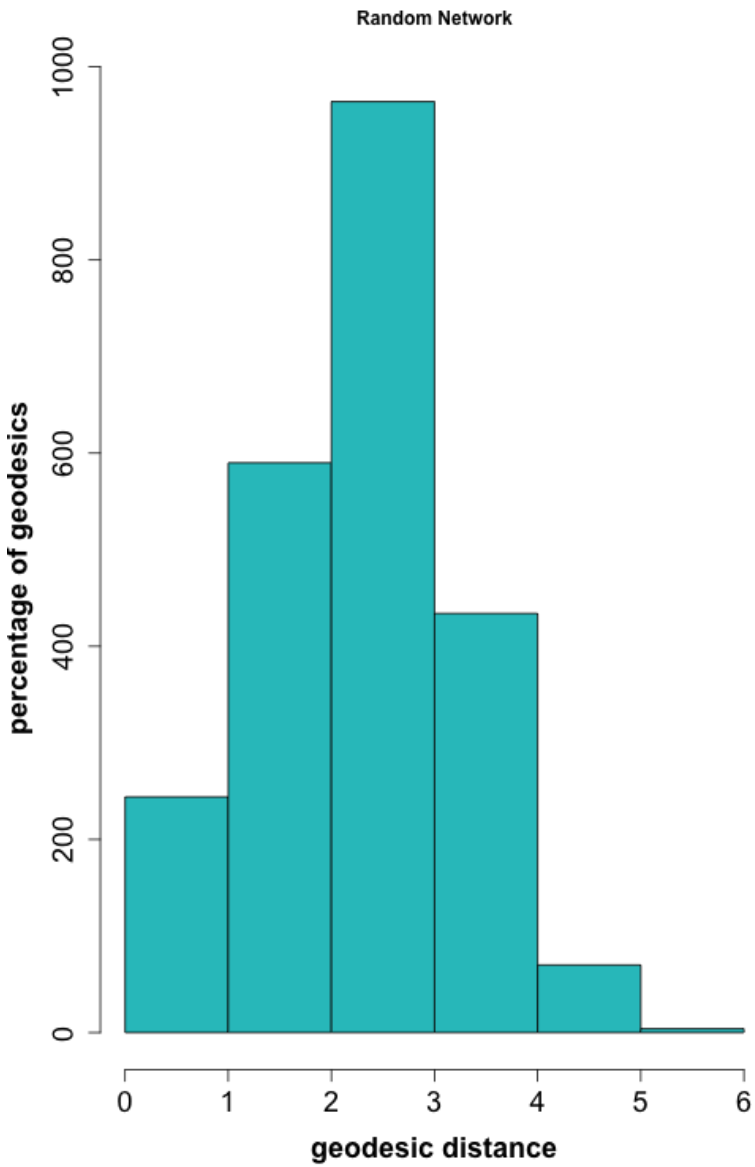


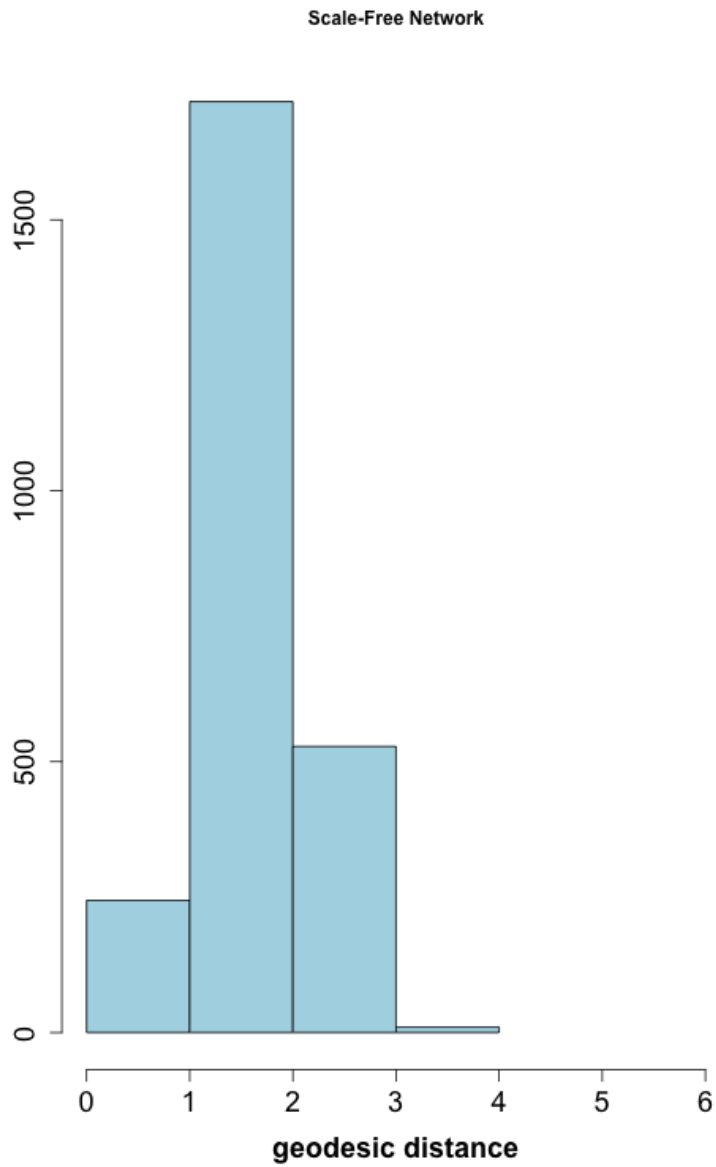
Scale-Free Networks: example

- $N = 50$
- $\gamma = 2.1$
- $E[k] = 3.88$
- $E(g) = 2.14$
- $d = 4$



Random Networks Vs Scale-Free Networks





Everywhere: power law

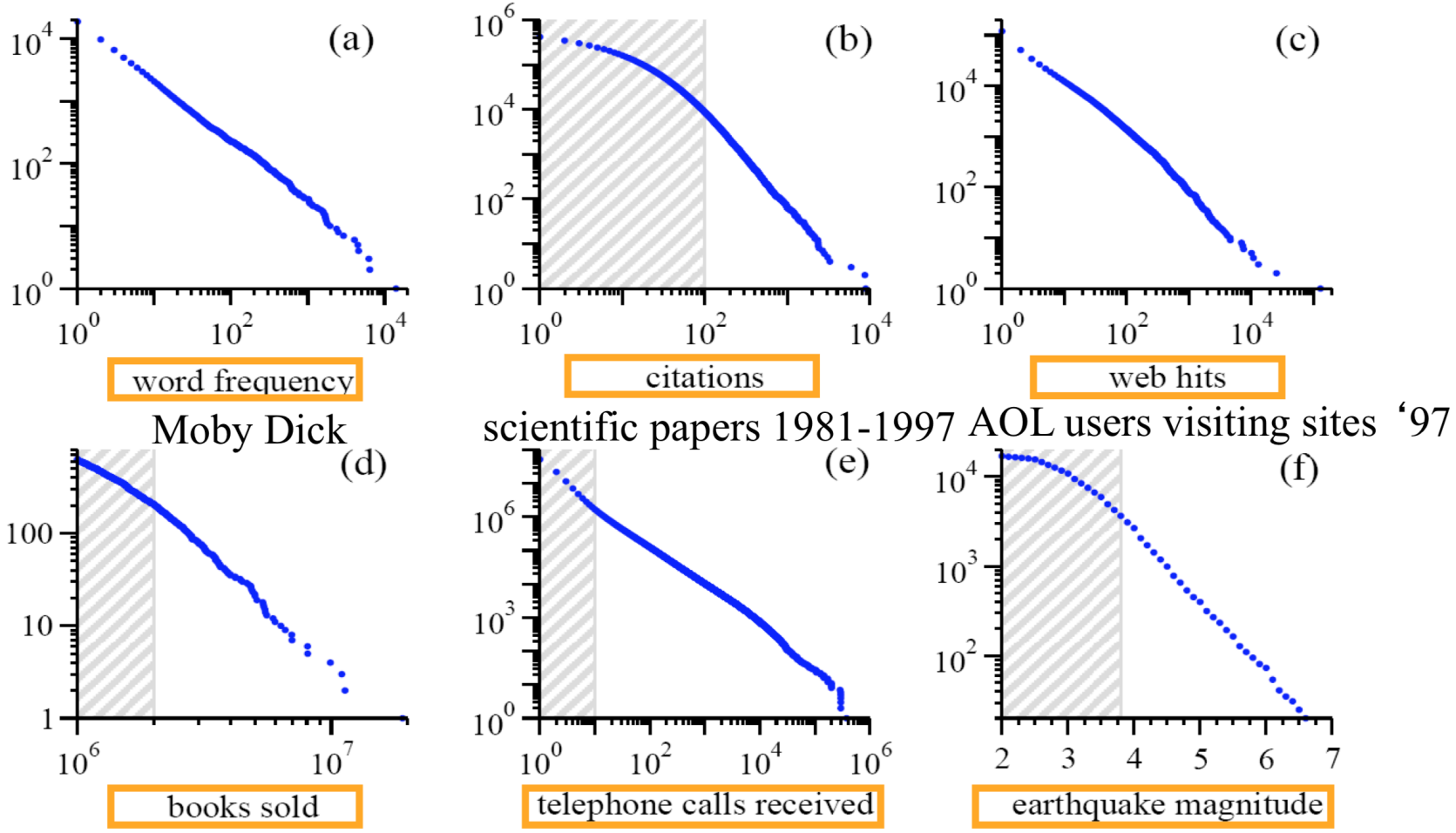
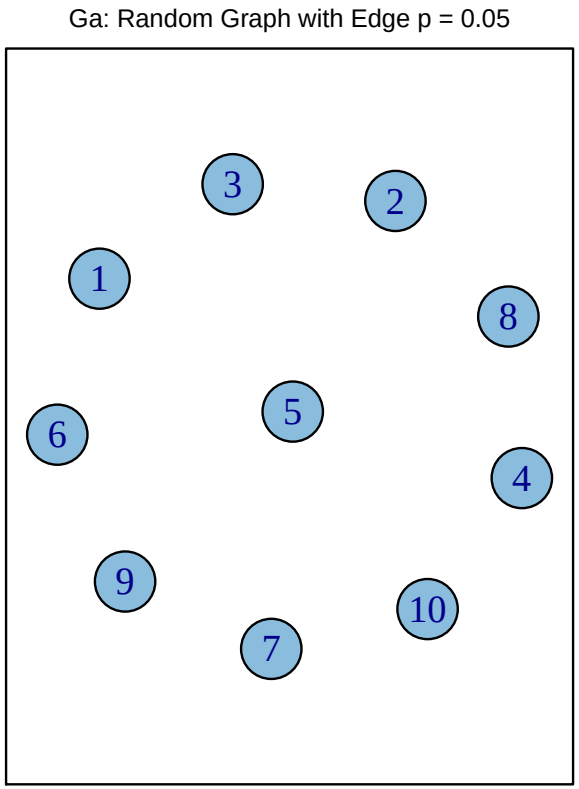
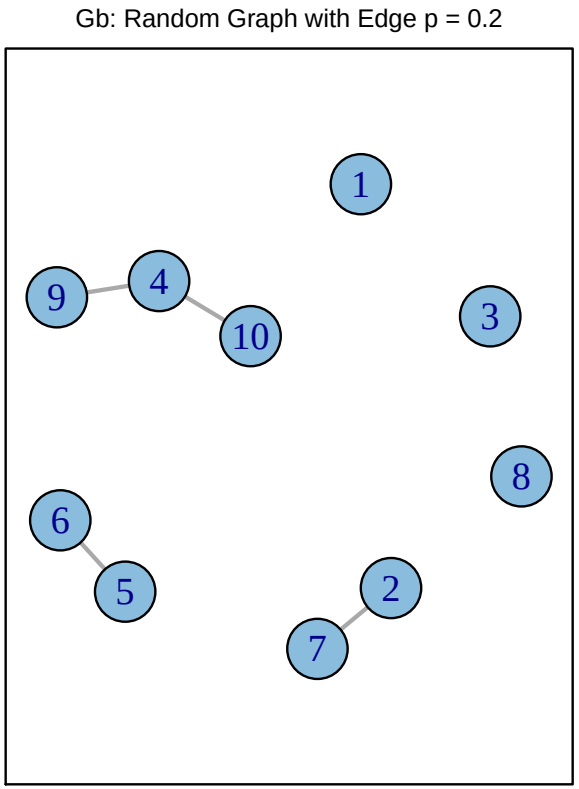


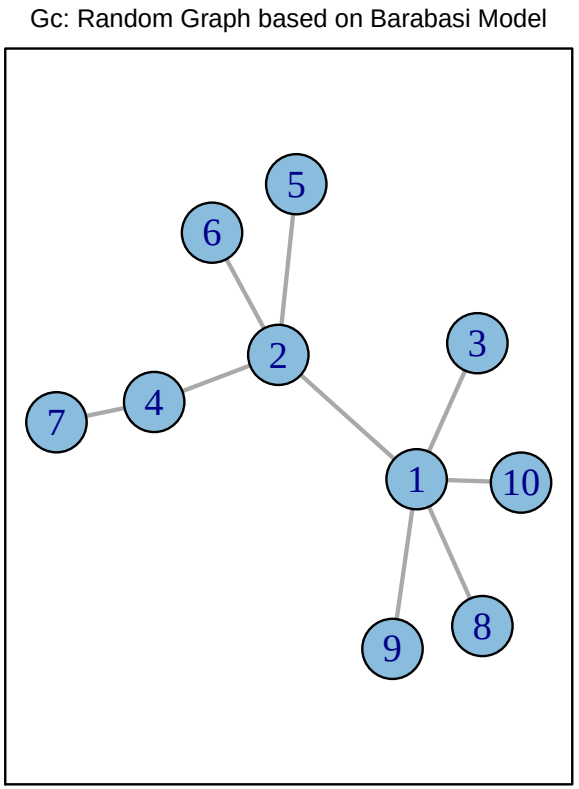
Figure 6: Scale-Free Networks (Lada, 2013)

Peer Effects

What do we mean when we talk about 'peer effects'? *Peer* is a person who is the same age or has the same position or the same abilities as other people in a group (Cambridge Dictionary).



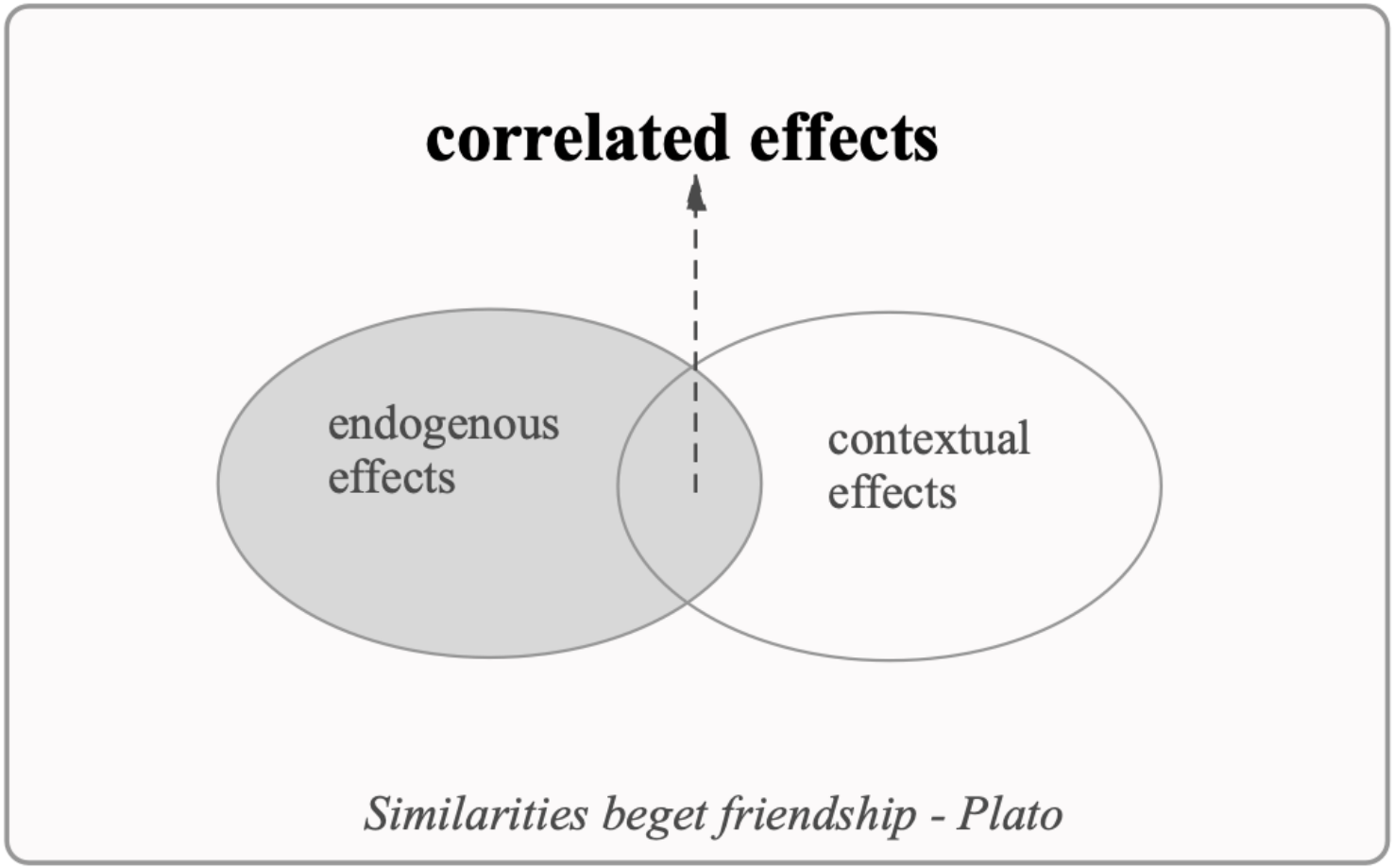




- Ga Vs Gb; or Ga Vs Gc ; or Gb Vs Gc

Peer Effects: Reflection Problem

According to Manski (1993), the reflection problem of peer effects can be explained by the following Venn diagram.



Peer Effects: Reflection Problem

Let's model all three effects into one equation:

$$y = \alpha + \underbrace{\beta E(y|g)}_{\text{Peer or Network Effects}} + \overbrace{E(z|g)'\gamma}^{\text{Contextutal Effects}} + z'\eta + u, \tag{7}$$

$$E(u|g,z) = g'\delta \quad \text{[Correlated Effects]} \tag{8}$$

$$\tag{9}$$

- y: scalar outcome (highschool achievement)
- g: attributes of the individual's reference **group**
- (z, u): attributes that directly affect outcome y, for example, z can be parents' education, u is unobservable factor (e.g. IQ).

Peer Effects: Reflection Problem

Taking expectation on both sides, we have (imaged setup):

$$E(y|g,z) = \alpha + \beta E(y|g) + E(z|g)'\gamma + g'\delta + z'\eta \tag{10}$$

In real world, we have the following social equilibrium:

$$E(y|g) = \alpha + \beta E(y|g) + E(z|g)'\gamma + g'\delta + E(z|g)'\eta \tag{11}$$

- Equation (10) means human being are *social animals*
- Now, do algebra and derive $E(y|g)$, we have:

$$E(y|g) = \frac{\alpha}{1-\beta} + E(z|g)'\frac{\gamma+\eta}{1-\beta} + \frac{g'\delta}{1-\beta} \tag{12}$$

Peer Effects: Reflection Problem

Substitute $E(y|g)$ from equation (12) into equation (10) to obtain:

$$E(y|g,z) = \frac{\alpha}{1-\beta} + E(z|g)'\frac{\gamma+\beta\eta}{1-\beta} + \frac{g'\delta}{1-\beta} + z'\eta \tag{13}$$

- Those above composed parameters are identified if $[1, E(z|g), z]$ are linearly independent in the population based on equality (12).
- It is impossible to separately identify the endogenous effect β.

So, what's the solution?

- Assume $E(u|g,z) = 0$, which means there is no correlated effects. (very artful or very sneaky?)

The Game

Consider the network with a finite agents $N = \{1, \cdots, n\}$. The network can be represented by the adjacency matrix $G = [g_{ij}]$, where $g_{ij} = 1$ if i and j have mutual friendship, and $g_{ij} = 0$, otherwise. Each player $i = 1, \cdots n$ selects an effort $x_i \geq 0$ and obtain the payoff

$$u_i(x_1, \cdots, x_n) = \alpha_i x_i - \frac{1}{2} x_i^2 + \lambda \sum_{j=1}^{n} g_{ij} x_i x_j \tag{14}$$

, which is strictly concave in her own effort. Now, decompose the effort x_i into two parts:

- y_i the effort of individual i absent of any peer influence;
- z_i the peer effort that corresponds to peer efforts from friends

Then, equation (14) becomes:

$$u_i(y, z; g) = \alpha_i(y_i + z_i) - \frac{1}{2}(y_i + z_i)^2 + \lambda \sum_{j=1}^{n} g_{ij} z_i z_j \tag{15}$$

$$= \alpha_i y_i - \frac{1}{2} y_i^2 + \alpha_i z_i - \frac{1}{2} z_i^2 - y_i z_i + \lambda \sum_{j=1}^{n} g_{ij} z_i z_j \tag{16}$$

Solve the above maximization problem, gives:

$$\frac{\partial u_i}{\partial y_i} = \alpha_i - y_i - z_i; \qquad \Rightarrow \quad y_i^* = \alpha_i - z_i^*$$

$$\frac{\partial u_i}{z_i} = \alpha_i - z_i - y_i + \lambda \sum_{j=1}^{n} g_{ij} z_j; \qquad \Rightarrow \quad z_i^* = \alpha_i - y_i^* + \lambda \sum_{j=1}^{n} g_{ij} z_j^*$$

$$\frac{\partial u_i}{\partial y_i \partial z_i} = -1;$$

The Game

In equlibirum, we have:

$$y_i^* = \alpha_i - z_i^*$$

$$z_i^* = \alpha_i - y_i^* + \lambda \sum_{j=1}^{n} g_{ij} z_j^*$$

- The game is symmetric in the way that every player faces the same group of FOCs;
- The game is not symmetric in the way that each player's BR depends on her position in the network.

Ballester et al. (2006) proved that the Nash Equilibrium exists and it is rooted in the vector of *Katz-Bonacich centralities* (Katz, 1953; Bonacich, 1987):

$$b(g, \phi) = (I - \phi G)^{-1} u \tag{17}$$

A network model of Peer Effects

A network model of peer effects with *Katz-Bonacich centralities* can be:

$$y = (I - \phi G)^{-1}(X\beta + \alpha u + \varepsilon) \quad \Rightarrow \quad (I - \phi G)y = X\beta + \alpha u + \varepsilon$$

- Assume $E(\varepsilon|g, z) = 0$, or $\varepsilon|A, X \sim N(0, \sigma^2)$ in matrix form, estimate:

$$y = \phi Gy + X\beta + \alpha u + \varepsilon \tag{18}$$

- Assume $E(\varepsilon|g, z) \neq 0$, estimate:

$$y = \phi Gy + X\beta + \alpha u + \varepsilon; \quad \varepsilon = \rho G\varepsilon + \nu \tag{19}$$

Dataset

Panel Study of Cologne Gymnasium Students (KGP) - `data_network2.csv` cleaned by Dr Livia Shkoza: 3391 observations and 81 variables.

Table B.1: Dataset Descriptions

Variables	Types	Descriptions
`ques_1_1`	Indicator	Include the network relationships,
$\cdots$		wherein each students list their
`ques_1_32`		friends based on ID numbers, such as $1, 2, 16, \cdots$
`id`	ID	Unique ID for each observation composed by interview wave, school, class and student IDs, e.g. 104102
`interview`	Indicator	Interview phase stage, all 1
`school`	ID	School ID, e.g. 13
`class`	Categorical	1 - class chosen based on school size 2 - class chosen based on school branch
`studentid`	ID	ID number represents each student, which is also used to indicate friendship
`classind`	ID	class ID number assigned within dataset
`classsize`	Discrete	Gives the size of each class
`idt`	ID	Tidy version of variable `id`
`st_V8`	Categorical	School branches, e.g. social science; 7 branches in total
`sch_score_68_3`	Discrete	Math grade in previous year (1968), 1 is the best
`sch_score_68_8`	Discrete	German grade in previous year (1968)
`sch_score_68_13`	Discrete	English grade in previous year (1968)

Dataset: network summary

	Directed Network		Undirected Network
	In	Out	Mutual
Average degree	5.826	5.826	2.754
Average density	0.0017		0.00077
Average geodesics			3.046

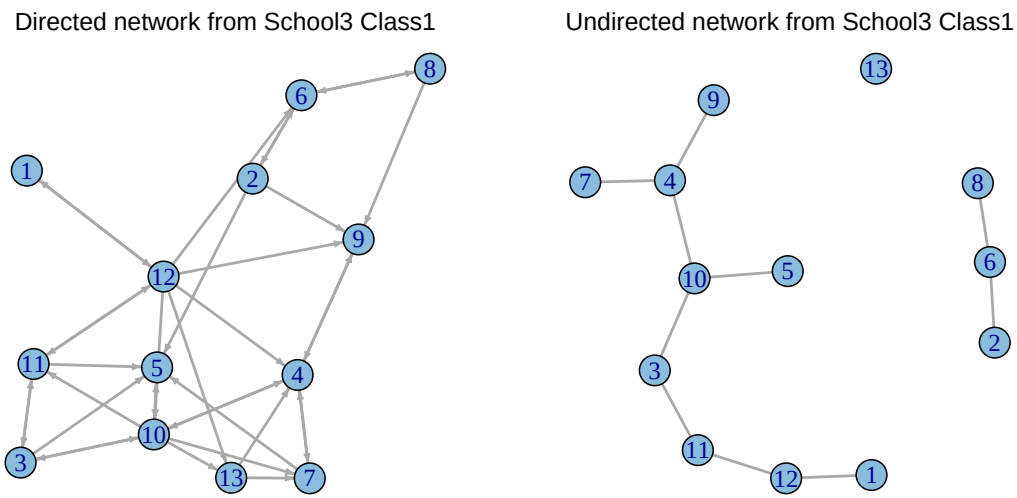


Figure 7: Network From School3 Class1

Dataset: Network fitted with exponential random graph model

The exponential random graph model (ERGM) has the distribution over networks

$$p(Y = y|\theta) = \frac{1}{Z} e^{\theta^T} \phi(y) \tag{20}$$

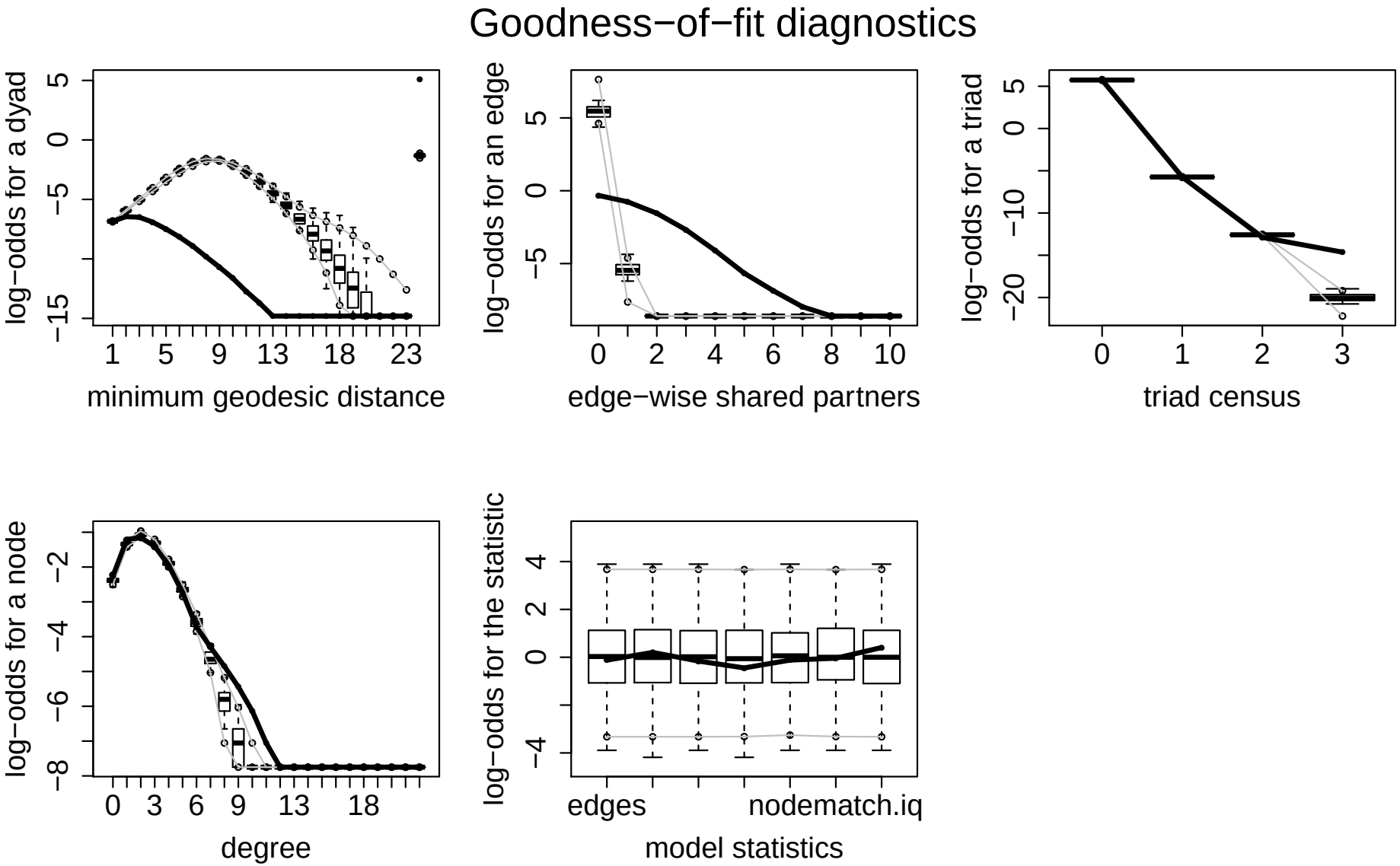
, where y is observed network adjacency matrix, $\phi(y)$ features the properties of the network and θ are parameters to be estimated, Z is normalized as the constant $\sum_y e^{\theta^t} \phi(y)$. Each component of the θ vector may be interpreted as the increase in the conditional log-odds of the network, per unit increase in the corresponding component of $\phi(y)$. Edges in the network are considered **conditionally independent** if they don't share a node.

Network: ERGM fit

Table 4.2: ERGM fit for the KGP network with covariates

	Undirected Network			Directed Network		
Terms	Estimate	SE	p value	Estimate	SE	p value
edges	-9.02945^{***}	0.07457	<1e-04	-7.78916^{***}	0.02927	<1e-04
match.age	0.23427^{***}	0.03814	<1e-04	0.17132^{***}	0.01885	<1e-04
match.gender	2.64159^{***}	0.07508	<1e-04	2.12883^{***}	0.02972	<1e-04
match.prestige	0.36691^{***}	0.06843	<1e-04	0.25575^{***}	0.03520	<1e-04
match.gpa_68	0.90769^{***}	0.08779	<1e-04	0.86141^{***}	0.04421	<1e-04
match.IQ	0.03712	0.10553	0.725	0.10175^{*}	0.05030	0.0431
match.gpa_70	1.06089^{***}	0.08692	<1e-04	0.99540^{***}	0.04411	<1e-04
BIC	42455			161263		

Network: ERGM fit - undirected network



Big picture on estimation

- Assume $E(\varepsilon|g,z)=0$, or $\varepsilon|A,X\sim N(0,\sigma^2)$ in matrix form, estimate:

$$y=\phi Gy+X\beta+\alpha u+\varepsilon$$

- Assume $E(\varepsilon|g,z)=0$ does not mean we can ignore them: using IVs A^2X
- **Problem**: weak identification issue even with IVs
- Assume $E(\varepsilon|g,z)\neq 0$, estimate:

$$y=\phi Gy+X\beta+\alpha u+\varepsilon;\quad \varepsilon=\rho G\varepsilon+\nu$$

- **Problem**: computation cost is extreme high (curse of dimensionality)

Big picture on estimation

Table 5.1: Summary of Estimated Models

Model	**Methods**	**Covariates**
Model 1a: $y = \phi Gy + X\beta + \alpha u + \varepsilon$	MLE without considering endogenous issue	previous gpa /previous score; gender dummy; age; IQ; prestige; father and mother education dummy (university)
Model 1b: $y = \phi Gy + X\beta + \alpha u + \rho\xi + \varepsilon$	MLE with considering endogenous issue	
Model 2: $y = \phi Gy + X\beta + GX\gamma + \alpha u + \varepsilon$	MLE with considering endogenous issue 2SLS with IVs 2SLS with IVs from estimated G	
Model 3a: $y = \phi Gy + X\beta + \alpha u + e;\ e = \rho Ge + \nu$ Model 3b: $y = \phi Gy + X\beta + GX\gamma + \alpha u + e; e = \rho Ge + \nu$	MLE	

Estimation Results

Table 5.2: Estimated Results with Average GPA as Dependent Variable

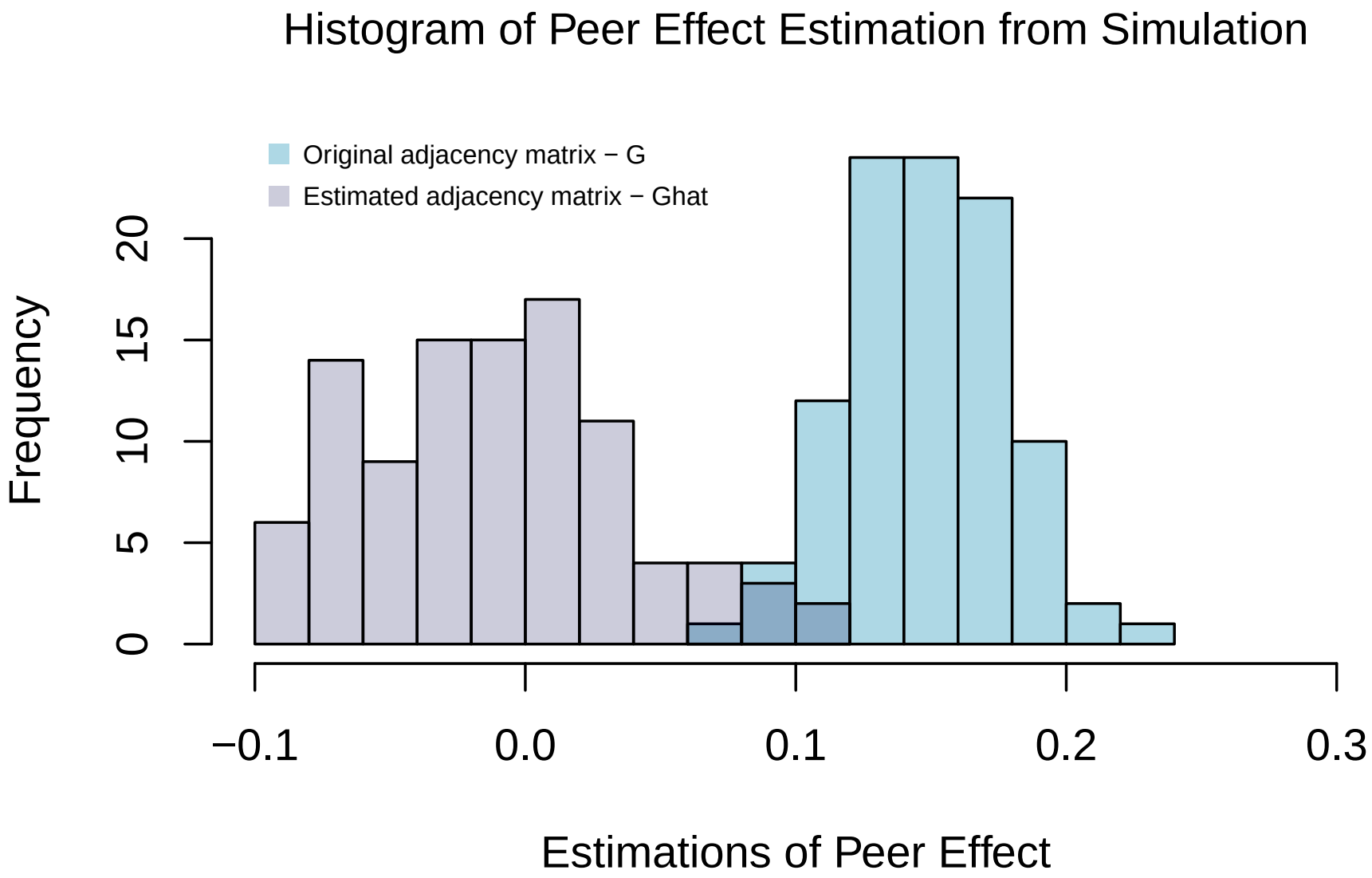
	Model						
	(1a) MLE	(1b) MLE	(2) MLE	(2) 2SLS(G)	(2)2SLS ($\hat{G}$)	(3a)	(3b)
ϕ	-0.0005	-0.0005	0.0368	0.1106***	-0.0151	0.0006	0.0034
	(0.0011)	(0.0011)	(0.0520)	(0.0286)	(0.0297)	(0.0025)	(0.0231)
α	1.0999***	1.0992***	1.1135***	1.065***	1.0564***	1.0076***	0.8903*
	(0.1287)	(0.1287)	(0.1344)	(0.2921)	(0.2979)	(0.2460)	(0.3591)
Previous GPA	0.6710***	0.6709***	0.6577***	0.6652***	0.6832***	0.6676***	0.6984***
	(0.0133)	(0.0133)	(0.0141)	(0.0147)	(0.0144)	(0.0256)	(0.0370)
Female	0.0102	0.0103	0.0153	0.0428	0.0581*	0.0255	0.1319*
	(0.0130)	(0.0130)	(0.0195)	(0.0273)	(0.0304)	(0.0283)	(0.0528)
age	0.0047	0.0047	0.0060	0.0043	0.0008	0.0156	0.0105
	(0.0073)	(0.0073)	(0.0077)	(0.0080)	(0.0080)	(0.0141)	(0.0204)
IQ	-0.0028***	-0.0028***	-0.0027***	-0.0022***	-0.0023***	-0.0053***	-0.0041*
	(0.0007)	(0.0007)	(0.0007)	(0.0008)	(0.0008)	(0.0009)	(0.0020)
prestige	-0.0010*	-0.0010*	−0.0009*	−0.0005	−0.0006	−0.0008	−0.0006
	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0008)	(0.0013)
Mother edu	-0.0003	-0.0003	-0.0005	0.0005	0.0031	0.0279	0.0026
	(0.0144)	(0.0144)	(0.0147)	(0.0151)	(0.0151)	(0.0271)	(0.040)
Father edu	0.0066	0.0067	0.0074	0.0211	0.0203	−0.0150	−0.006
	(0.0153)	(0.0153)	(0.0155)	(0.0160)	(0.0162)	(0.0291)	(0.045)
Unobservable		0.0012	0.0021				

Estimation Results

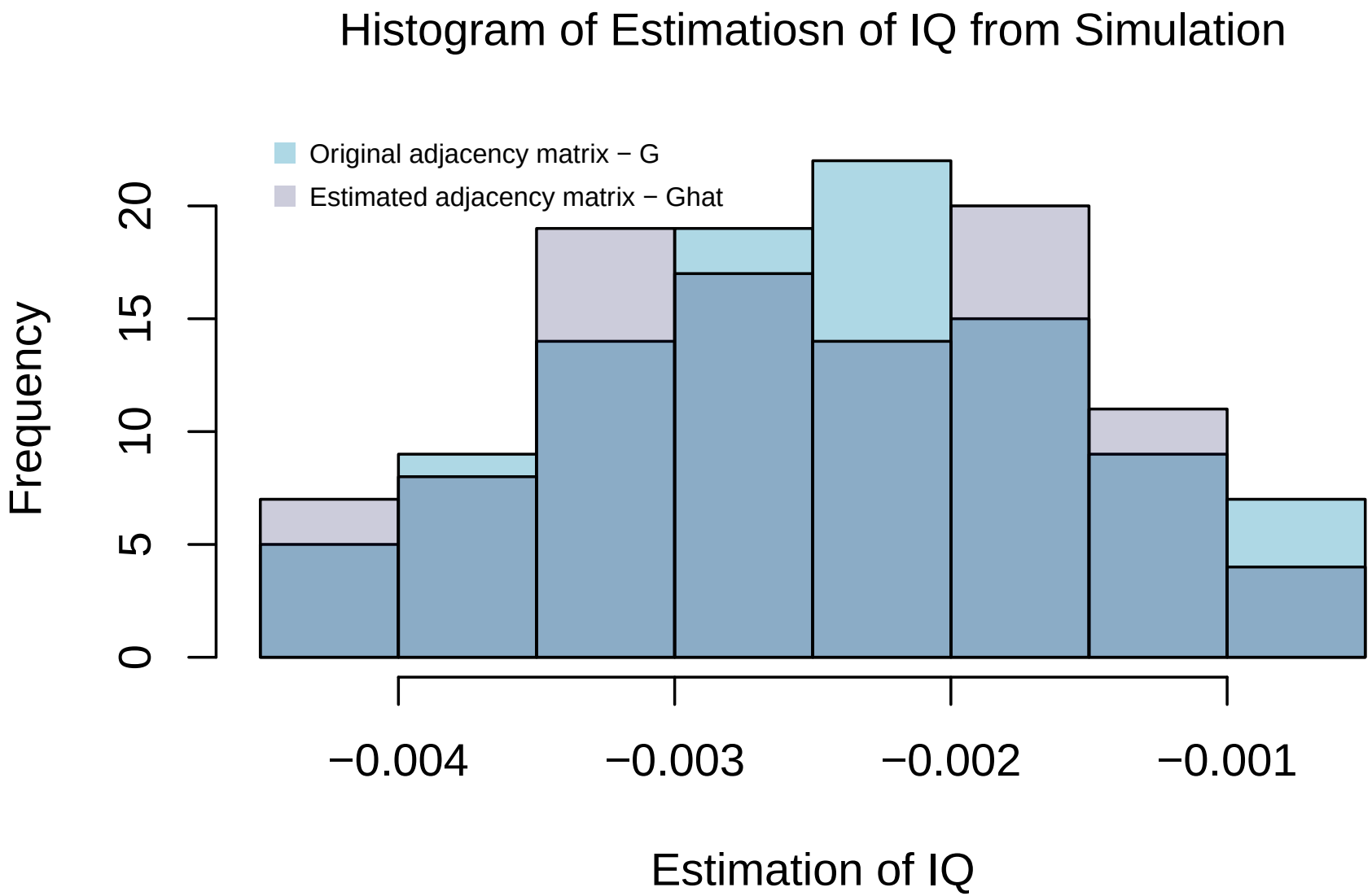
Table 5.3: Summary on selected estimated coefficients

Dependent Variables	Coefficients	Model (2)MLE	(2) 2SLS(G)	(2) 2SLS($\hat{G}$)	(3a)
Math	ϕ	0.0449	0.0932	−0.0164	0.0021
	α	2.1300^{***}	2.4368^{***}	2.4939^{***}	1.8590^{***}
	Previous score	0.5027^{***}	0.4950^{***}	0.5149^{***}	0.5733^{***}
	IQ	-0.0058^{***}	−0.0060	−0.0063	−0.0064
	ρ				0.0675^{***}
English	ϕ	−0.1250	0.0862^{***}	0.0170	0.0086
	α	1.0980^{***}	1.0659	0.9671	1.6046^{*}
	Previous score	0.5252^{***}	0.5363^{***}	0.5393^{***}	0.5087^{***}
	IQ	-0.0104^{***}	-0.0112^{***}	-0.0110^{***}	-0.0100^{***}
	ρ				0.0263
German	ϕ	0.1273^{***}	0.1291^{***}	−0.0187	−0.0004
	α	0.4663^{*}	0.8343	0.9530	−0.1806
	Previous score	0.5072^{***}	0.5044^{*}	0.5050^{***}	0.5018^{***}
	IQ	−0.0023	-0.0029^{*}	-0.0029^{*}	−0.0042

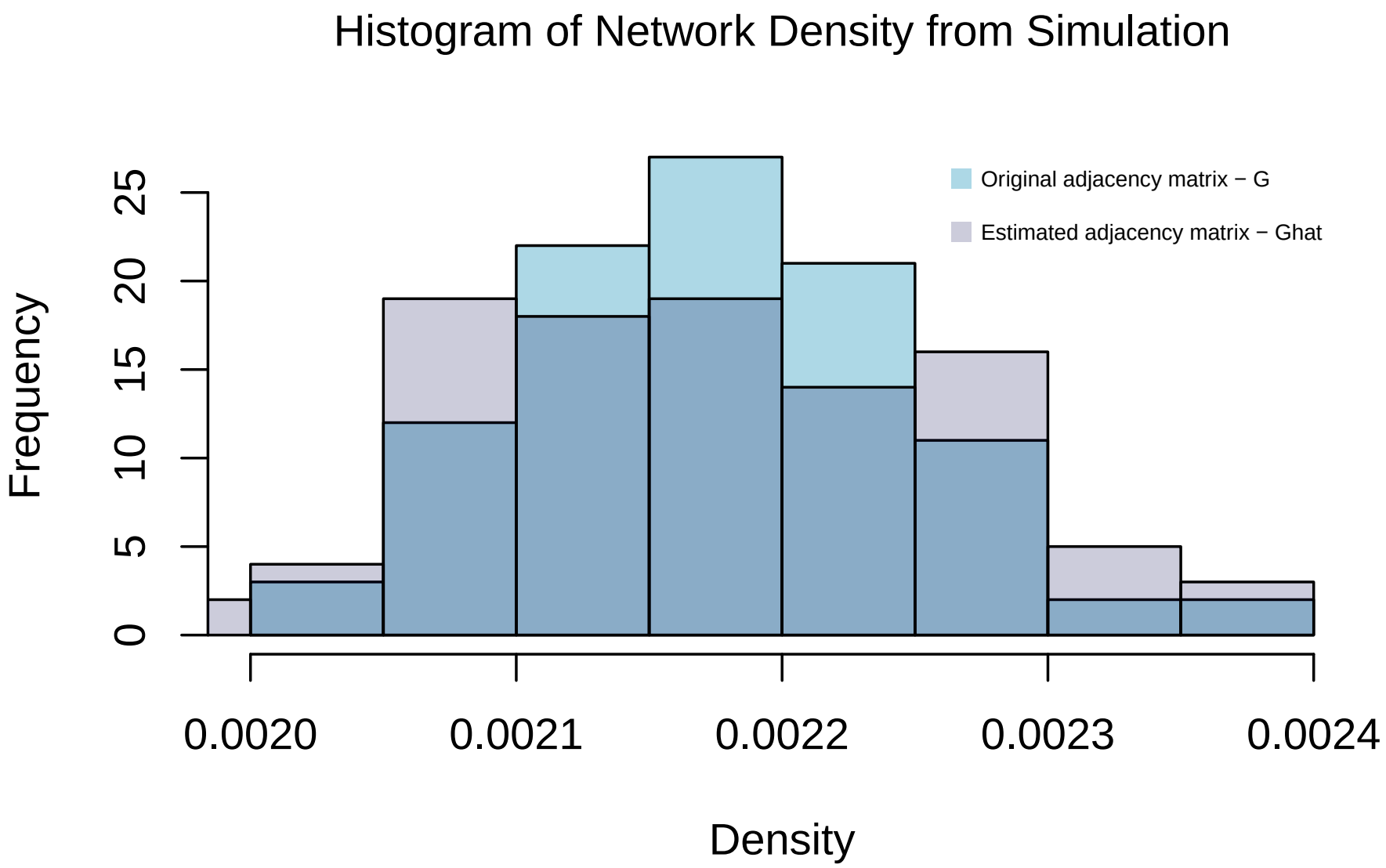
Simulations on 2SLS



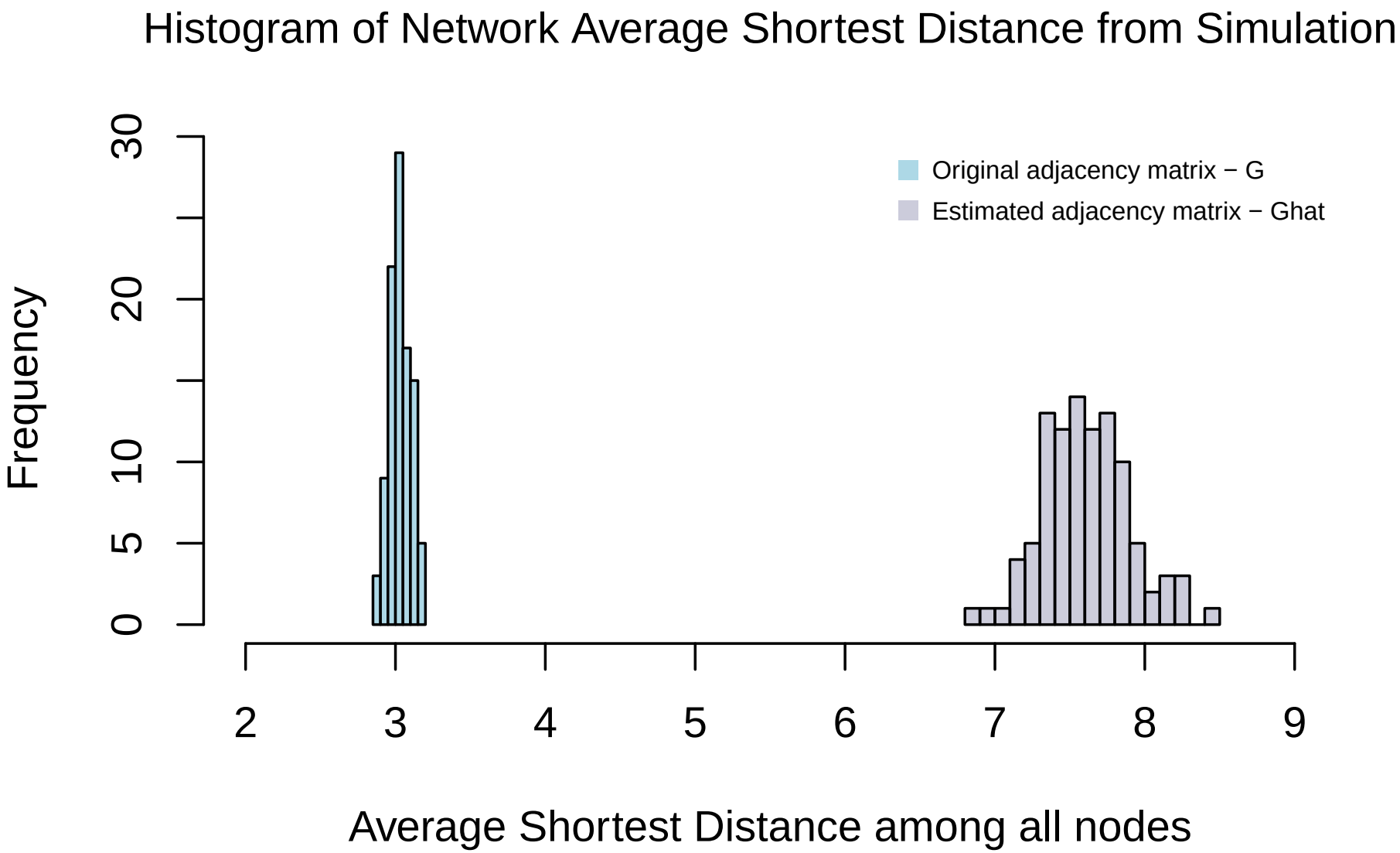
Simulations on 2SLS



Simulations on 2SLS



Simulations on 2SLS



Simulations on 2SLS: rejection rate

Table 1: Rejection rate for selected estimations from 100 simulations

Estimations	2SLS(G)	2SLS($\hat{G}$)
ϕ	0.97	0.06
α	1.00	1.00
iq	0.65	0.61
Previous GPA	1.00	1.00

H_0 : Estimated coefficients $= 0$; rejection at $p < 0.1$

Simulation on 2SLS: correlation between t-ratio and network properties

Table 2: Peer Effect Correlation Matrix for selected variables

	t-ratio(G)	t-ratio($\hat{G}$)	Density(G)	Density($\hat{G}$)	Distance(G)	Dista
t-ratio(G)	1	0.016	-0.272***	-0.222**	0.181*	0.
t-ratio($\hat{G}$)	0.016	1	-0.060	0.026	0.033	-0
Density(G)	-0.272***	-0.060	1	0.810	-0.338***	-0
Density($\hat{G}$)	-0.222**	0.026	0.810	1	-0.240**	-0
Distance(G)	0.181*	0.033	-0.338***	-0.240**	1	0.
Distance($\hat{G}$)	0.102	-0.009	-0.487	-0.720	0.127	

Distance is short for average shortest distance

Simulation on 2SLS: summary

- There is a weak identification/instruments issue with 2SLS when assuming $E(\varepsilon|g, z) = 0$
- The estimation of peer effects ϕ is highly inconsistent with various network densities, which is aligned with the study by Startz and Wood-Doughty (2017)
- Power of test is negatively correlated with the density of network

Limitations

- It didn't estimate spatial autocorrelation model with full sample size
- It didn't include MLE estimations in simulations
- It didn't apply Bayesian estimation on identifying and estimating peer effects (Hsieh and Lee, 2016)
- It didn't employ GMM

Conclusion

By analyzing and disentangling the social network formation, this paper shows:

- Preferential attachment mechanism brings the challenge for identifying and estimating the peer effects in social networks
  1. Assume $E(\varepsilon|g,z) = 0$ - Weak Identification Issue
  2. Assume $E(\varepsilon|g,z) \neq 0$ - spatial autocorrelation (high computational cost)
- Correlation does not imply causality, but identifying causality is difficult in the network

Reference I

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.

Andrew, B. (2018). Network of thrones. `https://networkofthrones.wordpress.com/the-series/season-2/`, Last accessed on 2019-04-09.

Ballester, C., Calvó-Armengol, A., and Zenou, Y. (2006). Who's who in networks. wanted: The key player. *Econometrica*, 74(5):1403–1417.

Bonacich, P. (1987). Power and centrality: A family of measures. *American journal of sociology*, 92(5):1170–1182.

Hsieh, C.-S. and Lee, L. F. (2016). A social interactions model with endogenous friendship formation and selectivity. *Journal of Applied Econometrics*, 31(2):301–319.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.

Reference II

Lada, A. (2013). Power-laws: Scale-free networks. https://cs.brynmawr.edu/Courses/cs380/spring2013/section02/slides/10_ScaleFreeNetworks.pdf, Lecture Notes, Last accessed on 2019-04-10.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542.

Startz, R. and Wood-Doughty, A. (2017). Improved estimation of peer effects using network data. Available at http://econ.ucsb.edu/~startz/Improved%20Estimation%20of%20Peer%20Effects%20Using%20Network%20Data.pdf, Unpublished.