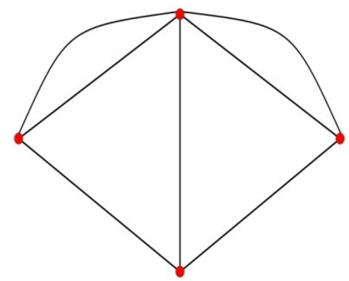
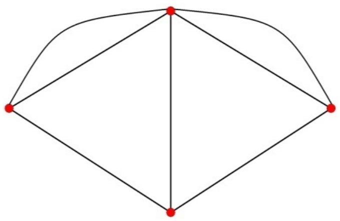
Social and Economic Networks: Models and Analysis

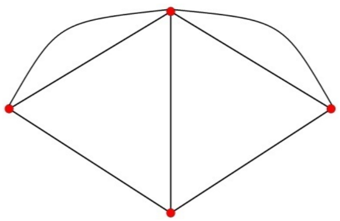


Matthew O. Jackson
Stanford University,
Santa Fe Institute, CIFAR,
www.stanford.edu\~jacksonm

3.1: Growing Random Networks

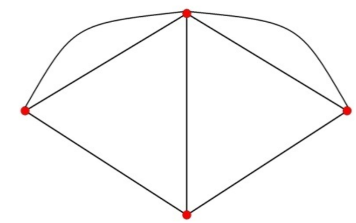


Outline



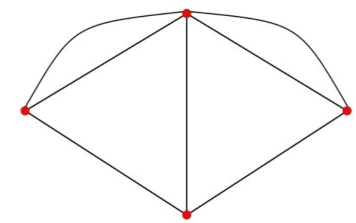
- Part I: Background and Fundamentals
  - Definitions and Characteristics of Networks (1,2)
  - Empirical Background (3)
- Part II: Network Formation
  - Random Network Models (4,5)
  - Strategic Network Models (6, 11)
- Part III: Networks and Behavior
  - Diffusion and Learning (7,8)
  - Games on Networks (9)

Growing Random Networks



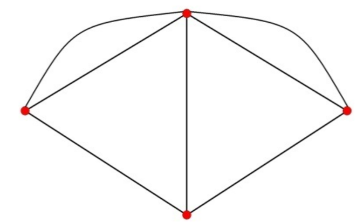
- Citation networks
- Web
- Scientific networks
- Societies...

What do they add?



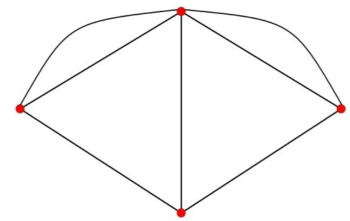
- Realism(?)
- Natural form of heterogeneity via age
- A form of dynamics
- Natural way of varying degree distributions
  - not pre-specified as in static models

Growing and Uniformly Random:



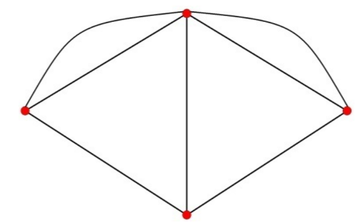
- Start with a simple benchmark model
- Given an idea of techniques
- Then we can enrich the model

Growing and Uniformly Random:



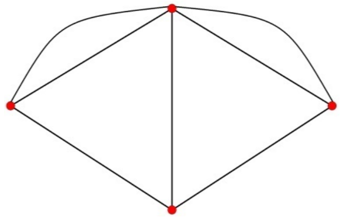
- Each date a new node is born
- Forms m links to existing nodes
- Each node is chosen with equal likelihood

Degree Distribution



- Start with m nodes fully connected
- New node forms m links to existing nodes
- An existing node has a probability m/t of getting new link each period
- No longer binomial, as probabilities vary with time

Distribution of Expected Degrees

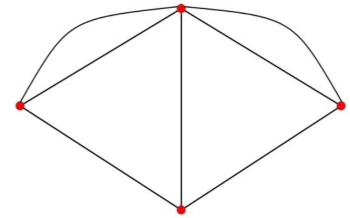


- Expected degree for node i born at $m<i<t$ is

$$m + m/(i+1) + m/(i+2) + \dots + m/t$$

formed at birth

Distribution of Expected Degrees

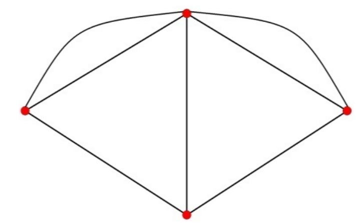


- Expected degree for node i born at m<i<t is

$$m + m/(i+1) + m/(i+2) + \ \dots \ + m/t$$

expected from next node born (time starts at 0, so there are i+1 nodes at time i)

Distribution of Expected Degrees

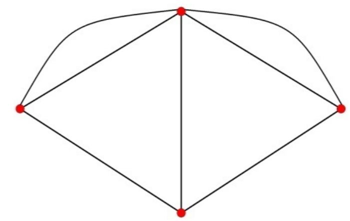


- Expected degree for node i born at m<i<t is

  $m + m/(i+1) + m/(i+2) + \ \dots\ + m/t$

  approx $= m(1+\log(t/i))$ (Harmonic numbers)

Distribution of Expected Degrees

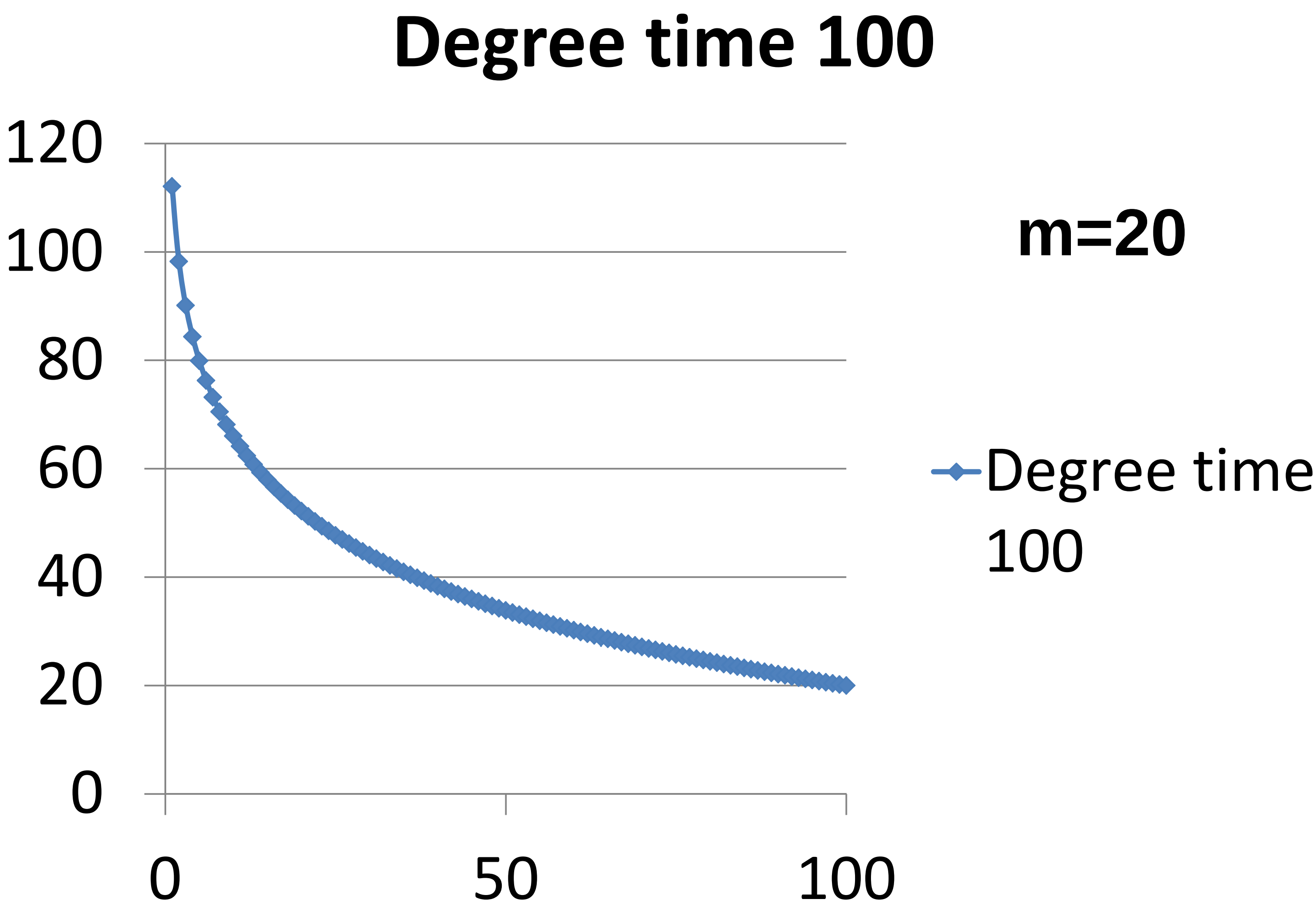


- Expected degree for node i born at m<i<t is

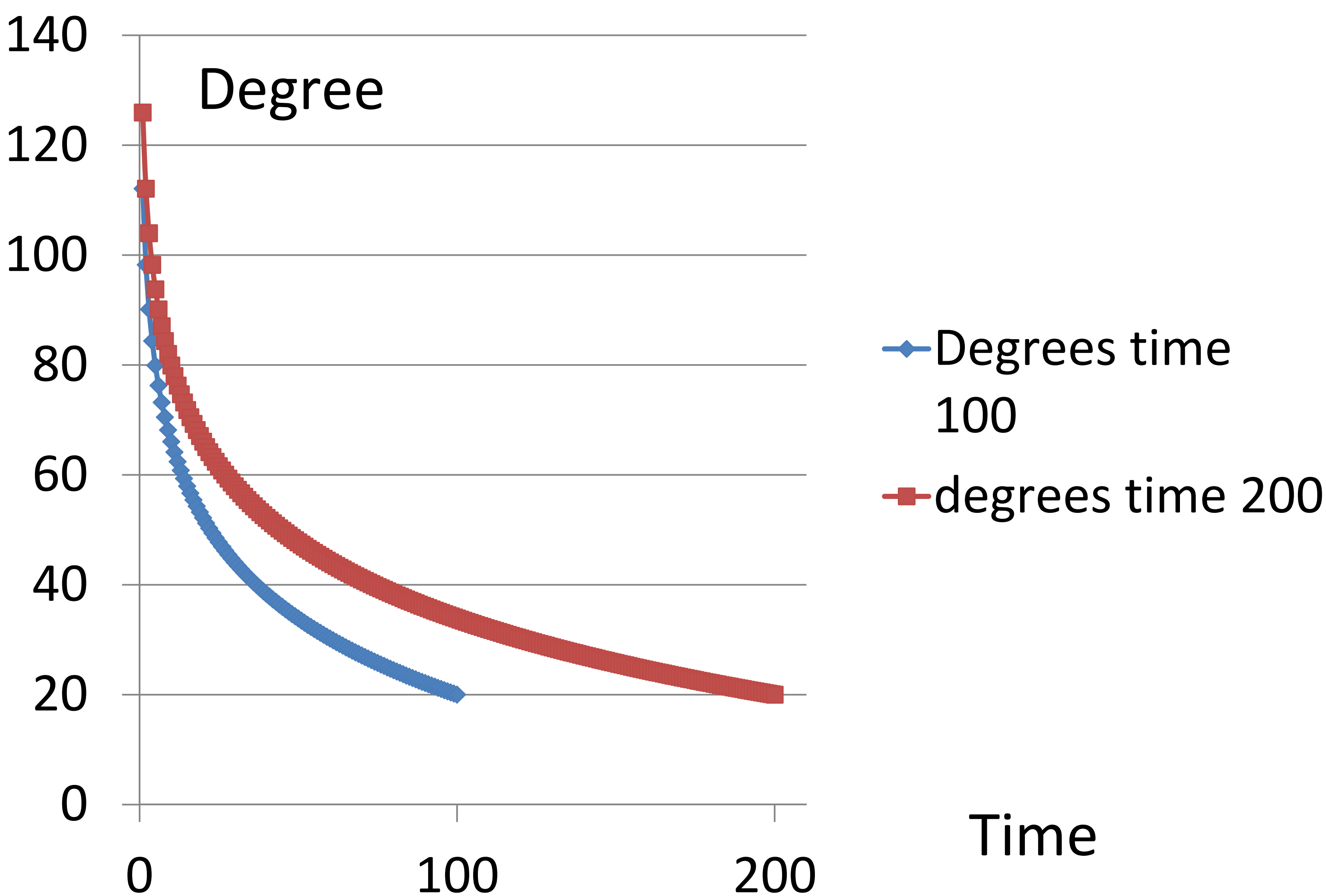
  $m + m/(i+1) + m/(i+2) + \ ... \ + m/ t$

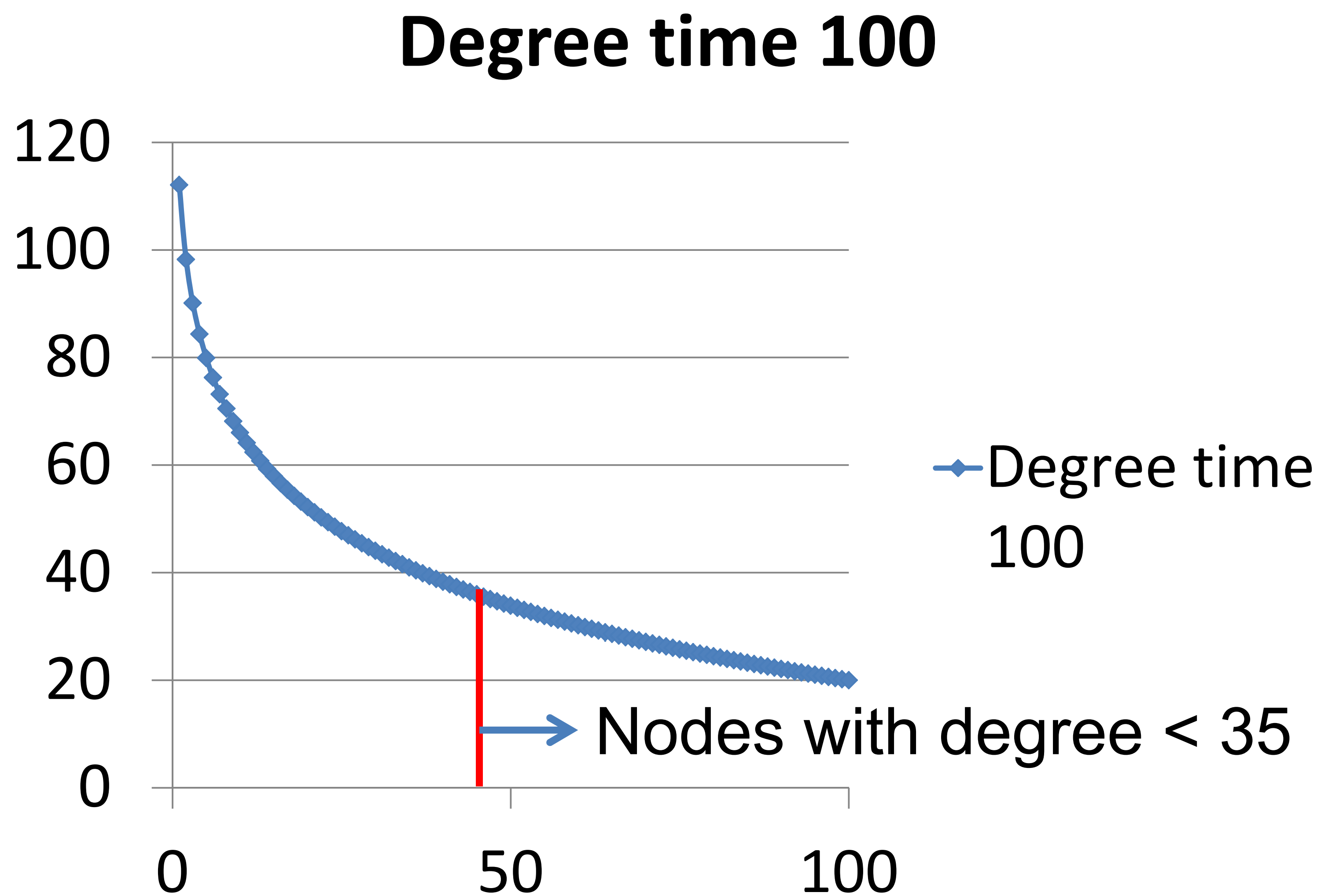
  or $m(1+\log(t/i))$ (Harmonic numbers)

- Nodes that have expected degree less than d at some time t are those such that $m(1+\log(t/i)) < d$

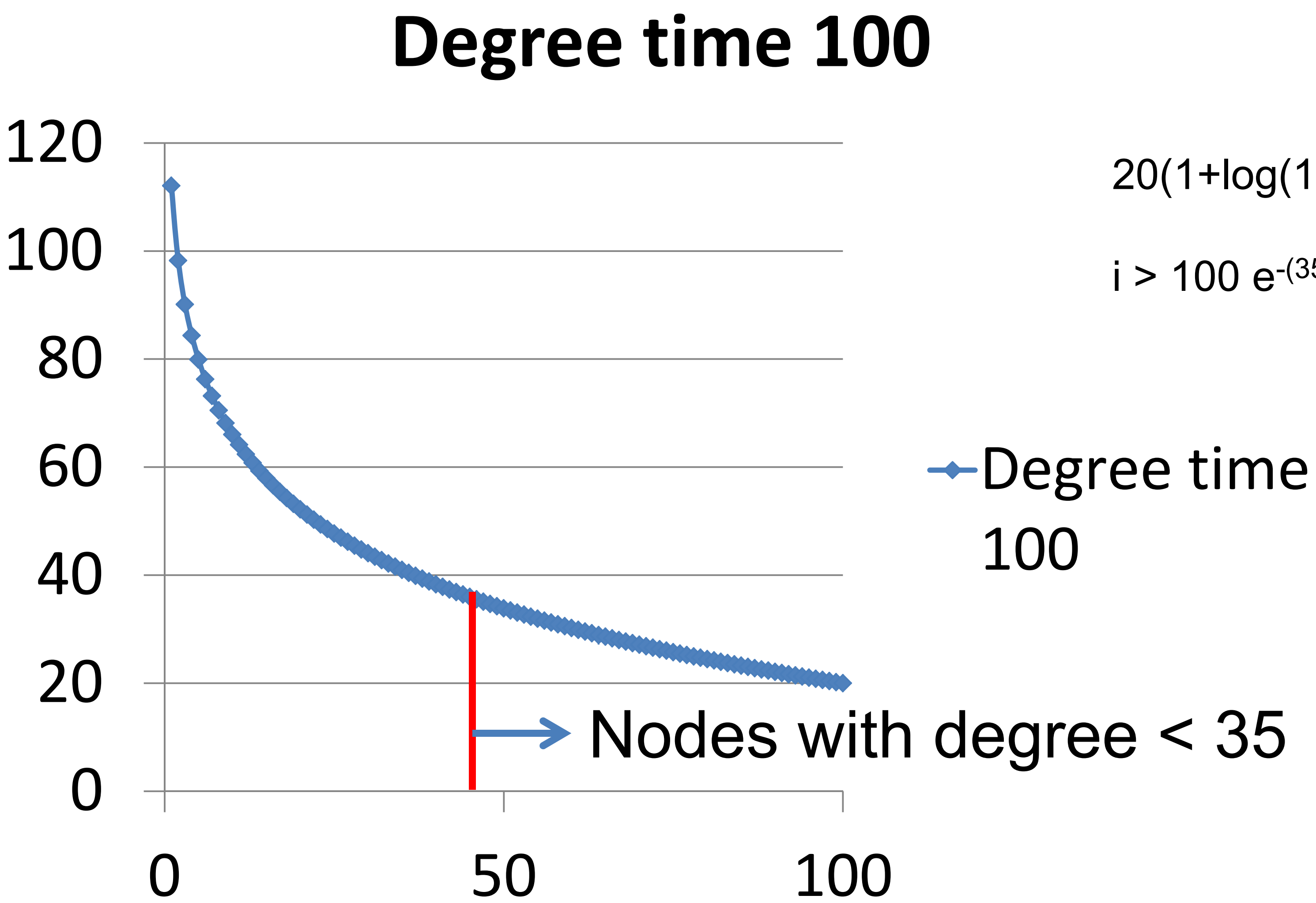


Degree time 100
120
100
80
60
40
20
0
0
50
100
m=20
Degree time 100


Degree
140
120
100
80
60
40
20
0
0
100
200
Time
Degrees time 100
degrees time 200



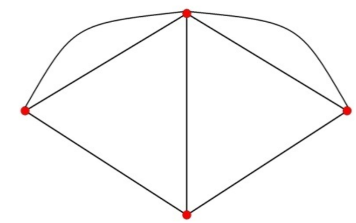
Degree time 100
120
100
80
60
40
20
0
0
50
100
Degree time 100
Nodes with degree < 35



$20(1+\log(100/i)) < 35$

$i > 100\ e^{-(35-20)/20} = 47.2$

Distribution of Expected Degrees



- Expected degree for node i born at m<i<t is

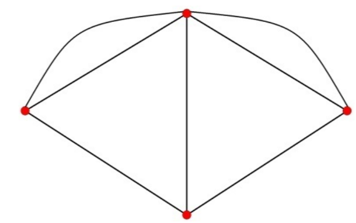
  $m + m/(i+1) + m/(i+2) + \dots + m/t$

  or $m(1+\log(t/i))$ (Harmonic numbers)

- Nodes that have expected degree less than d at some time t are those such that $m(1+\log(t/i)) < d$

  so it is those $i > t\, e^{-(d-m)/m}$

Distribution of Expected Degrees



- Expected degree for node i born at m<i<t is

  $m + m/(i+1) + m/(i+2) + \ ... \ + m/t$

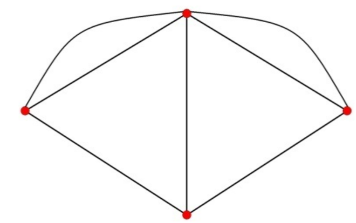
  or $m(1+\log(t/i))$ (Harmonic numbers)

- Nodes that have expected degree less than d at some time t are those such that $m(1+\log(t/i)) < d$

  $i > t\, e^{-(d-m)/m}$

  $F_t(d) = (t - t\, e^{-(d-m)/}m)/t = 1 - e^{-(d-m)/m}$

Degree distribution of growing random network

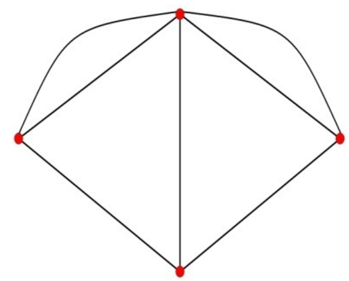


- Distribution of expected degrees is such that
  d-m is exponentially distributed (mean m)

- What about actual degrees?

- Good approximation for large t – need careful large numbers arguments

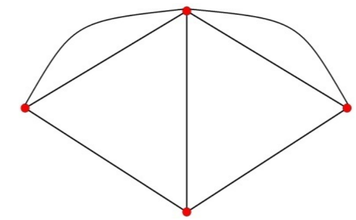
Social and Economic Networks: Models and Analysis



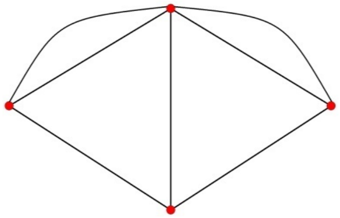
Matthew O. Jackson

Stanford University,
Santa Fe Institute, CIFAR,
www.stanford.edu\~jacksonm

3.2: Mean Field Approximations



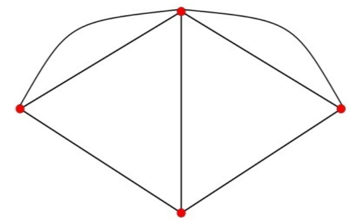
Continuous Time Approximation



- $dd_i(t)/dt = m/t$ and $d_i(i)=m$

new links gained per unit time, starting condition

Continuous Time Approximation

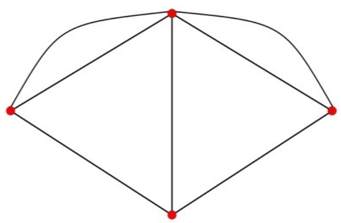


- $dd_i(t)/dt = m/t$ and $d_i(i)=m$

- $d_i(t) = m + m \log(t/i)$

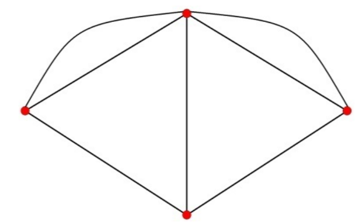
- Same equation as before, then the rest is the same

Growing Random Networks:



- Realism(?)
- Natural form of heterogeneity via age
- A form of dynamics
- Natural way of varying degree distributions

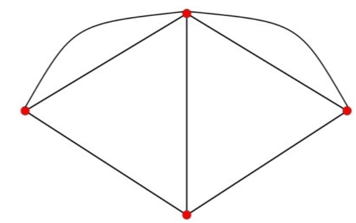
Preferential Attachment



- Other methods of linking than uniformly at random to existing nodes

- Can we get other degree distributions: ``Power laws''?

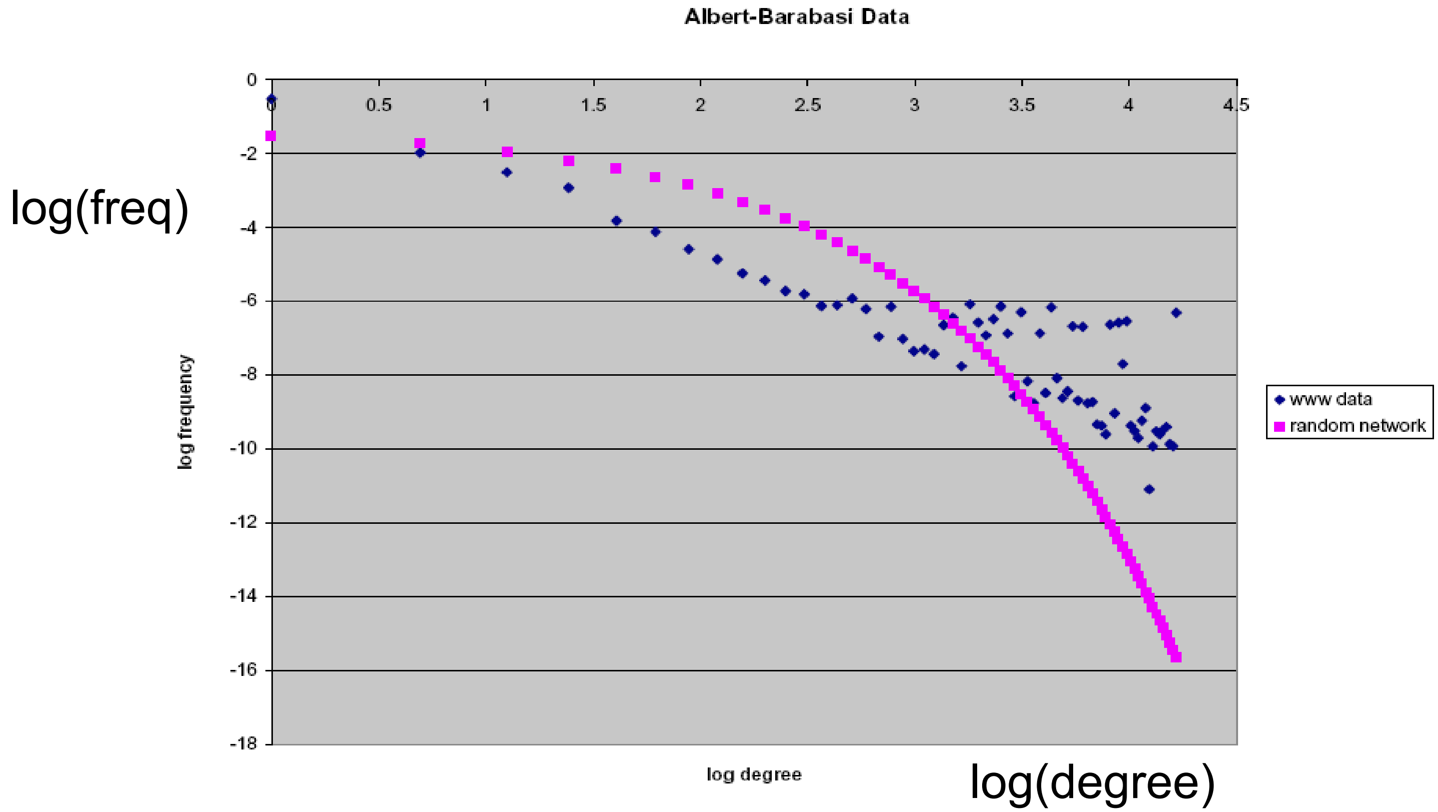
Distribution of links per node: Fat tails (Price 1965)



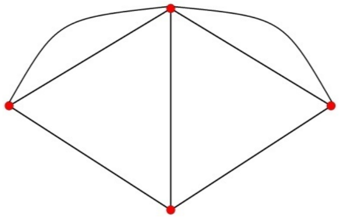
- More high and low degree nodes than predicted at random
  - Citation Networks - too many with 0 citations, too many with high numbers of citations to have citations drawn at random
  - ``Fat tails'' compared to random network

- Related to other settings (wealth, city size, word usage...): Pareto (1896), Yule (1925), Zipf (1949), Simon (1955),

Degree – ND www Albert, Jeong, Barabasi (1999)



Power Law Explanations

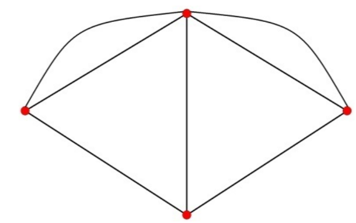


- Simon (1955):

- Rich get richer – growth of existing objects is proportional to size

- New objects enter over time

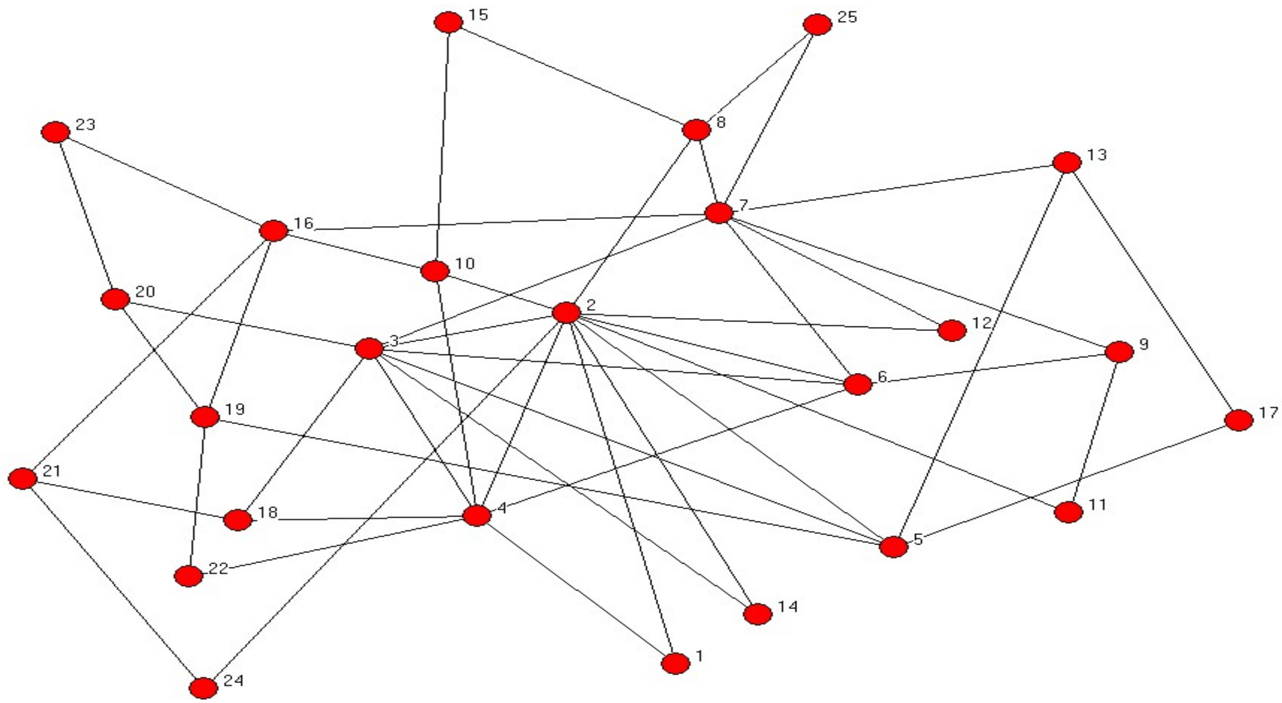
Preferential Attachment (Price (1976), Barabasi and Albert (2001))



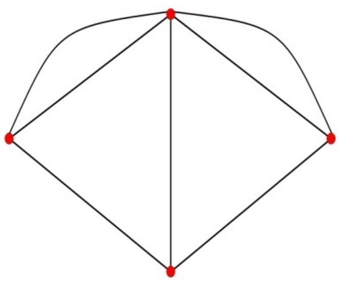
- Previous models don't have the ``fat tails'' of degree distributions

- Nodes born over time, form links at random with existing nodes
  - Form links with probability *proportional to number of links a node already has* - ``rich get richer''

Preferential Attachment (Price (1976), Barabasi and Albert (2001))



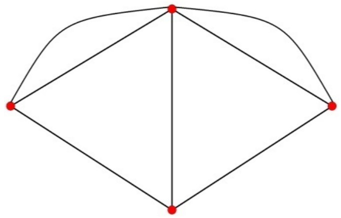
Social and Economic Networks: Models and Analysis



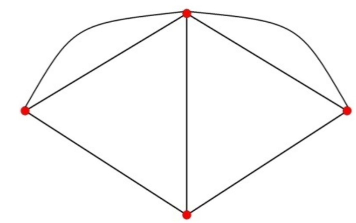
Matthew O. Jackson

Stanford University,
Santa Fe Institute, CIFAR,
www.stanford.edu\~jacksonm

3.3: Preferential Attachment

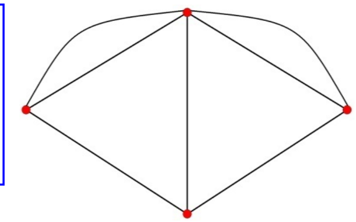


Preferential Attachment



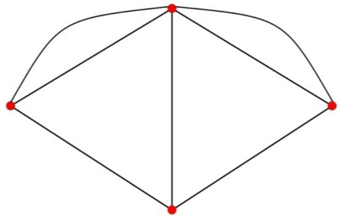
- Newborn nodes form m links to existing nodes
- tm links in total
- total degree is 2tm
- ***Probability of attaching to i is*** $d_i(t)/2tm$

Mean Field Approximation



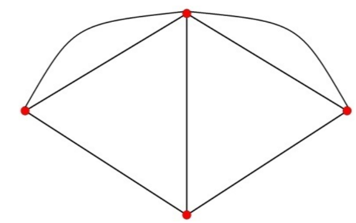
- Continuous time approximation
- Distribution of expected degrees
- Check by simulation??

Distribution of Expected Degrees

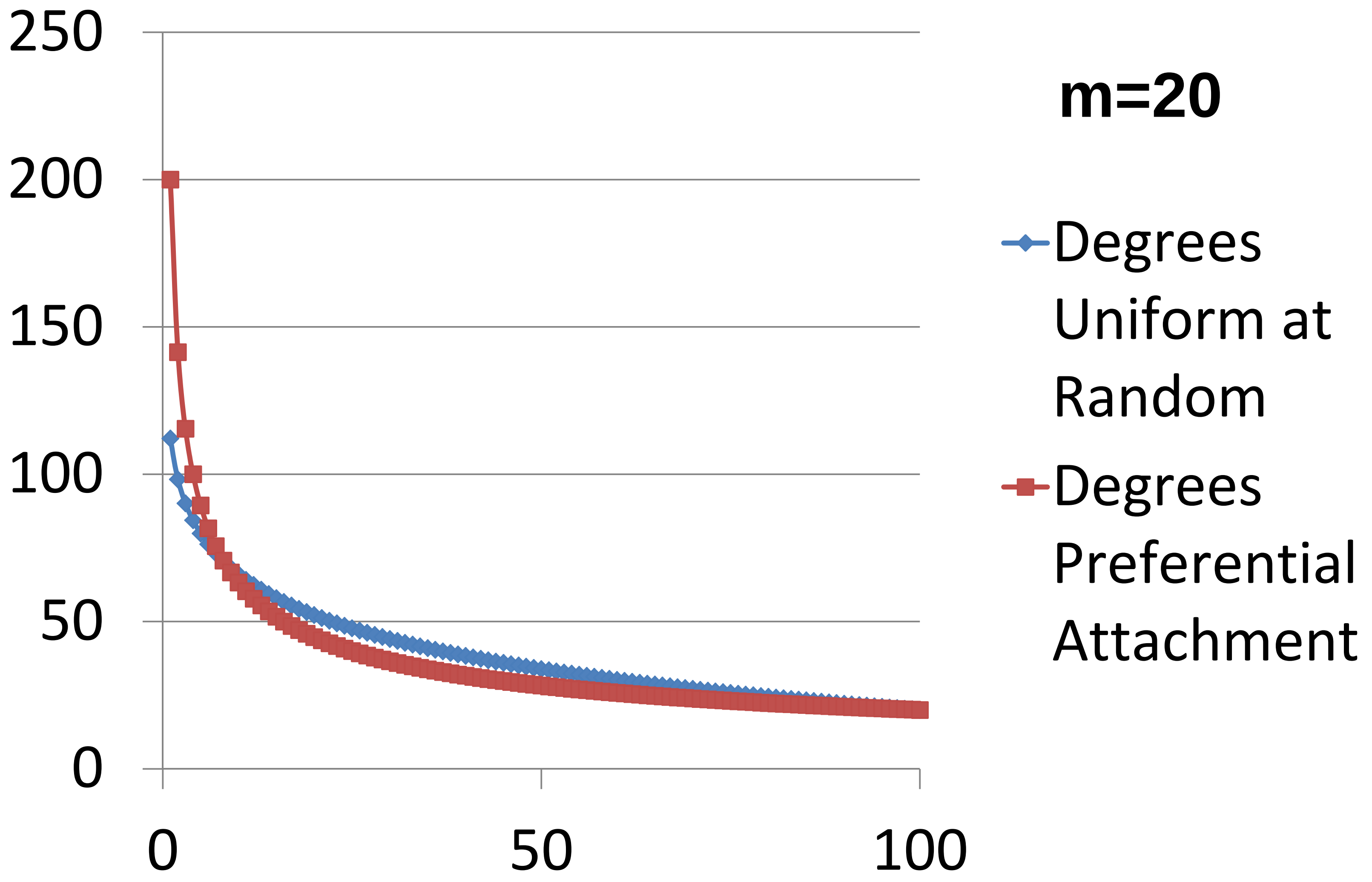


- $dd_i(t)/dt = m(d_i(t)/2tm)$ and $d_i(i)=m$

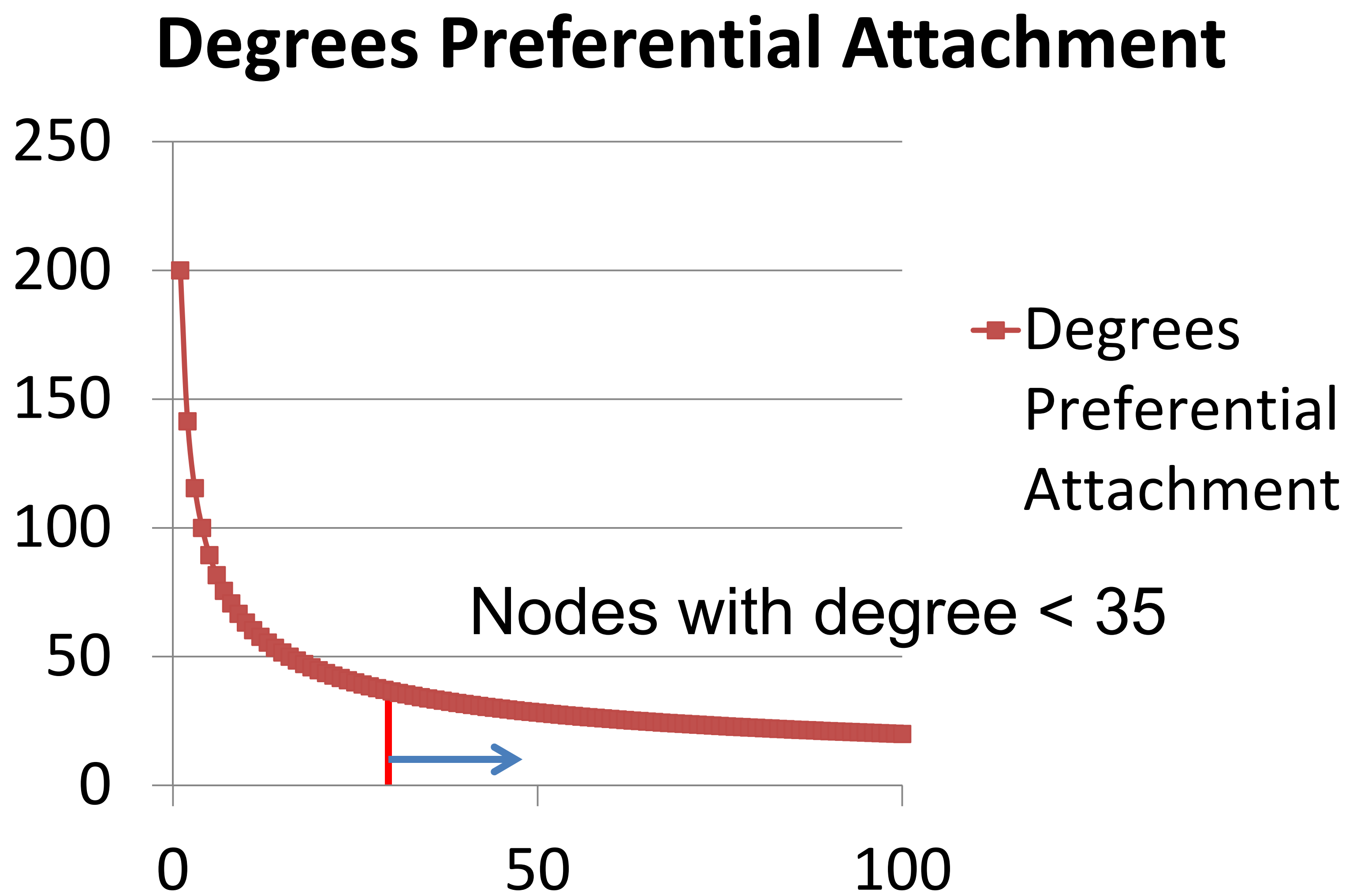
Distribution of Expected Degrees



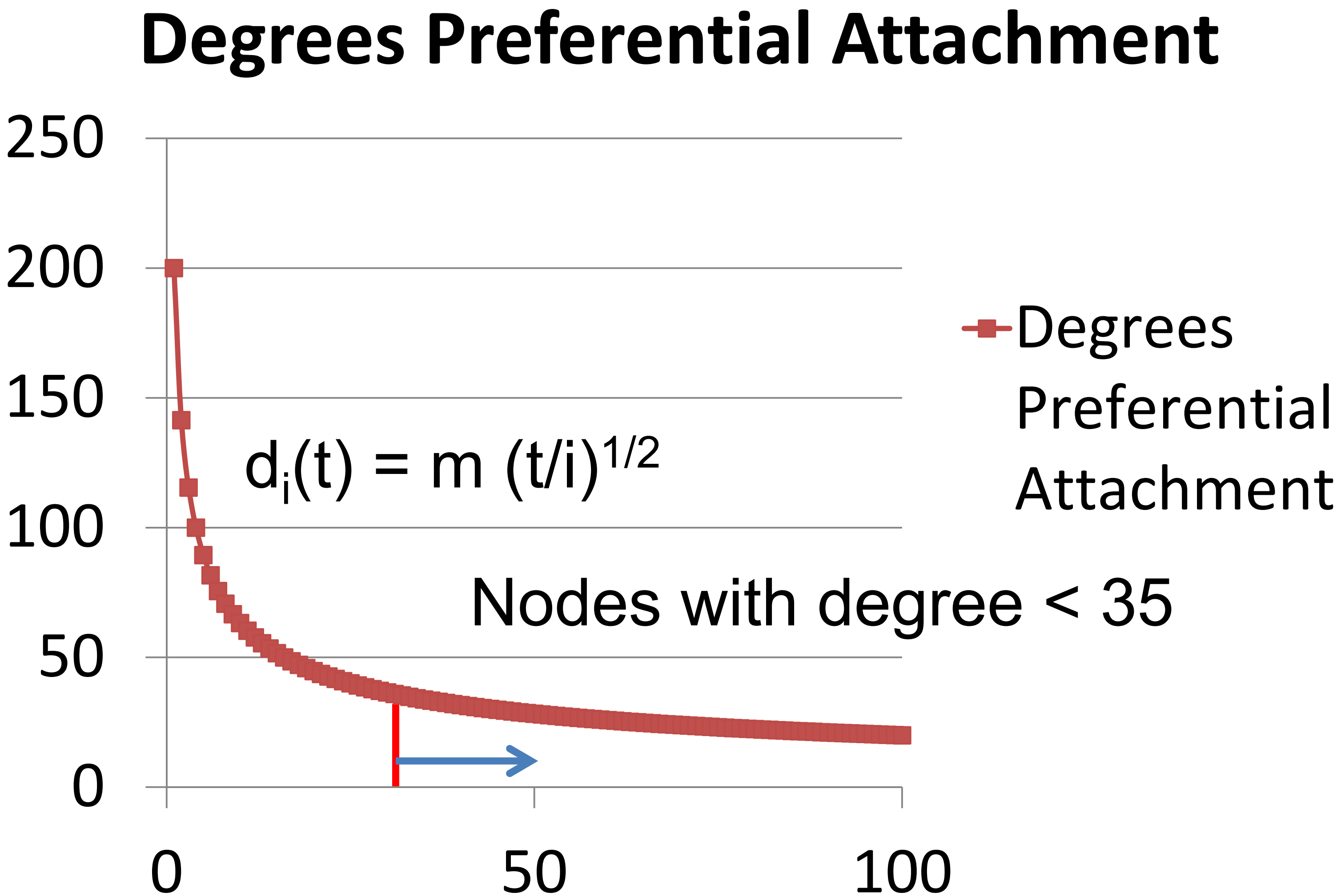
- $dd_i(t)/dt = md_i(t)/2tm = d_i(t)/2t$ and $d_i(i)=m$
- $d_i(t) = m\ (t/i)^{1/2}$



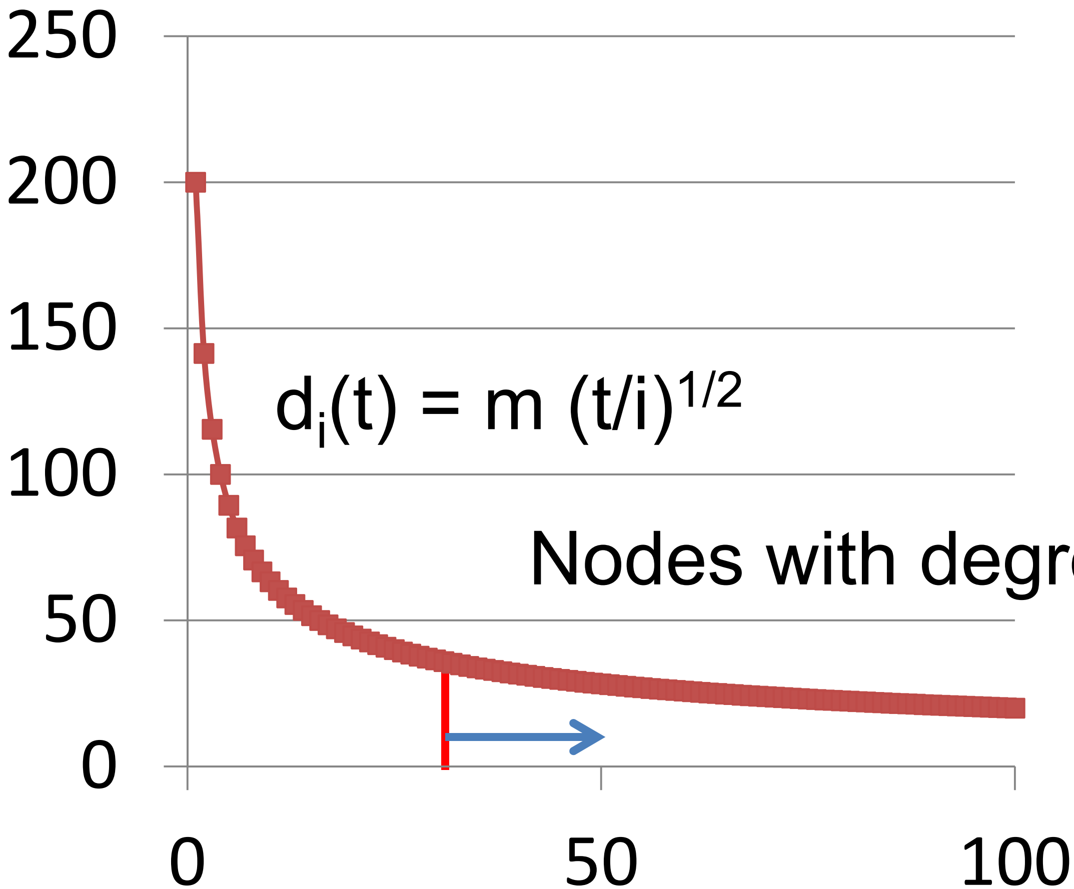
250
200
150
100
50
0
0
50
100
m=20
Degrees Uniform at Random
Degrees Preferential Attachment



Degrees Preferential Attachment
250
200
150
100
50
0
0
50
100
Degrees Preferential Attachment
Nodes with degree < 35


Degrees Preferential Attachment
250
200
150
100
50
0
0
50
100
$d_i(t) = m\ (t/i)^{1/2}$
Nodes with degree < 35
Degrees Preferential Attachment

Degrees Preferential Attachment



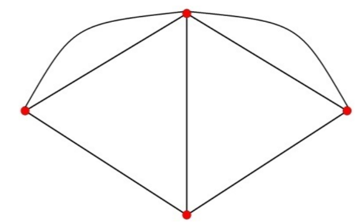
$d_i(t) = m\,(t/i)^{1/2}$

Nodes with degree < 35

$35 > 20\,(100/i)^{1/2}$

$i > 1600/49 = 32.7$

Distribution of Expected Degrees



- Expected degree for node i born at $m<i<t$ is

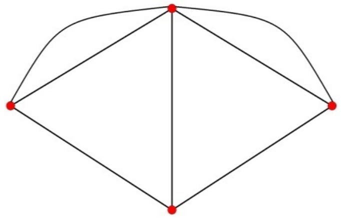
$$d_i(t) = m\,(t/i)^{1/2}$$

- Nodes that have expected degree less than d at some time t are those such that $m\,(t/i)^{1/2} < d$

$$i > t\,m^2/d^2$$

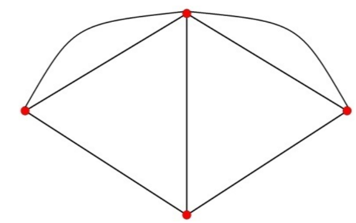
$$F_t(d) = (t - t\,m^2/d^2)/t = 1 - m^2/d^2$$

Distribution of Expected Degrees



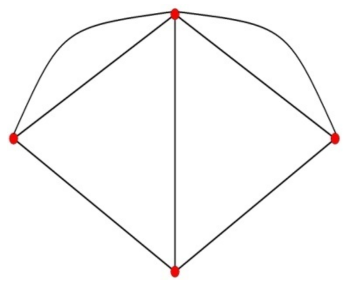
- $F_t(d) = 1 - (m/d)^2$ and $f_t(d) = 2m^2/d^3$

Power Law



- $f_t(d) = 2m^2/d^3$
- $\log(f(d)) = \log(2m^2) – 3 \log(d)$
- Why 3??
- Came from the $dd_i(t)/dt = d_i(t)/2t$

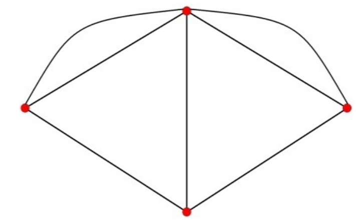
Social and Economic Networks: Models and Analysis



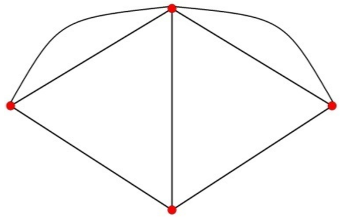
Matthew O. Jackson

Stanford University,
Santa Fe Institute, CIFAR,
www.stanford.edu\~jacksonm

3.4: Hybrid Models

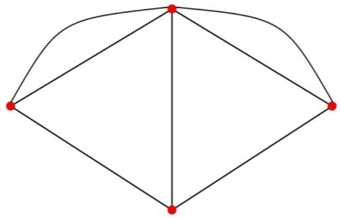


Outline



- Part I: Background and Fundamentals
  - Definitions and Characteristics of Networks (1,2)
  - Empirical Background (3)
- Part II: Network Formation
  - Random Network Models (4,5)
  - Strategic Network Models (6, 11)
- Part III: Networks and Behavior
  - Diffusion and Learning (7,8)
  - Games on Networks (9)

Hybrid Models

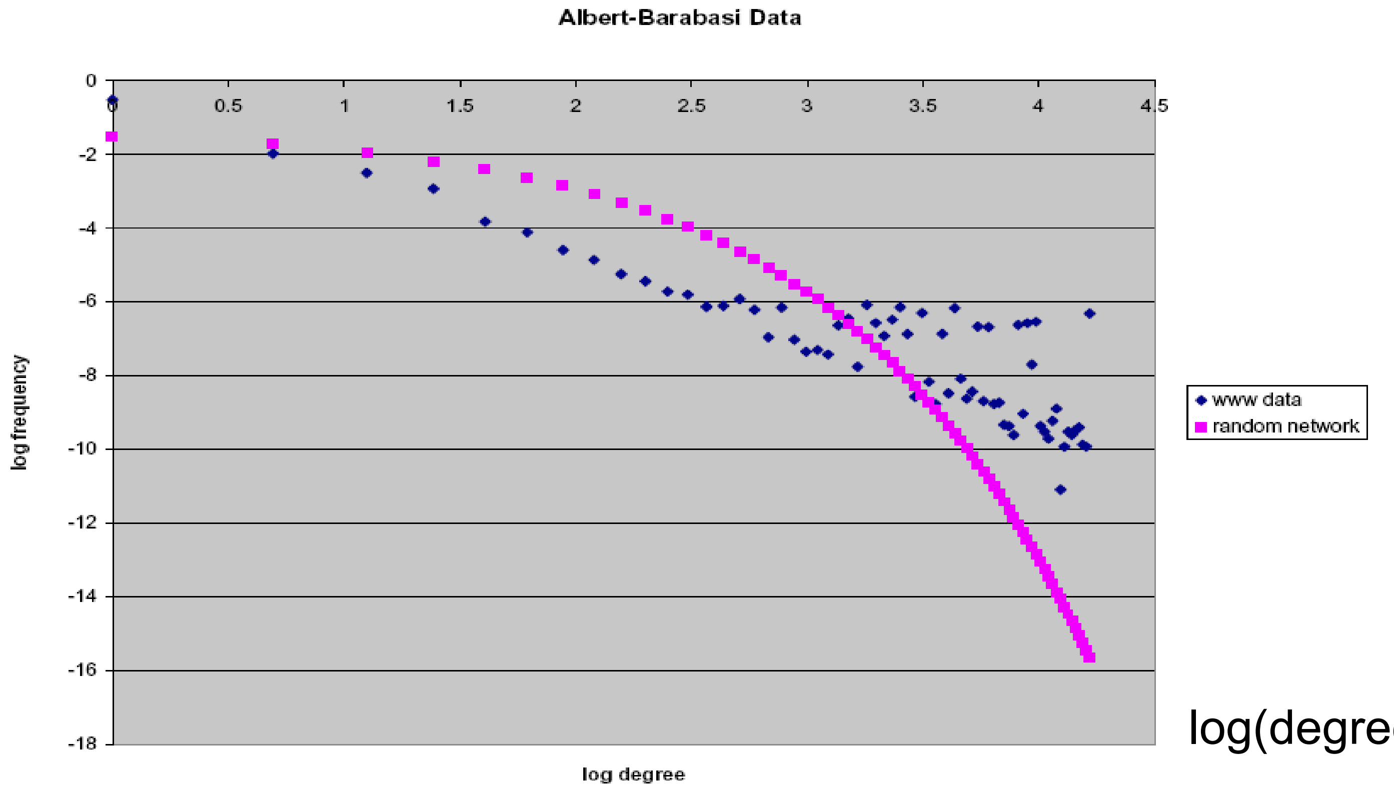


- More on Growing Random Network Models:
  - More general Degree Distributions
  - Other than extremes of random or preferential attachment

- Some fits of hybrid models

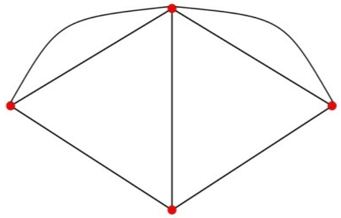
Degree – ND www Albert, Jeong, Barabasi (1999)

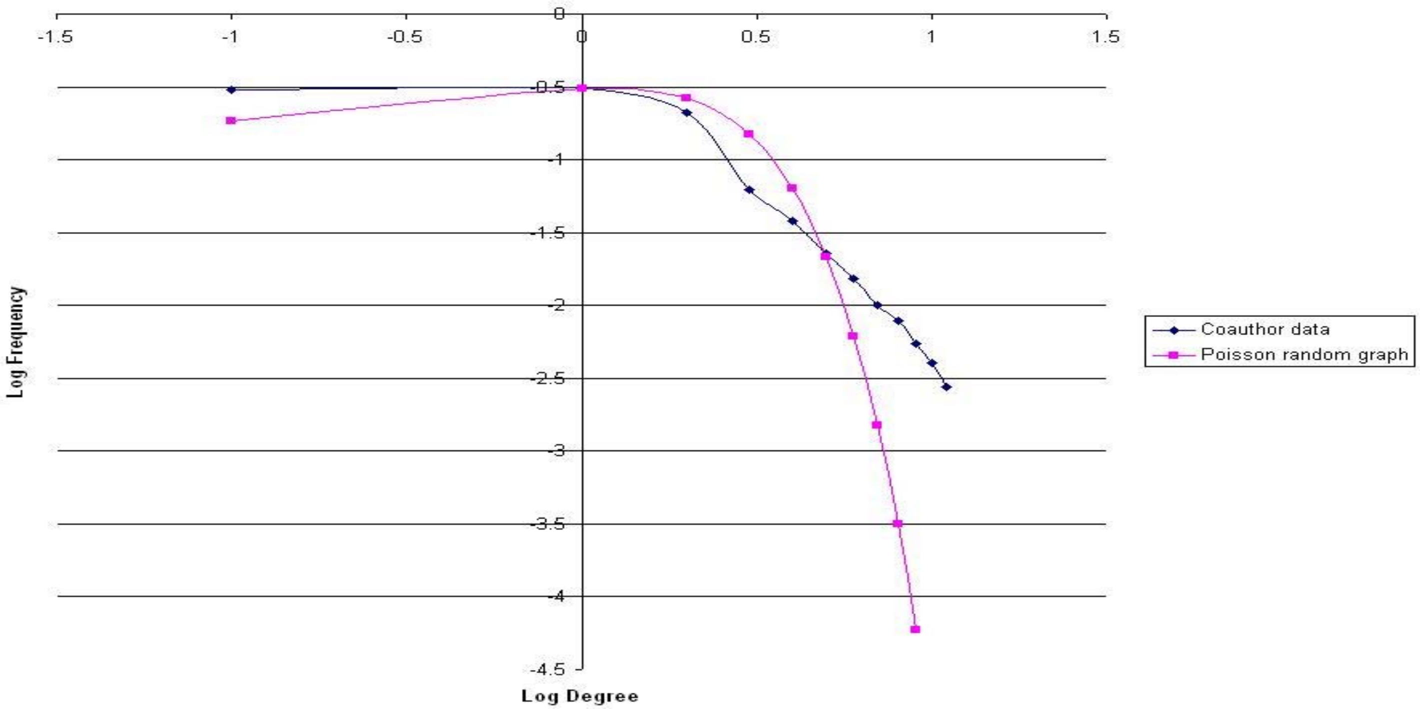
log(freq)



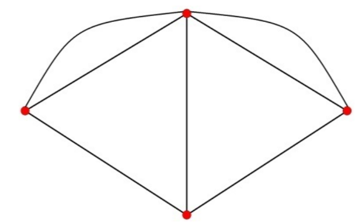
log(degree)

Hybrid Models



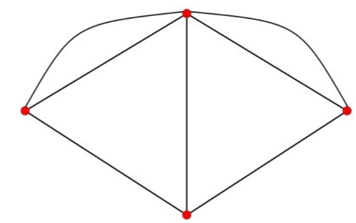


Simple Hybrid



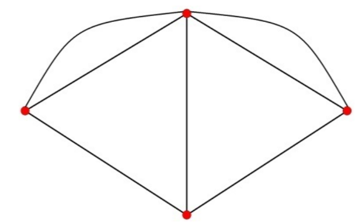
- Simple version of Jackson-Rogers (2007)
- Fraction *a* uniformly at random, 1-*a* via searching neighborhoods of friends

Meeting `Friends of Friends’



- Find new nodes via others: Network-based search
- New node meets *a*m nodes uniformly at random and directs links to them
- Meets (1-*a*)m of their neighbors and attaches to them too

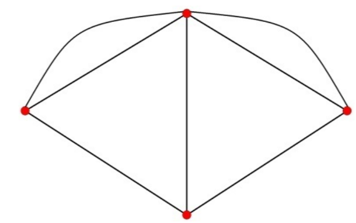
Friends of Friends



- The distribution of neighbors' nodes is not the same as the degree distribution – even with independent link formation

- A neighbor is more likely to be higher degree

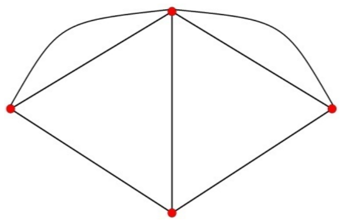
Relation to Preferential Attachment:



- In a network with *half* degree k and half degree 2k individuals:

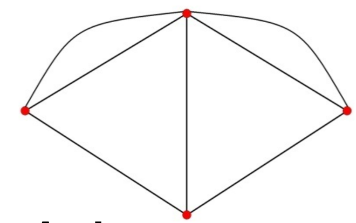
- randomly select a link and then a node on one end of it - *2/3* chance that it has degree 2k, *1/3* chance that it has degree k

Friends of Friends



- Randomly find a node
- Randomly pick one of the nodes it attached to
- Chance of finding some node via the second part of this procedure is proportional to its degree: find it if find one of its neighbors....

Simple Hybrid



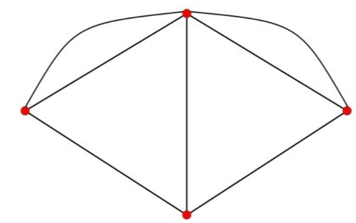
- Fraction *a* uniformly at random, 1-*a* via preferential attachment:

$$dd_i(t)/dt = am/t + (1-a)d_i(t)/2t$$

and $d_i(i)=m$

$$d_i(t) = (m + 2am/(1-a))(t/i)^{(1-a)/2} - 2am/(1-a)$$

Degree distribution



Nodes that have expected degree less than d at some time t are those i such that

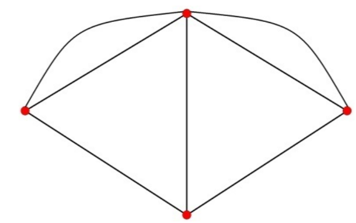
$$(m + xam)(t/i)^{1/x} - xam < d$$

where $x = 2/(1-a)$

critical i is such that

$$i/t = [(m + xam)/(d + xam)]^x$$

Degree Distribution

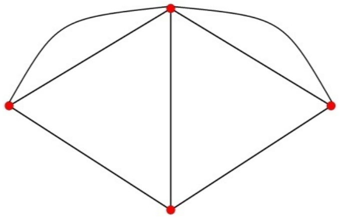


- $F(d) = (t - i)/t$

- $F(d) = 1 - ((m+amx)/(d+amx))^x$

  where $x = 2/(1-a)$

Spans Extremes



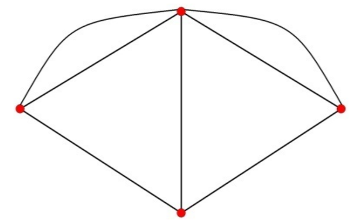
$F(d) = 1 - ((m+amx)/(d+amx))^x$

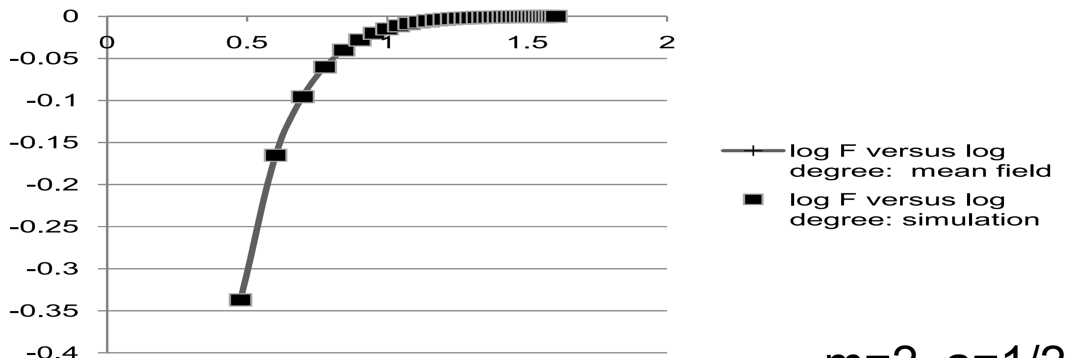
$x = 2/(1-a)$

a near 1 nearly exponential,

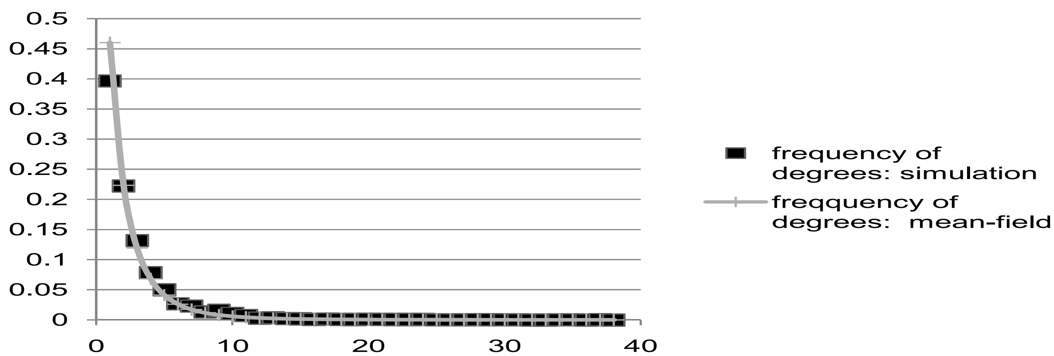
a near 0 nearly preferential

Simulate to check:

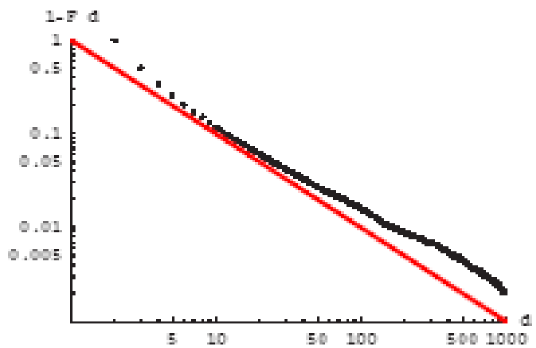




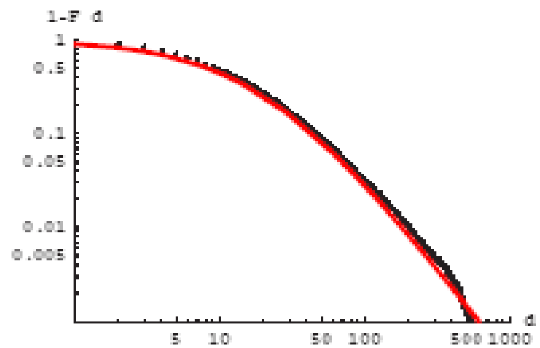
m=2, a=1/2, t=1000



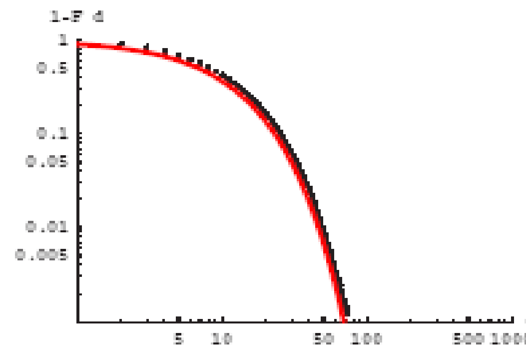
Degree Distributions as vary the random/search parameter



a=0

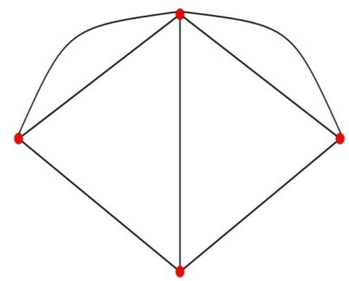


a=1/2



a= 1

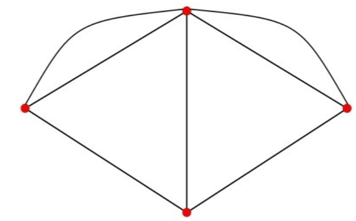
Social and Economic Networks: Models and Analysis



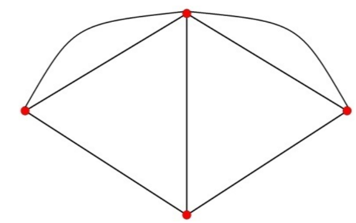
Matthew O. Jackson

Stanford University,
Santa Fe Institute, CIFAR,
www.stanford.edu\~jacksonm

3.5: Fitting Hybrid Models



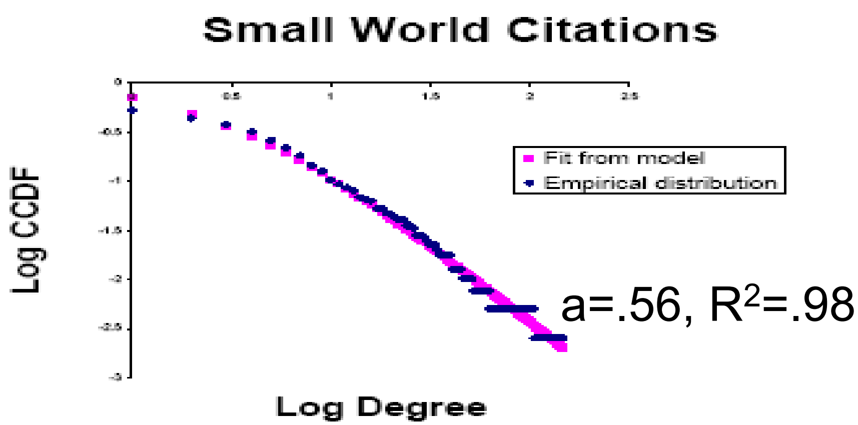
Fitting to data:

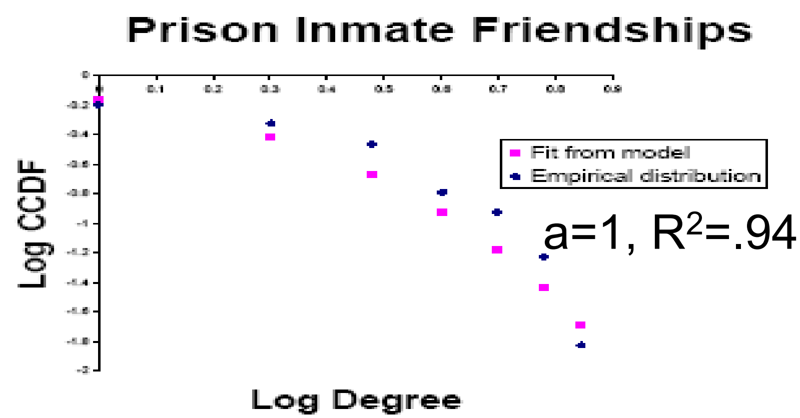


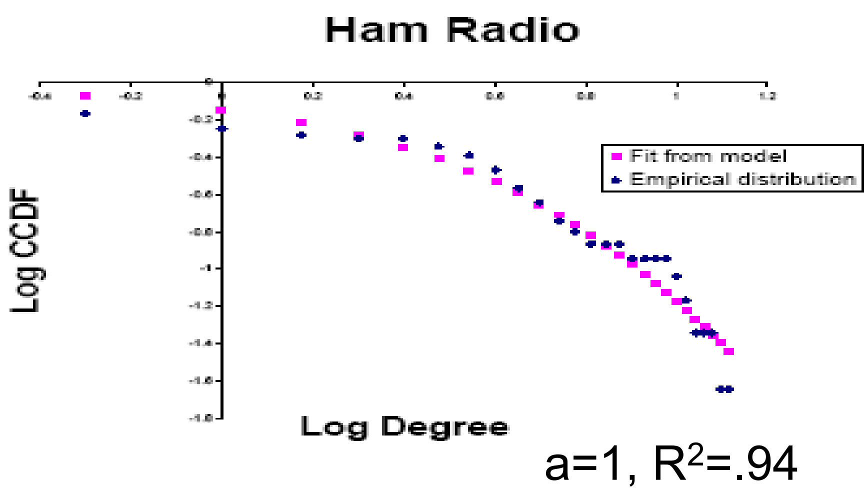
- $F(d) = 1 - ((m+amx)/(d+amx))^x \quad x = 2/(1-a)$

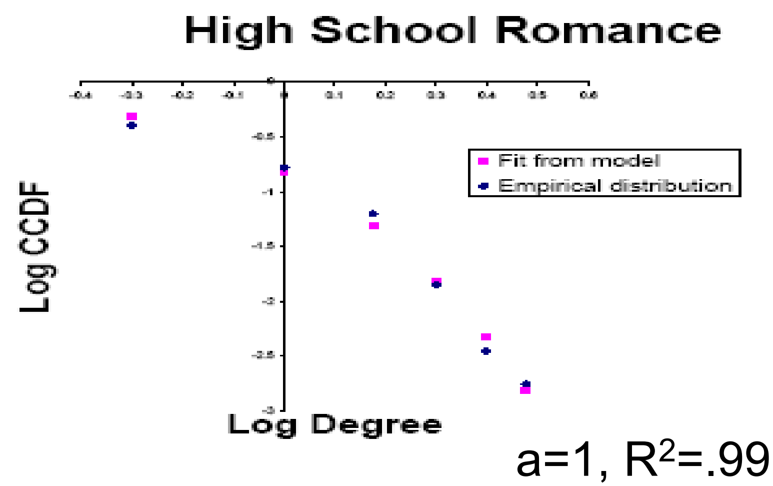
$$\log(1-F(d)) = c - x \log(d+amx)$$

- estimate m directly
- select a to minimize distance between actual distribution and model's distribution

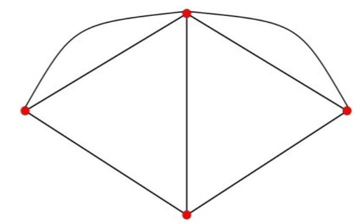

Small World Citations
Log CCDF
Log Degree
Fit from model
Empirical distribution
a=.56, R^2=.98


Prison Inmate Friendships
Log CCDF
Log Degree
Fit from model
Empirical distribution
a=1, R^2=.94


Ham Radio
Log CCDF
Log Degree
Fit from model
Empirical distribution
a=1, R^2=.94


High School Romance
Log CCDF
Log Degree
Fit from model
Empirical distribution
a=1, R^2=.99

Notes on Fits:



- Ham, Prison, Romance are even more curved than with a=1   :   random without growth fits even better (Poisson)

- Citations: too many with degree 0,   here start with some degree

Degree – ND www Albert, Jeong,

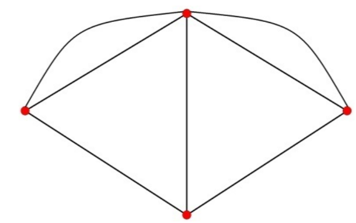
log(freq)

a=.36

R^2=.97

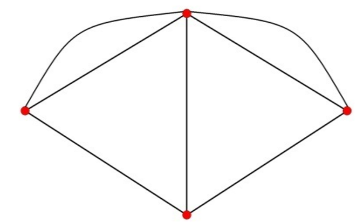
log(degree)

Preferential Attachment?



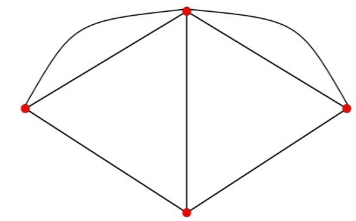
- Fit of Barabasi and Albert has a=.36
- More than 1/3 at random
- Eyeballing Log-Log plots can be misleading!!
- Fat Tails – Yes,   Actual Power law - No

Fitting the friends of friends model



- Fit to estimate ratio of random to network based links: $r = m_r / m_n$
- $r = a/(1-a)$ from before
- r ranges from 0 to infinity, while a ranges from 0 to 1

Friends of Friends/Hybrid Models:



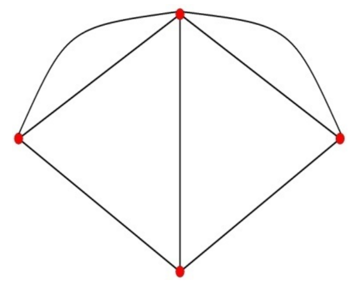
- Variety of degree distributions
- Tie degree distributions to way in which links formed:
  - fat tails from network meeting process
  - more likely to meet well-connected nodes
- *Clustering* from network meeting process
  - connecting to friends of friends
- *Diameter* naturally as small as E-R network
- *Assortativity* in degree based on age

$r = a/(1-a)$

TABLE 1—PARAMETER ESTIMATES ACROSS APPLICATIONS

Dataset:	WWW	Citations	Coauthor	Ham radio	Prison	High-school romance
Number of nodes:	325729	396	81217	44	67	572
Avg. in-degree: m	4.6	5.0	0.84	3.5	2.7	0.83
r from Fit	0.57	0.63	4.7	5.0	∞	∞
p from Fit	0.36	0.27	0.10	1	1	—
R^2 of Fit	0.97	0.98	0.99	0.94	0.94	0.99
Avg. clustering data	0.11	0.07	0.16	0.47	0.31	—
Avg. clustering fit	0.11	0.07	0.16	0.22	0.10	—
Diameter data	11.3 (avg)	4	26	5	7	—
Diameter fit	(6, 12)	(4, 8)	(19, 38)	(4, 8)	(5, 10)	(12, 24)

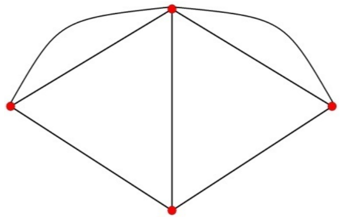
Social and Economic Networks: Models and Analysis



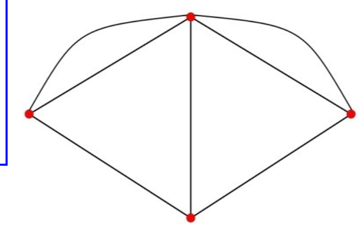
Matthew O. Jackson

Stanford University,
Santa Fe Institute, CIFAR,
www.stanford.edu\~jacksonm

3.6: Block Models

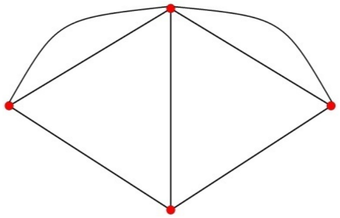


Random Network Models:



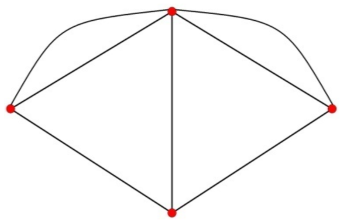
- Erdos-Renyi
  - Useful for understanding thresholds and how networks come to exhibit certain features
  - Miss many real-world features: e.g., clustering
- Other models link-by-link models
  - Watts and Strogatz, Barabasi and Albert, Jackson and Rogers....
  - Capture other features: clustering, degree distribution, correlation...
- Stochastic Block Models
  - Enrich Erdos-Renyi to allow for probabilities to depend on node characteristics, attributes (or on latent – unobserved characteristics)
- Popular set of models: ERGMs and new ones: SERGMs/SUGMs
  - flexible way to introduce various **local** features and dependencies
  - estimated statistically

Random Network Models:



- Erdos-Renyi
  - Useful for understanding thresholds and how networks come to exhibit certain features
  - Miss many real-world features: e.g., clustering
- Other models link-by-link models
  - Watts and Strogatz, Barabasi and Albert, Jackson and Rogers....
  - Capture other features: clustering, degree distribution, correlation...
- Stochastic Block Models
  - Enrich Erdos-Renyi to allow for probabilities to depend on node characteristics, attributes (or on latent – unobserved characteristics)
- Popular set of models: ERGMs and new ones: SERGMs/SUGMs
  - flexible way to introduce various **local** features and dependencies
  - estimated statistically

Block model



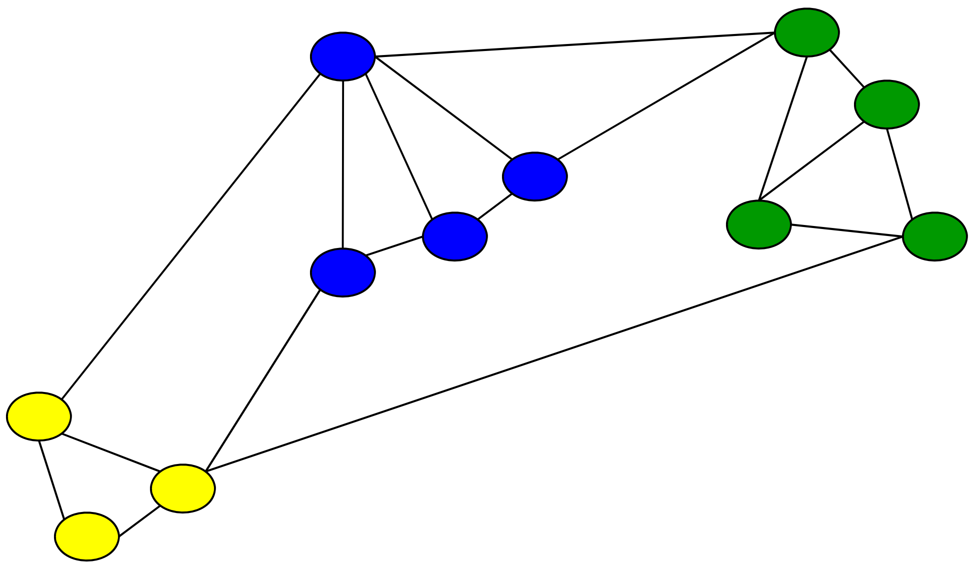
Extend the basic Erdos-Renyi G(n,p) model:

Nodes have characteristics:

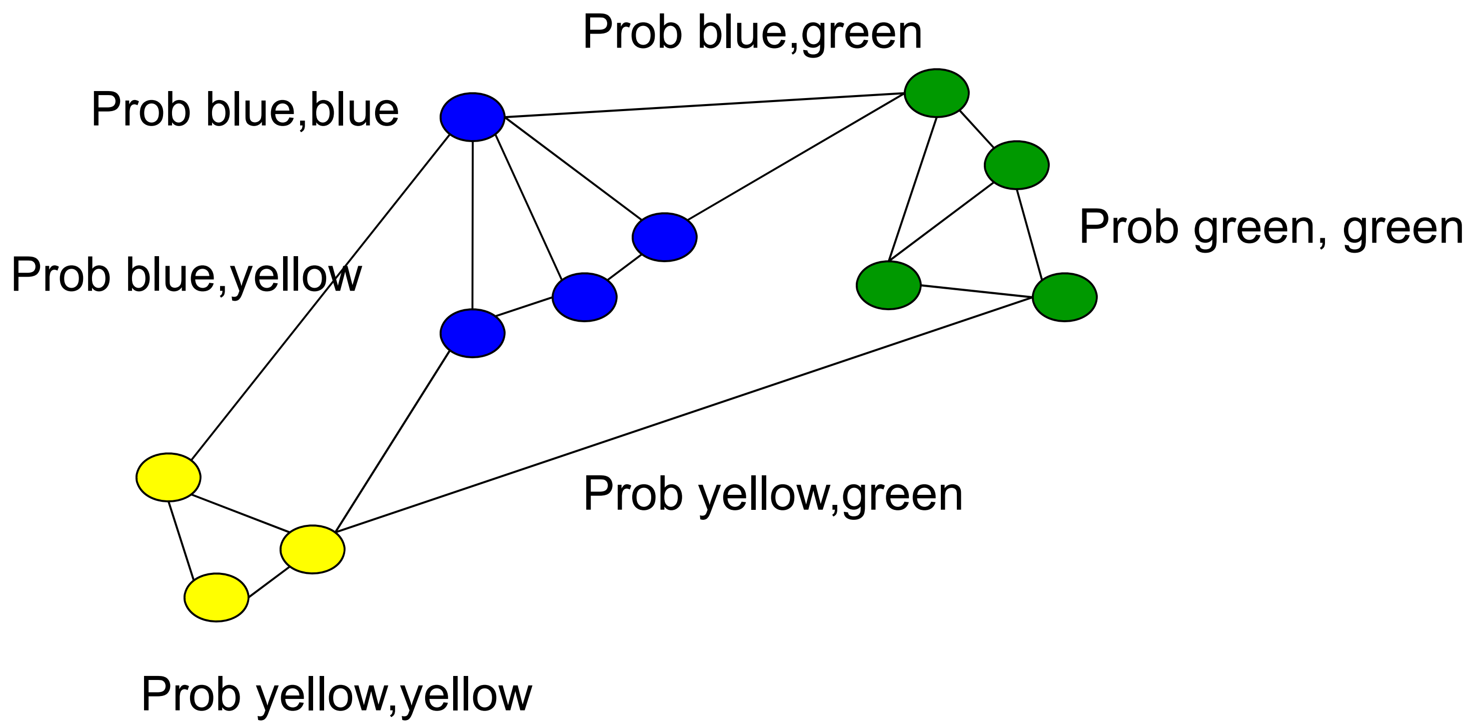
e.g., age, gender, religion, profession, etc.

links between nodes depend on the pairs' characteristics

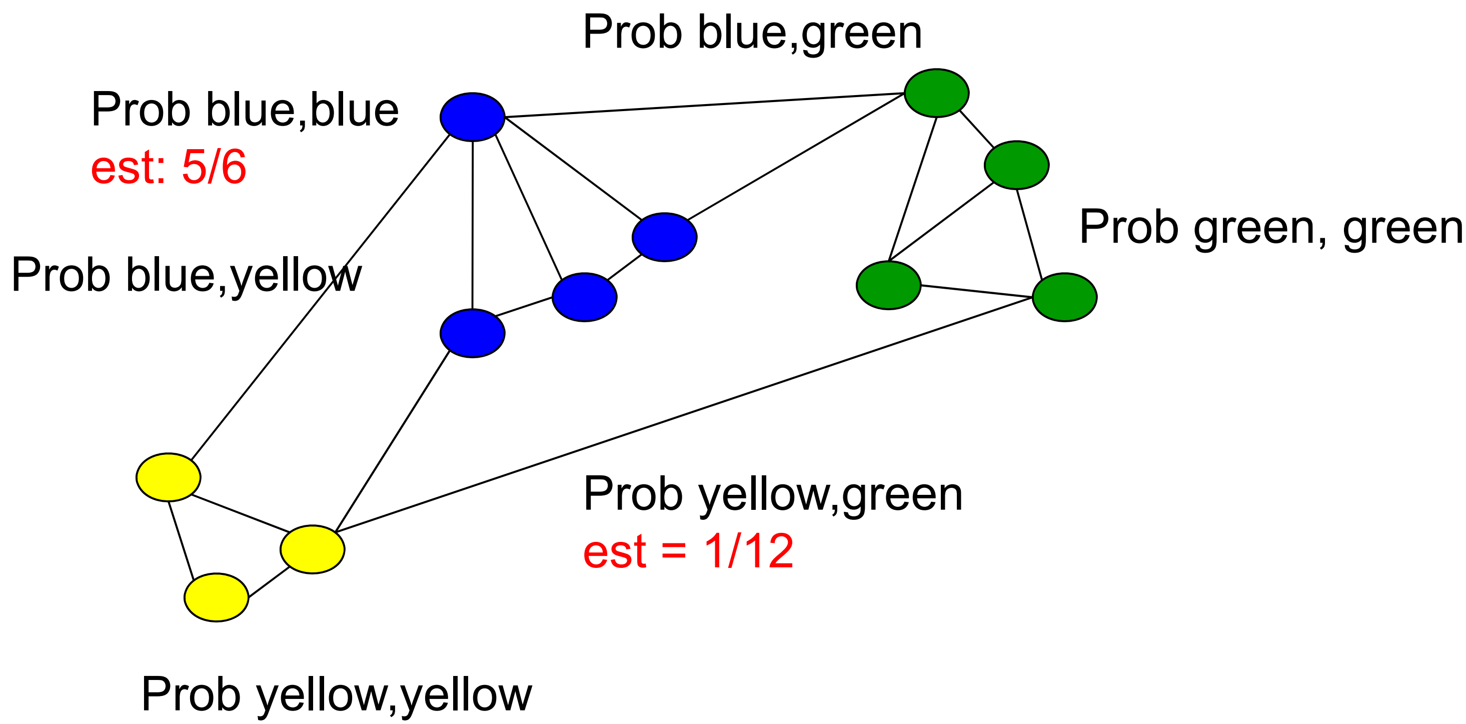
Networks with attributes



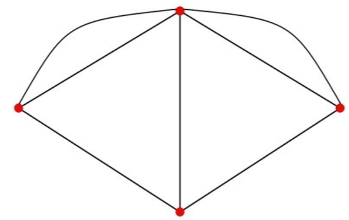
Networks with attributes



Networks with attributes



Block models



Continuous covariates:

Example: link between i and j depends on

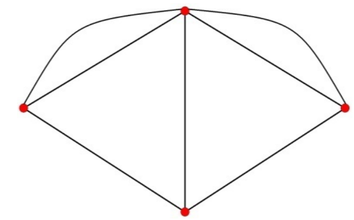
their characteristics:

$$\beta_i X_i + \beta_j X_j + \beta_{ij} |X_i - X_j|$$

E.g.,

$$\text{Log}(p_{ij} / (1-p_{ij})) = \beta_i X_i + \beta_j X_j + \beta_{ij} |X_i - X_j|$$

Block models

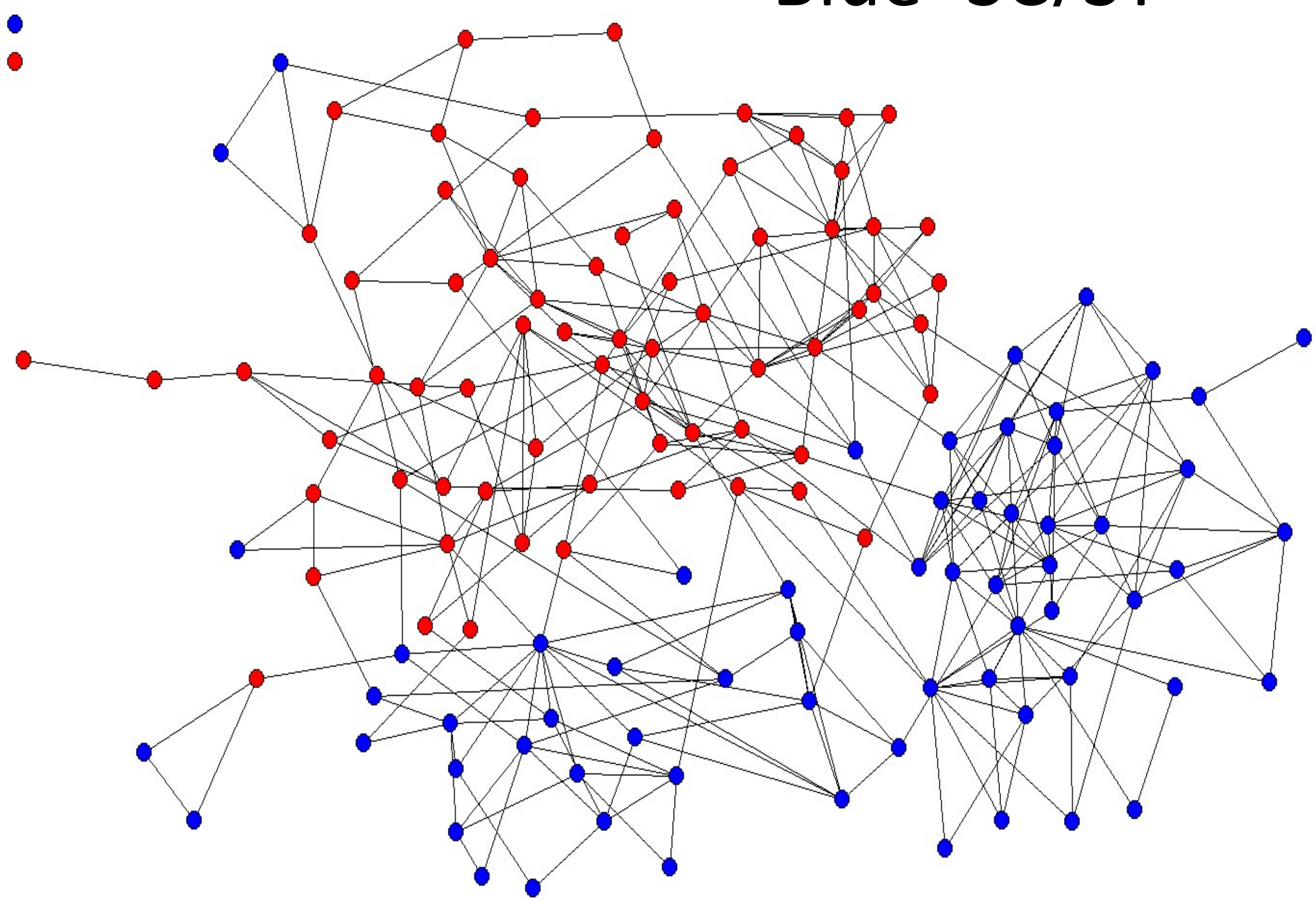


Could use this sort of model

to test for homophily...

Red=General/OBC BCDJ 2013

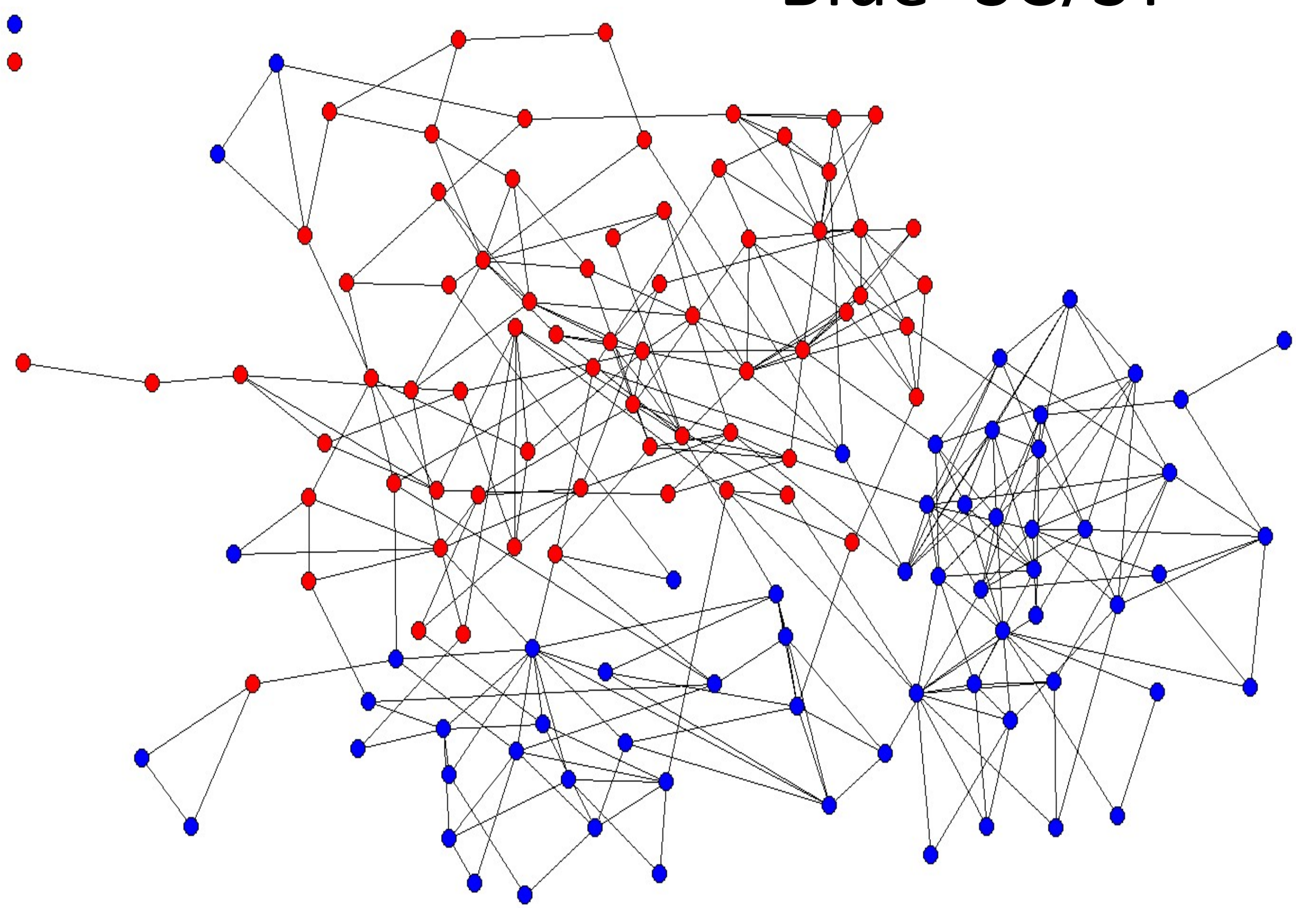
Blue=SC/ST V26 KeroRiceGo



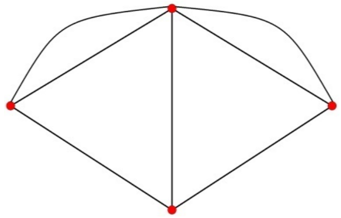
Pcross= .006
Pwithin=.089

Red=General/OBC BCDJ 2013

Blue=SC/ST V26 KeroRiceGo

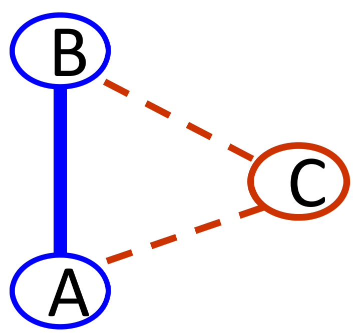


What is missed?

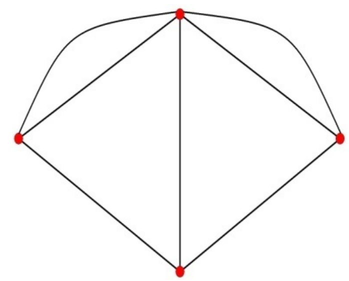


- Likelihood of link depends on node attributes (observed or latent)

- *also depends on whether nodes have friends in common*



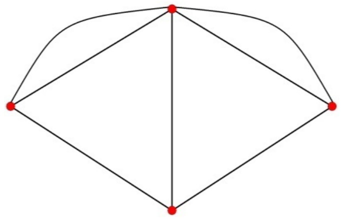
Social and Economic Networks: Models and Analysis



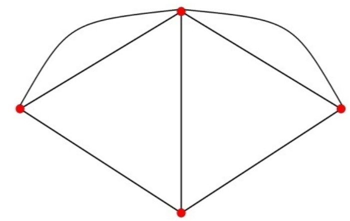
Matthew O. Jackson

Stanford University,
Santa Fe Institute, CIFAR,
www.stanford.edu\~jacksonm

3.7: ERGMs

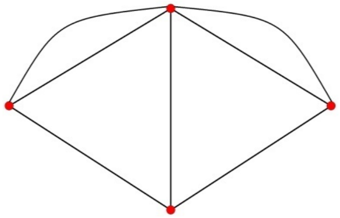


Random Network Models:



- Erdos-Renyi
  - Useful for understanding thresholds and how networks come to exhibit certain features
  - Miss many real-world features: e.g., clustering
- Other models link-by-link models
  - Watts and Strogatz, Barabasi and Albert, Jackson and Rogers....
  - Capture other features: clustering, degree distribution, correlation...
- Stochastic Block Models
  - Enrich Erdos-Renyi to allow for probabilities to depend on node characteristics, attributes (or on latent – unobserved characteristics)
- Popular set of models: ERGMs and new ones: SERGMs/SUGMs
  - flexible way to introduce various **local** features and dependencies
  - estimated statistically

Random Network Models:

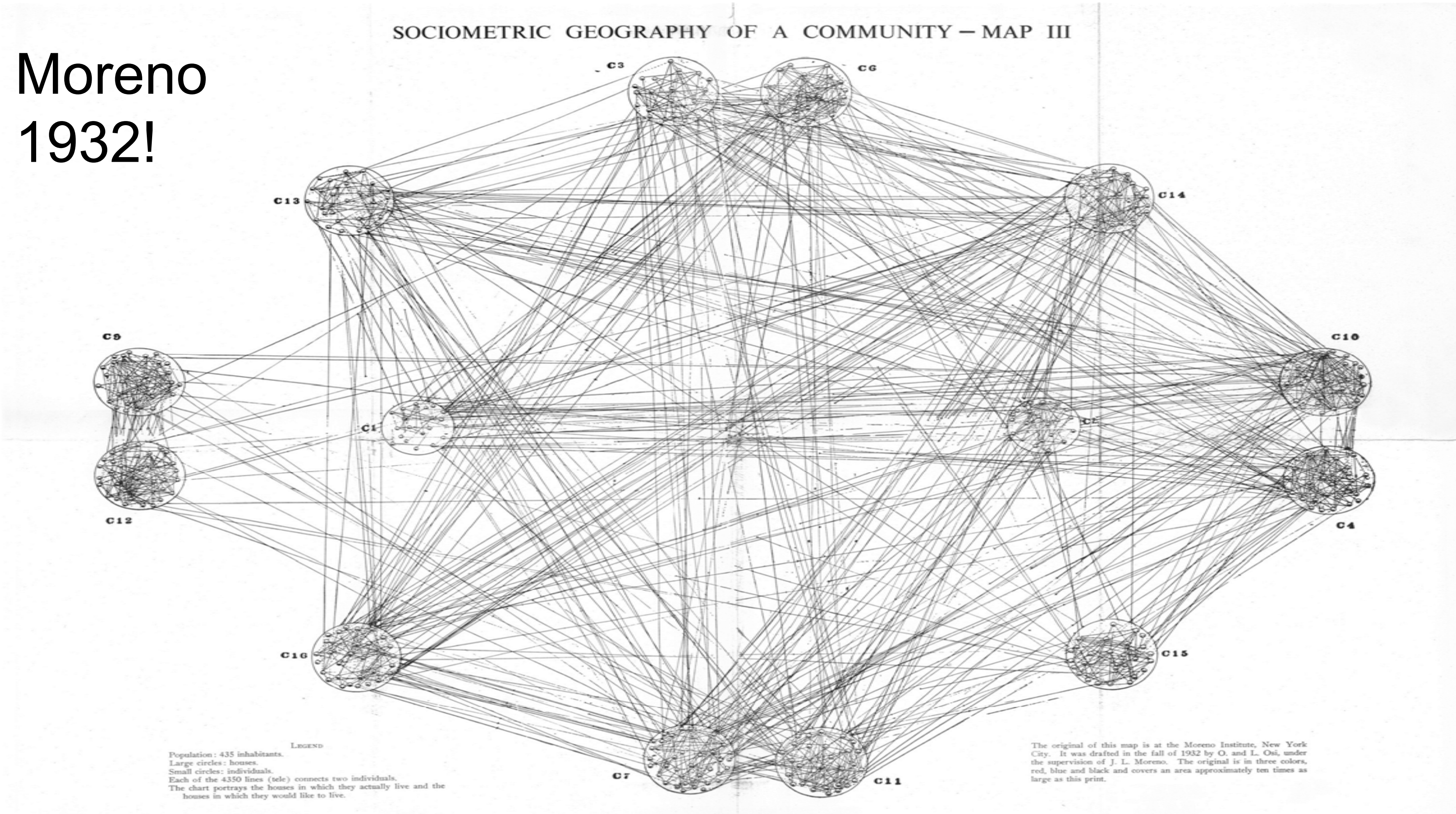


- Erdos-Renyi
  - Useful for understanding thresholds and how networks come to exhibit certain features
  - Miss many real-world features: e.g., clustering
- Other models link-by-link models
  - Watts and Strogatz, Barabasi and Albert, Jackson and Rogers....
  - Capture other features: clustering, degree distribution, correlation...
- Stochastic Block Models
  - Enrich Erdos-Renyi to allow for probabilities to depend on node characteristics, attributes (or on latent – unobserved characteristics)
- Popular set of models: ERGMs and new ones: SERGMs/SUGMs
  - flexible way to introduce various **local** features and dependencies
  - estimated statistically

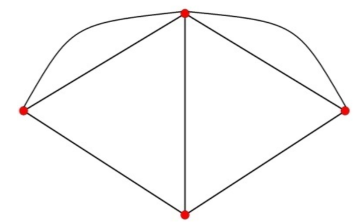
“A pertinent form of statistical treatment would be one which deals with social configurations as wholes, and not with single series of facts, more or less artificially separated from the total picture.”

Jacob Levy Moreno and Helen Hall Jennings, 1938.

Moreno
1932!


SOCIOMETRIC GEOGRAPHY OF A COMMUNITY – MAP III
C3
C6
C13
C14
C9
C10
C1
C2
C12
C4
C16
C15
C7
C11
LEGEND
Population: 435 inhabitants.
Large circles: houses.
Small circles: individuals.
Each of the 4350 lines (tele) connects two individuals.
The chart portrays the houses in which they actually live and the houses in which they would like to live.
The original of this map is at the Moreno Institute, New York City. It was drafted in the fall of 1932 by O. and L. Osi, under the supervision of J. L. Moreno. The original is in three colors, red, blue and black and covers an area approximately ten times as large as this print.

Markov, p* , ERGMs



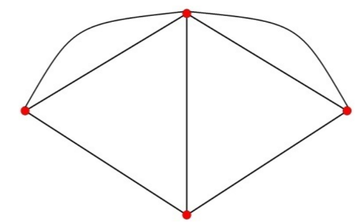
- Above models not sufficient for fitting data with clustering or other dependencies or *testing many social/economic theories*

- *Link ij's probability could depend on presence of jk and ik*

- But then things interlock and need to specify full interdependencies

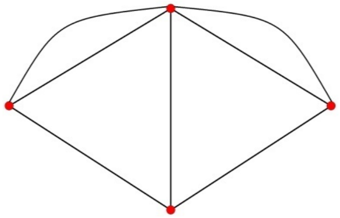
- Frank and Strauss (1986)) p* models (e.g., Wasserman and Pattison (1996)).

p* and ERG Models



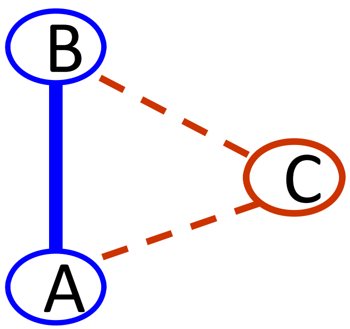
- **Example**: (studied extensively by Strauss (86), Park and Newman (04,05), Chatterjee, Diaconis (11)…)
  - Probability of a network depends on number of links
  - Probability of a network also depends on number of triangles.

Example

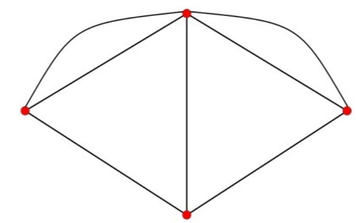


- Likelihood of link depends on node attributes

- *also depends on whether nodes have friends in common*



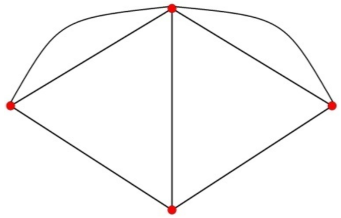
ERGMs



Example: probability depends on

$$\beta_L \ \#\mathrm{links}(g) + \beta_T \ \#\mathrm{triangles}(g)$$

ERGMs



Want probability *of network* to depend on

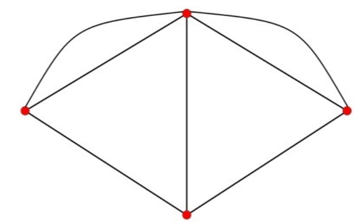
$$\beta_L\, L(g) + \beta_T\, T(g)$$

Set

$$\Pr(g) \sim \exp[\, \beta_L\, L(g) + \beta_T\, T(g) \,]$$

(now positive)

ERGMs



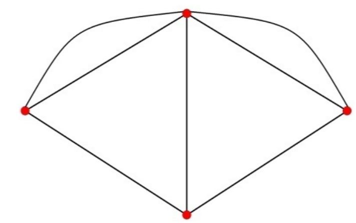
Want probability to depend on

$$\beta_L L(g) + \beta_T T(g)$$

Set $\Pr(g) \sim \exp[\, \beta_L L(g) + \beta_T T(g) \,]$

Theorem by Hammersly and Clifford (71): *any* network model can be expressed in the exponential family with counts of graph statistics

Example: Erdos-Renyi G(n,p)



- p – probability of a link, L(g) - number of links in g

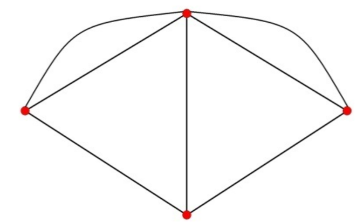
$$\Pr[(g)] = p^{L(g)}(1-p)^{n(n-1)/2-L(g)}$$

$$= [p/(1-p)]^{L(g)} (1-p)^{n(n-1)/2}$$

$$= \exp[\ \log(p/(1-p))\ L(g) - \log(1/(1-p))n(n-1)/2\ ]$$

$$= \exp[\qquad \beta_1 \qquad s_1(g) \quad - \quad c \quad]$$

ERGMs

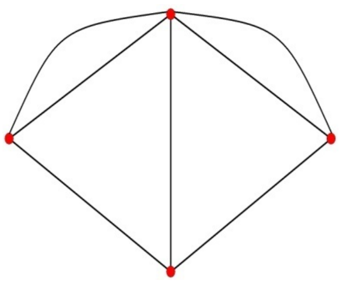


To be probability:

$$\Pr(g) = \frac{\exp[\ \beta_L\ L(g) + \beta_T\ T(g)\ ]}{\sum_{g'} \exp[\ \beta_L\ L(g') + \beta_T\ T(g')]}$$

$$\Pr(g) = \exp[\ \beta_L\ L(g) + \beta_T\ T(g) - c]$$

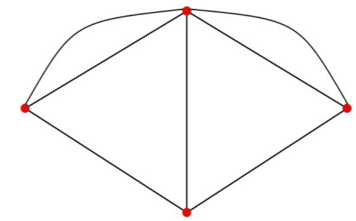
Social and Economic Networks: Models and Analysis



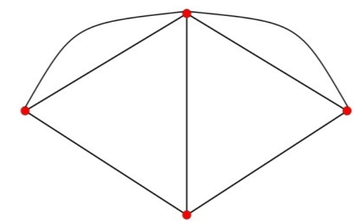
Matthew O. Jackson

Stanford University,
Santa Fe Institute, CIFAR,
www.stanford.edu\~jacksonm

3.8: Estimating ERGMs



Estimating p* / ERGMs

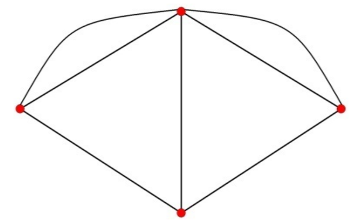


- $Pr(g) = \exp[\sum \beta_k s_k(g)] / \sum_{g'} \exp[\sum \beta_k s_k(g')]$

- Power of such models: can put all sorts of statistics in s_k - can have it depend on arbitrary shapes, be specific to certain nodes/links, etc.

- Weakness: how to estimate these?!

Example: Florentine Marriages

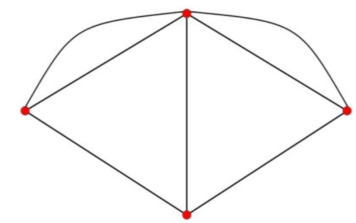


- Robins, Pattison, Kalish, Lusher (2007) fit an ERGM to Padgett and Ansel's Florentine data

- Business ties between the 16 major families

- Fit: #links, two stars, three stars, triangles

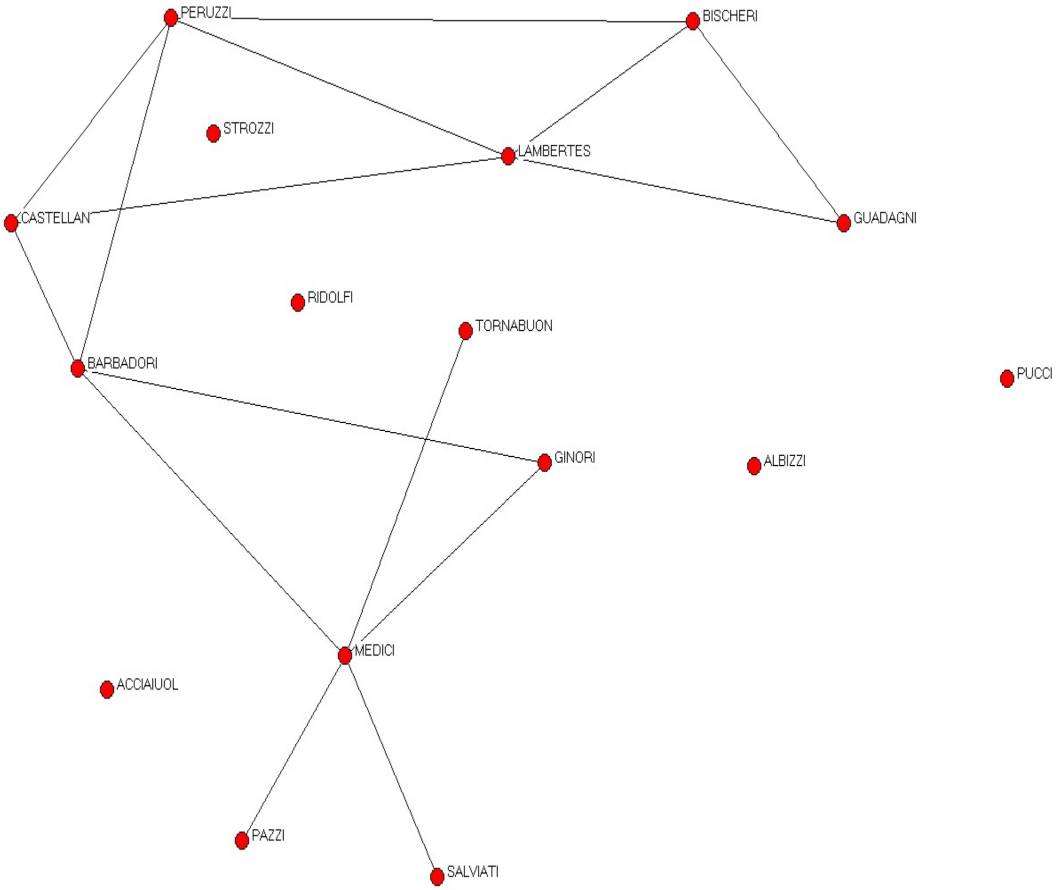
Example: Florentine Marriages



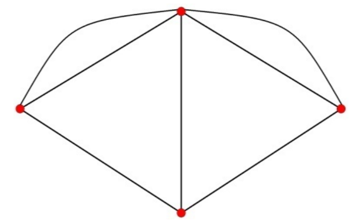
$$\Pr(g) = \exp[\ \beta_1 \text{ \#links} + \beta_2 \text{ \#two stars}$$

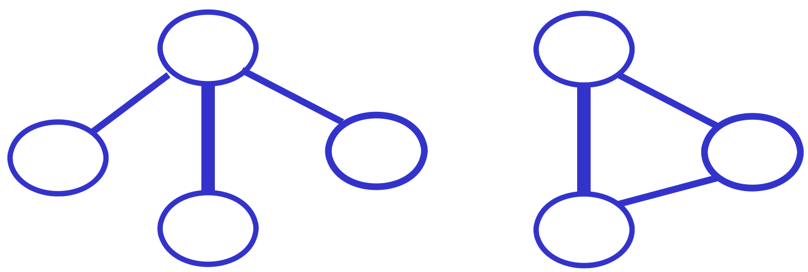
$$+ \beta_3 \text{ \#three stars} + \beta_4 \text{ \#triangles} - c\ ]$$

Business Only



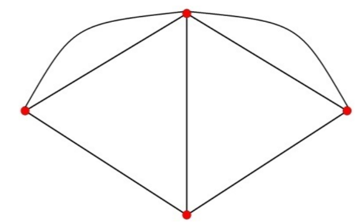
Example: Florentine Families Business Dealings





-4.27	1.09	-0.67	1.32
(1.13)	(0.65)	(0.41)	(0.65)

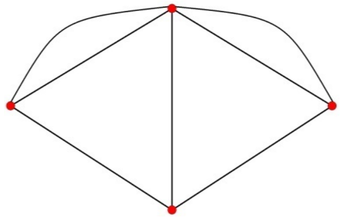
ERGMs



$$\Pr(g) = \frac{\exp[\,\beta_1 s_1(g) + \ldots + \beta_k s_k(g)\,]}{\sum_{g'} \exp[\,\beta_1 s_1(g') + \ldots + \beta_k s_k(g')\,]}$$

- MCMC techniques for estimation (Snijders 02, Handcock 03,...) have led to these becoming the standard

Issues:



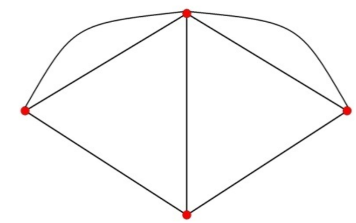
- $\Pr(g) = \dfrac{\exp[\ \beta_1\ s_1(g) + \ldots + \beta_k\ s_k(g)\ ]}{\sum_{g'}\ \exp[\ \beta_1\ s_1(g') + \ldots + \beta_k\ s_k(g')\ ]}$

- Recall: n=30 nodes, there are 2^{435} g's

(less than 2^{258} atoms in the universe...)

- ***Sampling g's will not lead to accurate estimates (not just MCMC limitation)***

ERGMs

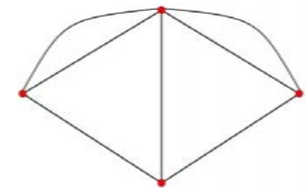


Bhamidi, Bresler and Sly (2008)
(see also Chatterjee and Diaconis (2011)):

For dense enough ERGMs, MCMC (Glauber dynamics - Gibbs sampling) estimates mix less than exponentially ***only if*** networks have approximately independent links

So, ERGMs that are interesting, cannot be estimated via techniques being used!

Simulations: also problems on sparse ones...

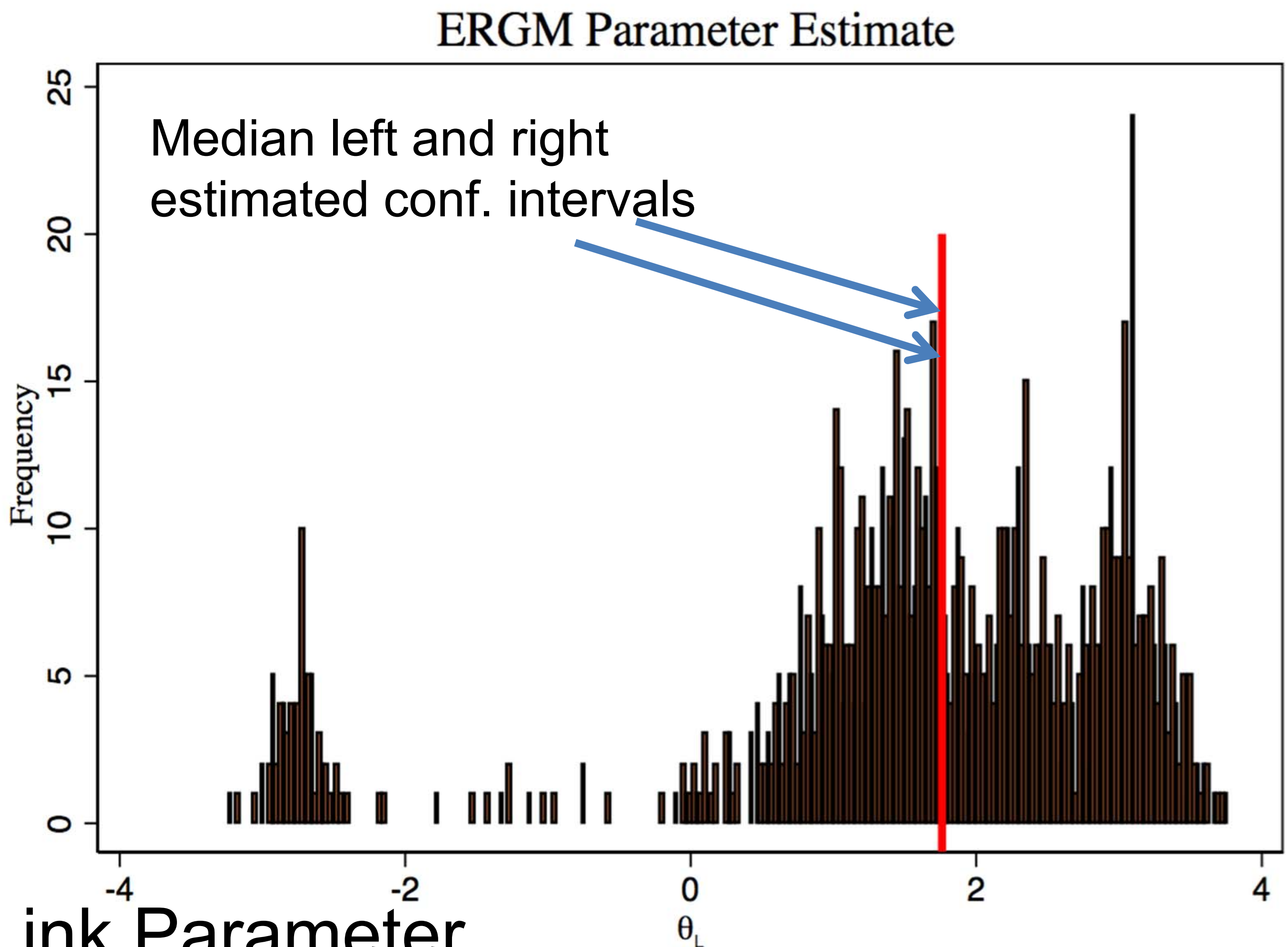


Example:

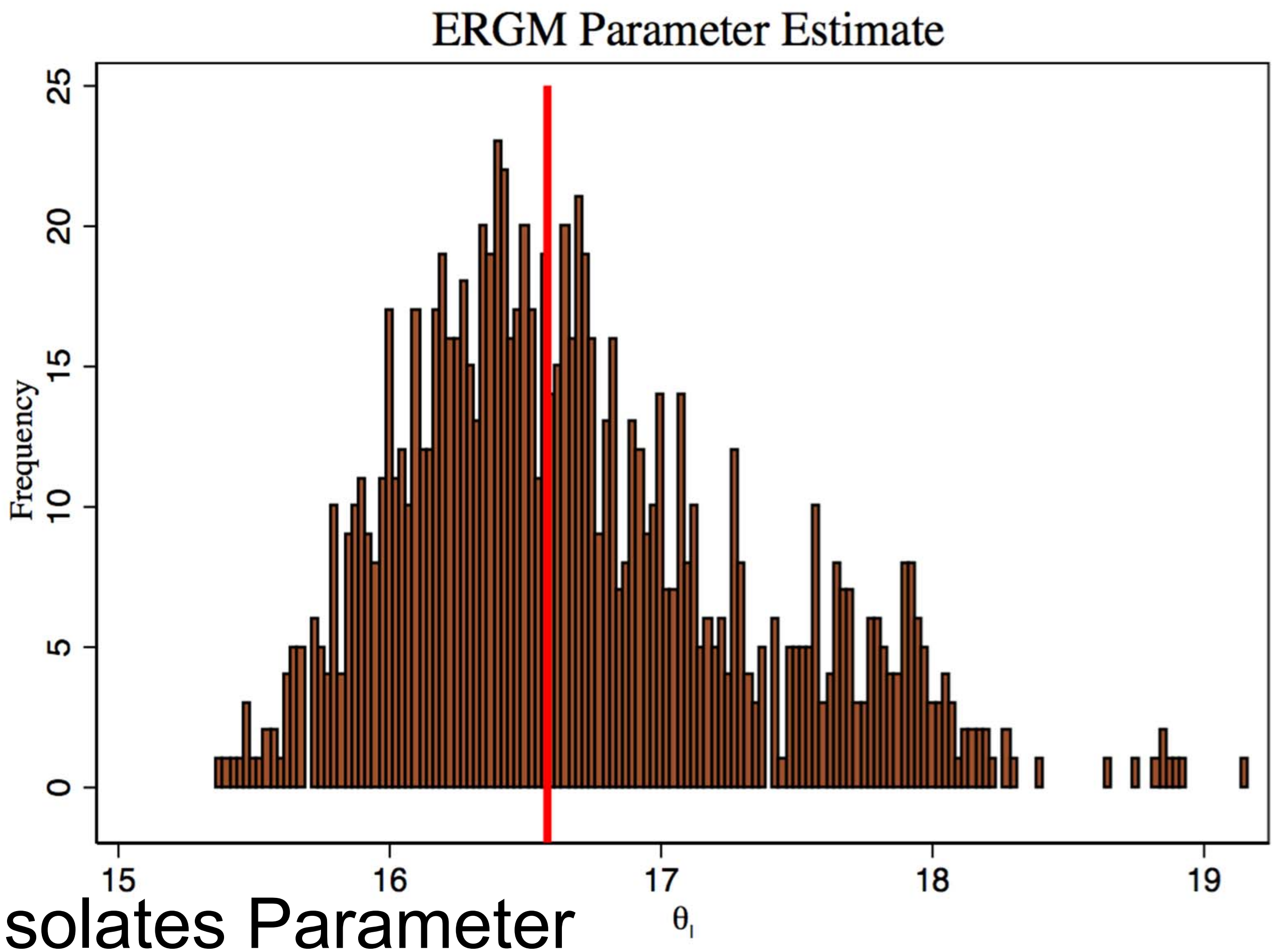
- Pr(g) = $$\frac{\exp[\beta_I\, I(g) + \beta_L\, L(g) + \beta_T\, T(g)\,]}{\sum_{g'} \exp[\beta_I\, I(g) + \beta_L\, L(g') + \beta_T\, T(g')\,]}$$

- I(g) = #isolates(g)
- L(g) = #links(g)
- T(g) = #triangles(g)

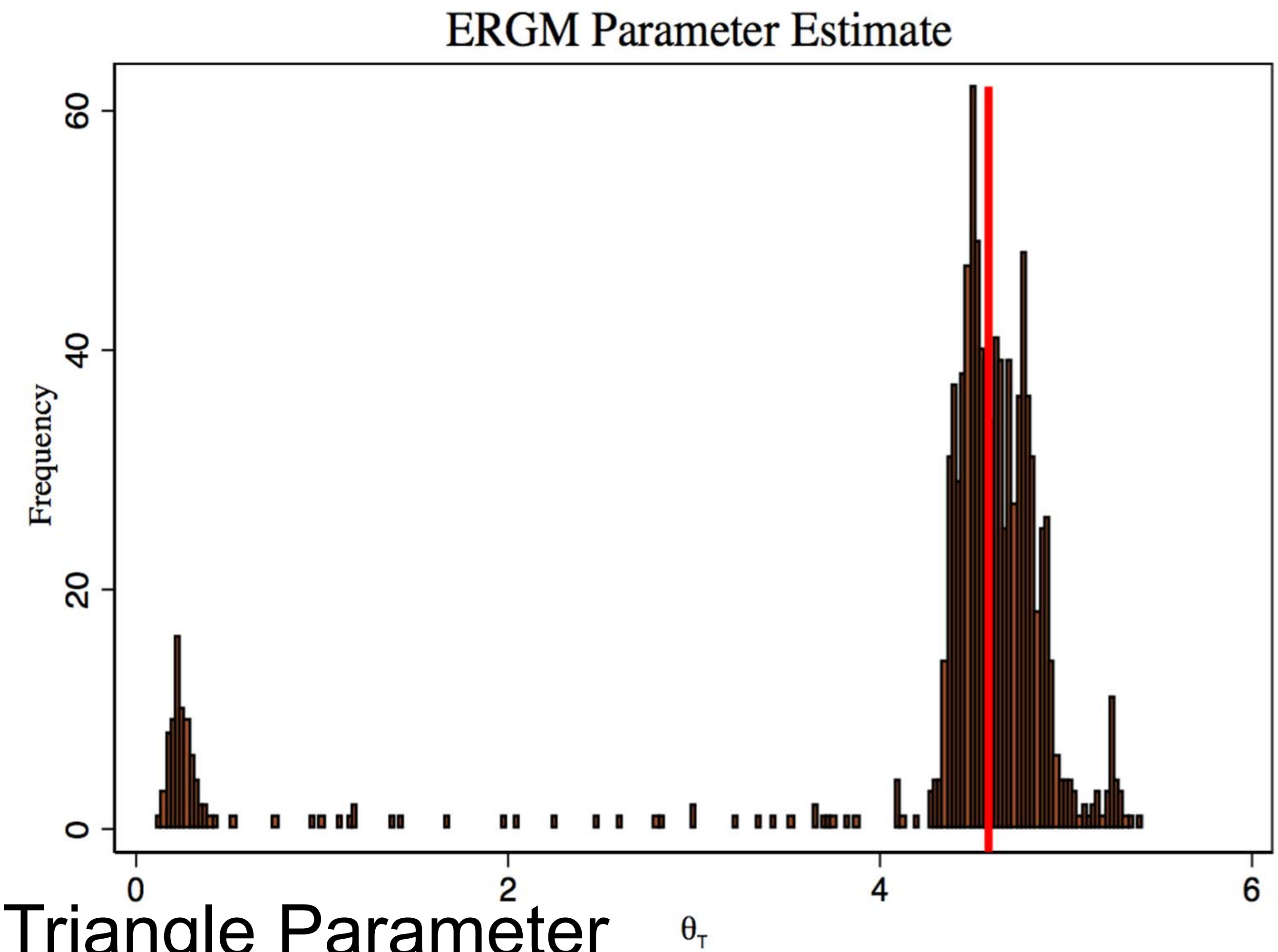
- n=50 nodes, 1000 iterations
- avg: 20 isol (so each isolated with prob .4), 10 triangles, 45 links



Link Parameter

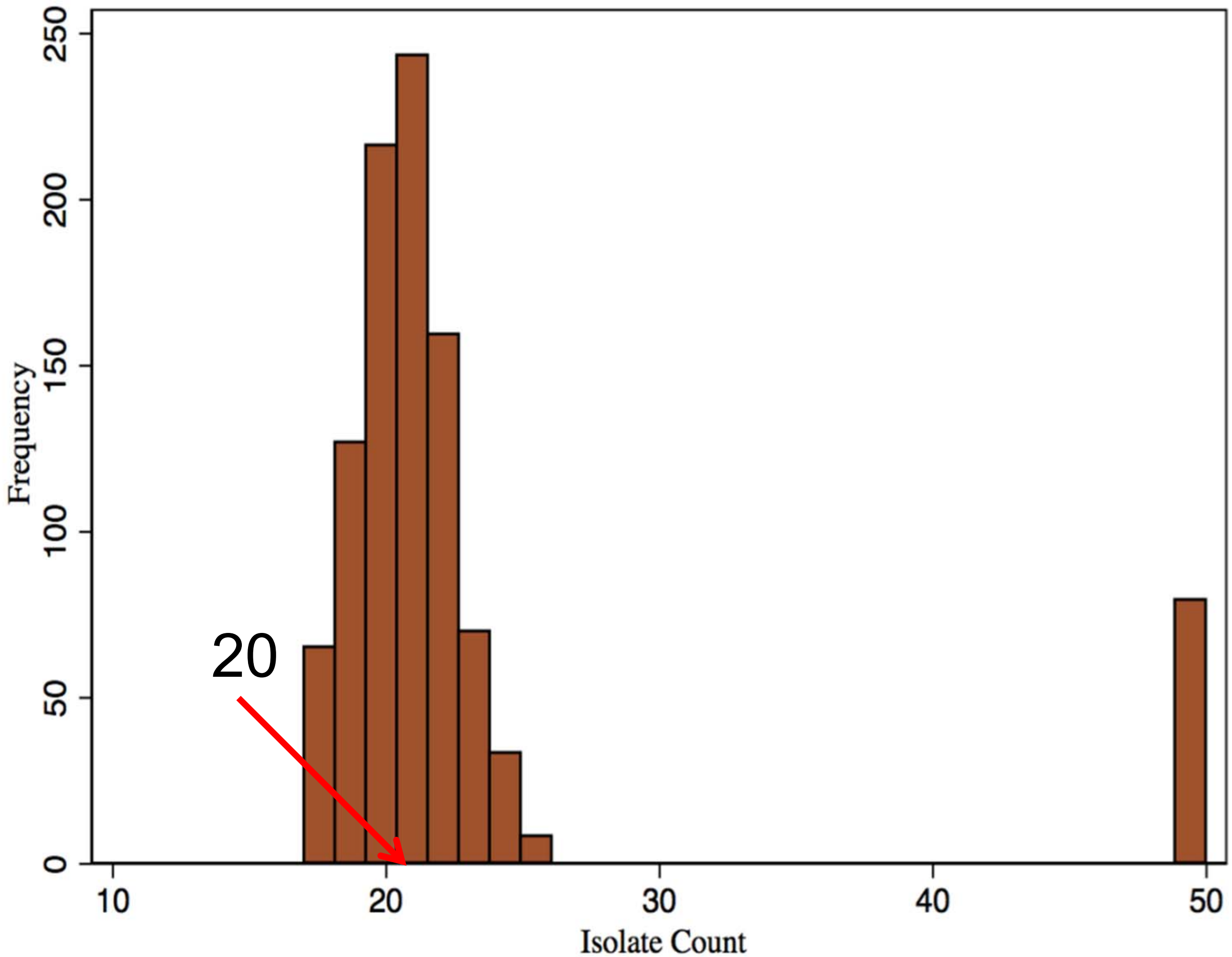


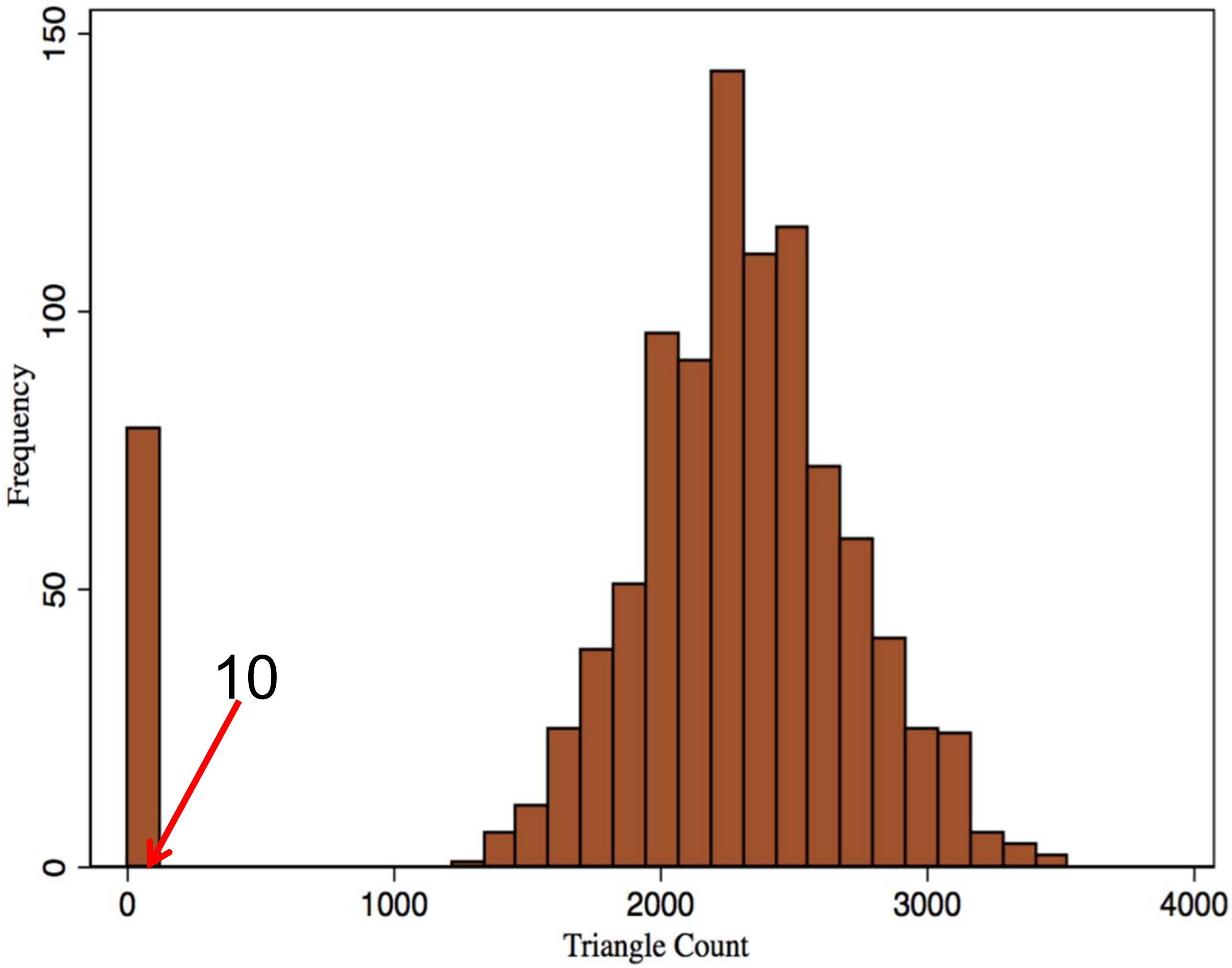
Isolates Parameter



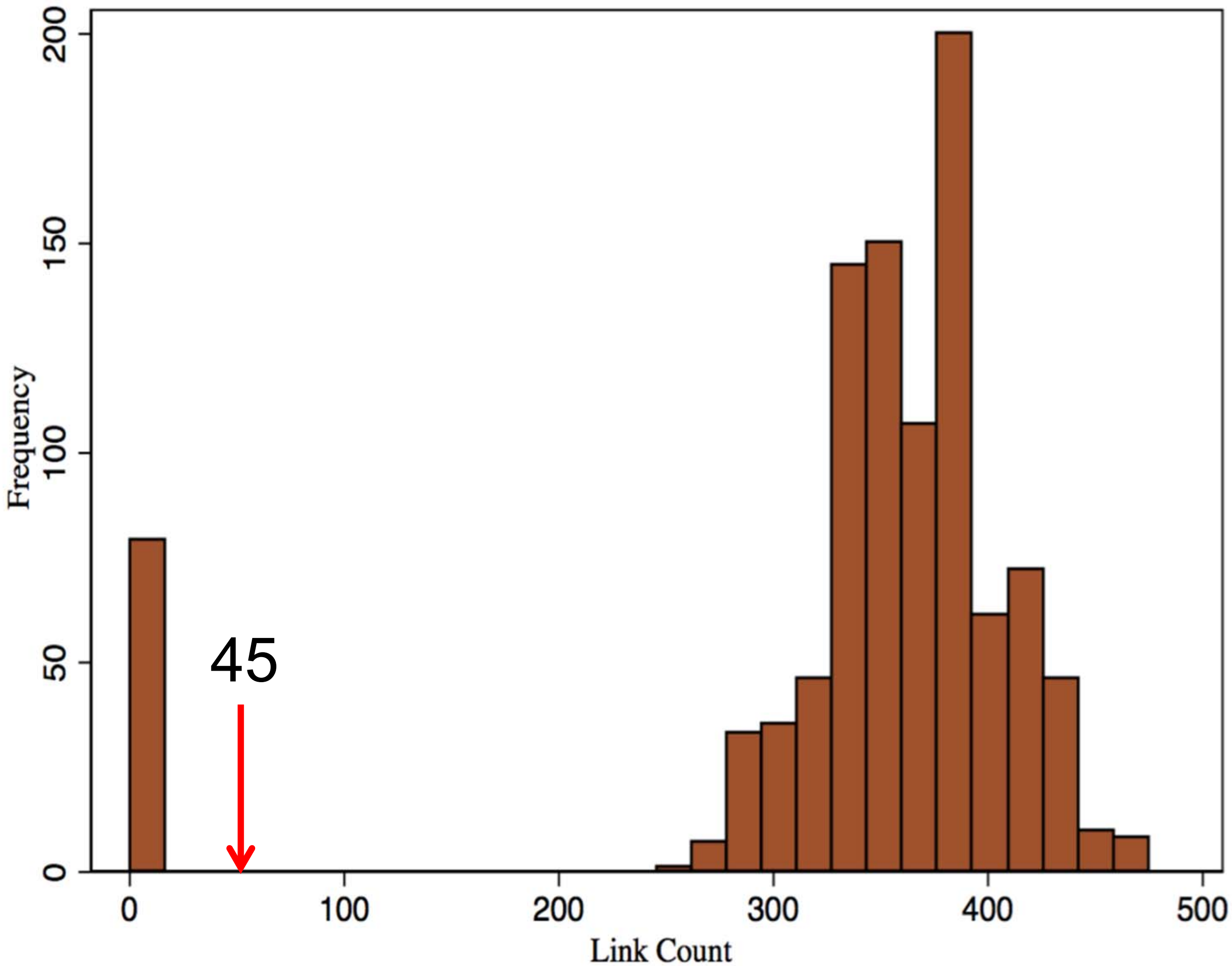
Triangle Parameter

Recreate Isolates



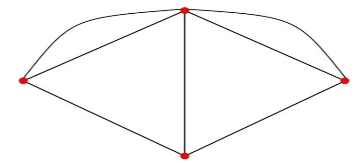


Recreate Triangles



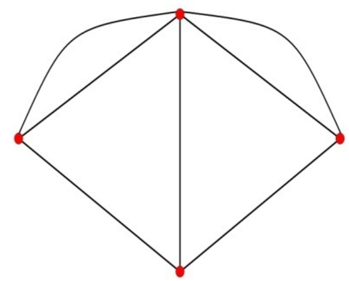
Recreate Links

Issues:



- MCMC estimation techniques are inaccurate:
  - Can one compute parameters?
- Consistency of estimators of ERGMs:
  - When are parameters accurate and how many nodes are needed?
- How to generate networks randomly?
  - Counterfactuals, validation…

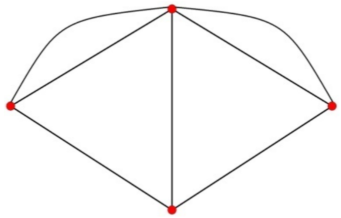
Social and Economic Networks: Models and Analysis



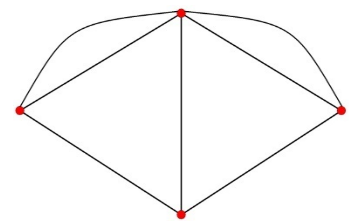
Matthew O. Jackson

Stanford University,
Santa Fe Institute, CIFAR,
www.stanford.edu\~jacksonm

3.9: SERGMs

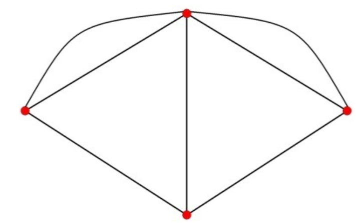


Random Network Models:



- Erdos-Renyi
  - Useful for understanding thresholds and how networks come to exhibit certain features
  - Miss many real-world features: e.g., clustering
- Other models link-by-link models
  - Watts and Strogatz, Barabasi and Albert, Jackson and Rogers....
  - Capture other features: clustering, degree distribution, correlation...
- Stochastic Block Models
  - Enrich Erdos-Renyi to allow for probabilities to depend on node characteristics, attributes (or on latent – unobserved characteristics)
- Popular set of models: ERGMs and new ones: SERGMs/SUGMs
  - flexible way to introduce various **local** features and dependencies
  - estimated statistically

SERGMs: Introduction

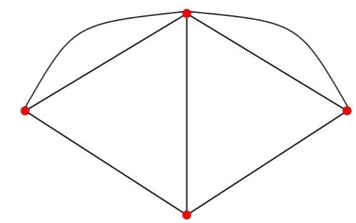


- ERGMs not accurately estimable in many cases....

- Too many alternative networks to consider...

- Way out: Many networks lead to the same statistics
  - Probabilities only depend on statistics
  - So, networks with same statistics are ``equivalent'' (equally likely)
- Collapse all equivalent networks

SERGMs: Chandrasekhar-Jackson 2012

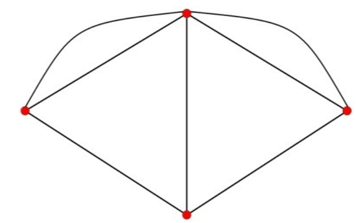


- Many g's   but many fewer possible statistics

- Many networks lead to the same statistics
  - Probabilities only depend on statistics
  - Thus, networks with same statistics are ``equivalent'' (equally likely)

- Collapse all equivalent networks

Sufficient Statistics

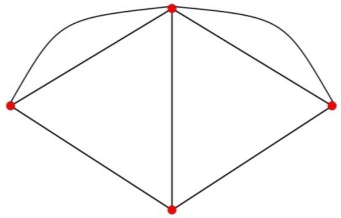


- $\Pr(g) = \dfrac{\exp[\ \beta\ S(g)\ ]}{\sum_{g'} \exp[\ \beta\ S(g')\ ]}$

Let $N(s')$ = number

networks with $S(g')=s'$

Sufficient Statistics



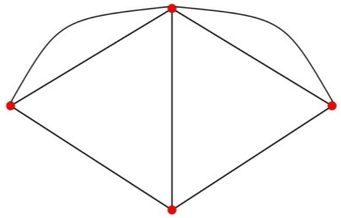
- $\Pr(g) = \dfrac{\exp[\ \beta\ S(g)\ ]}{\sum_{g'}\ \exp[\ \beta\ S(g')\ ]}$

  Let $N(s')$ = number networks

  s.t. $S(g')=s'$

- $\Pr(g) = \dfrac{\exp[\ \beta\ S(g)\ ]}{\sum_{s'} N(s')\ \exp[\ \beta\ s']}$

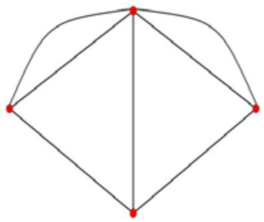
Statistical Form



- $\Pr(g) = \dfrac{\exp[\,\beta\, S(g)\,]}{\sum_{g'} \exp[\,\beta\, S(g')\,]}$

- $\Pr(s) = \dfrac{N(s)\ \exp[\,\beta\, s\,]}{\sum_{s'} N(s')\exp[\,\beta\, s'\,]}$

SERGMs – Chandrasekhar-Jackson (12)

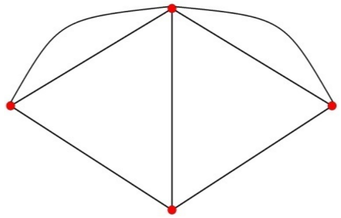


- $$\Pr(g) = \frac{\exp[\ \beta\ S(g)\ ]}{\sum_{g'} \exp[\ \beta\ S(g')\ ]}$$

- $$\Pr(s) = \frac{N(s)\ \exp[\ \beta\ s\ ]}{\sum_{s'} N(s')\ \exp[\ \beta\ s'\ ]}$$

Smaller space...

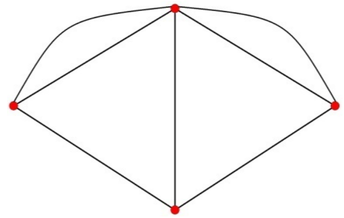
Statistical Form



Instead of asking what is the probability of a specific network:

Ask: What is the probability of observing a network that has density of links .1, clustering .3, and average path length of 2.7, etc.?

Statistical ERGMs: SERGMs

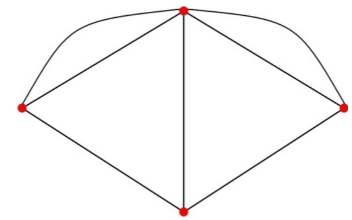


- $$\Pr(s) = \frac{N(s)\ \exp[\,\beta\, s\,]}{\sum_{s'} N(s')\ \exp[\,\beta\, s'\,]}$$

Why not some K(s) instead??

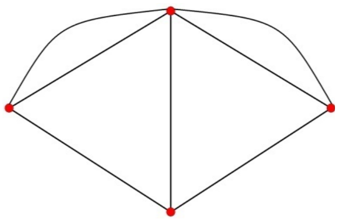
- $$\Pr(s) = \frac{K(s)\ \exp[\,\beta\, s\,]}{\sum_{s'} K(s')\ \exp[\,\beta\, s'\,]}$$

Emphasize: Idea Here



- *Many* g's, *many* fewer possible statistics
- Networks with same statistics equally likely
- Model based on *statistics* directly: general family of SERGMs that nest ERGMs
- $$Pr(s) = \frac{K(s)\ \exp[\ \beta\ s\ ]}{\sum_{s'} K(s')\ \exp[\ \beta\ s']}$$

SERGMs Include:



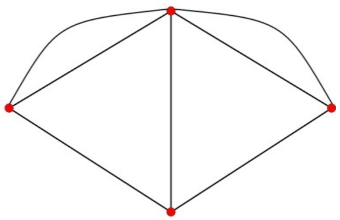
- $\Pr(s) = \dfrac{K(s)\ \exp[\ \beta\ s\ ]}{\sum_{s'} K(s')\ \exp[\ \beta\ s']}$

SERGMs

S can encode many things:

- Links, cliques, k-stars, subgraphs, friends in common per link, multi-graphs, adapt for degree distributions
- Can do preference based-models
- Allow for node characteristics…

Challenge:

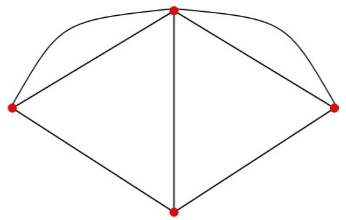


- ``*One*'' data point: often observe a single network

- But many observations of which links are present, which triangles, etc.

- These are *not independent* observations: do they still provide enough information?

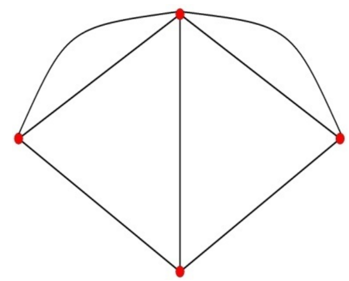
SERGM Estimation



- Chandrasekhar and Jackson (2012) provide results on
  - Classes of SERGMs for which maximum likelihood converge to true parameters as n grows
  - Simple ways of estimating those

- Look at a related set of models

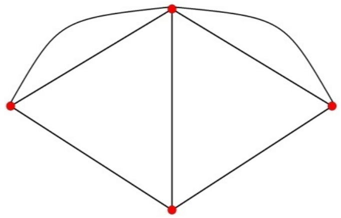
Social and Economic Networks: Models and Analysis



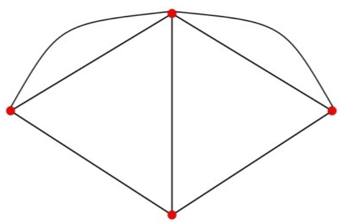
Matthew O. Jackson

Stanford University,
Santa Fe Institute, CIFAR,
www.stanford.edu\~jacksonm

3.10: SUGMs

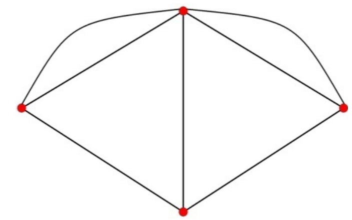


Random Network Models:



- Erdos-Renyi
  - Useful for understanding thresholds and how networks come to exhibit certain features
  - Miss many real-world features: e.g., clustering
- Other models link-by-link models
  - Watts and Strogatz, Barabasi and Albert, Jackson and Rogers....
  - Capture other features: clustering, degree distribution, correlation...
- Stochastic Block Models
  - Enrich Erdos-Renyi to allow for probabilities to depend on node characteristics, attributes (or on latent – unobserved characteristics)
- Popular set of models: ERGMs and new ones: SERGMs/SUGMs
  - flexible way to introduce various **local** features and dependencies
  - estimated statistically

SUGMs



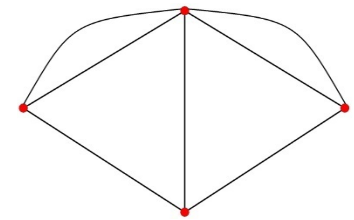
- Subgraph Generation Models

- *Subgraphs* are generated, network is

by-product

people form links, triangles,
some are anti-social (isolates),…

SUGMs

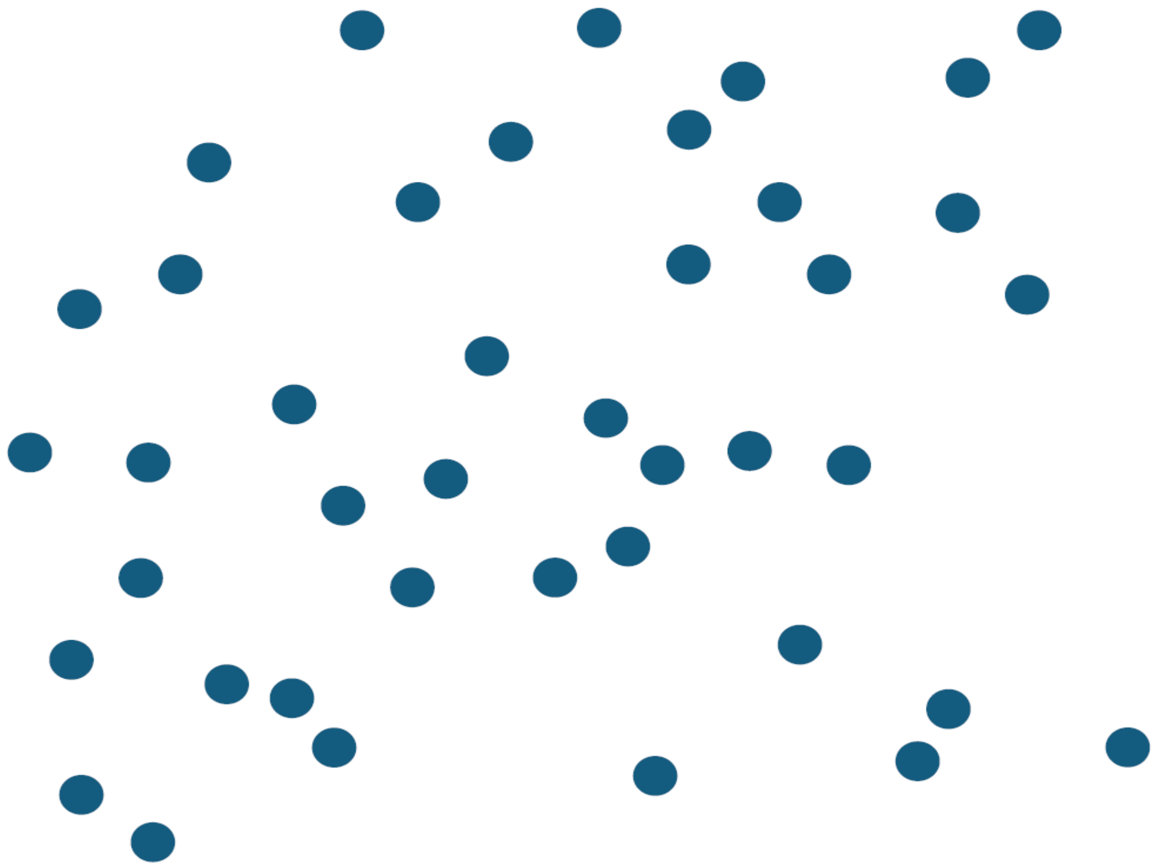


- Nature/people form S_j subnetworks of type j each indepedently with probability p_j

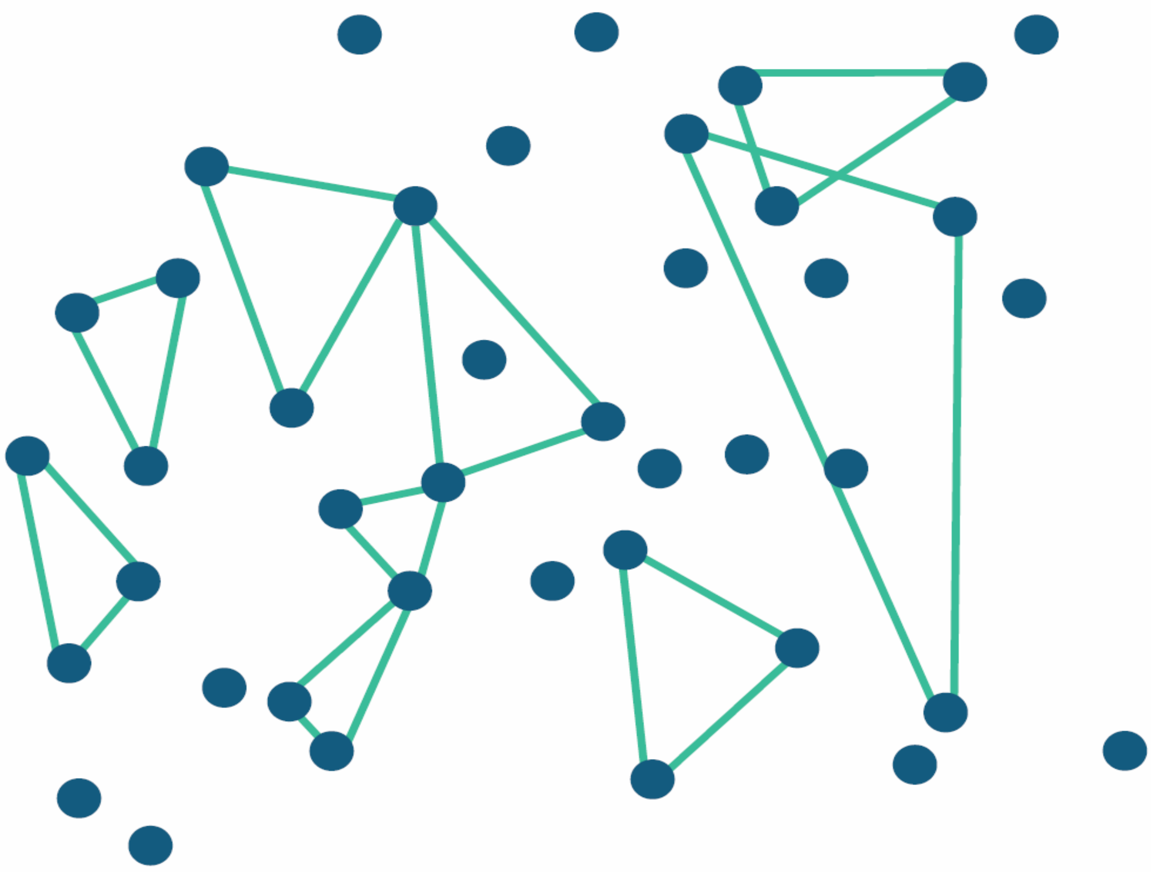
- May intersect and overlap

- We observe resulting network, infer the p_j's

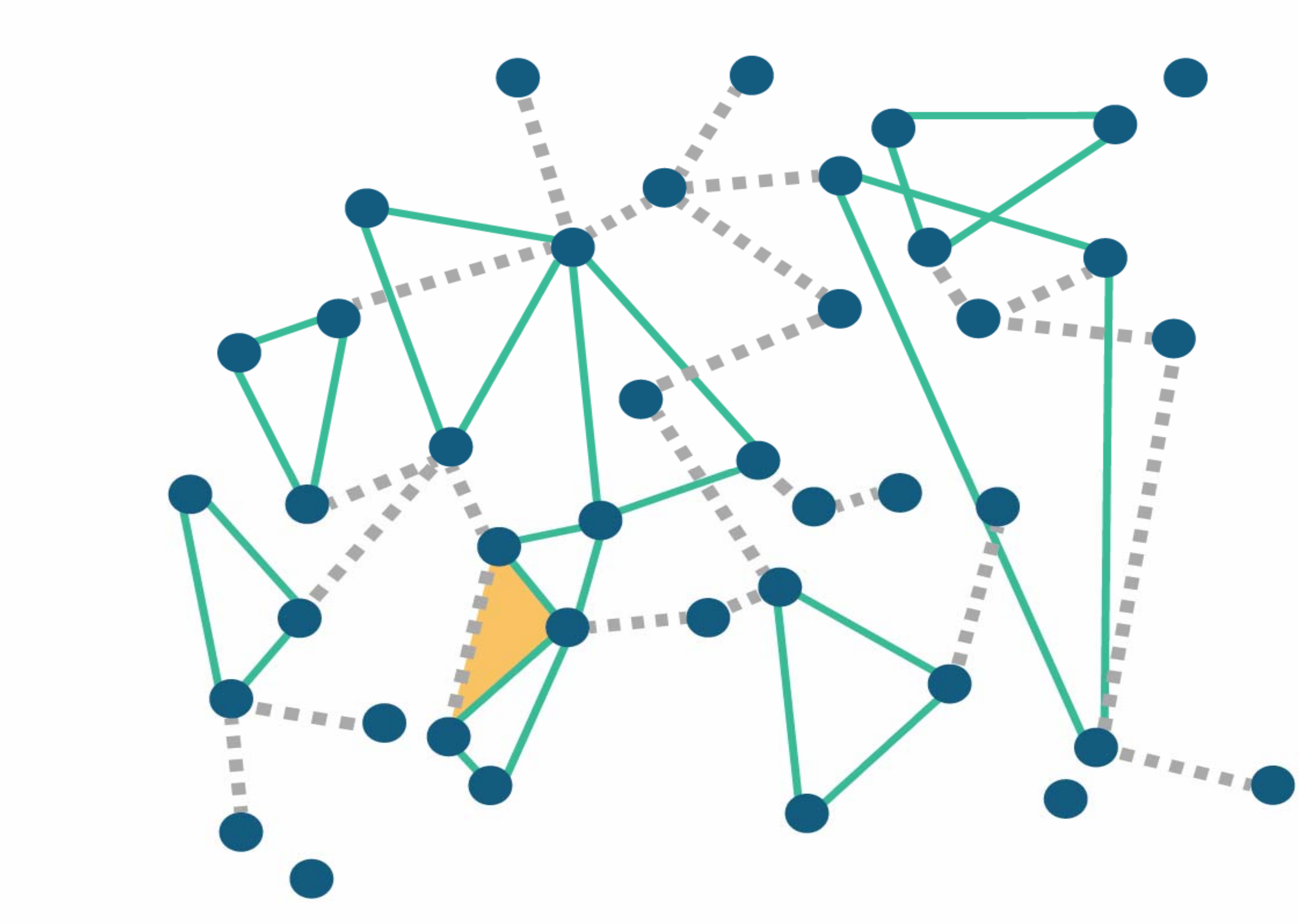
Example:



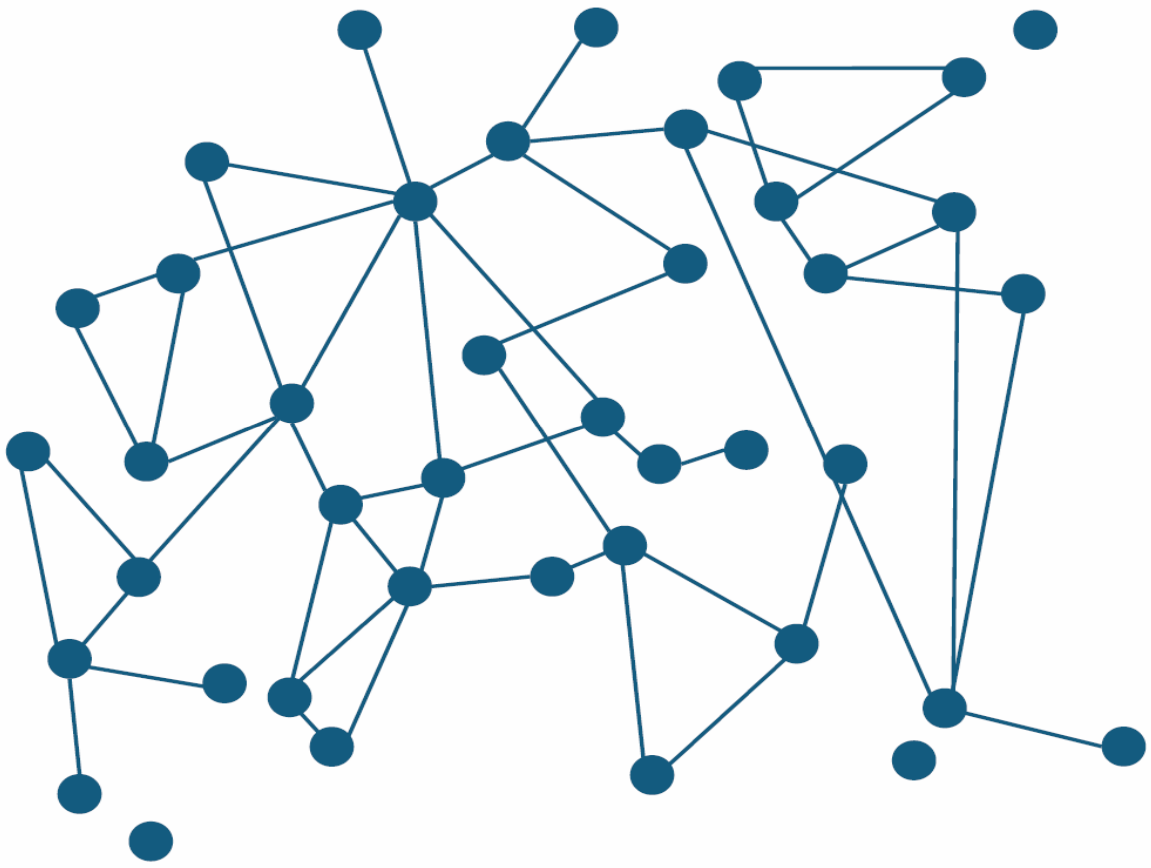
Example:



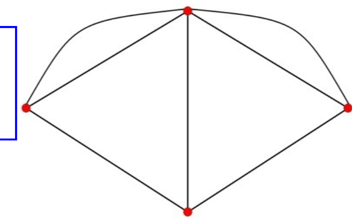
Links Form: Incidental Triangle



We See:

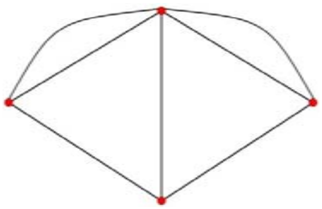


Relation: SERGMs/SUGMs



- Can we view SUGMs as SERGMs?
- Yes, and it motivates specific K's

Theorem: SUGMs and SERGMs

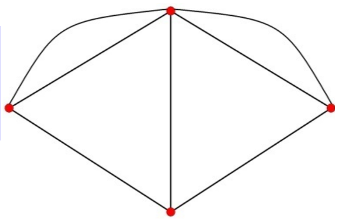


Consider a SUGM with parameters p_j and let S be the *true* counts of subgraphs.

$$\text{Then } \Pr(S) = \frac{K(S)\ \exp[\,\beta S\,]}{\sum_{s'} K(s')\ \exp[\,\beta s'\,]}$$

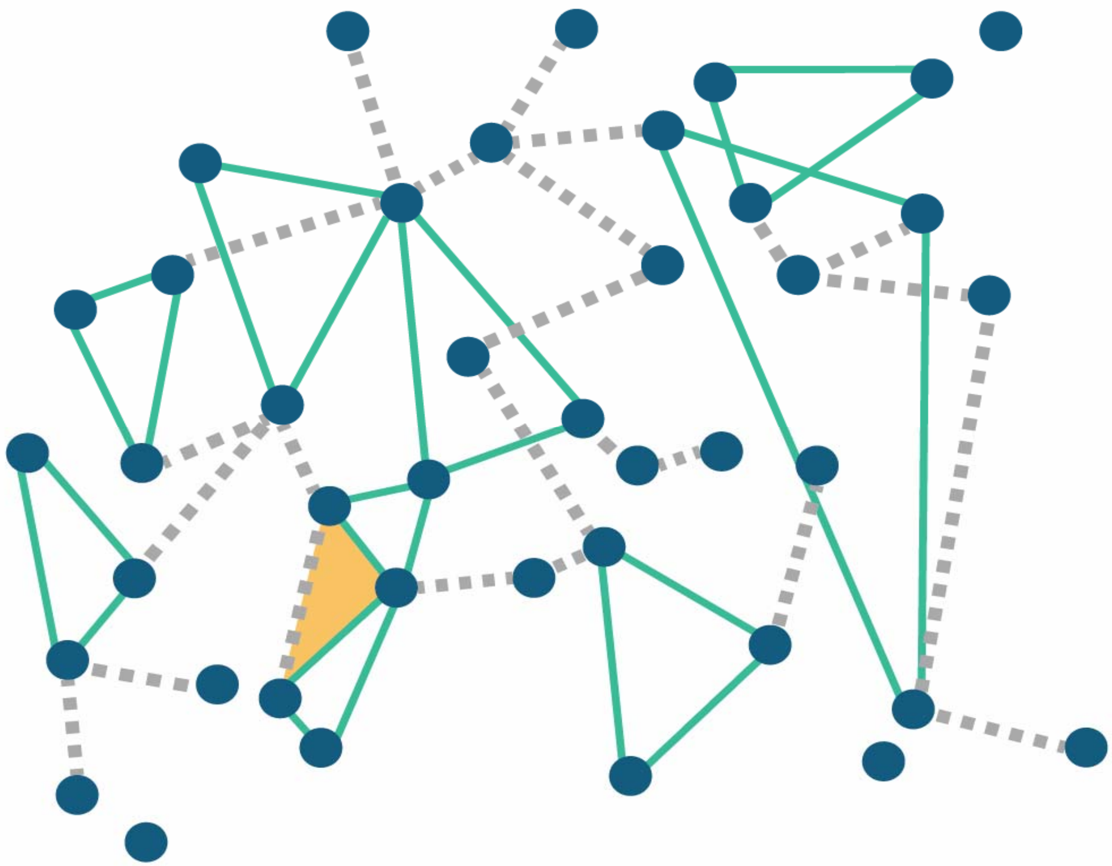
$\beta_j = \log(p_j/(1-p_j))$ and $K^n(s) = \prod_j \binom{\bar{S}_j^n}{s_j}$.

Relation: SERGMs/SUGMs

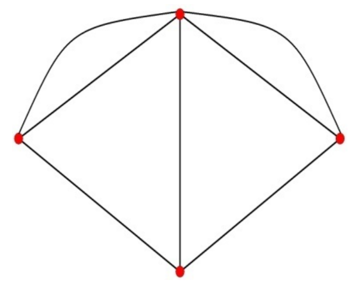


- True counts not observed
- However, can observe them when sparse!
- So, sparse SUGMs define easily estimated class of SERGMs…

Sparse: Few Incidental Triangles



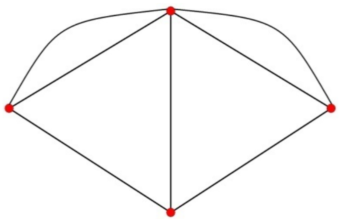
Social and Economic Networks: Models and Analysis



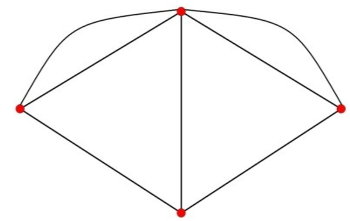
Matthew O. Jackson

Stanford University,
Santa Fe Institute, CIFAR,
www.stanford.edu\~jacksonm

3.11:Estimating SUGMs

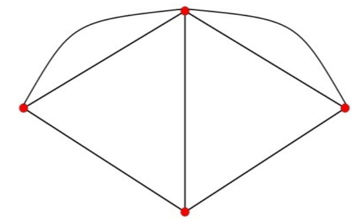


Random Network Models:



- Erdos-Renyi
  - Useful for understanding thresholds and how networks come to exhibit certain features
  - Miss many real-world features: e.g., clustering
- Other models link-by-link models
  - Watts and Strogatz, Barabasi and Albert, Jackson and Rogers....
  - Capture other features: clustering, degree distribution, correlation...
- Stochastic Block Models
  - Enrich Erdos-Renyi to allow for probabilities to depend on node characteristics, attributes (or on latent – unobserved characteristics)
- Popular set of models: ERGMs and new ones: SERGMs/SUGMs
  - flexible way to introduce various **local** features and dependencies
  - estimated statistically

SUGMs

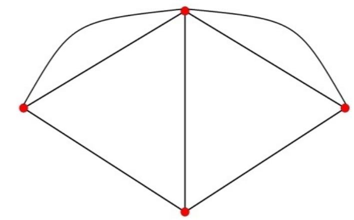


- Nature/people form S_j subnetworks of type j each independently with probability p_j

- May intersect and overlap

- We observe resulting network, infer the p_j's

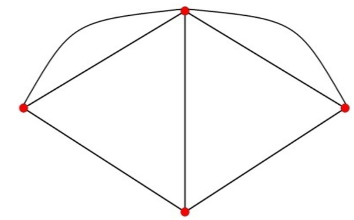
Estimation – Two Approaches



- Sparse graphs: rare incidentals, Direct estimation is valid/consistent

- Algorithm: corrects for small n, and provides estimates for non-sparse (see CJ paper…)

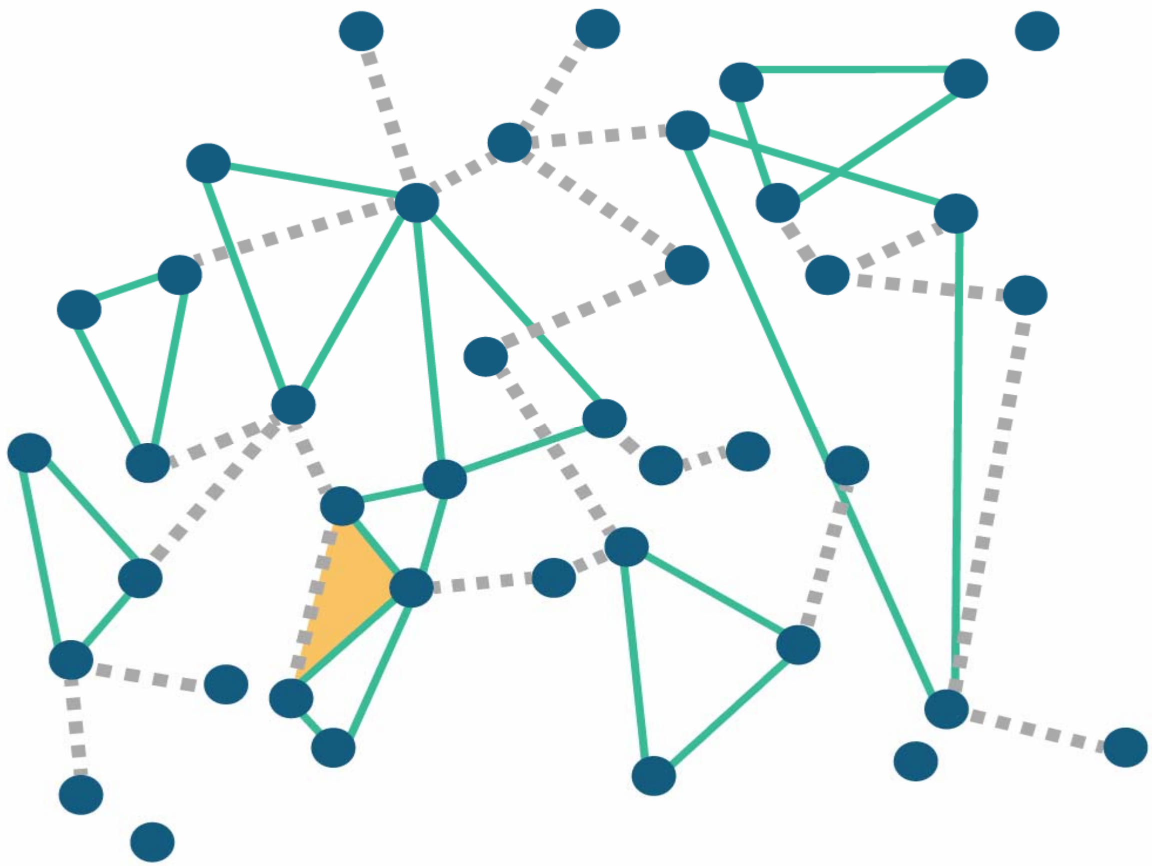
Sparseness



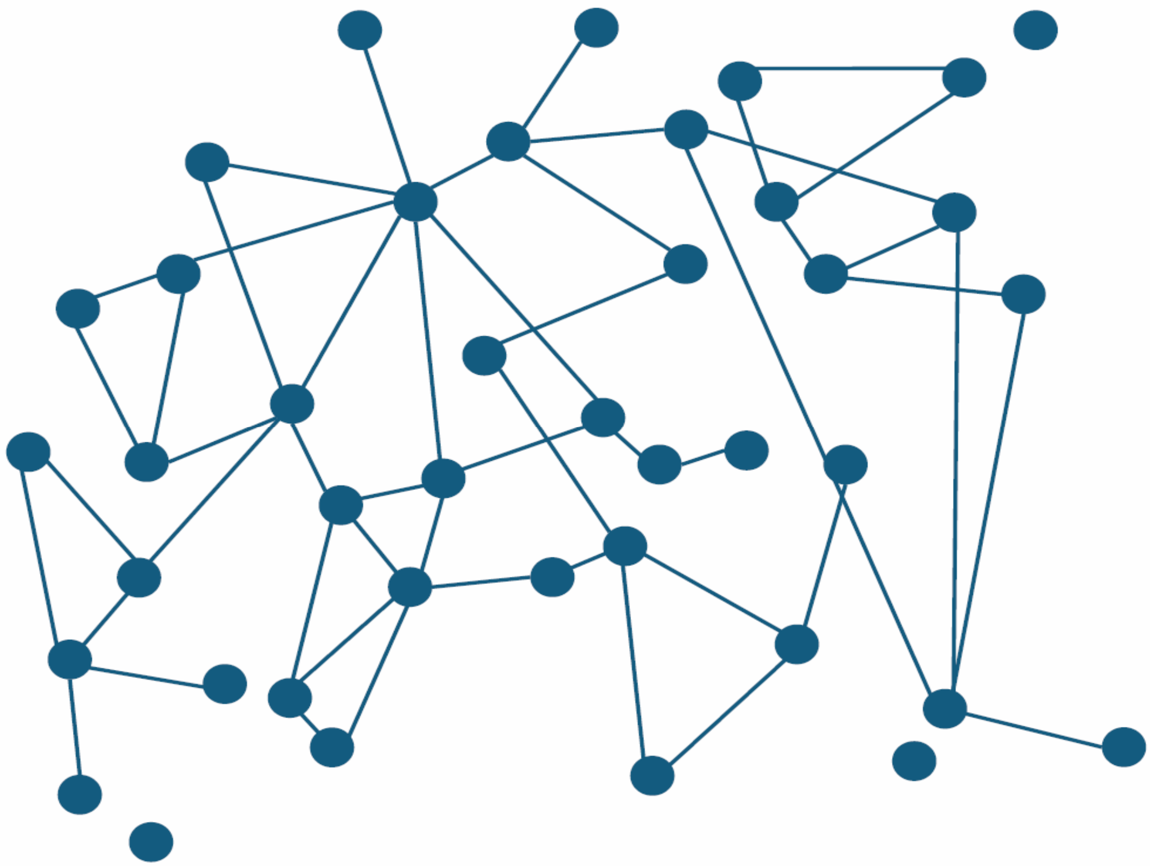
- Incidentals generated by combinations of other subgraphs
- Sparsity definition relates rates of all subgraphs to each other (none grow too quickly)
- Intuitive example: links and triangles
  - $p_L = o(n^{-1/2})$, $p_T = o(n^{-3/2})$
  - Typical node involved in less than

    $n^{1/2}$ links, $n^{1/2}$ triangles

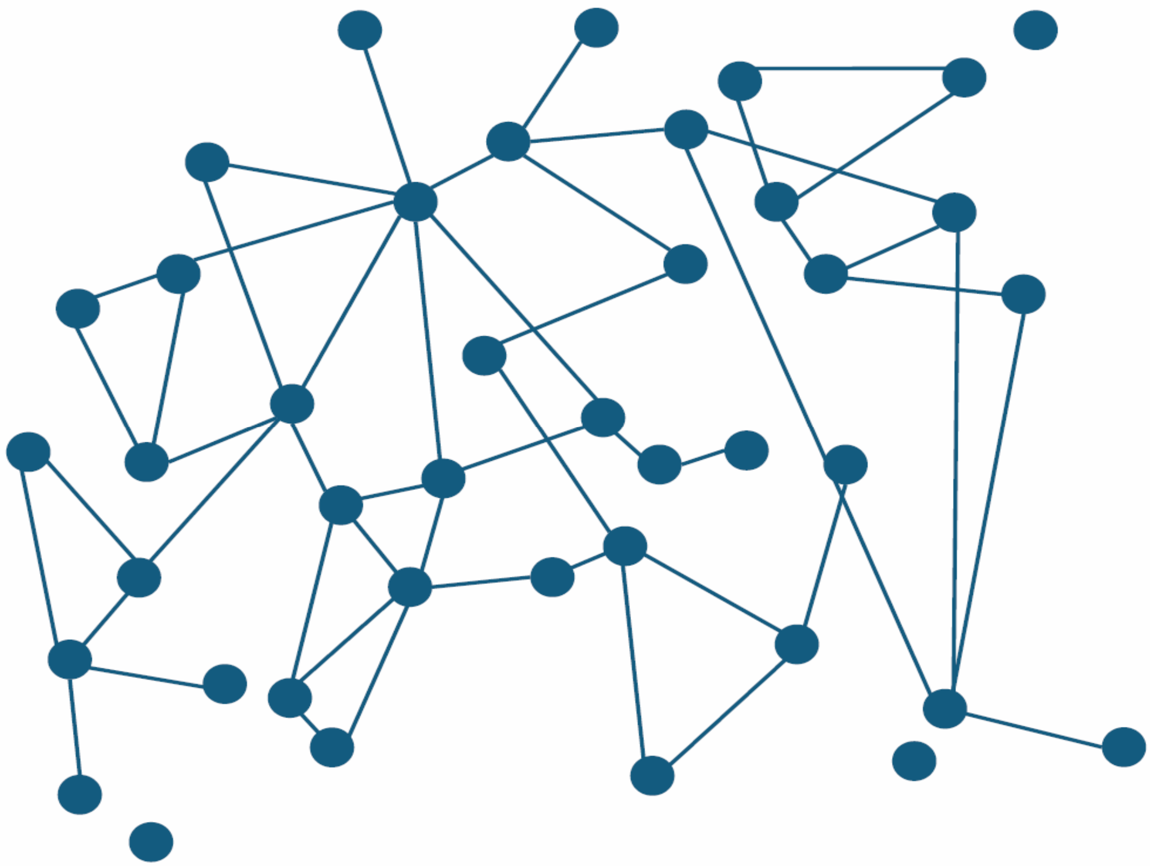
Links Form: Incidental Triangle



Estimation:

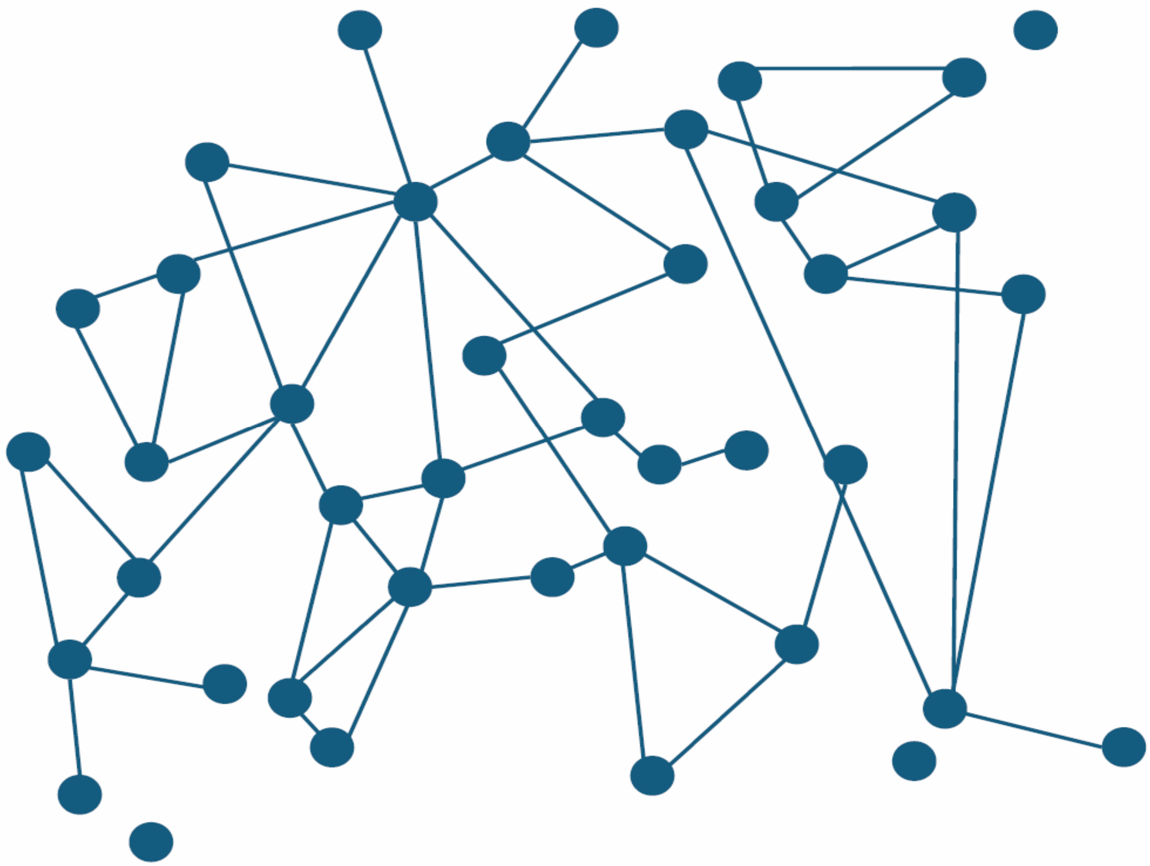


Estimation: Triangles



n = 42
T = 10

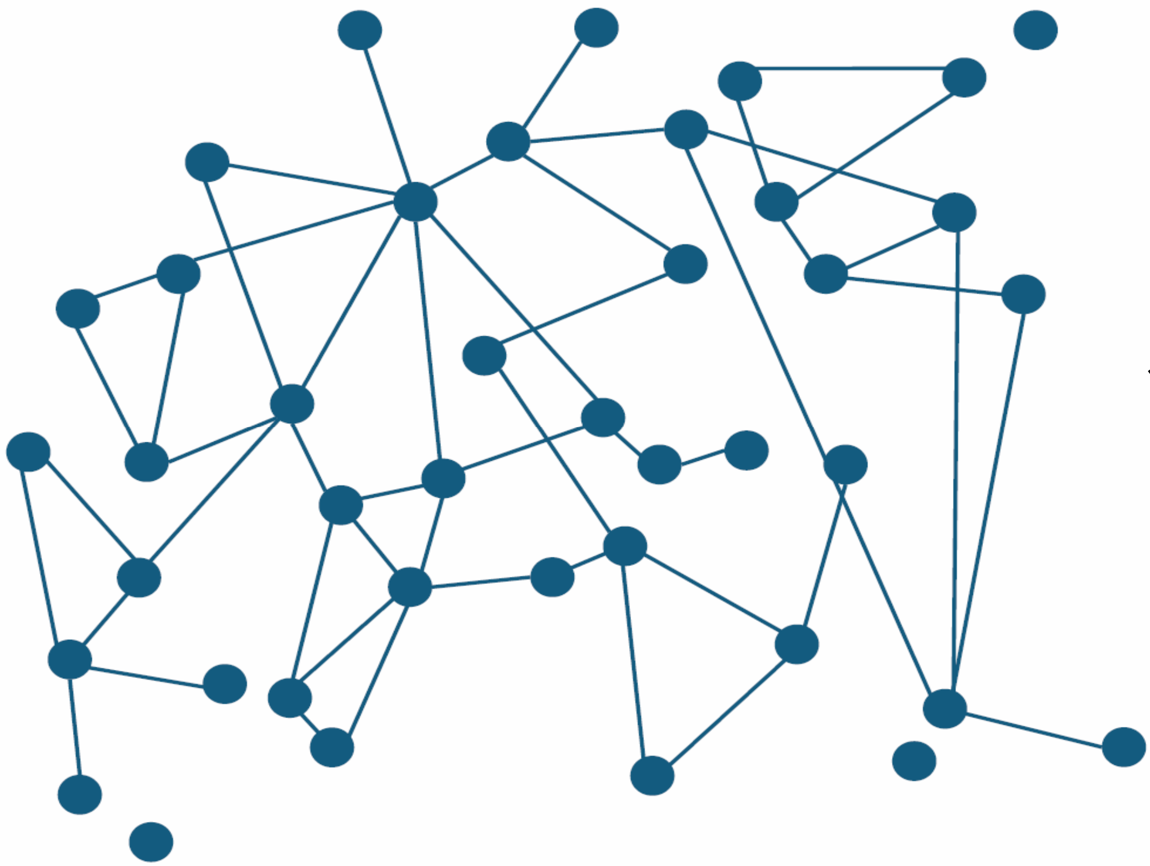
Estimation: Triangles



n = 42
T = 10

n choose 3
=
11480

Estimation: Triangles



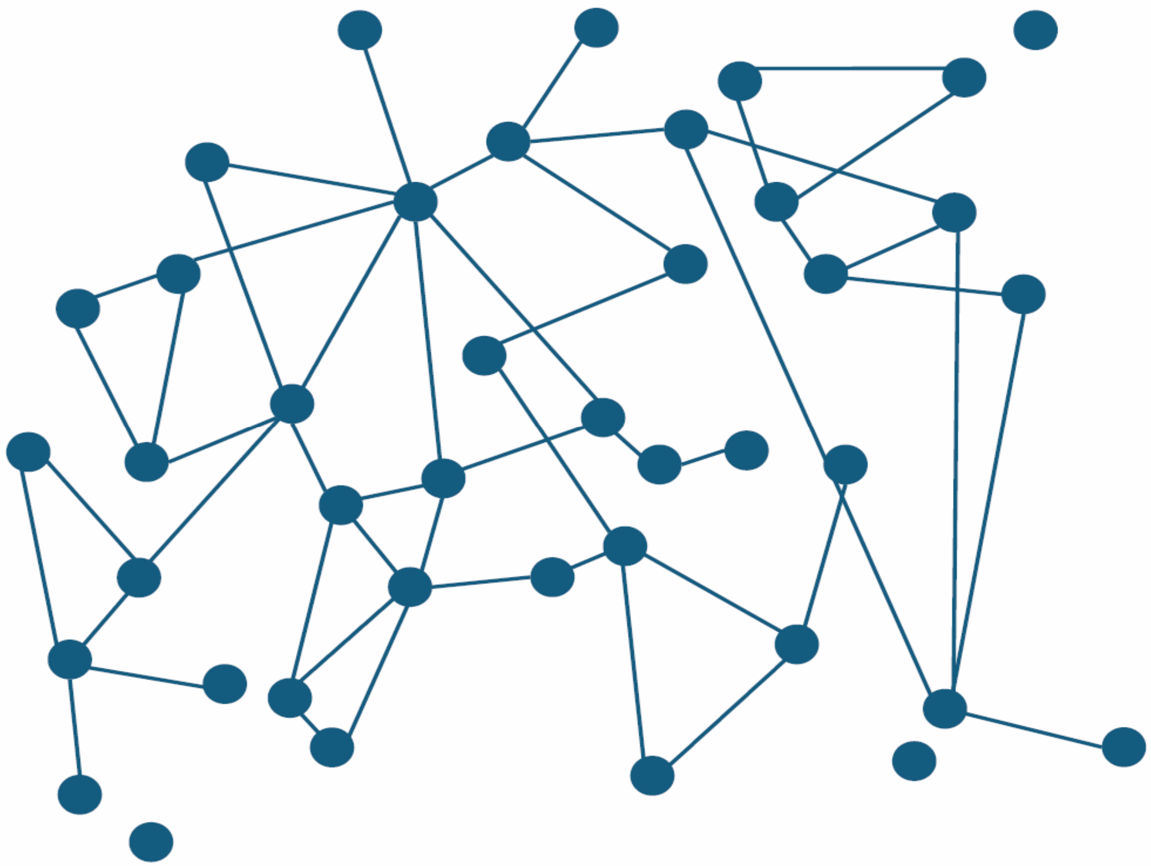
n = 42
T = 10

n choose 3
=
11480

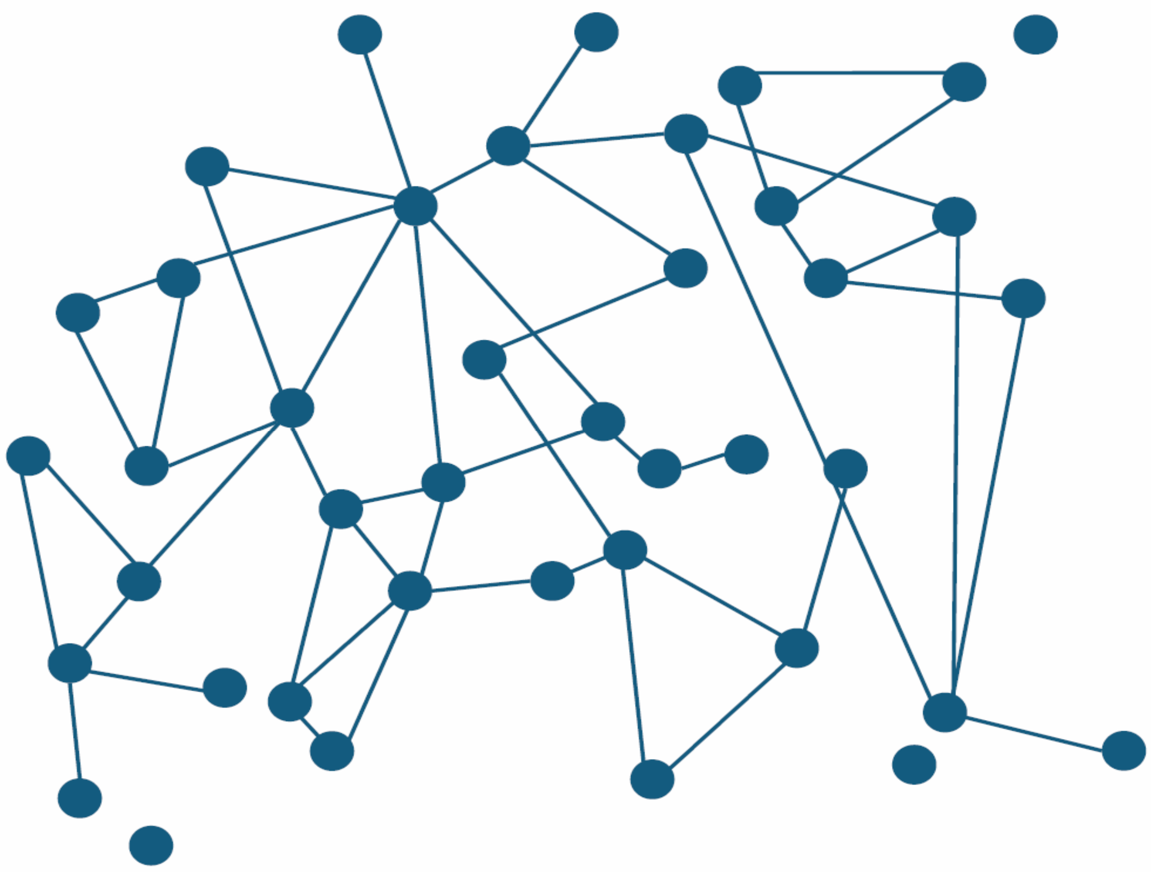
$\hat{p}_T$
=
.00087

Estimation: Links

n = 42
L = 23 (not in triangles)



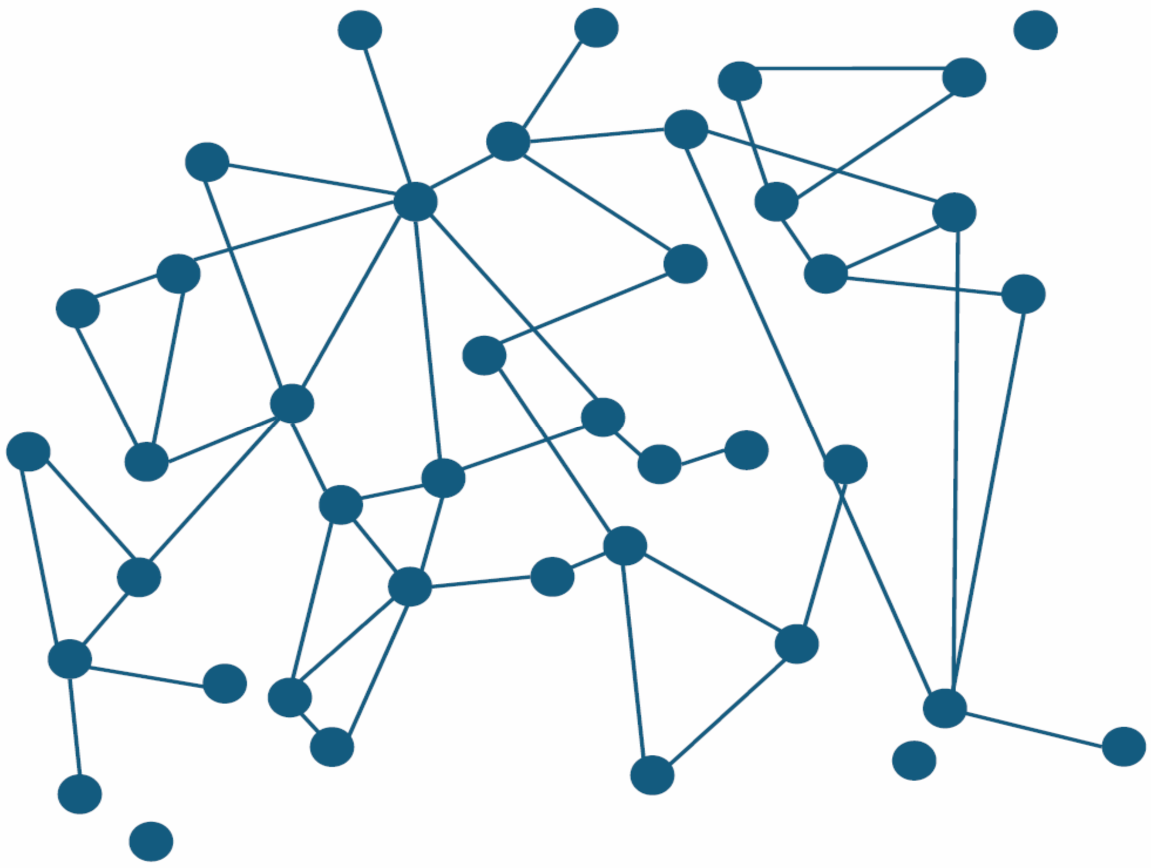
Estimation: Links



n = 42
L = 23 (not in triangles)

n choose 2
=
861

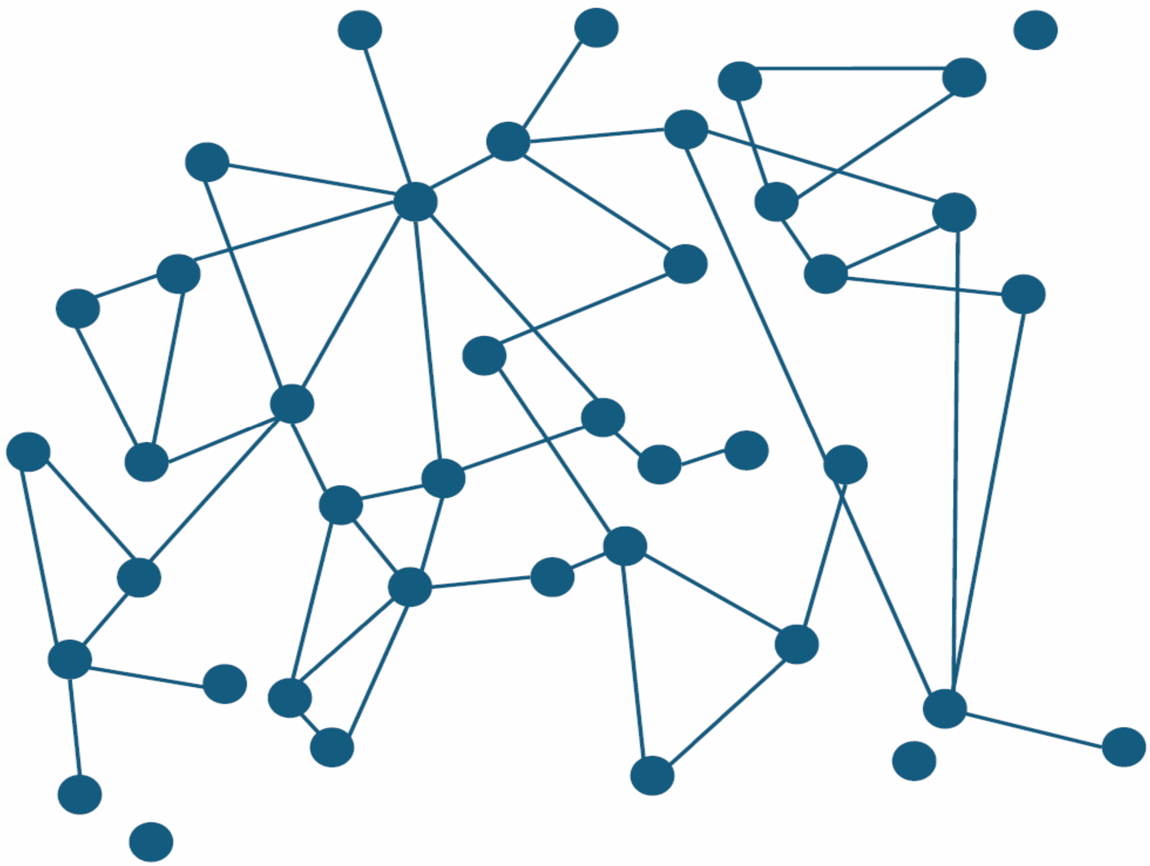
Estimation: Links



n = 42
L = 23 (not in triangles)

n choose 2
=
861 - 28 links in triangles =
833 possible links not in triangles

Estimation: Links

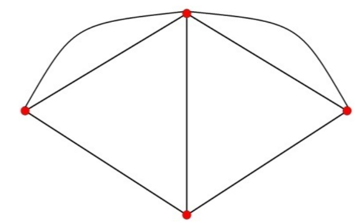


n = 42
L = 23 (not in triangles)

n choose 2
=
861 - 28 links in triangles =
833 possible links

$\hat{p}_L$ = 23 / 833
= .0276

Theorem: Consistency and Distribution

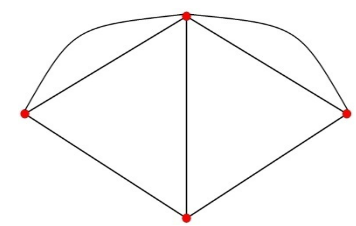


Consider a sequence of sparse SUGMs

The empirical frequency $\hat{p}^n_j = S^n_j / \bar{S}^n_j$ is (ratio) consistent: $\hat{p}^n_j / p^n_j \rightarrow 1$ and

$D^{1/2} (\hat{p}^n - p^n) \rightarrow N(0,I)$

Need for SERGMs/SUGMs:

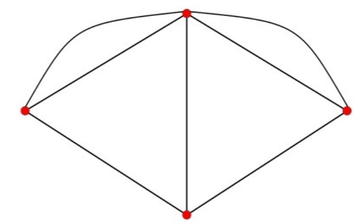


- Examine data from 75 Indian villages from BCDJ '13

- Estimate a model and then use it to generate networks

- How well do the model-recreated networks match real networks on ***non-modeled characteristics***

Need for SERGMs/SUGMs:

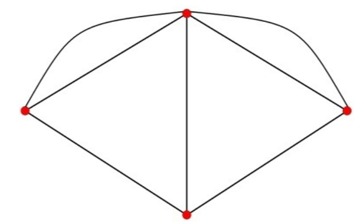


- Estimate SUGM based on covariates, allowing for triangle counts

- Estimate standard link-based (block) model based on covariates

- Does SUGM do better than block model at recreating networks?

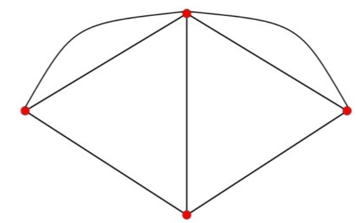
Need for SERGMs/SUGMs:



- Two nodes are either same or different

- Same if

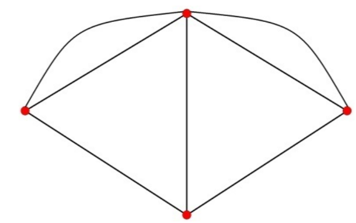
Need for SERGMs/SUGMs:



- Two nodes are either same or different

- Same if
  - same caste

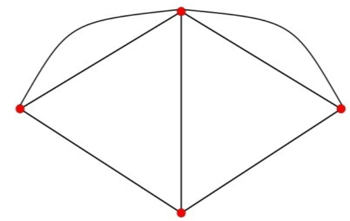
Need for SERGMs/SUGMs:



- Two nodes are either same or different

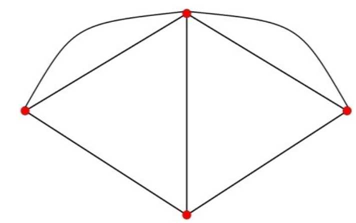
- Same if
  - same caste
  - and gps distance between homes is less than median distance

Need for SERGMs/SUGMs:



- Two nodes are either same or different
- Same if
  - same caste
  - and gps distance between homes is less than median distance
- Different otherwise

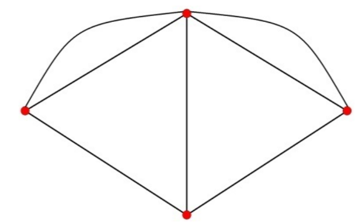
Need for SERGMs/SUGMs:



- Block model
  - prob of link if both same
  - prob of link if different

- SUGM add in
  - prob of triangle if all same
  - prob of triangle if some different

Need for SERGMs/SUGMs:

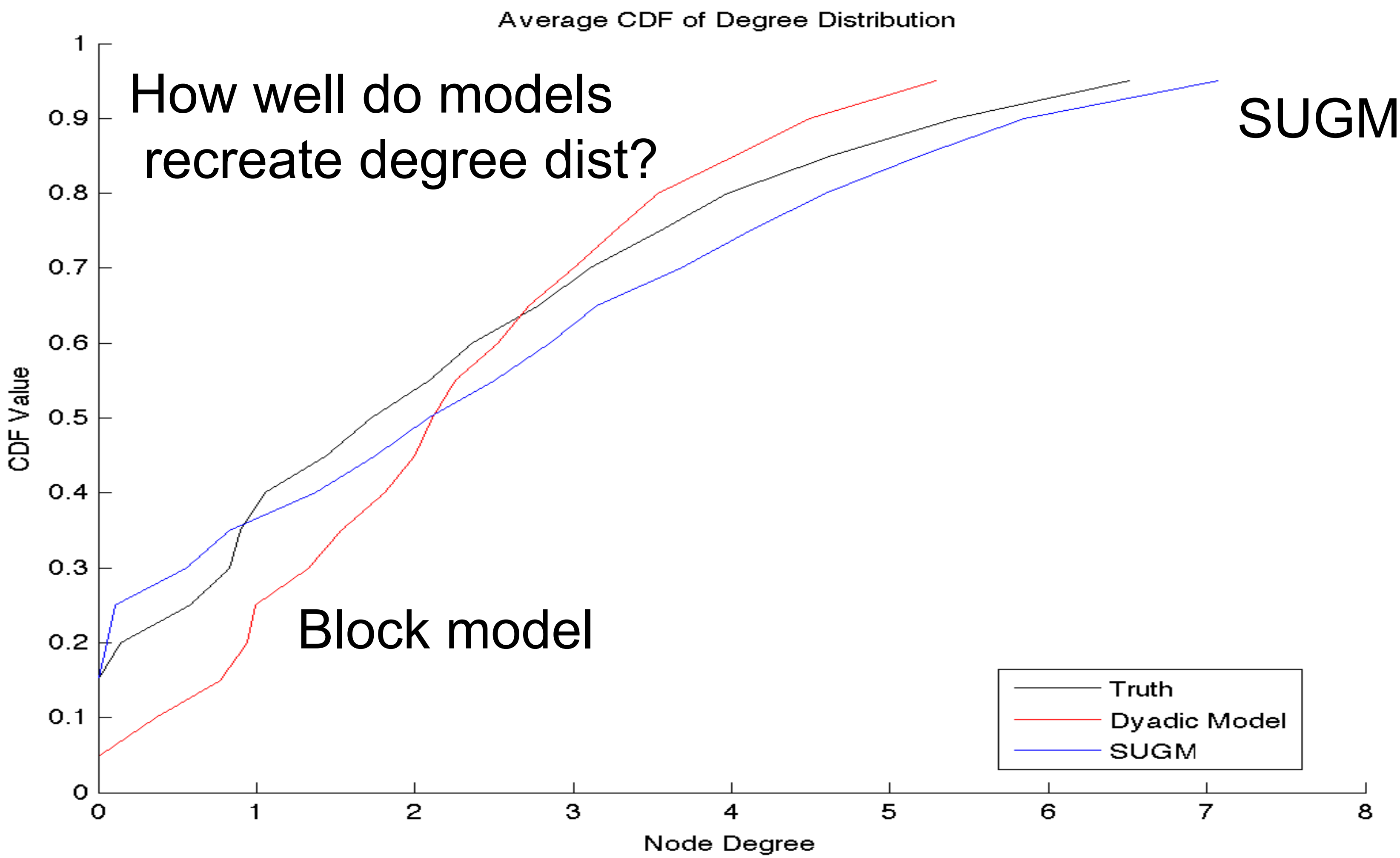


- Step 1: Estimate models
  - Block model, estimate $p_{LinkSame}$ $p_{LinkDiff}$
  - SUGM, estimate $p_{LinkSame}$ $p_{LinkDiff,}$ $p_{TriadSame}$ $p_{TriadDiff}$

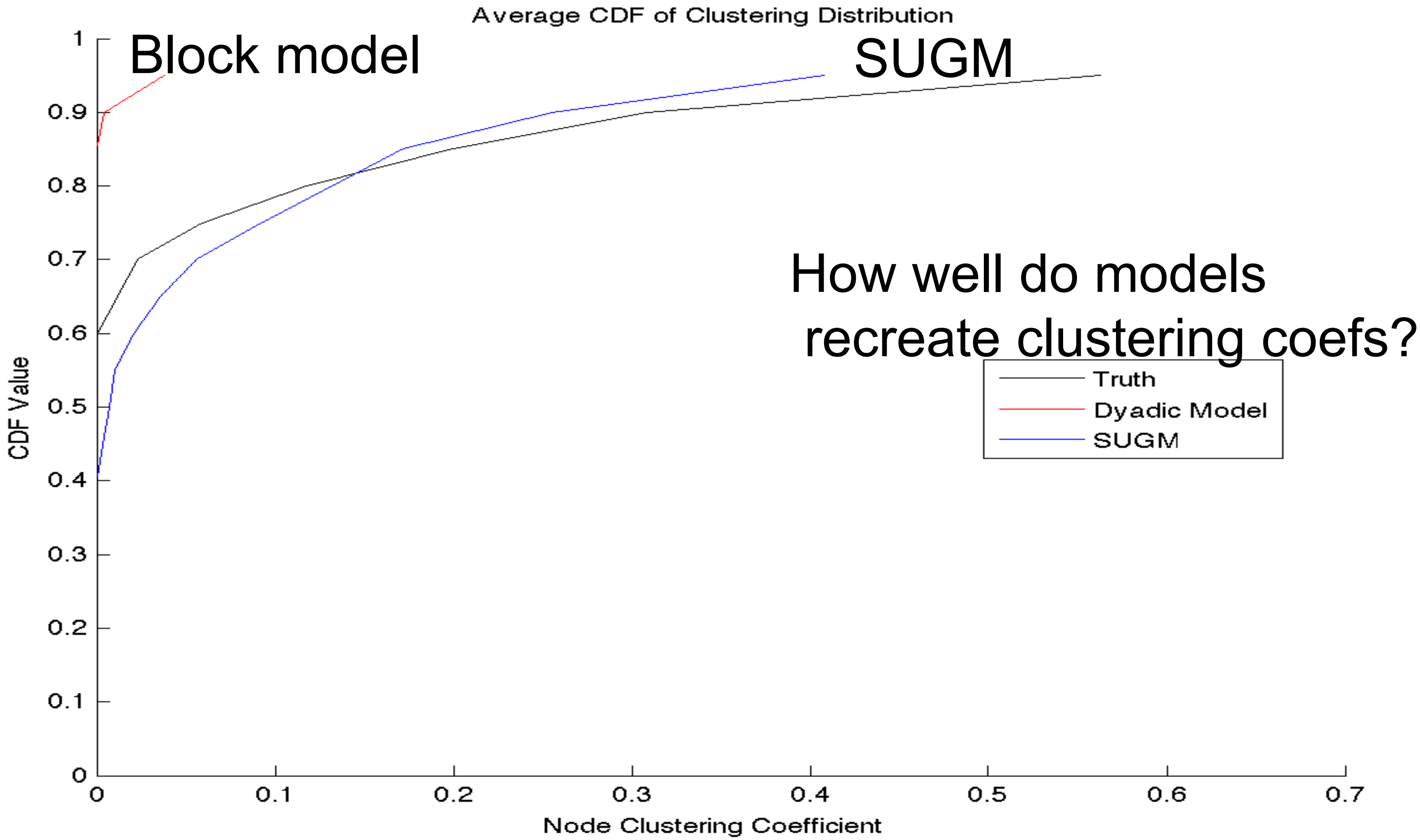
- Step 2: randomly generate networks
  - Block model – randomly generate links
  - SUGM – randomly generate links, triangles...

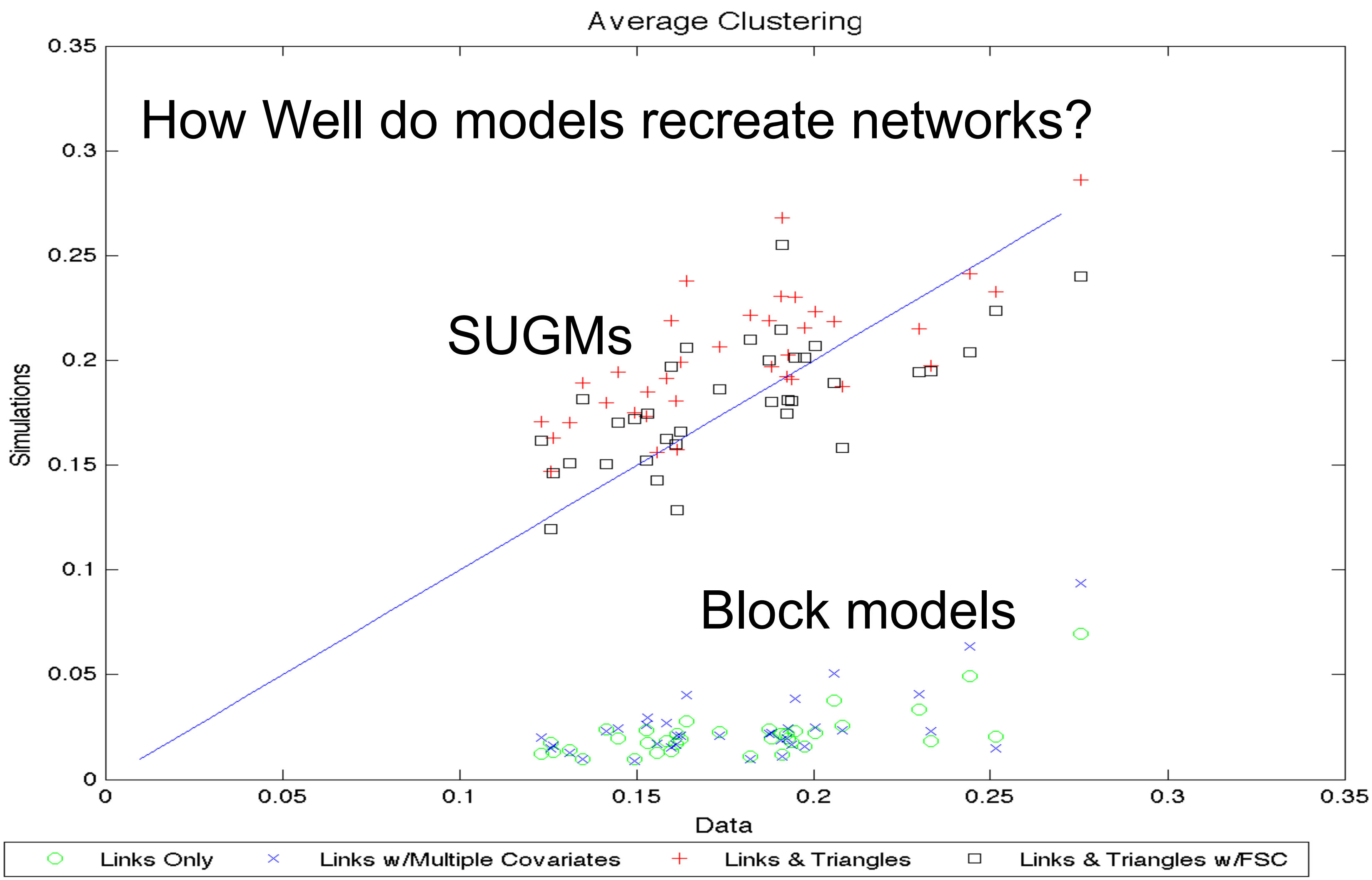
Recreate Networks

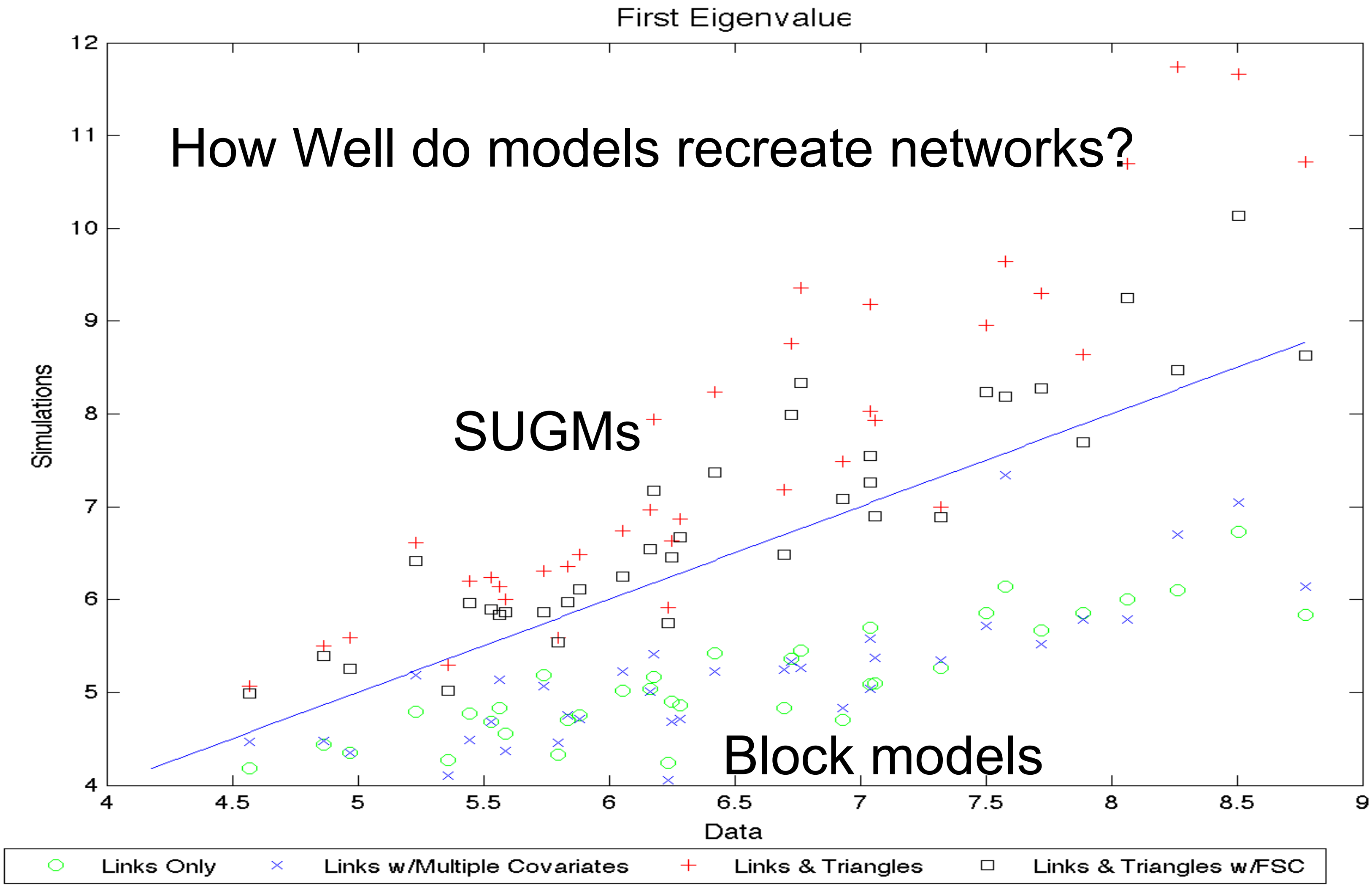
	Data	Block Model	SUGM	SUGM ISOL
# Unsupp. Links	161	236	161	162
# Triangles	39	3	40	39
Avg. Degree	2.3	2.3	2.6	2.5
Isolates	55	26	31	66
Clustering	0.09	0.01	0.13	.09
Frac. Giant Comp.	0.71	0.83	0.79	.67
First Eigvalue	5.5	3.9	4.7	5.3
Second Eigvalue	0.96	0.96	0.96	.91
Avg Path Length	4.7	5.7	5.1	4.1

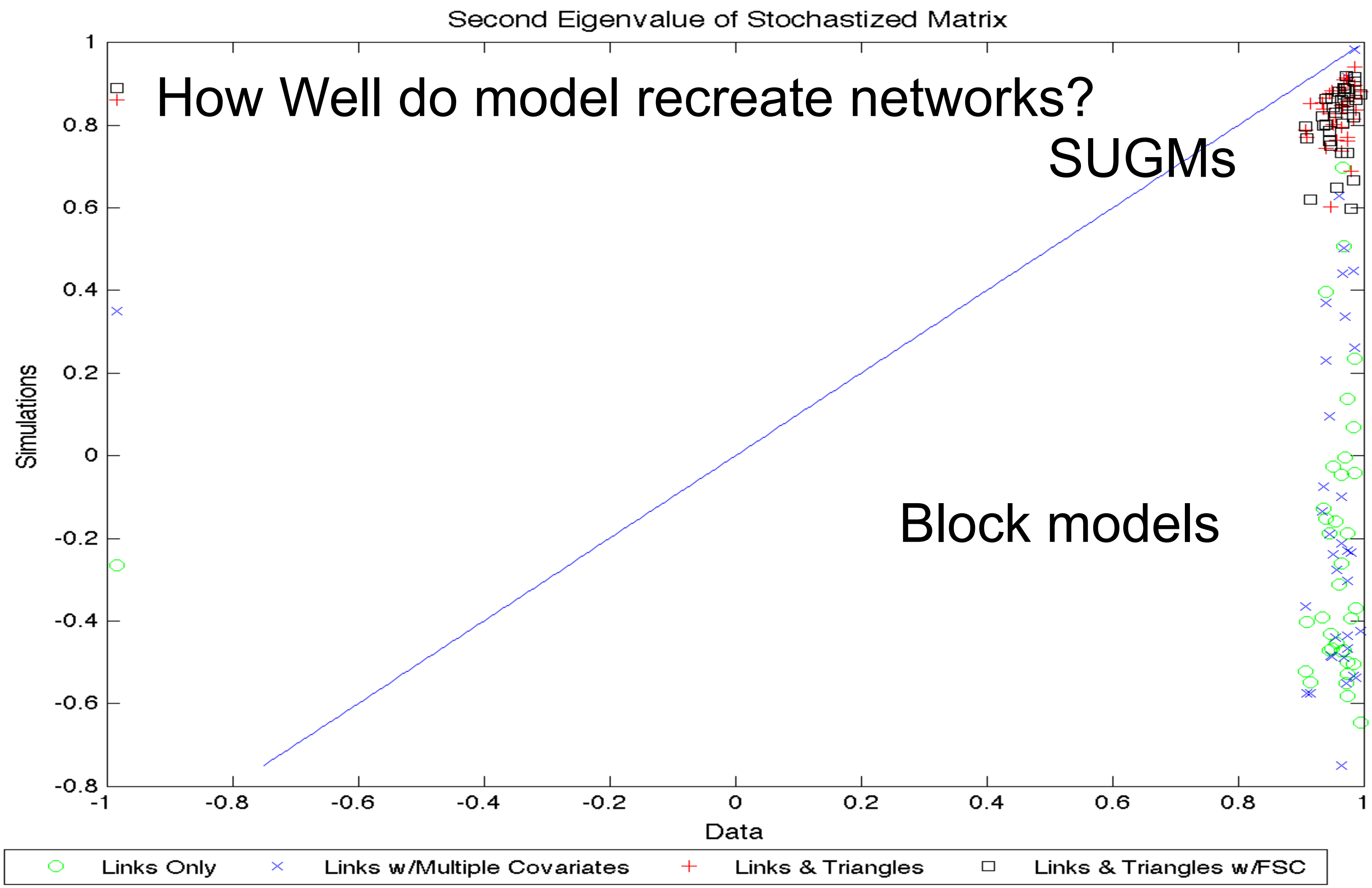


Average CDF of Degree Distribution
How well do models recreate degree dist?
SUGM
Block model
CDF Value
Node Degree
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
1
2
3
4
5
6
7
8
Truth
Dyadic Model
SUGM


Average CDF of Clustering Distribution
Block model
SUGM
How well do models recreate clustering coefs?
CDF Value
Node Clustering Coefficient
Truth
Dyadic Model
SUGM
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

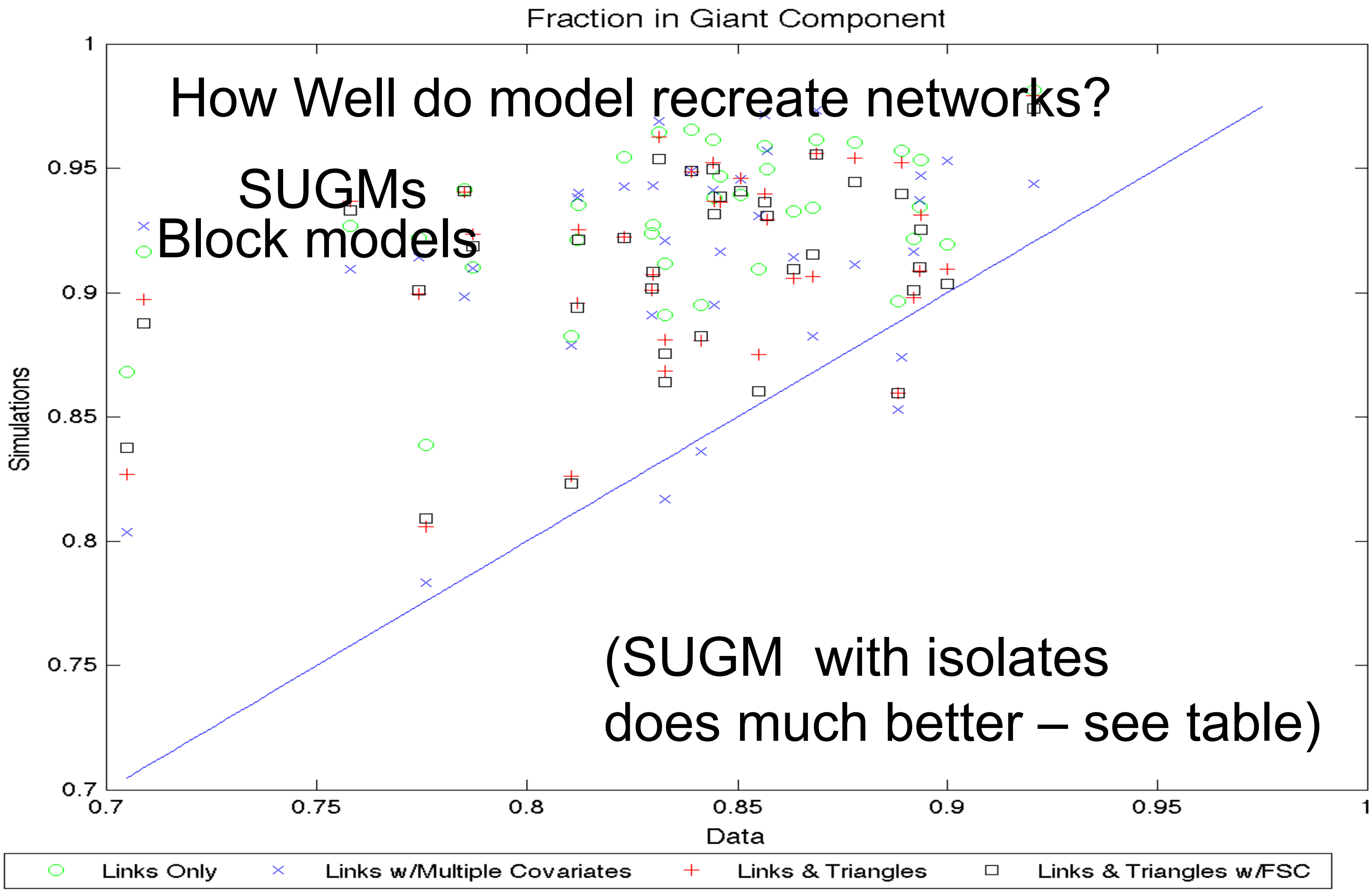

Average Clustering
How Well do models recreate networks?
SUGMs
Block models
Simulations
Data
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
Links Only
Links w/Multiple Covariates
Links & Triangles
Links & Triangles w/FSC


First Eigenvalue
How Well do models recreate networks?
SUGMs
Block models
Simulations
Data
4
4.5
5
5.5
6
6.5
7
7.5
8
8.5
9
10
11
12
Links Only
Links w/Multiple Covariates
Links & Triangles
Links & Triangles w/FSC


Second Eigenvalue of Stochastized Matrix
How Well do model recreate networks?
SUGMs
Block models
Simulations
Data
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
Links Only
Links w/Multiple Covariates
Links & Triangles
Links & Triangles w/FSC

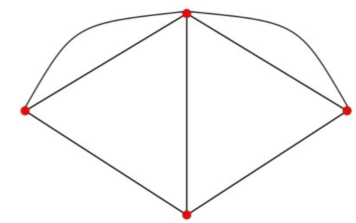
How Well do model recreate networks?

SUGMs
Block models



(SUGM with isolates
does much better – see table)

Dependencies



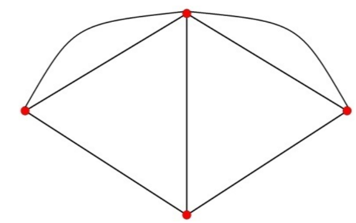
- ``Social'' by definition generates dependencies

- Need tractable models to capture/test these

- ERGMs are rich family, but not always accurately estimable

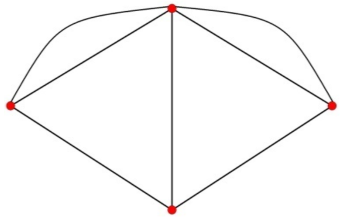
- SERGMs and SUGMs offer easy and consistent estimation

Network Models



- Statistical models offer a medium, but also need models *in context*
- Understand dependencies? Friends of friends, social enforcement....
- What should we be testing for?
- Example, see lecture 4.9....

Strengths Random Networks:

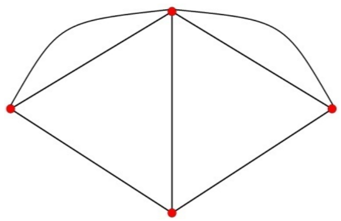


- Generate large networks with well identified properties

- Mimic real networks (at least in some characteristics)

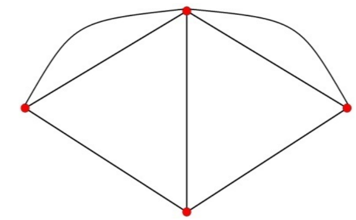
- ***Tie specific properties to specific processes***

Weaknesses of Random Network Models



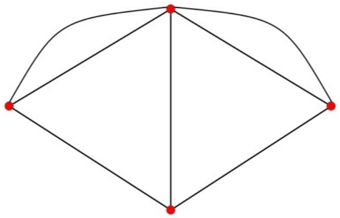
- Missing the ``Why''
  - Why this process? (lattice, preferential attach...)
- Missing implications of network structure: context or relevance
  - welfare, efficiency?
- Literature is missing careful empirical analysis of many ``stylized facts'' (small worlds, power laws, clustering...)
  - ERGMs have been filling that niche, but need estimable models
  - New models are emerging!

Week 3 Wrap



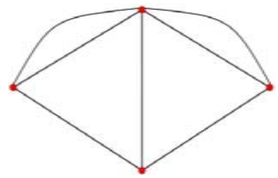
- Growing random networks: provide heterogeneity based on age - analyze via mean field...
- Can lead to power laws with pure preferential attachment
- Many networks lie between the extremes: friends of friends
- Class of models providing statistical fits: ERGMS
  - Allows formation based on structures beyond links – correlations
  - Can be challenging to estimate: need to calculate relative probabilities
  - New techniques/variations based on direct statistic counts offer alternatives

Week 3: References in order Mentioned



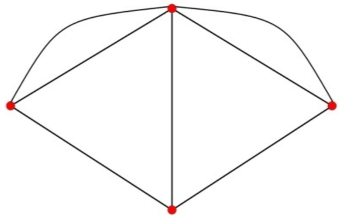
- Price, D.J.S. (1965) “Networks of Scientific Papers,” *Science 149:510–515.*
- Albert, R., H. Jeong, and A.L. Barab´asi (1999) “Diameter of theWorldWideWeb,” *Nature 401:130–131.*
- Simon, H. (1955) “On a Class of Skew Distribution Functions,” *Biometrika 42(3*–4):425–440.
- Price, D.J.S. (1976) “A General Theory of Bibliometric and Other Cumulative Advantage Processes,” *Journal American Society of Information Science 27:292–306.*
- Barab´asi A., and R. Albert (1999) “Emergence of Scaling in Random Networks,” *Science 286:509–512.*
- Jackson, M.O., and B.W. Rogers (2007) “Meeting Strangers and Friends of Friends: How Random Are Social Networks?” *American Economic Review 97(3):890–915.*
- Frank, O. and D. Strauss (1986): “Markov graphs,” *Journal of the American Statistical Association, 832–842.*
- Wasserman, S. and P. Pattison (1996): “Logit models and logistic regressions for social networks: I. An introduction to Markov graphs andp,” *Psychometrika, 61,* 401–425.
- Strauss, D. (1986). On a general class of models for interaction. SIAM Rev., 28 513{527.
- Park, J. and Newman, M. E. J. (2004). Solution of the two-star model of a network. Phys. Rev. E (3), 70 066146, 5.
- Park, J. and Newman, M. E. J. (2005). Solution for the properties of a clustered network. Phys. Rev. E (3), 72 026136, 5.
- Chatterjee, S. and P. Diaconis (2011): “Estimating and Understanding Exponential Random Graph Models,” *Arxiv preprint arXiv:1102.2650. 4*
- Hammersley, J. and P. Clifford (1971): “Markov fields on finite graphs and lattices,” . 1
- Robins, G., P. Pattison, Y. Kalish, D. Lusher (2007) An introduction to exponential random graph (p^*) models for social networks, Social Networks, 29:2, 173-191.
- Snijders, T. (2002): “Markov Chain Monte Carlo Estimation of Exponential Random Graph Models,” *Journal of Social Structure, 3, 2-40*
- Handcock, M. S. (2003). Assessing degeneracy in statistical models of social networks, Working Paper 39. Tech. rep., Center for Statistics and the Social Sciences, University of Washington.

Week 3: References Cont'd



- Bhamidi, S., G. Bresler, and A. Sly (2008): "Mixing time of exponential random graphs," *Arxiv preprint arXiv:0812.2265*
- Chandrasekhar, A.G. and M.O. Jackson (2012) ``Tractable and Consistent Exponential Random Graph Models'' SSRN 2150428, ArXiv: http://arxiv.org/abs/1210.7375

Summary



- ERGMs face estimation challenges, and no general results on consistency

- Work with statistics/subgraphs rather than networks simplifies substantially: ***SERGMs / SUGMs***

- Consistency and fast estimation theorems