Overview of the ”Growing Random Network” literature

The Questions

- Large socially-generated networks share many features
- By what processes are social networks formed?
  - How do changes in formation relate to structure?
- What behavioral or welfare implications might differences in structure have?

Specific Goals

- Model the formation of social networks
  - Nodes meet each other completely at random
  - Nodes meet through connections: “network-based’’ or “friends of friends”
- Establish the properties implied by such a process
  - Match observed “stylized’’ facts
- Use the model to infer random/local search rates
  - Fit model to various data sets
- Implications of fit (time-permitting)
  - Efficiency properties
  - Diffusion processes on a network

Outline

- Background on social networks
- Network formation Model
- Results on implied structural properties
- Fit model to various network data
- Implications of fit:
  - Efficiency
  - Diffusion

Representation of network

- Set of agents identified as nodes
  T={1,2,3,…}
- Pairwise relationships modeled as binary links ij
- A network is a graph: a set of nodes and set of links between pairs of nodes

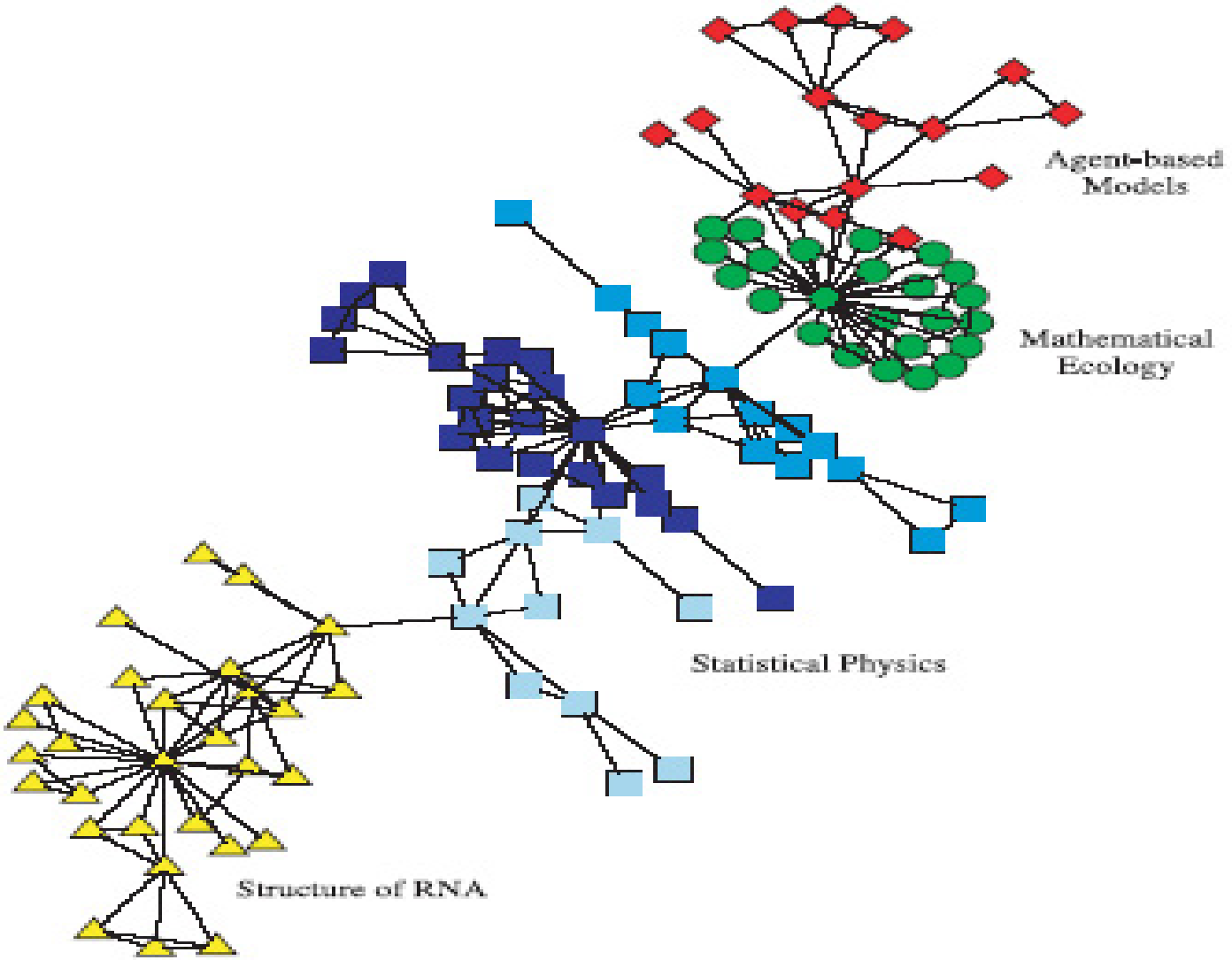
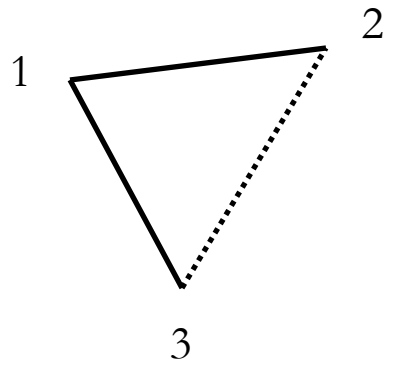


Fig. 1. An example of a small coauthorship network depicting collaborations among scientists at a private research institution. Nodes in the network represent scientists, and a line between two of them indicates they coauthored a paper during the period of study. This particular network appears to divide into a number of subcommunities, as indicated by the shapes of the nodes, and these subcommunities correspond roughly to topics of research, as discussed by Girvan and Newman (37).

Illustration of Concepts



- Degree:
  - *With* link 23, degrees are (2, 2, 2)
  - *Without* link 23, degrees are (2, 1, 1)
- Clustering: What is the probability of link 23?
- Diameter:
  - *With* link 23, diameter = 1
  - *Without* link 23, diameter = 2

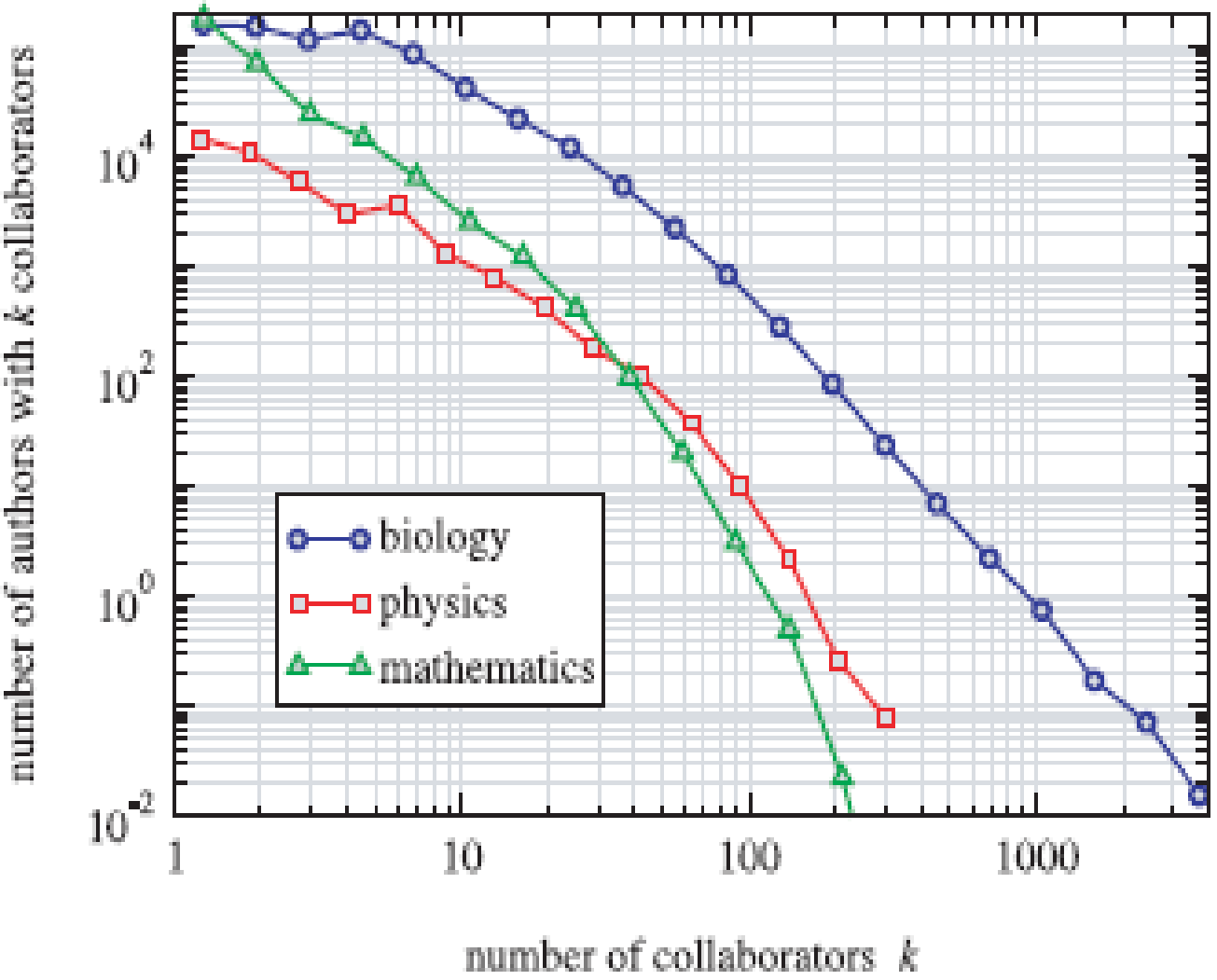
Social Networks Share Many Features

1. Degree distributions: “Scale-free”– heavy tails

   - Many more nodes with very high and very low degrees relative to what one would find in a completely random network

Degree Distributions

Co-Authorship Data (Newman and Grossman)



Features of Social Networks

1. Degree distributions: “Scale-free’’– heavy tails
2. High clustering – “cliquishness”
   - Proportion of triads out of possible triads in a (sub)graph
   - Any two of a given node’s neighbors are likely to themselves be neighbors

Movie Actors	Math Co-authors	Physics Co-authors	WWW
0.79	0.43	0.15	0.11

Features of Social Networks

1. Degree distributions: “Scale-free’’– heavy tails
2. High clustering – “cliquishness”
3. Low diameter and average path length

	Movie Actors	Math Co-authors	Physics Co-authors	WWW
Average	3.7	7.6	5.9	3.1
Diamet		27	20	

Features of Social Networks

1. Degree distributions: “Scale-free”– heavy tails
2. High clustering – “cliquishness”
3. Low diameter and average path length
4. Assortativity
   - Positive correlation in the degrees of linked nodes
   - Special to socially-generated networks

Assortativity

- Correlation in Degree (Newman 2003):
  - Socially generated networks:
    - .12 math co-authorship, .13 biology, .36 physics
    - .09 emails
    - .21 film actors
  - Technologically generated networks
    - -.19 internet
    - -.003 power grid
    - -.23 neural network

Features of Social Networks

1. Degree distributions: “Scale-free”– heavy tails
2. High clustering – “cliquishness”
3. Low diameter and average path length
4. Assortativity
5. Negative clustering-degree relationship
   - Individual clustering coefficients
   - Degree-weighted avg. is smaller than simple avg.

Inverse Degree - Clustering

- Lower clustering on higher degree nodes – compare weighted average to (unweighted) avg clustering:
  - .20/.78 film actors
  - .15/.34 math co-authorship
  - .45/.56 physics
  - .09/.60 bio

(From Newman 2003, Table II)

Related Classes of Models

- Random Graphs (Erdos and Renyi (1960), …)
  - Diameter: log(n)
- Preferential Attachment (Price (1965), Barabasi-Albert (1999), …)
  - Scale-free; Diameter: log(n) or smaller
- Our model fits in between
- A range of others (Vazquez (2003), Pennock et al (2002), …)

Introduction to the ”Growing Random Networks” Literature

- New nodes are born over time and form attachments to existing nodes when they are born.
- Directed networks
- Time introduces a natural heterogeneity to nodes based on their age in a growing network.

Growing Erdös-Réyni networks

- Newborn nodes pick nodes to link to uniformly at random.
- Nodes are indexed by the order of their birth.
- Node *i* is born at date *i*, where *i*={0,1,2,...}
- $d_i(t)$ degree of node *i* (born at time *i*) at a time *t*.
- $d_i(i)$ number of links formed at a node's birth
- $d_i(t) - d_i(i)$ number of links that node *i* gets from the new nodes that were born between time *i* and time *t*.

Network formation

- Each newborn node randomly selects m of the existing nodes and links to them.
- Start the network with m+1 nodes born at time {0,1,....,m}, each connected to each other.
- The first newborn node that we consider is the one born at time m+1.

Network formation

- At the end of time *m*+1, *m* of the older nodes will have new links and one older will not. The newest node will have *m* links.
- At time *t*, a node *i* born at time $m \leq i < t$ will have an *expected degree* at time *t* of:

$$m + \frac{m}{i+1} + \frac{m}{i+2} + \ldots + \frac{m}{t}$$

$$\Leftrightarrow \quad m\left(1 + \frac{1}{i+1} + \frac{1}{i+2} + \ldots + \frac{1}{t}\right)$$

Network formation

- For large *t*, this is approximately:

$$m\left(1+\log\left(t/i\right)\right)$$

Network formation

- Distribution of *expected* degrees
- For large *t*, the nodes that have expected degree less than *d* are (approximately) those such that:

$$m\left(1+\log(\,t\,/\,i\,)\right)<d$$

which is equivalent to:

$$i\;>\;t\;e^{1-d\;/\;m}\;\Leftrightarrow\;\frac{t\;-\;i}{t}<\;1\;-\;e^{1-d\;/\;m}$$

Thus the nodes with expected degree less than *d* are those born after time $te^{1-d/m}$. This is a fraction of $1-e^{1-d/m}$

Network formation

- Thus, for $d < m(1 + \log(t / m))$, the fraction of nodes with expected degrees less than *d* is:

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

- This is a variation of the exponential distribution.
- Each node starts out with *m* links and then the expected links that a random node expects to gain over over time has an exponential distribution with expected value *m*.

Mean-field Approximations

- We have calculated the distribution of *expected degrees* after time *t*.
- How close is this to the distribution of *actual degrees* after time *t* ?
- The full randomness of the process is quite complex. Difficult to deduce the degree distribution of the process directly.
- Approximation of the distribution of expected degrees.

Continuous Time Mean-field Approximations of Degree Distributions

- Standard technique derived from statistical physics.
- One assumes that the system evolves so that things occur at the *average level* rather than *randomly*.

Mean-field approximations

- Example: Assume that there are already 100 nodes and a new node appears and is supposed to form links to existing nodes independently with proba 1/10.
- Under MFA, we suppose that the node forms exactly 10 links.
- With continuous-time approximation: Change in time of a given node's degree occurs at a *fixed rate* rather than at a *stochastic rate*.

Continuous-time approximation of degree distributions

- Go back to the model of growing Erdös-Réyni network formation.
- Alternative technique: Continuous-time mean-field approximation
- As before: New node born at time *t* and forms *m* links by uniformly randomly picking *m* out of the *t* existing nodes at time *t*.

Continuous-time mean-field approximation

- Starting condition: $d_i(i) = m$
- Approximate change over time for $t > i$:

$$\frac{dd_i(t)}{dt} = \frac{m}{t}$$

- New node born at each time is spreading it's m new links randomly over the t existing nodes at time t.

Continuous-time mean-field approximation

- Solution to this differential equation:

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

- Approximate degree distribution
- If we ask how many nodes have degree of no more than *d* ?
- And we see that a node born at time τ has degree of exactly *d*
- Then equivalent to ask how many nodes were born on or after time τ
- So, at time *t*, the fraction of nodes having a degree of no more than *d* would be $(t - \tau)/t$

Continuous-time mean-field approximation

- For any *d* and time *t* , we find the node *i*(*d*) such that

$$d_{i(d)} = d$$

- The nodes that have degree of less than *d* are those born after *i*(*d*). The resulting cummulative distribution function is:

$$F_t(d) = 1 - \frac{i(d)}{t}$$

Continuous-time mean-field approximation

- Apply technique to this random network process, solve for $i(d)$ such that:

$$d = m + m \log\left(\frac{t}{i(d)}\right)$$

- This implies that:

$$\frac{i(d)}{t} = e^{-(d-m)/m}$$

Continuous-time mean-field approximation

- Such a network would have a distribution function described by

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

- Negative exponential distribution with support from *m* to infinity and a mean of degree 2*m*

Continuous-time mean-field approximation: Preferential attachment

- Nodes are born over time and indexed by their date of birth $i = \{0,1,2,\ldots,t,\ldots\}$
- Upon birth each new node forms m links with pre-existing nodes.
- Instead of selecting m of the nodes uniformly at random, it attaches to nodes with probabilities proportional to their degrees.

Preferential attachment

- Probability that an existing node *i* receives a new link to the newborn node at time *t* is *m* times *i*'s degree relative to the overall degree of all existing nodes at time t, i.e.

$$m \frac{d_i(t)}{\sum_{j=1}^{t} d_j(t)}$$

Preferential attachment

- As there are *t m* total links in the system at time *t*, it follows that

$$\sum_{j=1}^{t} d_j(t) = 2tm$$

- Thus, proba that node *i* gets a new link in period *t* is:

$$\frac{d_i(t)}{2t}$$

Preferential attachment

- Start with a pre-existing group of *m* nodes all connected to one another.
- Starting condition:

$$d_i(i) = m$$

- Approximate change over time for $t > i$:

$$\frac{dd_i(t)}{dt} = \frac{d_i(t)}{2t}$$

Preferential attachment

- Solution to this differential equation:

$$d_i(t) = m \log\left(\frac{t}{i}\right)^{1/2}$$

- To find the fraction of nodes with degrees that exceed some given level *d* at some time *t*, we just need to identify which node is at exactly level *d* at time *t*.

Preferential attachment

- Let $i_t(d)$ be the node that has degree *d* at time *t*, or such that

$$d_{i_t(d)}(t) = d$$

- Then, solving the last equation leads to: $\frac{i_t(d)}{t} = \left(\frac{m}{d}\right)^2$
- The fraction of nodes that have degree of less than *d* at time *t* are those born after $i_t(d) = t(m/d)^2$
- The resulting cummulative distribution function is:

$$F_t(d) = 1 - m^2 d^{-2}$$

Preferential attachment

- Such a network would have a distribution function described by

$$F_t(d) = 1 - m^2 d^{-2}$$

- Corresponding density or frequency distribution

(for $d \geq m$) of

$$f_t(d) = 2m^2 d^{-3}$$

Thus the degree distribution (of expected degrees) is a power distribution with an exponent of -3.

Preferential attachment

- Distribution has time independence
- Because the relative degrees of nodes are determined by their relative birth dates.

Basic techniques underlying mean-field analyses

- Consider any growing network in which nodes are indexed in the order of their birth
- Node *i*'s degree at time *t* can be represented as

$$d_i(t) = \phi_t(i)$$

where $\phi_i(t)$ is a decreasing function of *i*. This indicates that younger nodes have lower degrees. It also means that

$\phi_i(t)$ is an invertible function, so that if some degree *d* is specified, then we can determine which node has degree *d* at time *t*.

Basic techniques underlying mean-field analyses

- Because degree increases with age, the fraction of nodes with degree at least *d* are precisely those older than the node *i* satisfying

$$\phi_i(t) = d$$

That is those nodes older than $\phi_t^{-1}(d)$

Basic techniques underlying mean-field analyses

- Thus the degree distribution at time *t* is:

$$F_t(d) = 1 - \frac{\phi_t^{-1}(d)}{t}$$

- This if we can derive an expression for $d_i(t)$ that is decreasing in *i*, so that older nodes have more links, then we can easily derive the associated degree distribution.

Jackson-Rogers AER 2008 Key Intuitions

- Network-based meetings: likely to find nodes with many links
  - » variation on scale-free: growth partly proportional to size
- Random aspect: not entirely scale-free
  - » fit lower tail too
- Local search: may connect to two nodes that are already linked
  - » high clustering

Key Intuitions (cont.)

- Search aspect generates hub-like nodes
  - » low diameter
- Randomness connects different neighborhoods
  - » even lower diameter
- Nodes enter sequentially and connect to existing nodes
  - » Assortativity: age is correlated with both degree and links
- High-degree nodes get most connections from network-based meetings
  - » Lower clustering in high-degree nodes

Mean-field approximation

- Network formation is stochastic and path-dependent
  - Difficult to analyze directly
- Use a deterministic continuous-time system
  - All changes occur at the mean rate of underlying stochastic system
- Apply, e.g., to expected changes in degrees of nodes
- Can analytically solve for steady state

The Network Formation Model

- Agents indexed by date of birth t ={1,2,3,...}
- Upon birth, node t indentifies m_r nodes uniformly at random: “Parent nodes”
- Node t also meet other nodes in its parents’ immediate neighborhoods to find m_n more nodes
  - Like entering at a random web page and following links
  - These nodes are picked uniformly at random.
- Link to a given node if net utility is positive (probability p)
  - (Extensions discussed in the paper)

The Network Formation Model

- Intuition
- A newborn node links to existing nodes through two different processes:

A combination of linking by uniformly at random and preferential attachment.

Each newborn node forms *m* links, with a fraction $\alpha < 1$ of them formed to existing nodes selected uniformly at random and $1- \alpha$ of them formed to existing nodes by preferential attachment.

Mean-field approximation

Change in the degree of a node over time (mean-field):

$$\frac{dd_i(t)}{dt} = \alpha \frac{m}{t} + (1-\alpha)\frac{md_i(t)}{2mt}$$

Solution:

$$d_i(t) = \phi_t(i) = \left(d_0 + \frac{2\alpha m}{1-\alpha}\right)\left(\frac{t}{i}\right)^{(1-\alpha)/2} - \frac{2\alpha m}{1-\alpha}$$

d_0 initial number of links that a node has when it is born.

Mean-field approximation

- We deduce

$$\phi_t^{-1}(d) = t\left(\frac{d_0 + \frac{2\alpha m}{1-\alpha}}{d + \frac{2\alpha m}{1-\alpha}}\right)^{2/(1-\alpha)}$$

- Setting $d_0 = m$, we have

$$F_t(d) = 1 - \left(\frac{m + \frac{2\alpha m}{1-\alpha}}{d + \frac{2\alpha m}{1-\alpha}}\right)^{2/(1-\alpha)}$$

Intuition

$$F_t(d) = 1 - \left(\frac{m + \frac{2\alpha m}{1-\alpha}}{d + \frac{2\alpha m}{1-\alpha}} \right)^{2/(1-\alpha)}$$

- When α = 0, preferential attachment:

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

- When $\alpha \to 1$, purely random:

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

Important Parameters in more general model

- Start with an initial network on a set of at least $m_r + m_n + 1$ nodes, where each node has at least $m_r + m_n$ neighbors

- r = m_r/m_n = ratio of the number of links formed at random vs. by network-based meetings

- m = p($m_r + m_n$) = average number of links per node

More general expression

Then the proba that a given existing node *i* with in-degree $d_i(t)$ gets a new link in period t+1 is roughly:

$$p_r \frac{m_r}{t} + p_n \left(\frac{m_r d_i(t)}{t} \right) \left(\frac{m_n}{m_r \left(p_r m_r + p_n m_n \right)} \right)$$

Network formation

- Let

$$m = p_r m_r + p_n m_n$$

be the expected number of links that a new node forms.

Then the proba that a given existing node i with in degree $d_i(t)$

gets a new link in period t+1 is roughly:

$$\frac{p_r m_r}{t} + \frac{p_n m_n d_i(t)}{mt}$$

Theorem: Degree Distribution

The in-degree distribution of the mean-field process has a degree distribution with cdf

$$F(d) = 1 - [r\,m/(d + r\,m)]^{1+r}$$

- Approximates a power distribution for large d
- Lower tail is thinner

Sketch of Proof

1. Mean-field:

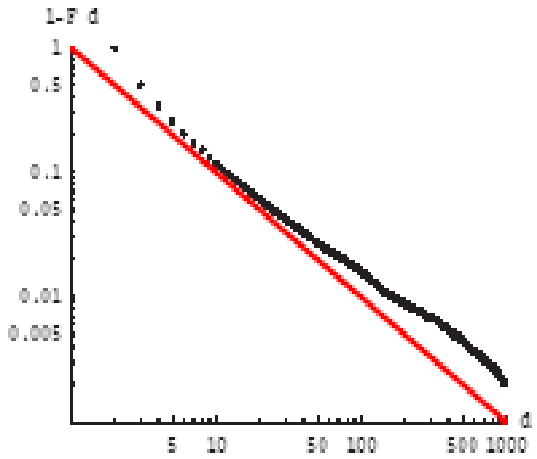
$$\partial d_i(t)/\partial t = E[\text{change}(d_i(t))]; \quad d_i(i) = 0$$

2. Solve the differential equation:

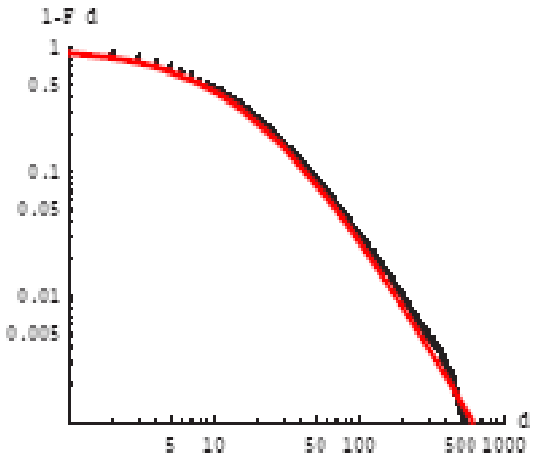
$$d_i(t) = rm(t/i)^{[1/(1+r)]-rm}$$

3. $1-F_t(d)$ = proportion of nodes with degree > d at time t
   - Let $i^*(d)$ be such that $d_{i^*(d)}(t)=d$
   - $\Rightarrow 1-F_t(d)=i^*(d)/t$
4. Solve for $i^*(d)$ from Step 2 and plug in

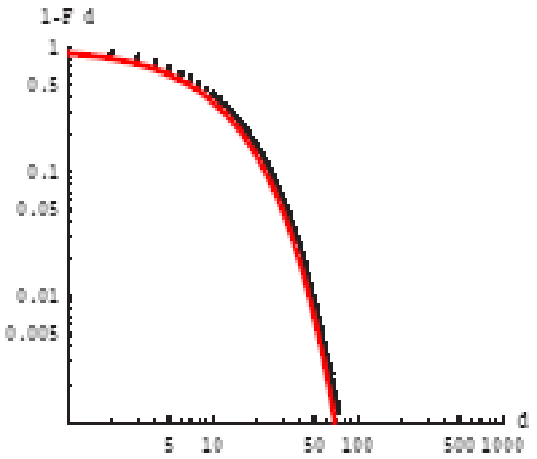
Varying r: the Relative Rate of Random Meetings



r=0

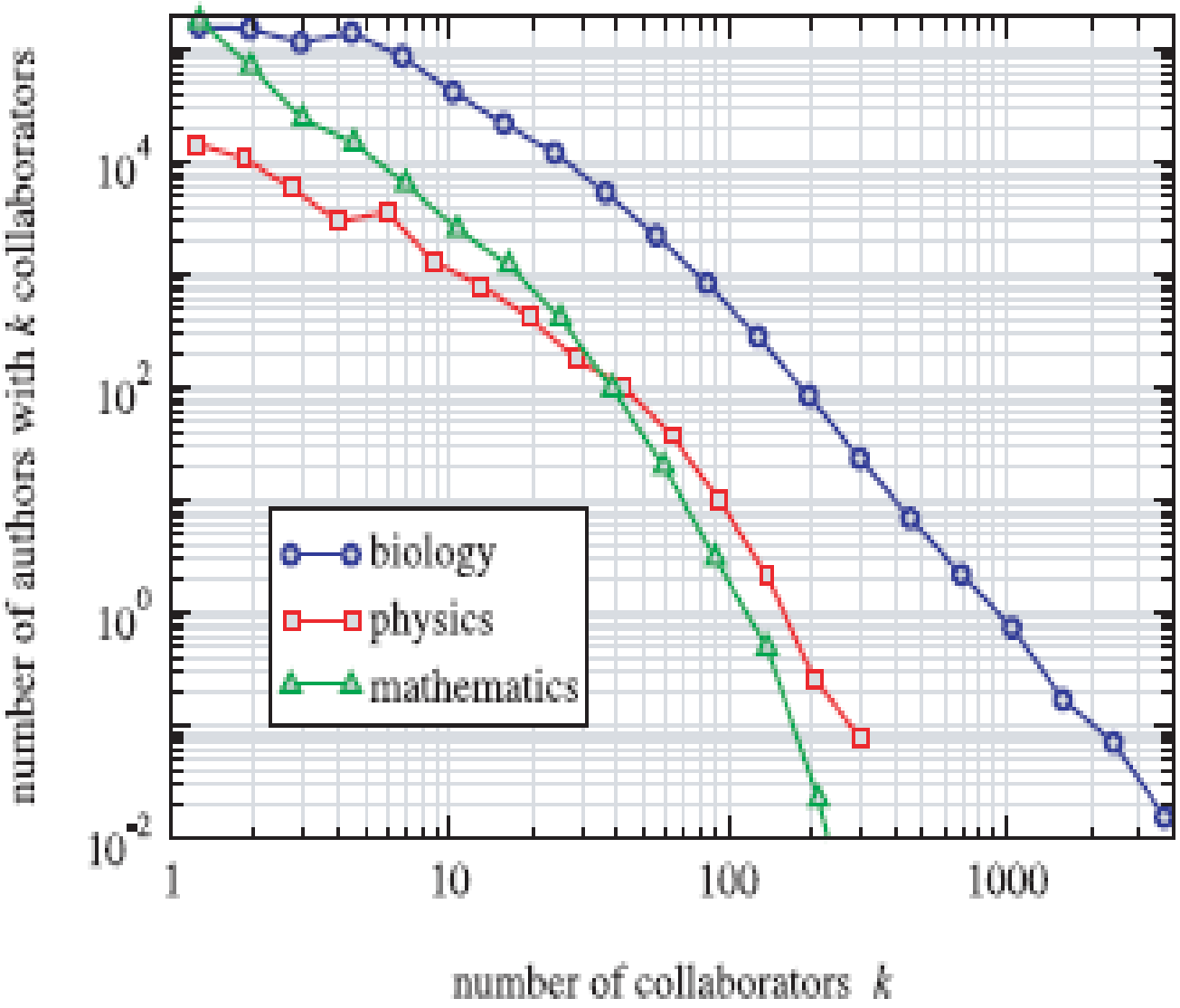


r=1



r= ∞

Co-Authorship is not entirely scale-free (same with www – Pennock et al (2002))



Clustering: Definitions

Total Clustering (weighted avg.):

$$C(g) = (\sum_i \text{links among neighbors of } i) / (\sum_i \#\text{ pairs of neighbors of } i)$$

Average Clustering (unweighted avg.):

$$C^{Avg}(g) = 1/n \sum_i [(\text{links among neighbors of } i) / (\#\text{ pairs of neighbors of } i)]$$

Theorem: Clustering

Under the mean-field approximation,

Total clustering:

- goes to 0 if $r \leq 1$
- goes to $6p / (1+r)[(3m-2)(r-1)+2mr]$ otherwise

Average clustering:

- Bounded away from 0
- $2p\int(m+d+(1/r-1)rm\log(1+d/(rm)))/[(d+m)(d+m-1)]dF(d)$

Clustering: Remarks

- Clustering decreasing in r:

  Higher clustering corresponds to more search

- Total clustering vanishes if r is too small:
  - High degree nodes have low clustering and have many of the links

- Average clustering does not vanish
  - average is not weighted by degree

Sketch of Proof

- $C^{Avg}(g)$ tends to $\int f(d)C(d)$, where C(d) is the clustering coefficient for a node of degree d
- Denominator of C(d) = (d+m)(d+m-1)/2
- Numerator of C(d) = summing up all possible combinations of links among i's neighbors
  - E.g. # links pointing to i from random meeting is $d_i^r(t)=p_r m_r \log(t/i)$
  - Compute t/i
  - Each of these nodes has $p_n m_n/m_r$ links to i's neighbors
  - Solving, one gets $rm[\log(d/rm + 1)]p_n m_n/m_r$ triads
- Substitute expressions and simplify

Theorem: Diameter

If $m_r \geq 2$ and $m_n = 1$, then the network consists of a single component with diameter proportional to log(t)/log(log(t)) almost surely.

Proof: Bollabas and Riordan (2002), very involved

Lower than diameter of random graph or rewired graph (log(t))

Theorem: Assortativity

Under the mean-field approximation (with nontrivial network-based meetings),

if $d_i(t) > d_j(t)$,

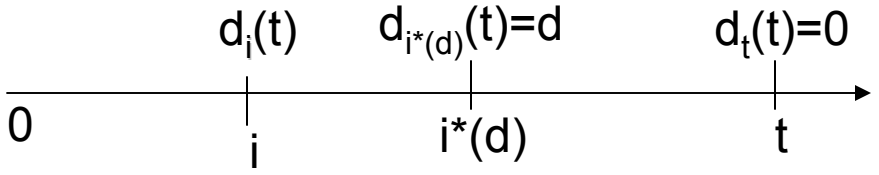
then $1\text{-}F_i^t(d) > 1\text{-}F_j^t(d)$ for all $d < d_i(t)$

- FOSD is stronger than positive correlation

Sketch of Proof

1. Obvious for $d_j(t) \le d < d_i(t)$

2. $1\text{-}F_i^t(d) = d_i(i_t^*(d))/d_i(t)$



3. Consider $d < d_j(t)$.

   Sufficient to show that for all $i < j < t' < t$,

   $$d_i(t')/d_i(t) > d_j(t')/d_j(t)$$

- Verified from solution for $d_i(t)$

Theorem: Inverse Clustering-Degree

Under the mean-field approximation (with nontrivial network-based meetings), there exists $d^* > 0$ such that for all $d > d^*$ there exists a $D > 0$ so that $C(d) > C(d')$ for all $d' > d + D$.

- Proof: Approximate $C(d)$ for large d as $2p/dm^2$
- Conjecture: $C(d)$ is strictly decreasing in d, but not able to prove this

Sketch of Proof

- Have expression for C(d) from clustering proof
- Approximate C(d) for large d as $2p/dm^2$ to simplify
- Approximation is proportional to 1/d, so result follows

Fitting the Model to Data

- Fix m by direct calculation from data
- Estimate r by fitting the degree distribution
- Use clustering coefficients to estimate p
- Simulations give accurate estimates for diameter

Parameters, Clustering and Diameter from the fits

Data Set:	WWW	Citations	Co-author	Ham Radio	Prison	High School Romance
Number of Nodes:	325729	396	81217	44	67	572
Avg. In-Degree: m	4.5	5	1.7	3.5	2.7	.84
r from Fit	0.5	0.62	3.5	5.0	590	1000
p from Fit	.33	.26	.17	1	1	-
R^2 of Fit	.97	.98	.99	.94	.94	.99
Avg. Clustering Data	.11	.07	.16	.47	.31	-
Avg. Clustering Fit	.11	.07	.16	.22	.10	-
Diameter Data	11.3 (avg)	4	26	5	7	-
Diameter Fit	(16,32)	(4,5)	(18,36)	(4,5)	5	(14,24)

Consequences of network structure

- Efficiency
- Diffusion

- Rely on stochastic dominance relationships of the degree distribution

Theorem: FOSD and MPS

- Consider a distribution F with parameters (m,r) and a distribution F’ with parameters (m’,r’). Then:
  - If r’ = r > 0 and m’ > m, then F’ strictly FOSD F
  - If m’ = m > 0 and r’ < r, then F’ is a strict MPS of F

Sketch of Proof

- FOSD: easy. F(d) is decreasing in m
- MPS:
  1. $\int_0^X [F(d)-F(d')] > 0$
  2. $-m\left([(X+r'm)/r'm - 1] - [(X+rm)/rm - 1]\right)$
  3. Show $(X/r'm + 1)^{r'} > (X/rm + 1)^{r}$
  4. Take logs and differentiate w.r.t. r

Corollary: Efficiency

- If utility to a node (information, employment, etc.) is increasing in degree, then average utility increases with m.
- If utility to a node is concave in degree, then average utility increases with r.

Concluding remarks

- Random + network-based meetings imply the features:
  - Scale-free (upper tail)
  - Clustering
  - Low diameter
  - Assortativity
  - Degree/cluster relation
- Parameters vary dramatically across applications
- Diffusion
  - Stochastic dominance is a useful and general tool here
  - Implications for long-run behavior

Concluding remarks

- Random + network-based meetings imply the features:
  - Scale-free (upper tail)
  - Clustering
  - Low diameter
  - Assortativity
  - Degree/cluster relation
- Parameters vary dramatically across applications
- Diffusion
  - Stochastic dominance is a useful and general tool here
  - Implications for long-run behavior

Extensions:

- Degree-dependent utility (c cost, u is per-link utility bound):

  $1\text{-}F(d) = [\ (d_0+rm-c/u)/(d+rm-c/u)\ ]^{m/(pm\_n)}$

- Exponential growth in number of nodes at rate g

  $1\text{-}F(d) = [\ (d_0+rm)/(d+rm)\ ]^{m \log(1+g)/(gpm\_s)}$

- Larger neighborhood search

  Lowers prob of being found by search proportionally

- Non-directed networks

  similar results to non-directed case, probability of being found under search complicated by correlation in

Utility maximization

- Utility is node specific and iid
  - Link formation probability by p
  - (Non-directed case is p^2)
- Utility is proportional to how connected a node is
  - Link formation probability is degree-dependent
  - Benefits from indirect connections
  - Correlation in values
- Start with first case, second works with some complications

Newman and Grossman's

Table 1. Summary statistics for the three coauthorship networks analyzed here

	Biology	Physics	Mathematics
Number of authors	1,520,251	52,909	253,339
Number of papers	2,163,923	98,502	—
Papers per author	6.4	5.1	6.9
Authors per paper	3.75	2.53	1.45
Average collaborators	18.1	9.7	3.9
Largest component	92%	85%	82%
Average distance	4.6	5.9	7.6
Largest distance	24	20	27
Clustering coefficient	0.066	0.43	0.15
Assortativity	0.13	0.36	0.12

The statistics are, from top to bottom, total number of authors appearing in the corresponding databases; total number of papers appearing; mean number of papers published by an author; mean number of coauthors on a paper; mean number of different individuals an author collaborated with; largest connected group of individuals in the network; mean vertex–vertex distance between connected individuals in the network; largest such distance; the clustering coefficient, which is the mean probability that two coauthors will also be coauthors of one another; and the degree assortativity coefficient, which is the Pearson correlation coefficient of the degrees (i.e., number of collaborators) of adjacent vertices in the network. The material shown here is after Newman (12) and Grossman (9).

Why does lognormal (preferential attach.) +growth work?

- $d(d_i)/dt = k\, d_i/t$
- Implies $d_i = d_0(t/i)^k$
- Growing system: $1\text{-}F(d) = i(d)/t$
- Here $i(d)/t = (d_0/d)^{1/k}$
- So $1\text{-}F(d) = (d_0/d)^{1/k}$

Other Characteristics

- m=5 on average in data
- Estimate r = .5 (R^2 is .97)
- Clustering coefficient: p=1/3 gives 0.11
  - Data: 0.11 (Adamic)
- Diameter: bracketed 16 to 32
  - Data: 20

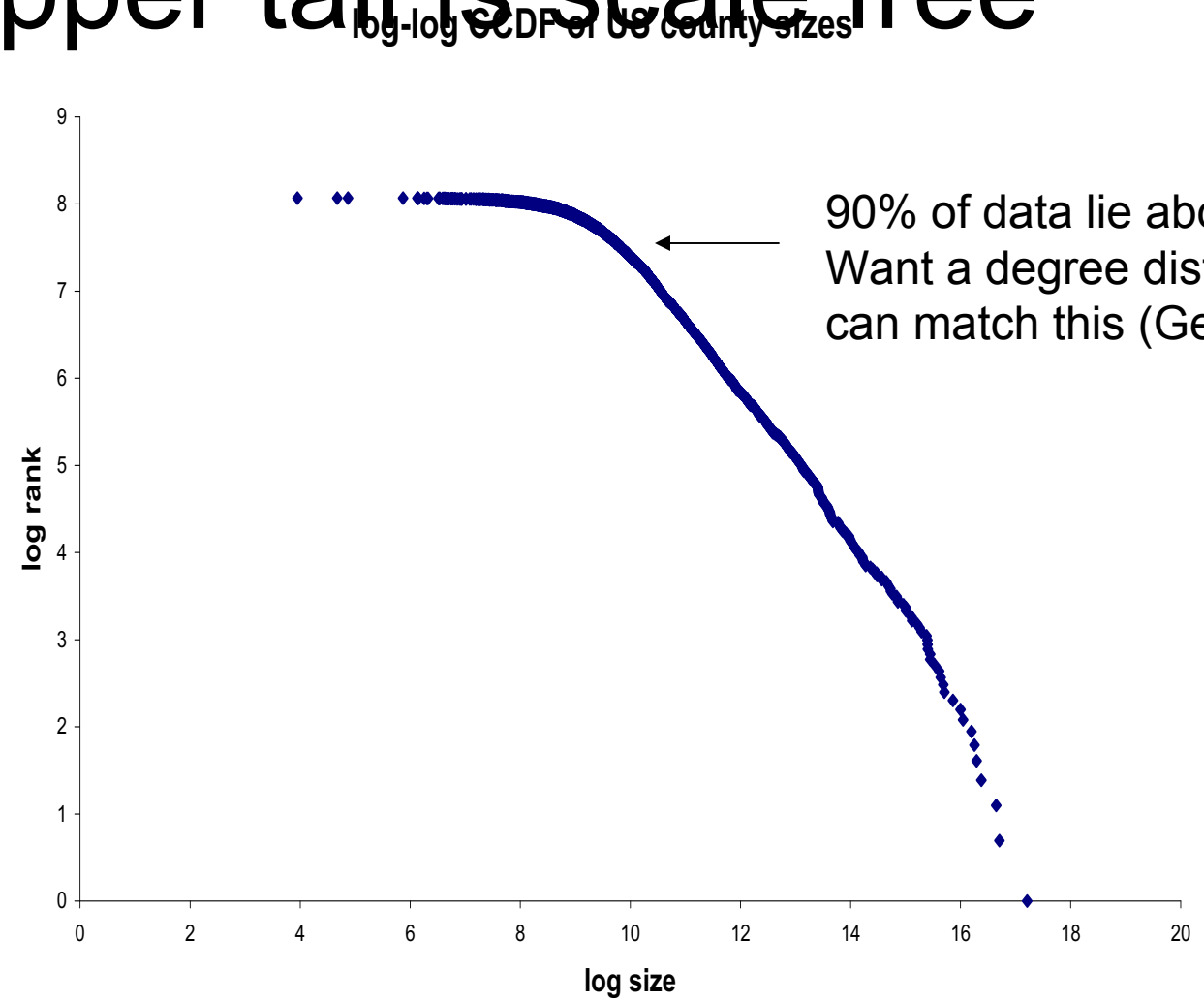
Fitting Economics Co-author data

- m=1 in data (non-directed, links initiated)
- Estimate r=3.54 (R^2 is .99)
- Clustering coefficient: p=.17 gives 0.17
  - Data: .167 overall (.149 largest component)
- Diameter: bracketed 18 to 36
  - Data: 26

Other data sets:

- Random/Search:
  - WWW links: r=.5
  - Small World Citation: r=.62
  - Econ co-authors: r=3.5
  - Ham radio: r=5
  - Prison Friendships: r=590
  - High School Romances: r=1000

City Sizes (Zipf's Law) – only the upper tail is scale free



90% of data lie above here
Want a degree distribution that can match this (Gell-Mann)

Proposition: Degree/Clustering

Under the mean field approximation to the search model (with nontrivial search), C(d) is approximated by $2p/dm^2$ for large d and t, and so C(d) > C(d'') for large enough d'' > d

Previous Models:

- Erdos and Renyi (1960) – random graphs
  - low clustering, not scale free, diameter: log(n)
- Watts-Strogatz (1998) - rewire lattice
  - high clustering, not scale free, diameter: log(n)
- Barabasi-Albert (1999) ... - preferential attach
  - low clustering, scale free, diameter: log(n)/log(log(n))
- Carlson-Doyle (1999) ... – HOT (optimized)
  - clustering?, scale free, diameter?(tech. not social)
- Growing models (Krapivsky-Redner; Callaway et al):
  - assortative
- Mixed Models (Kleinberg et al, Kumar et al)
  - random or copying - scale free, but no clustering, diameter...

Previous Models:

- Random Graphs (Erdos and Renyi (1960)...)
  - diameter: log(n)
- Perturbed Lattices (Watts-Strogatz (1998)...)
  - high clustering, diameter: log(n)
- Preferential Attachment (Barabasi-Albert (1999) ...)
  - scale free, diameter: log(n) or smaller
- Mixed Models (Kleinberg et al (1999), Kumar et al (2000), Vazquez(2003),...)
  - scale free(?), sometimes clustering

Power Laws: Rediscovering the Wheel

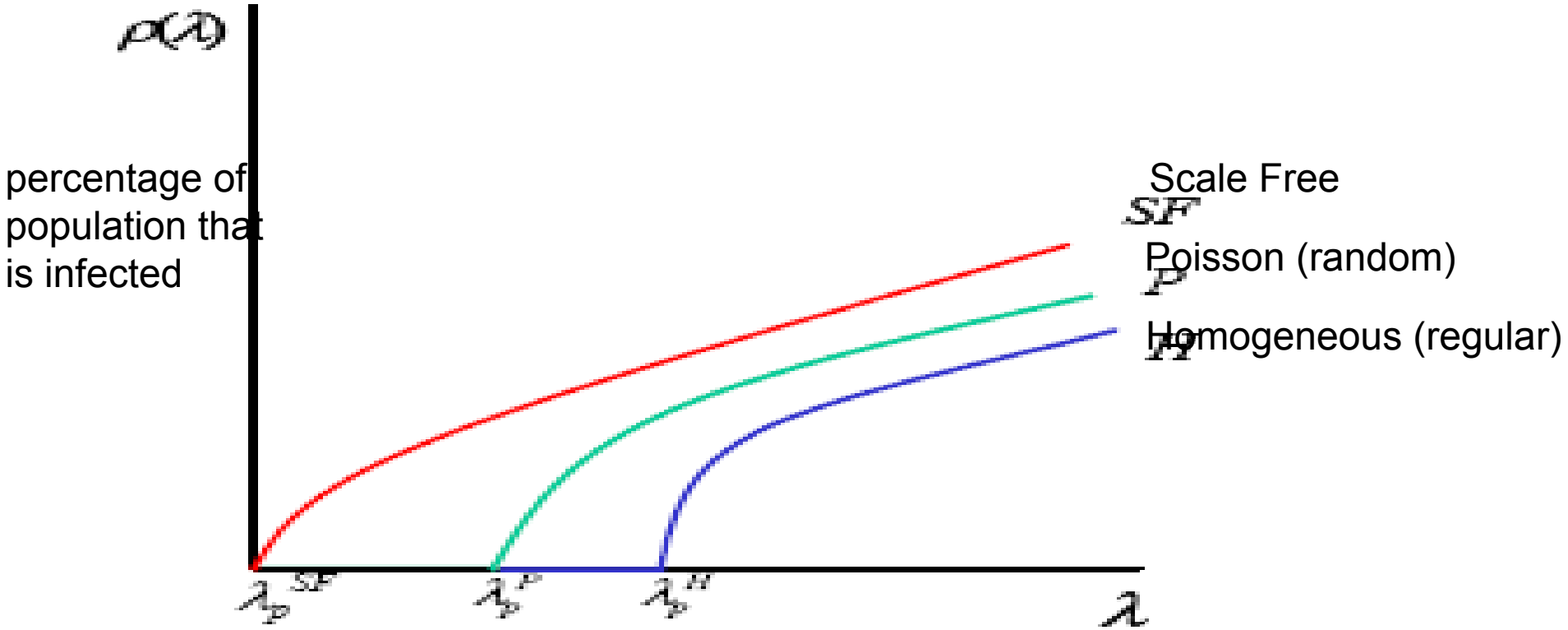
- Pareto (1896), Yule (1925), Zipf (1949), Simon (1955), Price (1965), Albert, Jeong, Barabasi (1999), ...

- Key Ingredients (Simon):
  - Growing system
  - Lognormal growth of objects (e.g.,degree of nodes)
    - nodes gain links with probability proportional to current size
      - exponential growth – linear on log-log graph

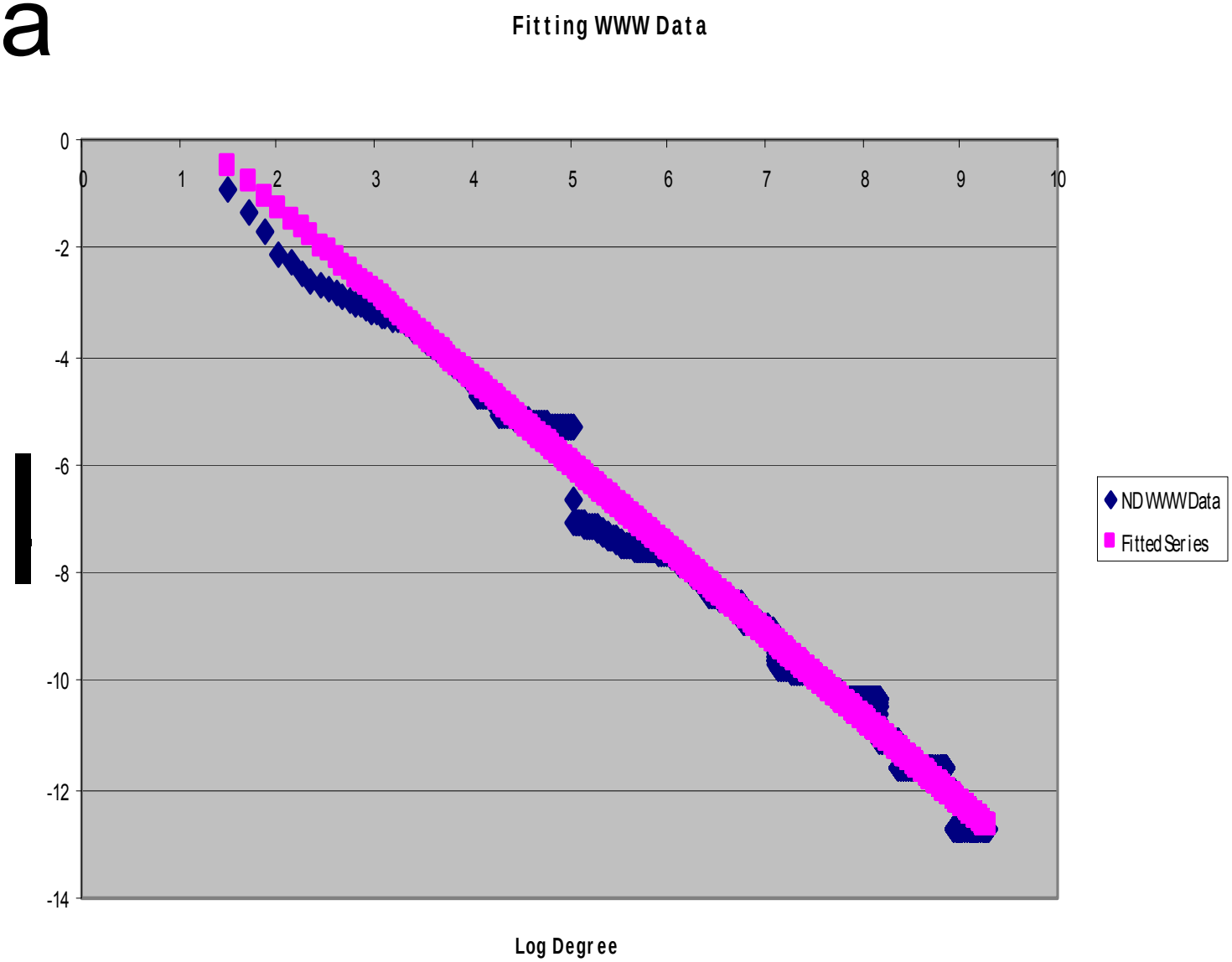
Why Should We care?

- Diffusion of viruses, information, behavior...
- Pastor-Satorras and Vespignani (2001), Lopez-Pintado (2003), ..., SIS models:
  - Probability a sick neighbor infects a healthy one
  - Probability a sick node gets healthy
  - When does a virus die out without infecting a significant portion of the population?
  - For viruses that survive, what portion of population is infected at a given time?

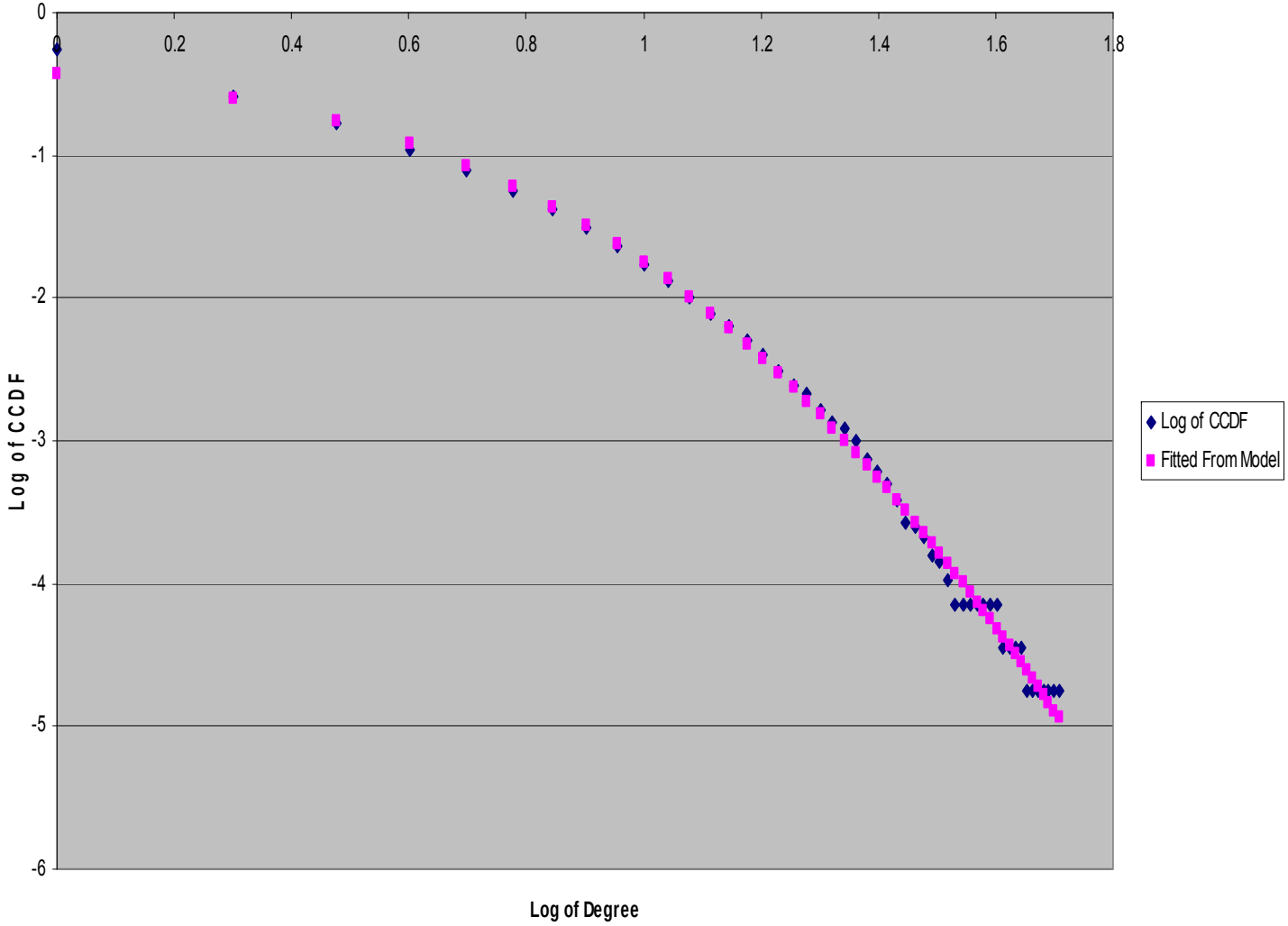
Why Should we care: Lopez-Pintado - infection spreading



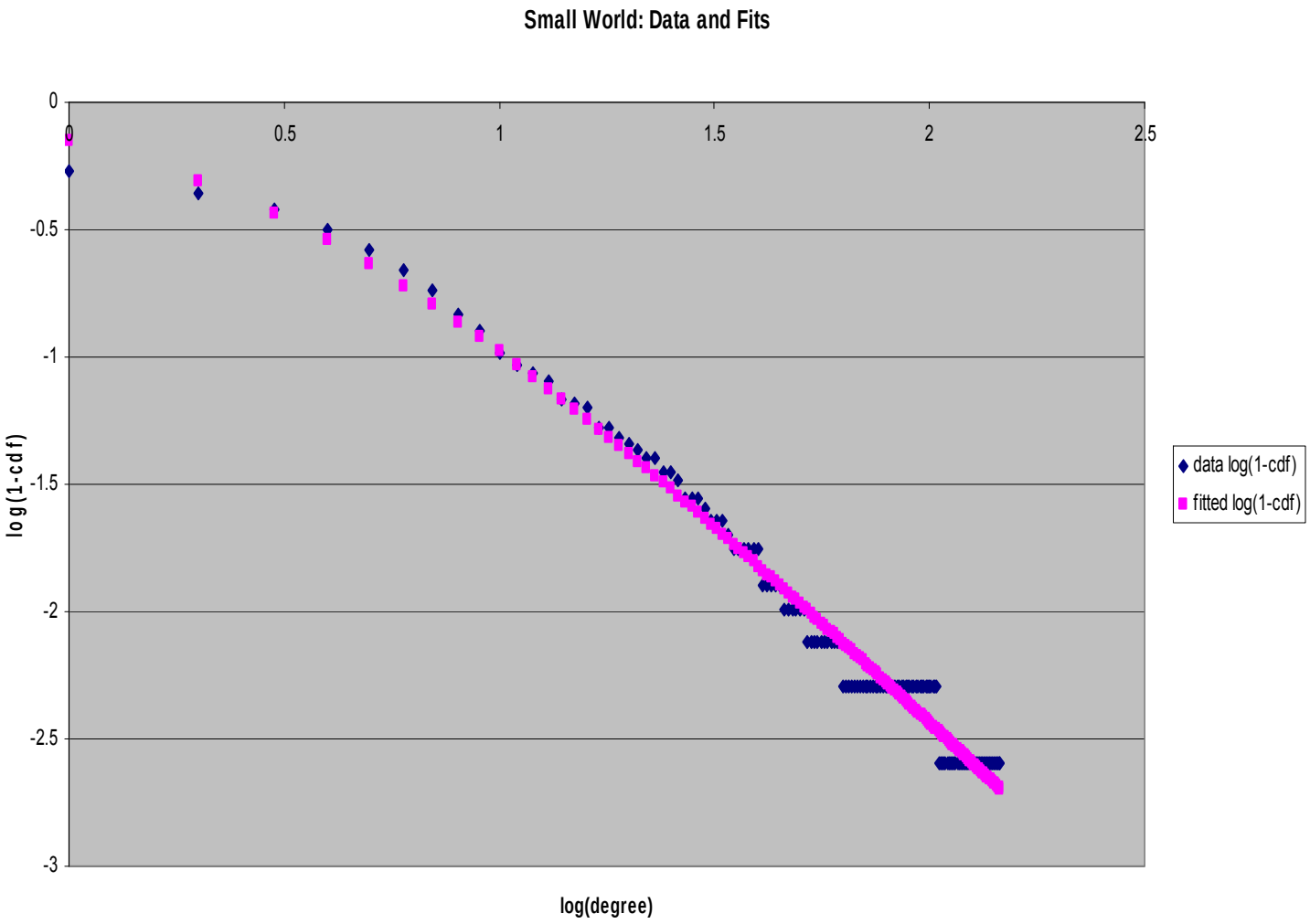
Comparison: fitting the www data



Co-author data of Goyal et al



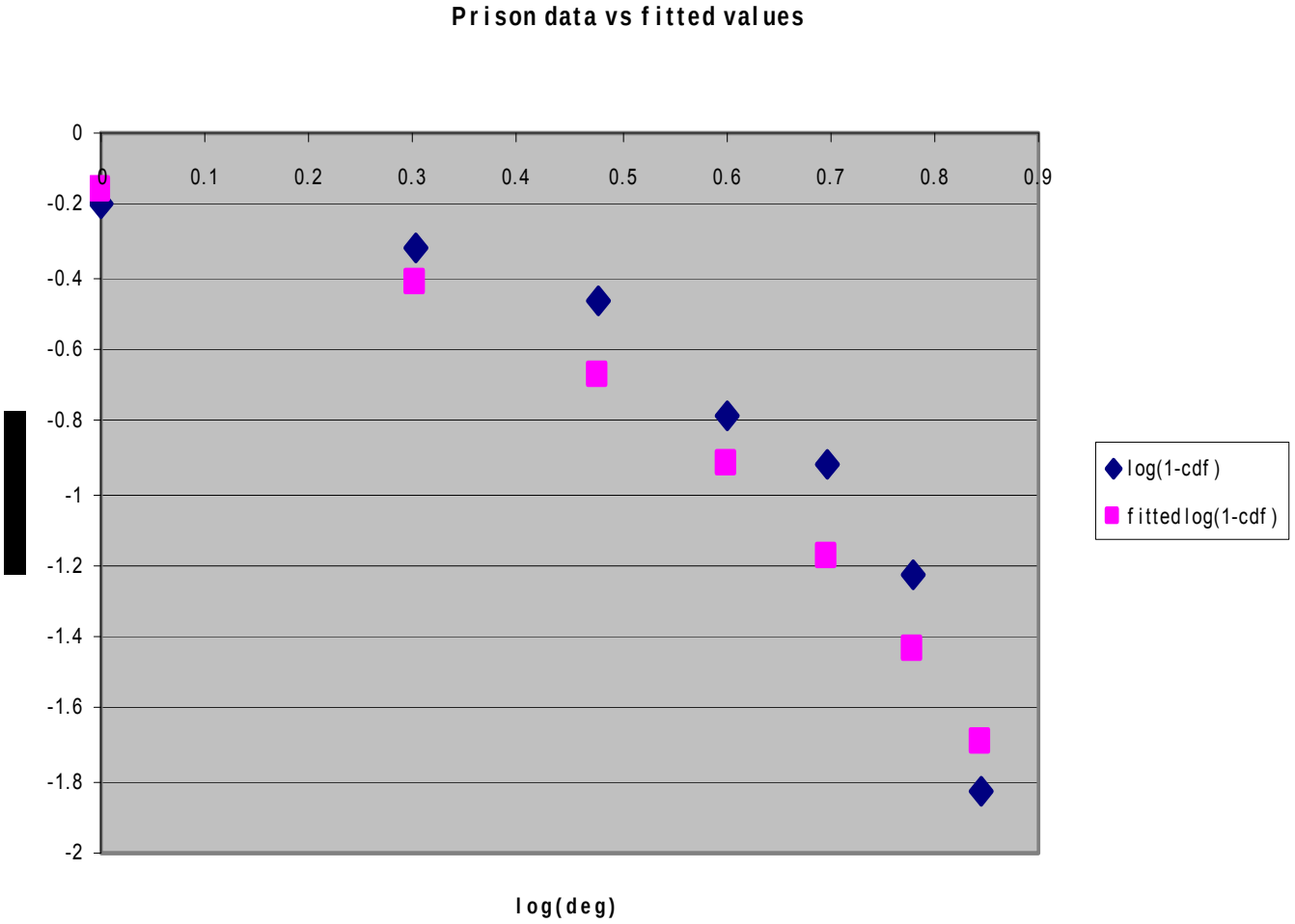
Other data sets we have fit:



citation data: Milgram's small worlds. (Mcrae (2003)

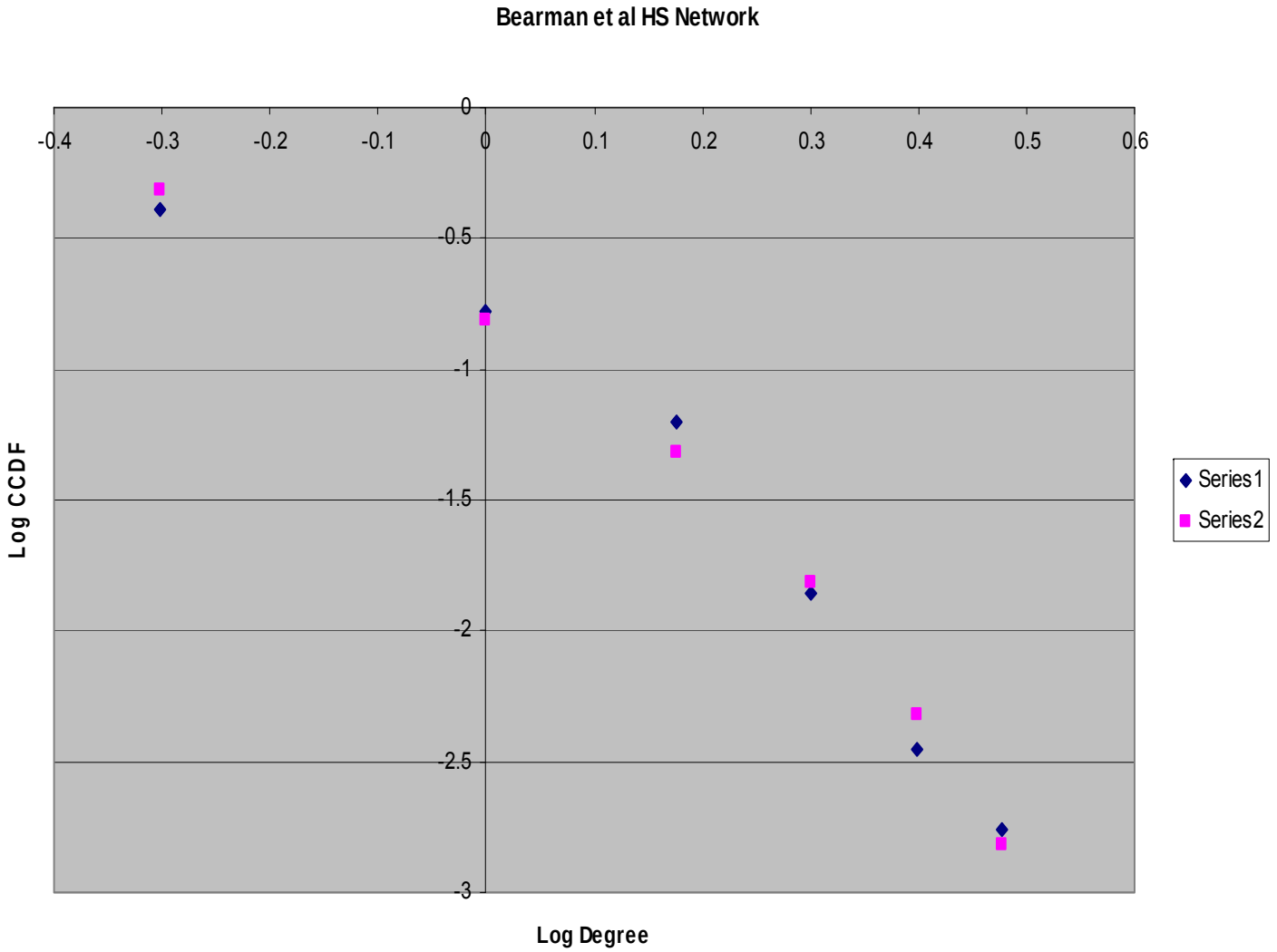
fit: m=5, r=.62
R-sq .98
n=396

Garfields (1960) Inmate network



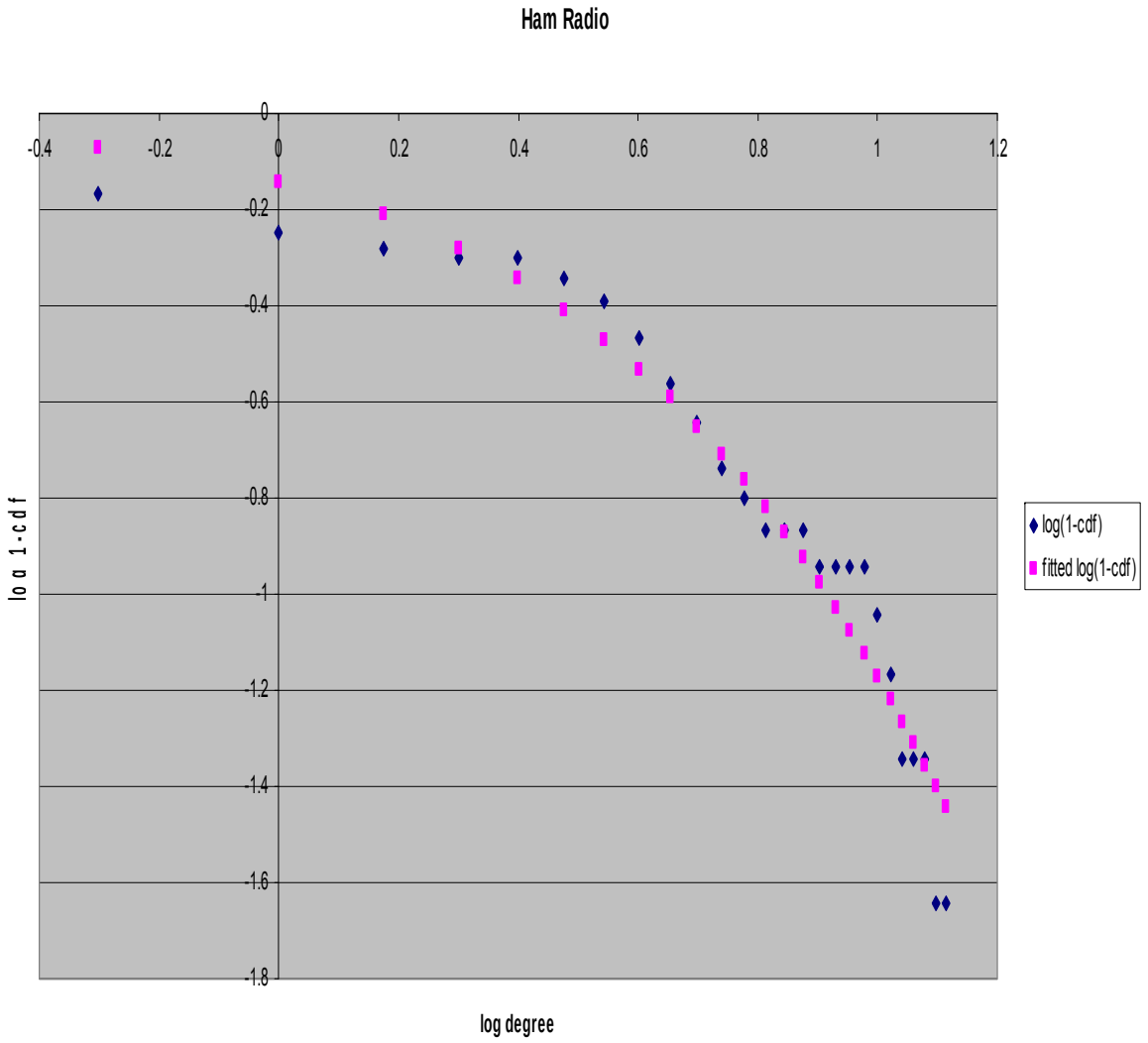
Fit: r=590
m=2.7
R-sq ..94
n=67

High School Romance Network



Bearman et al (2004) midwest high school romance network
m=.84
r=1000
R-sq .99
n=572

Ham Radio



r=5

m=3.5

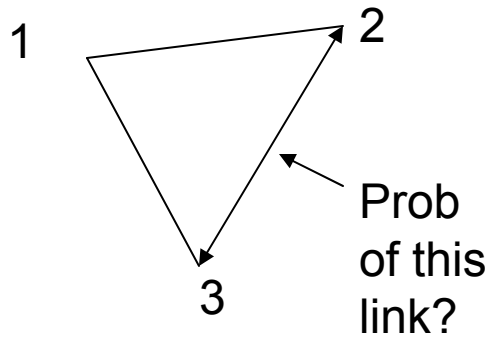
n=44

R-sq=.94

Stylized Facts: Small diameter

- Milgram (1967) letter experiments
  - median 5 for the 25% that made it
- Examples:
  - Actors in same movie (Kevin Bacon Oracle)
    - Watts and Strogatz (1998) – mean 3.7
  - Co-Authorship studies
    - Grossman (1999) Math mean 7.6, max 27,
    - Newman (2001) Physics mean 5.9, max 20
    - Goyal et al (2004) Economics mean 9.5, max 29
  - WWW
    - Adamic, Pitkow (1999) – mean 3.1 (85.4% possible of 50M pages)

Stylized Facts: High Clustering Coefficients



- Watts and Strogatz (1998)
  - .79 for movie acting
- Newman (2001) co-authorship
  - .496 CS, .43 physics, .15 math, .07 biomed
- Adamic (1999)
  - .11 for web links (versus .0002 for random graph of same size)

Power Laws: scale-free

- Plot of log(frequency) versus log(degree) is approximately linear

- prob(degree) = c degree^{-a}
  - log[prob(degree)] = log[c] – a log[degree]
  - 1-cdf(degree) = c' degree^{1-a}

- Fat tails compared to random network

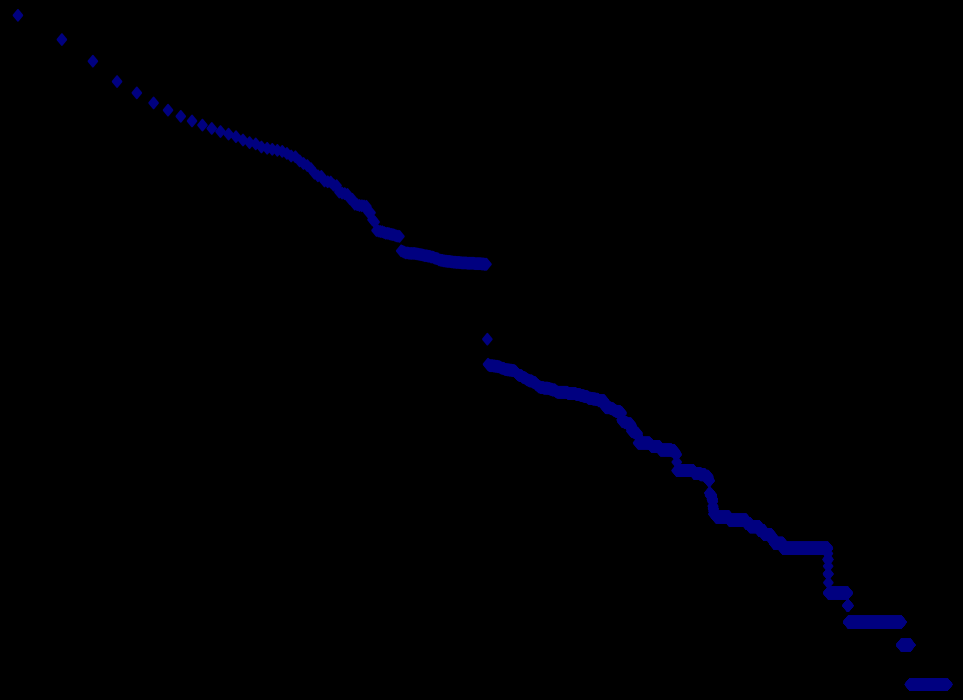
Comparison:

- Co-author calibration: 3.5/1 random to search
- WWW calibration: 1/2 random to search
- Co-author network is 7 times more “random’’

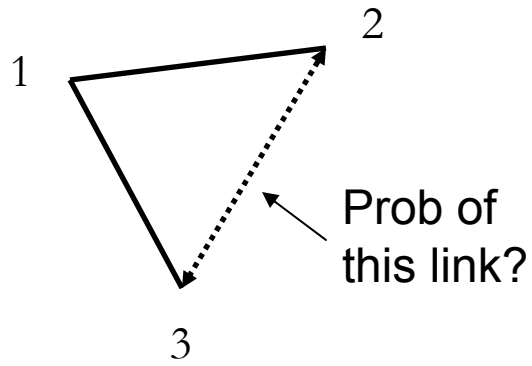
A Few Analytic Results

- Degree distribution:
  - $F(d) = 1 - (rm)^{1+r} (d + rm)^{-(1+r)}$
  - Approximates a power distribution for large d
  - Thinner lower tail
- Clustering:
  - Bounded away from 0
  - Decreasing in r: Less search → lower clustering
- Diameter:
  - If $m_r \geq 2$ and $m_s = 1$, proportional to log(t)/log(log(t))

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



Clustering: A Definition



Average Clustering:

$$C^{Avg}(g) = \sum_i (\text{links among neighbors of i}) / (\text{\# pairs of neighbors of i})\ n$$

Theorem: Clustering

Under the mean-field approximation, average clustering tends to

$$2p\int(m+d+(1/r-1)rm\log(1+d/(rm)))/[(d+m)(d+m-1)]dF(d)$$

- Bounded away from 0
- Decreasing in r:
  - Less search – lower clustering