UNIVERSITY OF TWENTE.

Uncovering disassortativity in large scale-free networks

Nelly Litvak

University of Twente,
Stochastic Operations Research group

Joint work with Remco van der Hofstad

Supported by EC FET Open project NADINE

Trento, Italy, 23-07-2012

Power laws

- degree of the node = # links, [fraction nodes degree k] = p_k,

Power laws

- degree of the node = # links, [fraction nodes degree k] = p_k,
- Power law: $p_k \approx const \cdot k^{-\alpha}$, $\alpha > 1$.

Power laws

- degree of the node = # links, [fraction nodes degree k] = p_k,
- Power law: $p_k \approx const \cdot k^{-\alpha}$, $\alpha > 1$.
- Power laws: Internet, WWW, social networks, biological networks, etc...

Power laws

- degree of the node = # links, [fraction nodes degree k] = p_k,
- Power law: $p_k \approx const \cdot k^{-\alpha}$, $\alpha > 1$.
- Power laws: Internet, WWW, social networks, biological networks, etc...
- Model for high variability, scale-free graph

Power laws

- degree of the node = # links, [fraction nodes degree k] = p_k,
- Power law: $p_k \approx const \cdot k^{-\alpha}$, $\alpha > 1$.
- Power laws: Internet, WWW, social networks, biological networks, etc...
- Model for high variability, scale-free graph
- signature log-log plot: $\log p_k = log(const) - \alpha \log k$

Power laws

- degree of the node = # links, [fraction nodes degree k] = p_k,
- Power law: $p_k \approx const \cdot k^{-\alpha}$, $\alpha > 1$.
- Power laws: Internet, WWW, social networks, biological networks, etc...
- Model for high variability, scale-free graph
- signature log-log plot: $\log p_k = log(const) - \alpha \log k$
- Faloutsos, Faloutsos, Faloutsos (1999): power laws in Internet

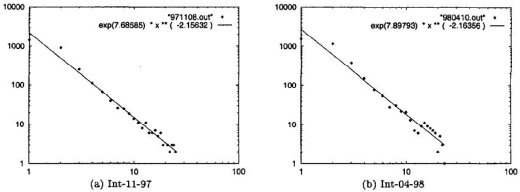


Figure 5: The outdegree plots: Log-log plot of frequency f_d versus the outdegree d.

But Power Law is not everything!

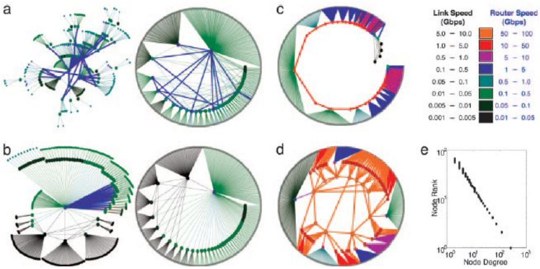
Example: Robustness of the Internet.

- Albert, Jeong and Barabasi (2000): Achille's heel of Internet: Internet is sensitive to targeted attack

But Power Law is not everything!

Example: Robustness of the Internet.

- Albert, Jeong and Barabasi (2000): Achille's heel of Internet: Internet is sensitive to targeted attack
- Doyle et al. (2005): Robust yet fragile nature of Internet: Internet is not a random graph, it is designed to be robust



But Power Law is not everything! (cont.)

Example: Spread of infections

- Classical epidemiology, e.g. Adnerson and May (1991): epidemic only if infection rate exceeds a critical value

But Power Law is not everything! (cont.)

Example: Spread of infections

- Classical epidemiology, e.g. Adnerson and May (1991): epidemic only if infection rate exceeds a critical value
- Vespignani et al. (2001): power law networks have a zero critical infection rate!

But Power Law is not everything! (cont.)

Example: Spread of infections

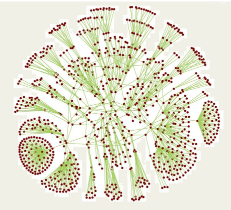
- Classical epidemiology, e.g. Adnerson and May (1991): epidemic only if infection rate exceeds a critical value
- Vespignani et al. (2001): power law networks have a zero critical infection rate!
- Eguiluz et al. (2002): a specially wired highly clustered network is resistant up to a certain critical infection rate.

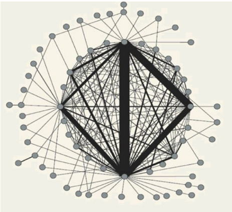
But Power Law is not everything! (cont.)

Example: Spread of infections

- Classical epidemiology, e.g. Adnerson and May (1991): epidemic only if infection rate exceeds a critical value
- Vespignani et al. (2001): power law networks have a zero critical infection rate!
- Eguiluz et al. (2002): a specially wired highly clustered network is resistant up to a certain critical infection rate.

Example: Technological versus economical networks





Degree-degree correlations

- It is clearly important how the network is wired

Degree-degree correlations

- It is clearly important how the network is wired
- To start with: do hubs connect to each other?

Degree-degree correlations

- It is clearly important how the network is wired
- To start with: do hubs connect to each other?
  YES for banks, NO for Internet

Degree-degree correlations

- It is clearly important how the network is wired
- To start with: do hubs connect to each other?
  YES for banks, NO for Internet
- Assortative networks: nodes with similar degree connect to each other.
- Disassortative networks: nodes with large degrees tend to connect to nodes with small degrees.

Assortativity coefficient

- $G = (V, E)$ undirected graph of n nodes
- d_i degree of node $i = 1, 2, \ldots, n$

Assortativity coefficient

- $G = (V, E)$ undirected graph of n nodes
- d_i degree of node $i = 1, 2, \ldots, n$
- We are interested in correlations between degrees of neighboring nodes

Assortativity coefficient

- $G = (V, E)$ undirected graph of n nodes
- d_i degree of node $i = 1, 2, \dots, n$
- We are interested in correlations between degrees of neighboring nodes
- Newman (2002): assortativity measure ρ_n

$$\rho_n = \frac{\frac{1}{|E|}\sum_{ij\in E} d_i d_j - \left(\frac{1}{|E|}\sum_{ij\in E}\frac{1}{2}(d_i + d_j)\right)^2}{\frac{1}{|E|}\sum_{ij\in E}\frac{1}{2}(d_i^2 + d_j^2) - \left(\frac{1}{|E|}\sum_{ij\in E}\frac{1}{2}(d_i + d_j)\right)^2}$$

- Statistical estimation of the correlation coefficient between degrees on two ends of a random edge

Assortativity coefficient

- $G = (V, E)$ undirected graph of n nodes
- d_i degree of node $i = 1, 2, \ldots, n$
- We are interested in correlations between degrees of neighboring nodes
- Newman (2002): assortativity measure ρ_n

$$\rho_n = \frac{\frac{1}{|E|}\sum_{ij\in E} d_i d_j - \left(\frac{1}{|E|}\sum_{ij\in E}\frac{1}{2}(d_i + d_j)\right)^2}{\frac{1}{|E|}\sum_{ij\in E}\frac{1}{2}(d_i^2 + d_j^2) - \left(\frac{1}{|E|}\sum_{ij\in E}\frac{1}{2}(d_i + d_j)\right)^2}$$

- Statistical estimation of the correlation coefficient between degrees on two ends of a random edge
- Very popular measure of assortativity!

Is there something wrong with ρ_n?

- Preferential Attachment graph appears to be assortatively neutral (Newman 2003, Dorogovtsev et al. 2010)
- Recent criticism: ρ_n depends on the size of the networks (Raschke et al. 2010; Dorogovtsev et al. 2010)

What IS assortativity measure?

- ρ_n is a statistical estimation for the coefficient of variation
$$\rho = \frac{E(XY) - [E(X)]^2}{Var(X)},$$
- X and Y are the degrees of the nodes on the two ends of a randomly chosen edge

What IS assortativity measure?

- ρ_n is a statistical estimation for the coefficient of variation
$$\rho = \frac{E(XY) - [E(X)]^2}{Var(X)},$$
- X and Y are the degrees of the nodes on the two ends of a randomly chosen edge
- Problems?

What IS assortativity measure?

- ρ_n is a statistical estimation for the coefficient of variation

$$\rho = \frac{E(XY) - [E(X)]^2}{Var(X)},$$

- X and Y are the degrees of the nodes on the two ends of a randomly chosen edge
- Problems? YES!!!

What IS assortativity measure?

- ρ_n is a statistical estimation for the coefficient of variation

$$\rho = \frac{E(XY) - [E(X)]^2}{Var(X)},$$

- X and Y are the degrees of the nodes on the two ends of a randomly chosen edge
- Problems? YES!!!
- X and Y are power law r.v.'s, exponent $\alpha - 1$

$$P(X = k) = kp_k / E(\text{degree}).$$

- In real networks (WWW) we often have $2 < \alpha < 3$, so
$$E(X) = \sum_k k \frac{kp_k}{E(\text{degree})} = \infty$$

What IS assortativity measure?

- ρ_n is a statistical estimation for the coefficient of variation
$$\rho = \frac{E(XY) - [E(X)]^2}{Var(X)},$$
- X and Y are the degrees of the nodes on the two ends of a randomly chosen edge
- Problems? YES!!!
- X and Y are power law r.v.'s, exponent $\alpha - 1$
$$P(X = k) = kp_k / E(\text{degree}).$$
- In real networks (WWW) we often have $2 < \alpha < 3$, so
$$E(X) = \sum_k k \frac{kp_k}{E(\text{degree})} = \infty$$
- ρ is not defined in the power law model! Then: what are we measuring?

Assortative and disassortative graphs

- Newman(2003)

	network	type	size n	assortativity r	error σ_r	ref.
social	physics coauthorship	undirected	52 909	0.363	0.002	a
	biology coauthorship	undirected	1 520 251	0.127	0.0004	a
	mathematics coauthorship	undirected	253 339	0.120	0.002	b
	film actor collaborations	undirected	449 913	0.208	0.0002	c
	company directors	undirected	7 673	0.276	0.004	d
	student relationships	undirected	573	−0.029	0.037	e
	email address books	directed	16 881	0.092	0.004	f
technological	power grid	undirected	4 941	−0.003	0.013	g
	Internet	undirected	10 697	−0.189	0.002	h
	World-Wide Web	directed	269 504	−0.067	0.0002	i
	software dependencies	directed	3 162	−0.016	0.020	j
biological	protein interactions	undirected	2 115	−0.156	0.010	k
	metabolic network	undirected	765	−0.240	0.007	l
	neural network	directed	307	−0.226	0.016	m
	marine food web	directed	134	−0.263	0.037	n
	freshwater food web	directed	92	−0.326	0.031	o

Assortative and disassortative graphs

- Newman(2003)

	network	type	size n	assortativity r	error σ_r	ref.
social	physics coauthorship	undirected	52 909	0.363	0.002	a
	biology coauthorship	undirected	1 520 251	0.127	0.0004	a
	mathematics coauthorship	undirected	253 339	0.120	0.002	b
	film actor collaborations	undirected	449 913	0.208	0.0002	c
	company directors	undirected	7 673	0.276	0.004	d
	student relationships	undirected	573	−0.029	0.037	e
	email address books	directed	16 881	0.092	0.004	f
technological	power grid	undirected	4 941	−0.003	0.013	g
	Internet	undirected	10 697	−0.189	0.002	h
	World-Wide Web	directed	269 504	−0.067	0.0002	i
	software dependencies	directed	3 162	−0.016	0.020	j
biological	protein interactions	undirected	2 115	−0.156	0.010	k
	metabolic network	undirected	765	−0.240	0.007	l
	neural network	directed	307	−0.226	0.016	m
	marine food web	directed	134	−0.263	0.037	n
	freshwater food web	directed	92	−0.326	0.031	o

- Technological and biological networks are disassortative, $\rho_n < 0$
- Social networks are assortative, $\rho_n > 0$

Assortative and disassortative graphs

- Newman(2003)

	network	type	size n	assortativity r	error σ_r	ref.
social	physics coauthorship	undirected	52 909	0.363	0.002	a
	biology coauthorship	undirected	1 520 251	0.127	0.0004	a
	mathematics coauthorship	undirected	253 339	0.120	0.002	b
	film actor collaborations	undirected	449 913	0.208	0.0002	c
	company directors	undirected	7 673	0.276	0.004	d
	student relationships	undirected	573	−0.029	0.037	e
	email address books	directed	16 881	0.092	0.004	f
technological	power grid	undirected	4 941	−0.003	0.013	g
	Internet	undirected	10 697	−0.189	0.002	h
	World-Wide Web	directed	269 504	−0.067	0.0002	i
	software dependencies	directed	3 162	−0.016	0.020	j
biological	protein interactions	undirected	2 115	−0.156	0.010	k
	metabolic network	undirected	765	−0.240	0.007	l
	neural network	directed	307	−0.226	0.016	m
	marine food web	directed	134	−0.263	0.037	n
	freshwater food web	directed	92	−0.326	0.031	o

- Technological and biological networks are disassortative, $\rho_n < 0$
- Social networks are assortative, $\rho_n > 0$
- Note: large networks are never strongly disassortative...

ρ_n in terms of moments of the degrees

- Write

$$\sum_{ij\in E} \tfrac{1}{2}(d_i + d_j) = \sum_{i\in V} d_i^2, \qquad \sum_{ij\in E} \tfrac{1}{2}(d_i^2 + d_j^2) = \sum_{i\in V} d_i^3$$

ρ_n in terms of moments of the degrees

- Write

$$\sum_{ij\in E} \tfrac{1}{2}(d_i + d_j) = \sum_{i\in V} d_i^2, \qquad \sum_{ij\in E} \tfrac{1}{2}(d_i^2 + d_j^2) = \sum_{i\in V} d_i^3$$

- Then

$$\rho_n = \frac{\sum_{ij\in E} d_i d_j - \frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2}{\sum_{i\in V} d_i^3 - \frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2}.$$

Extreme value theory

Theorem (Extreme value theory)

$D_1, D_2, \ldots, D_n$ are i.i.d. with $1 - F(x) = P(D > x) = Cx^{-\alpha+1}$. Then

$$\lim_{n\to\infty} P\left(\frac{\max\{D_1, D_2, \ldots, D_n\} - b_n}{a_n} \leqslant x\right) = \exp(-(1+\delta x)^{-1/\delta}),$$

with $\delta = 1/(\alpha - 1)$, $a_n = \delta C^\delta n^\delta$, $b_n = C^\delta n^\delta$.
(Therefore, the maximum is 'of the order' $n^{1/(\alpha-1)}$)

CLT for heavy tails

Theorem (CLT for heavy tails)

$D_1, D_2, \ldots, D_n$ are i.i.d. with $1 - F(x) = P(D > x) = Cx^{-\alpha+1}$. If $p > \alpha - 1$ then

$$\frac{1}{a_n}\sum_{i=1}^{n} X_i^p \xrightarrow{d} Z,$$

where $a_n = [1 - F]^{-1}(1/n^p) = C^{1/(\alpha-1)} n^{p/(\alpha-1)}$ and Z has a stable distribution with parameter $(\alpha - 1)/p$. (Therefore, the sum is 'of the order' $n^{p/(\alpha-1)}$)

In the empirical setting

- $P(d_1 \geqslant x) \approx Cx^{-\alpha+1}$

In the empirical setting

- $P(d_1 \geqslant x) \approx Cx^{-\alpha+1}$
- $\max\{d_1, d_2, \ldots, d_n\} = O(n^{1/(\alpha-1)})$
- Alternative interpretation for the maximum:
  $P(d \geqslant x) = 1/n \Rightarrow x = O(n^{1/(\alpha-1)})$

In the empirical setting

- $P(d_1 \geqslant x) \approx Cx^{-\alpha+1}$
- $\max\{d_1, d_2, \ldots, d_n\} = O(n^{1/(\alpha-1)})$
- Alternative interpretation for the maximum:
  $P(d \geqslant x) = 1/n \Rightarrow x = O(n^{1/(\alpha-1)})$
- $\mathbb{P}(d_i = k) = p_k = const \cdot k^{-\alpha}$, usually $\alpha \in (2, 4)$

In the empirical setting

- $P(d_1 \geqslant x) \approx Cx^{-\alpha+1}$
- $\max\{d_1, d_2, \ldots, d_n\} = O(n^{1/(\alpha-1)})$
- Alternative interpretation for the maximum:
  $P(d \geqslant x) = 1/n \Rightarrow x = O(n^{1/(\alpha-1)})$
- $\mathbb{P}(d_i = k) = p_k = const \cdot k^{-\alpha}$, usually $\alpha \in (2, 4)$
- If $p > \alpha - 1$ then $\mathbb{E}(D^p) = \infty$

In the empirical setting

- $P(d_1 \geqslant x) \approx Cx^{-\alpha+1}$
- $\max\{d_1, d_2, \dots, d_n\} = O(n^{1/(\alpha-1)})$
- Alternative interpretation for the maximum: $P(d \geqslant x) = 1/n \Rightarrow x = O(n^{1/(\alpha-1)})$
- $\mathbb{P}(d_i = k) = p_k = const \cdot k^{-\alpha}$, usually $\alpha \in (2, 4)$
- If $p > \alpha - 1$ then $\mathbb{E}(D^p) = \infty$
- CLT: for $p > \alpha - 1$ holds

$$\frac{1}{n} \sum_{i \in V} d_i^p \sim c_p n^{p/(\alpha-1)-1},$$

- But we get the same result just by adding up $k^p p_k$ from $k = 1$ to $k = n^{1/(\alpha-1)}$.

Assumptions

$$cn \leqslant |E| \leqslant Cn, (SLLN)$$
$$cn^{1/(\alpha-1)} \leqslant \max_{i\in[n]} d_i \leqslant Cn^{1/(\alpha-1)},$$
$$cn^{\max\{p/(\alpha-1),1\}} \leqslant \sum_{i\in[n]} d_i^p \leqslant Cn^{\max\{p/(\alpha-1),1\}}, \quad p = 2, 3,$$

where $C, c > 0$.

Assumptions

$$cn \leqslant |E| \leqslant Cn, (SLLN)$$
$$cn^{1/(\alpha-1)} \leqslant \max_{i\in[n]} d_i \leqslant Cn^{1/(\alpha-1)},$$
$$cn^{\max\{p/(\alpha-1),1\}} \leqslant \sum_{i\in[n]} d_i^p \leqslant Cn^{\max\{p/(\alpha-1),1\}}, \quad p = 2, 3,$$

where $C, c > 0$.
Very natural and non-restrictive assumptions for power law graphs.

Back to ρ_n

$$\rho_n = \frac{\textit{crossproducts} - \textit{expectation}^2}{\textit{variance}} \geqslant - \frac{\textit{expectation}^2}{\textit{variance}} = \rho_n^-$$

Back to ρ_n

$$\rho_n = \frac{crossproducts - expectation^2}{variance} \geqslant -\frac{expectation^2}{variance} = \rho_n^-$$

$$\rho_n^- = -\frac{\frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2}{\sum_{i\in V} d_i^3 - \frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2}.$$

Back to ρ_n

$$\rho_n = \frac{\textit{crossproducts} - \textit{expectation}^2}{\textit{variance}} \geqslant -\frac{\textit{expectation}^2}{\textit{variance}} = \rho_n^-$$

$$\rho_n^- = -\frac{\frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2}{\sum_{i\in V} d_i^3 - \frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2}.$$

- We have $\sum_{i\in V} d_i^3 \geqslant cn^{3/(\alpha-1)}$
- But also

$$\frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2 \leqslant (C^2/c)n^{\max\{4/(\alpha-1)-1,1\}}.$$

- When $\alpha \in (2,4)$ we have $\max\{4/(\alpha-1)-1,1\} < 3/(\alpha-1)$, so that the denominator of ρ_n^- outweighs its numerator.

No disassortative scale-free random graphs

$$\rho_n \geqslant \rho_n^- = -\frac{\frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2}{\sum_{i\in V} d_i^3 - \frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2}.$$

- Take e.g. $\alpha = 2.5$

No disassortative scale-free random graphs

$$\rho_n \geqslant \rho_n^- = -\frac{\frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2}{\sum_{i\in V} d_i^3 - \frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2}.$$

- Take e.g. $\alpha = 2.5$
- $4/(\alpha-1) - 3/(\alpha-1) = -1/3$

No disassortative scale-free random graphs

$$\rho_n \geqslant \rho_n^- = -\frac{\frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2}{\sum_{i\in V} d_i^3 - \frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2}.$$

- Take e.g. $\alpha = 2.5$
- $4/(\alpha-1) - 3/(\alpha-1) = -1/3$
- $\rho_n^- = O(n^{-1/3})$

No disassortative scale-free random graphs

$$\rho_n \geqslant \rho_n^- = -\frac{\frac{1}{|E|}\left(\sum_{i\in V} d_i^2\right)^2}{\sum_{i\in V} d_i^3 - \frac{1}{|E|}\left(\sum_{i\in V} d_i^2\right)^2}.$$

- Take e.g. $\alpha = 2.5$
- $4/(\alpha-1) - 3/(\alpha-1) = -1/3$
- $\rho_n^- = O(n^{-1/3})$
- ρ_n^- converges to zero as $n \to \infty$ in ANY power law graph

No disassortative scale-free random graphs

$$\rho_n \geqslant \rho_n^- = -\frac{\frac{1}{|E|}\left(\sum_{i\in V} d_i^2\right)^2}{\sum_{i\in V} d_i^3 - \frac{1}{|E|}\left(\sum_{i\in V} d_i^2\right)^2}.$$

- Take e.g. $\alpha = 2.5$
- $4/(\alpha - 1) - 3/(\alpha - 1) = -1/3$
- $\rho_n^- = O(n^{-1/3})$
- ρ_n^- converges to zero as $n \to \infty$ in ANY power law graph
- Large scale-free graphs are never disassortative!

No disassortative scale-free random graphs

$$\rho_n \geqslant \rho_n^- = -\frac{\frac{1}{|E|}\left(\sum_{i\in V} d_i^2\right)^2}{\sum_{i\in V} d_i^3 - \frac{1}{|E|}\left(\sum_{i\in V} d_i^2\right)^2}.$$

- Take e.g. $\alpha = 2.5$
- $4/(\alpha - 1) - 3/(\alpha - 1) = -1/3$
- $\rho_n^- = O(n^{-1/3})$
- ρ_n^- converges to zero as $n \to \infty$ in ANY power law graph
- Large scale-free graphs are never disassortative!
- Reason: high variability in values $\Rightarrow$ dependence on n

Alternative: rank correlations

- $((X_i, Y_i))_{i=1}^n$ random variables

Alternative: rank correlations

- $((X_i, Y_i))_{i=1}^n$ random variables
- r_i^X and r_i^Y the rank of X_i and Y_i, respectively

Alternative: rank correlations

- $((X_i, Y_i))_{i=1}^n$ random variables
- r_i^X and r_i^Y the rank of X_i and Y_i, respectively
- Spearman's rho:

$$\rho_n^{\mathrm{rank}} = \frac{\sum_{i=1}^n (r_i^X - (n+1)/2)(r_i^Y - (n+1)/2)}{\sqrt{\sum_{i=1}^n (r_i^X - (n+1)/2)^2 \sum_i^n (r_i^Y - (n+1)/2)^2}}$$

Alternative: rank correlations

- $((X_i, Y_i))_{i=1}^n$ random variables
- r_i^X and r_i^Y the rank of X_i and Y_i, respectively
- Spearman's rho:

$$\rho_n^{\text{rank}} = \frac{\sum_{i=1}^n (r_i^X - (n+1)/2)(r_i^Y - (n+1)/2)}{\sqrt{\sum_{i=1}^n (r_i^X - (n+1)/2)^2 \sum_i^n (r_i^Y - (n+1)/2)^2}}$$

- Correlation coefficient for r_i^X and r_i^Y
- r_i^X and r_i^Y are from uniform distribution: $n \cdot \mathit{Uniform}(0, 1)$

Alternative: rank correlations

- $((X_i, Y_i))_{i=1}^n$ random variables
- r_i^X and r_i^Y the rank of X_i and Y_i, respectively
- Spearman's rho:

$$\rho_n^{\text{rank}} = \frac{\sum_{i=1}^n (r_i^X - (n+1)/2)(r_i^Y - (n+1)/2)}{\sqrt{\sum_{i=1}^n (r_i^X - (n+1)/2)^2 \sum_i^n (r_i^Y - (n+1)/2)^2}}$$

- Correlation coefficient for r_i^X and r_i^Y
- r_i^X and r_i^Y are from uniform distribution: $n \cdot Uniform(0,1)$
- Factor n cancels, no influence of high dispersion

Classical approach!

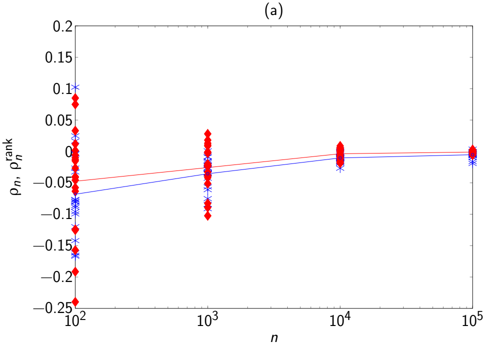
H. Hotelling and M.R. Pabst (1936):
‘Certainly where there is complete absence of knowledge of the form of the bivariate distribution, and especially if it is believed not to be normal, the rank correlation coefficient is to be strongly recommended as a means of testing the existence of relationship.’

Configuration model (CM)

- Nodes with i.i.d. power law distributed number of half-edges are created
- The half-edges connected to each other in a random fashion. Self-loops and double edges are removed.

Configuration model (CM)

- Nodes with i.i.d. power law distributed number of half-edges are created
- The half-edges connected to each other in a random fashion. Self-loops and double edges are removed.
- ρ_n (blue), ρ_n^{rank} (red), and mean ρ_n^- (black) in 20 simulations for different n



Configuration model with intermediate edge (CMIE)

- Nodes are connected randomly. Then each edge broken in two by adding one intermediate node. Strong negative correlation: all original nodes are connected to nodes of degree 2

Configuration model with intermediate edge (CMIE)

- Nodes are connected randomly. Then each edge broken in two by adding one intermediate node. Strong negative correlation: all original nodes are connected to nodes of degree 2
- Clearly strongly disassortative graph

Configuration model with intermediate edge (CMIE)

- Nodes are connected randomly. Then each edge broken in two by adding one intermediate node. Strong negative correlation: all original nodes are connected to nodes of degree 2
- Clearly strongly disassortative graph
- d_i's are original degrees in the CM, $\ell_n = \sum_i d_i$. In CMIE we obtain:

$$\rho_n = \frac{2\sum_{i\in V} 2d_i - \frac{1}{2\ell_n}\Big(\sum_{i\in V} d_i^2 + 2\ell_n\Big)^2}{\sum_{i\in V} d_i^3 + 4\ell_n - \frac{1}{2\ell_n}\Big(\sum_{i\in V} d_i^2 + 2\ell_n\Big)^2}.$$

Configuration model with intermediate edge (CMIE)

- Nodes are connected randomly. Then each edge broken in two by adding one intermediate node. Strong negative correlation: all original nodes are connected to nodes of degree 2
- Clearly strongly disassortative graph
- d_i's are original degrees in the CM, $\ell_n = \sum_i d_i$. In CMIE we obtain:

$$\rho_n = \frac{2\sum_{i\in V} 2d_i - \frac{1}{2\ell_n}\Big(\sum_{i\in V} d_i^2 + 2\ell_n\Big)^2}{\sum_{i\in V} d_i^3 + 4\ell_n - \frac{1}{2\ell_n}\Big(\sum_{i\in V} d_i^2 + 2\ell_n\Big)^2}.$$

- One can see that $\rho_n^- \to 0$

Configuration model with intermediate edge (CMIE)

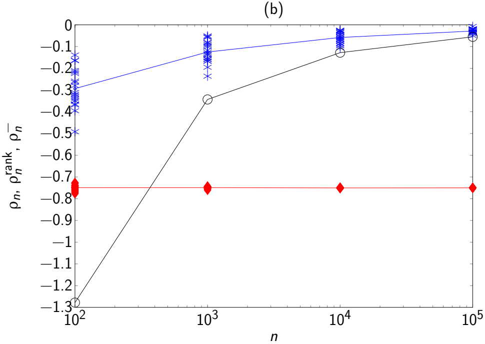
- Nodes are connected randomly. Then each edge broken in two by adding one intermediate node. Strong negative correlation: all original nodes are connected to nodes of degree 2
- Clearly strongly disassortative graph
- d_i's are original degrees in the CM, $\ell_n = \sum_i d_i$. In CMIE we obtain:

$$\rho_n = \frac{2\sum_{i\in V} 2d_i - \frac{1}{2\ell_n}\Big(\sum_{i\in V} d_i^2 + 2\ell_n\Big)^2}{\sum_{i\in V} d_i^3 + 4\ell_n - \frac{1}{2\ell_n}\Big(\sum_{i\in V} d_i^2 + 2\ell_n\Big)^2}.$$

- One can see that $\rho_n^- \to 0$

Configuration model with intermediate edge: results

- Nodes are connected randomly. Then each edge broken in two by adding one intermediate node. Strong negative correlation: all original nodes are connected to nodes of degree 2.
- ρ_n (blue), ρ_n^{rank} (red), and mean ρ_n^- (black) in 20 simulations for different n



Preferential Attachment (PA) graph

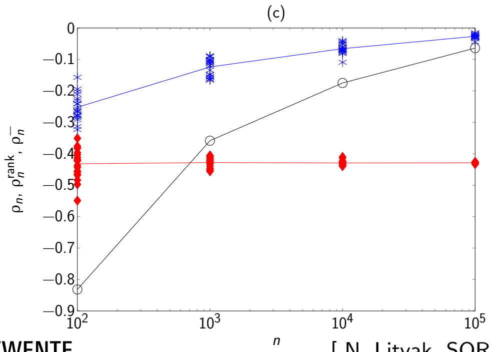
- Albert and Barabási (1999), simplest version with one outgoing edge per node.
- Nodes arrive one at a time. A new node connects to a node i with probability proportional to current degree of i.

Preferential Attachment (PA) graph

- Albert and Barabási (1999), simplest version with one outgoing edge per node.
- Nodes arrive one at a time. A new node connects to a node i with probability proportional to current degree of i.
- $\rho_n \to 0$ (Newman, 2003; Dorogovtsev et al. 2010). Assortatively neutral?

Preferential Attachment (PA) graph

- Albert and Barabási (1999), simplest version with one outgoing edge per node.
- Nodes arrive one at a time. A new node connects to a node i with probability proportional to current degree of i.
- $\rho_n \to 0$ (Newman, 2003; Dorogovtsev et al. 2010). Assortatively neutral?



Assortative networks

$$\rho_n = \frac{\sum_{ij\in E} d_i d_j - \frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2}{\sum_{i\in V} d_i^3 - \frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2}.$$

Two possible scenarios:

- Denominator outweighs numerator, $\rho_n \to 0$
- Denominator and numerator are of the same order of magnitude. Limit?

Collection of bipartite graphs

- $((X_i, Y_i))_{i=1}^n$ i.i.d.
$$X = bU_1 + bU_2, \quad Y = bU_1 + aU_2, \quad b > 0, a > 1$$
U_1, U_2 i.i.d. random variables with power law tail, exponent α.
- For $i = 1, \ldots, n$, we create a complete bipartite graph of X_i and Y_i vertices, respectively.
- These n complete bipartite graphs are not connected to one another.
- Extreme scenario of a network consisting of highly connected clusters of different size. Such networks can serve as models for physical human contacts and are used in epidemic modelling (Eubank et al. 2004).
- Disassortative for $n = 1$ but positive dependence between X and Y prevails for larger n.

Collection of bipartite graphs: analysis

- $|V| = \sum_{i=1}^{n}(X_i + Y_i)$, $|E| = 2\sum_{i=1}^{n} X_i Y_i$,

$$\sum_{i\in V} d_i^p = \sum_{i=1}^{n}(X_i^p Y_i + Y_i^p X_i) \qquad \sum_{ij\in E} d_i d_j = 2\sum_{i=1}^{n}(X_i Y_i)^2.$$

Collection of bipartite graphs: analysis

- $|V| = \sum_{i=1}^{n}(X_i + Y_i)$, $|E| = 2\sum_{i=1}^{n} X_i Y_i$,

$$\sum_{i \in V} d_i^p = \sum_{i=1}^{n}(X_i^p Y_i + Y_i^p X_i) \qquad \sum_{ij \in E} d_i d_j = 2\sum_{i=1}^{n}(X_i Y_i)^2.$$

- Take $\mathbb{P}(U_j > x) = c_0 x^{-\alpha+1}$, where $c_0 > 0$, $x \geqslant x_0$, and $\alpha \in (4, 5)$, so that $E[U^3] < \infty$, but $E[U^4] = \infty$.

Collection of bipartite graphs: analysis

- $|V| = \sum_{i=1}^{n}(X_i + Y_i)$, $|E| = 2\sum_{i=1}^{n} X_i Y_i$,

$$\sum_{i \in V} d_i^p = \sum_{i=1}^{n}(X_i^p Y_i + Y_i^p X_i) \qquad \sum_{ij \in E} d_i d_j = 2\sum_{i=1}^{n}(X_i Y_i)^2.$$

- Take $\mathbb{P}(U_j > x) = c_0 x^{-\alpha+1}$, where $c_0 > 0$, $x \geqslant x_0$, and $\alpha \in (4, 5)$, so that $E[U^3] < \infty$, but $E[U^4] = \infty$.
- Then $|E|/n \xrightarrow{p} 2E[XY] < \infty$ and $\frac{1}{n}\sum_{i \in V} d_i^2 \xrightarrow{p} E[XY(X+Y)] < \infty$.

Collection of bipartite graphs: analysis

Theorem (L& van der Hofstad, 2012)

$$n^{-4/(\alpha-1)} b^{-4} \sum_{i=1}^{n} (X_i^3 Y_i + Y_i^3 X_i) \xrightarrow{d} (a^3 + a) Z_1 + 2 Z_2,$$

$$n^{-4/(\alpha-1)} b^{-4} \sum_{i=1}^{N} (X_i Y_i)^2 \xrightarrow{d} a^2 Z_1 + Z_2,$$

where Z_1 and Z_2 and two independent stable distributions with parameter $(\alpha - 1)/4$.

Collection of bipartite graphs: analysis

Theorem (L& van der Hofstad, 2012)

$$n^{-4/(\alpha-1)} b^{-4} \sum_{i=1}^{n} (X_i^3 Y_i + Y_i^3 X_i) \xrightarrow{d} (a^3 + a) Z_1 + 2Z_2,$$

$$n^{-4/(\alpha-1)} b^{-4} \sum_{i=1}^{N} (X_i Y_i)^2 \xrightarrow{d} a^2 Z_1 + Z_2,$$

where Z_1 and Z_2 and two independent stable distributions with parameter $(\alpha - 1)/4$.

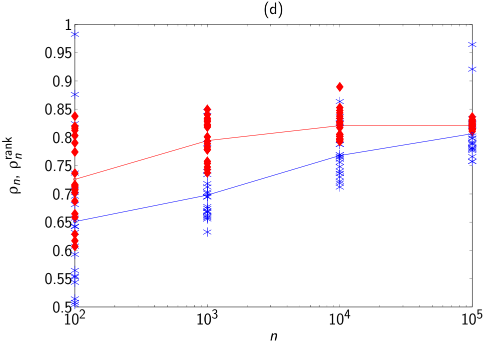
Result:

$$\rho_n \xrightarrow{d} \frac{2a^2 Z_1 + 2Z_2}{(a + a^3) Z_1 + 2Z_2}, \qquad \text{as } n \to \infty,$$

which is a random variable taking values in $(2a/(1 + a^2), 1)$, $a > 1$.

Collection of bipartite graphs: results

ρ_n (blue), ρ_n^{rank} (red), and mean ρ_n^- (black) in 20 simulations for different n



Web and social networks

Dataset	Description	# nodes	max d	ρ_n	ρ_n^{rank}	ρ_n^-
stanford-cs	web domain	9,914	340	-0.1656	-0.1627	-0.4648
eu-2005	.eu web crawl	862,664	68,963	-0.0562	-0.2525	-0.0670
uk@100,000	.uk web crawl	100,000	55,252	-0.6536	-0.5676	-1.117
uk@1,000,000	.uk web crawl	1,000,000	403,441	-0.0831	-0.5620	-0.0854
enron	e-mailing	69,244	1,634	-0.1599	-0.6827	-0.1932
dblp-2010	co-authorship	326,186	238	0.3018	0.2604	-0.7736
dblp-2011	co-authorship	986,324	979	0.0842	0.1351	-0.2963
hollywood-2009	co-starring	1,139,905	11,468	0.3446	0.4689	-0.6737

- Data from the Laboratory of Web Algorithms (LAW) at the Università degli studi di Milano
- All graphs are made undirected

Web and social networks

Dataset	Description	# nodes	max d	ρ_n	ρ_n^{rank}	ρ_n^-
stanford-cs	web domain	9,914	340	-0.1656	-0.1627	-0.4648
eu-2005	.eu web crawl	862,664	68,963	-0.0562	-0.2525	-0.0670
uk@100,000	.uk web crawl	100,000	55,252	-0.6536	-0.5676	-1.117
uk@1,000,000	.uk web crawl	1,000,000	403,441	-0.0831	-0.5620	-0.0854
enron	e-mailing	69,244	1,634	-0.1599	-0.6827	-0.1932
dblp-2010	co-authorship	326,186	238	0.3018	0.2604	-0.7736
dblp-2011	co-authorship	986,324	979	0.0842	0.1351	-0.2963
hollywood-2009	co-starring	1,139,905	11,468	0.3446	0.4689	-0.6737

- Data from the Laboratory of Web Algorithms (LAW) at the Università degli studi di Milano
- All graphs are made undirected
- Spearman's rho is able to reveal strong negative correlations in large networks

Web and social networks

Dataset	Description	# nodes	max d	ρ_n	ρ_n^{rank}	ρ_n^-
stanford-cs	web domain	9,914	340	-0.1656	-0.1627	-0.4648
eu-2005	.eu web crawl	862,664	68,963	-0.0562	-0.2525	-0.0670
uk@100,000	.uk web crawl	100,000	55,252	-0.6536	-0.5676	-1.117
uk@1,000,000	.uk web crawl	1,000,000	403,441	-0.0831	-0.5620	-0.0854
enron	e-mailing	69,244	1,634	-0.1599	-0.6827	-0.1932
dblp-2010	co-authorship	326,186	238	0.3018	0.2604	-0.7736
dblp-2011	co-authorship	986,324	979	0.0842	0.1351	-0.2963
hollywood-2009	co-starring	1,139,905	11,468	0.3446	0.4689	-0.6737

- Data from the Laboratory of Web Algorithms (LAW) at the Università degli studi di Milano
- All graphs are made undirected
- Spearman's rho is able to reveal strong negative correlations in large networks
- 'Infinite variance' is not a formality, it affects the results

Web and social networks

Dataset	Description	# nodes	max d	ρ_n	ρ_n^{rank}	ρ_n^-
stanford-cs	web domain	9,914	340	-0.1656	-0.1627	-0.4648
eu-2005	.eu web crawl	862,664	68,963	-0.0562	-0.2525	-0.0670
uk@100,000	.uk web crawl	100,000	55,252	-0.6536	-0.5676	-1.117
uk@1,000,000	.uk web crawl	1,000,000	403,441	-0.0831	-0.5620	-0.0854
enron	e-mailing	69,244	1,634	-0.1599	-0.6827	-0.1932
dblp-2010	co-authorship	326,186	238	0.3018	0.2604	-0.7736
dblp-2011	co-authorship	986,324	979	0.0842	0.1351	-0.2963
hollywood-2009	co-starring	1,139,905	11,468	0.3446	0.4689	-0.6737

- Data from the Laboratory of Web Algorithms (LAW) at the Università degli studi di Milano
- All graphs are made undirected
- Spearman's rho is able to reveal strong negative correlations in large networks
- 'Infinite variance' is not a formality, it affects the results

Conclusions and discussion

Conclusions and discussion

- The assortativity coefficient ρ_n is not suitable for measuring dependencies in power law data with $\alpha < 4$.
  - ρ_n depends on n

Conclusions and discussion

- The assortativity coefficient ρ_n is not suitable for measuring dependencies in power law data with $\alpha < 4$.
  - ρ_n depends on n
  - For disassortative networks, ρ_n goes to zero as n grows

Conclusions and discussion

- The assortativity coefficient ρ_n is not suitable for measuring dependencies in power law data with $\alpha < 4$.
  - ρ_n depends on n
  - For disassortative networks, ρ_n goes to zero as n grows
  - For assortative networks, ρ_n converges either to zero or to a random variable.

Conclusions and discussion

- The assortativity coefficient ρ_n is not suitable for measuring dependencies in power law data with $\alpha < 4$.
  - ρ_n depends on n
  - For disassortative networks, ρ_n goes to zero as n grows
  - For assortative networks, ρ_n converges either to zero or to a random variable.
- Assortativity can be used in the network analysis ONLY if $\alpha > 4$.

Conclusions and discussion

- The assortativity coefficient ρ_n is not suitable for measuring dependencies in power law data with $\alpha < 4$.
  - ρ_n depends on n
  - For disassortative networks, ρ_n goes to zero as n grows
  - For assortative networks, ρ_n converges either to zero or to a random variable.
- Assortativity can be used in the network analysis ONLY if $\alpha > 4$.
- Spearman's rho is a good alternative.
  - Resolving ties (Mesfioui, M. and Tajar 2005; Nevslehova 2007)
  - Consistency: proved for i.i.d. continuous (X_i, Y_i), variance $O(1/n)$ (Borkowf 2002).

Conclusions and discussion

- The assortativity coefficient ρ_n is not suitable for measuring dependencies in power law data with $\alpha < 4$.
  - ρ_n depends on n
  - For disassortative networks, ρ_n goes to zero as n grows
  - For assortative networks, ρ_n converges either to zero or to a random variable.
- Assortativity can be used in the network analysis ONLY if $\alpha > 4$.
- Spearman's rho is a good alternative.
  - Resolving ties (Mesfioui, M. and Tajar 2005; Nevslehova 2007)
  - Consistency: proved for i.i.d. continuous (X_i, Y_i), variance $O(1/n)$ (Borkowf 2002).
  - In a graph the degrees on the ends of random edges are in general dependent. Can we analyse Spearman's rho? Work in progress.

Thank you!