UNIVERSITY OF TWENTE.

Analysis of centrality measures based on network structure

Nelly Litvak,

University of Twente,
Stochastic Operations Research group

NADINE review 14-11-2013, Toulouse

Publications

- N. Litvak, and R. van der Hofstad, "Uncovering disassortativity in large scale-free networks." *Phys.Rev.E* v.87, p.022801, 2013
- N. Litvak, and R. van der Hofstad, "Degree-degree correlations in random graphs with heavy-tailed degrees." Accepted in *Internet Mathematics*, 2013
- P. van der Hoorn and N. Litvak, "Degree-degree correlations in directed networks with heavy-tailed degrees." arXiv:1310.6528, 2013
- K. Avrachenkov, N. Litvak, V. Medyanikov, and M. Sokol, "Alpha current flow betweenness centrality." Accepted In: *WAW2013*, Harvard University, 2013
- K. Avrachenkov, N. Litvak, M. Sokol, and D.Towsley, "Quick detection of nodes with large degrees." *WAW2012*, Halifax, NS, Canada, pp. 54-65, 2013
- L. Ostroumova, K. Avrachenkov and N. Litvak. "Quick detection of popular entities in large directed networks." Submitted

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: $p_k \approx const \cdot k^{-\gamma-1}$, $\gamma > 1$.

Power laws

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

Power laws

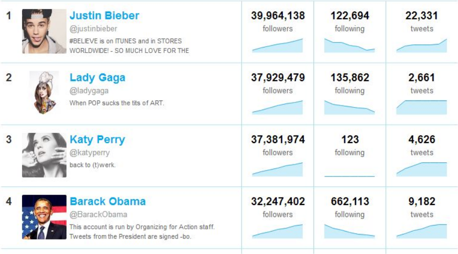
- degree of the node = # links, [fraction nodes degree k] = p_k,
- Power law: $p_k \approx const \cdot k^{-\gamma-1}$, $\gamma > 1$.
- Power laws: Internet, WWW, social networks,etc...
- Model for high variability, scale-free graphs

Power laws

- degree of the node = # links, [fraction nodes degree k] = p_k,
- Power law: $p_k \approx const \cdot k^{-\gamma-1}$, $\gamma > 1$.
- Power laws: Internet, WWW, social networks,etc...
- Model for high variability, scale-free graphs
- Model for hubs: nodes with extremely large number of connections

Power laws

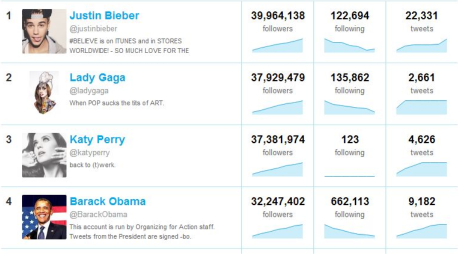
- degree of the node = # links, [fraction nodes degree k] $= p_k$,
- Power law: $p_k \approx const \cdot k^{-\gamma-1}$, $\gamma > 1$.
- Power laws: Internet, WWW, social networks,etc...
- Model for high variability, scale-free graphs
- Model for hubs: nodes with extremely large number of connections



1	Justin Bieber @justinbieber #BELIEVE is on ITUNES and in STORES WORLDWIDE! - SO MUCH LOVE FOR THE	39,964,138 followers	122,694 following	22,331 tweets
2	Lady Gaga @ladygaga When POP sucks the tits of ART.	37,929,479 followers	135,862 following	2,661 tweets
3	Katy Perry @katyperry back to (t)werk.	37,381,974 followers	123 following	4,626 tweets
4	Barack Obama @BarackObama This account is run by Organizing for Action staff. Tweets from the President are signed -bo.	32,247,402 followers	662,113 following	9,182 tweets

Power laws

- degree of the node = # links, [fraction nodes degree k] = p_k,
- Power law: $p_k \approx const \cdot k^{-\gamma-1}$, $\gamma > 1$.
- Power laws: Internet, WWW, social networks,etc...
- Model for high variability, scale-free graphs
- Model for hubs: nodes with extremely large number of connections



1	Justin Bieber @justinbieber #BELIEVE is on ITUNES and in STORES WORLDWIDE! - SO MUCH LOVE FOR THE	39,964,138 followers	122,694 following	22,331 tweets
2	Lady Gaga @ladygaga When POP sucks the tits of ART.	37,929,479 followers	135,862 following	2,661 tweets
3	Katy Perry @katyperry back to (t)werk.	37,381,974 followers	123 following	4,626 tweets
4	Barack Obama @BarackObama This account is run by Organizing for Action staff. Tweets from the President are signed -bo.	32,247,402 followers	662,113 following	9,182 tweets

- Hubs play a crucial role in the analysis of networks

Formal view on the hubs

Let D be a degree of a random node. Regular varying distribution:

$$P(D > x) = L(x)x^{-\gamma} \quad (1)$$

$L(x)$ is slowly varying, i.e. $\lim_{t\to\infty} L(tx)/L(t) = 1$, $x \geqslant 0$

Formal view on the hubs

Let D be a degree of a random node. Regular varying distribution:

$$P(D > x) = L(x)x^{-\gamma} \tag{1}$$

$L(x)$ is slowly varying, i.e. $\lim_{t\to\infty} L(tx)/L(t) = 1$, $x \geqslant 0$

Extreme value theory. Let $F_1 \geqslant F_2 \geqslant \cdots \geqslant F_N$ be the order statistics of the i.i.d. r.v.'s $D_1, D_2, \ldots, D_N$ as in (1). Then there are (a_N), (b_N) such that for finite k

$$\left(\frac{F_1 - b_N}{a_N}, \cdots \frac{F_k - b_N}{a_N}\right) \xrightarrow{d} \left(\frac{E_1^{-\delta} - 1}{\delta}, \cdots, \frac{\left(\sum_{i=1}^{k} E_i\right)^{-\delta} - 1}{\delta}\right),$$

where $\delta = 1/\gamma$ and E_i'are i.i.d. exponential(1) r.v.'s.

Formal view on the hubs

Let D be a degree of a random node. Regular varying distribution:

$$P(D > x) = L(x)x^{-\gamma} \tag{1}$$

$L(x)$ is slowly varying, i.e. $\lim_{t\to\infty} L(tx)/L(t) = 1$, $x \geqslant 0$

Extreme value theory. Let $F_1 \geqslant F_2 \geqslant \cdots \geqslant F_N$ be the order statistics of the i.i.d. r.v.'s $D_1, D_2, \ldots, D_N$ as in (1). Then there are (a_N), (b_N) such that for finite k

$$\left(\frac{F_1 - b_N}{a_N}, \cdots \frac{F_k - b_N}{a_N}\right) \xrightarrow{d} \left(\frac{E_1^{-\delta} - 1}{\delta}, \cdots, \frac{\left(\sum_{i=1}^{k} E_i\right)^{-\delta} - 1}{\delta}\right),$$

where $\delta = 1/\gamma$ and E_i'are i.i.d. exponential(1) r.v.'s.

Example. $P(D > x) = Cx^{-\gamma}$, then $a_N = \delta C^{\delta} N^{\delta}$, $b_N = C^{\delta} N^{\delta}$.
The largest degrees are 'of the order' $N^{1/\gamma}$.

Finding most popular entities in social networks

- Social networks are large

Finding most popular entities in social networks

- Social networks are large
- The complete graphs structure is only available to the owners

Finding most popular entities in social networks

- Social networks are large
- The complete graphs structure is only available to the owners
- Many companies maintain network statistics (*twittercounter.com*, *followerwonk.com*, *twitaholic.com*, *www.insidefacebook.com*, *yavkontakte.ru*)

Finding most popular entities in social networks

- Social networks are large
- The complete graphs structure is only available to the owners
- Many companies maintain network statistics (*twittercounter.com*, *followerwonk.com*, *twitaholic.com*, *www.insidefacebook.com*, *yavkontakte.ru*)
- The network can be accessed only via API, with limited access

Finding most popular entities in social networks

- Social networks are large
- The complete graphs structure is only available to the owners
- Many companies maintain network statistics (*twittercounter.com*, *followerwonk.com*, *twitaholic.com*, *www.insidefacebook.com*, *yavkontakte.ru*)
- The network can be accessed only via API, with limited access
- Twitter API allows one access per minute. We need 950 years to crawl the current Twitter graph!

Finding most popular entities in social networks

- Social networks are large
- The complete graphs structure is only available to the owners
- Many companies maintain network statistics (*twittercounter.com*, *followerwonk.com*, *twitaholic.com*, *www.insidefacebook.com*, *yavkontakte.ru*)
- The network can be accessed only via API, with limited access
- Twitter API allows one access per minute. We need 950 years to crawl the current Twitter graph!

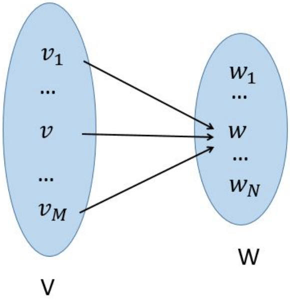
Goal: Find top-k most popular entities in social (directed) networks (nodes with highest in/out-degrees, largest interest groups, largest user categories), using the minimal number of API requests.

Problem formulation

- Consider a bi-partite graph (V, W, E)
- V and W are sets of entities, $|V| = M$, $|W| = N$.
- A directed edge $(v, w) \in E$ represents a relation between $v \in V$ and $w \in W$.
- Goal: Quickly find entities in W with highest degrees.

Problem formulation

- Consider a bi-partite graph (V, W, E)
- V and W are sets of entities, $|V| = M$, $|W| = N$.
- A directed edge $(v, w) \in E$ represents a relation between $v \in V$ and $w \in W$.
- Goal: Quickly find entities in W with highest degrees.



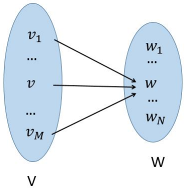
Example. $V = W$ is a set of Twitter users, (v, w) means that v follows w.

Example. V is a set of users, W is a set of interest groups, (v, w) means that user v is a member of an interest group w.

Algorithm for finding top-k most popular entities

Algorithm for finding top-k most popular entities

1. Choose a set $A \subset V$ of n_1 nodes sampled from V at random.
2. For each $v \in A$ retrieve the id's of nodes in W that have an edge from v.
3. Compute S_w – the number of edges of $w \in W$ from A.
4. Retrieve the actual degrees for the n_2 nodes w with the largest values of S_w.
5. Return the identified top-k list of most popular entities in W.



In total, we use $n = n_1 + n_2$ requests to API (Step 2 and Step 4).

Example: finding most followed users on Twitter

- Huge network (more than 500M users)

Example: finding most followed users on Twitter

- Huge network (more than 500M users)
- Network accessed only through Twitter API

Example: finding most followed users on Twitter

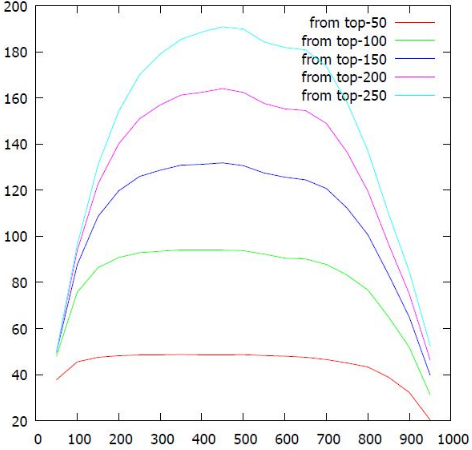
- Huge network (more than 500M users)
- Network accessed only through Twitter API
- The rate of requests is limited
- One request:
  - ID's of at most 5000 followers of a node, or
  - the number of followers of a node
- In a randomly chosen set of n_1 Twitter users only a few users follow more than 5000 people. Thus, we retrieve at most 5000 followees of each node. This does not affect the results.

Example: finding most followed users on Twitter

- Huge network (more than 500M users)
- Network accessed only through Twitter API
- The rate of requests is limited
- One request:
  - ID's of at most 5000 followers of a node, or
  - the number of followers of a node
- In a randomly chosen set of n_1 Twitter users only a few users follow more than 5000 people. Thus, we retrieve at most 5000 followees of each node. This does not affect the results.
- Make a guess: We use 1000 requests to API. For which k can we identify a top-k list of most followed Twitter users with 90% precision?

Results

$N = 500M$, $n = 1000$



Interest groups VKontakte

- Popular social network in Russian, more than 200M users.

Rank	Number of participants	Topic
1	4,35M	humor
2	4,1M	humor
3	3,76M	movies
4	3,69M	humor
5	3,59M	humor
6	3,58M	facts
7	3,36M	cookery
8	3,31M	humor
9	3,14M	humor
10	3,14M	movies
100	1,65M	success

- With $n_1 = 700$, $n_2 = 300$, our algorithm identifies on average 73.2 from the top-100 interest groups (averaged over 25 experiments). The standard deviation is 4.6.

Sublinear complexity

- $1, \dots, k$ – top-k nodes in W; $F_1, \dots, F_k$ – their degrees

Sublinear complexity

- $1, \ldots, k$ – top-k nodes in W; $F_1, \ldots, F_k$ – their degrees
- $S_j \sim Binomial(n_1, F_j/N)$

Sublinear complexity

- $1, \ldots, k$ – top-k nodes in W; $F_1, \ldots, F_k$ – their degrees
- $S_j \sim Binomial(n_1, F_j/N)$
- With normal approximation, and error pr-ty α we need that

$$\sqrt{\frac{n_1}{N}} \frac{F_k - F_{n_2}}{\sqrt{F_k + F_{n_2}}} > z_{1-\alpha}$$

Sublinear complexity

- $1, \ldots, k$ – top-k nodes in W; $F_1, \ldots, F_k$ – their degrees
- $S_j \sim Binomial(n_1, F_j/N)$
- With normal approximation, and error pr-ty α we need that

$$\sqrt{\frac{n_1}{N}} \frac{F_k - F_{n_2}}{\sqrt{F_k + F_{n_2}}} > z_{1-\alpha}$$

- $F_k >> F_{n_2}$

Sublinear complexity

- $1, \ldots, k$ – top-k nodes in W; $F_1, \ldots, F_k$ – their degrees
- $S_j \sim Binomial(n_1, F_j/N)$
- With normal approximation, and error pr-ty α we need that

$$\sqrt{\frac{n_1}{N}} \frac{F_k - F_{n_2}}{\sqrt{F_k + F_{n_2}}} > z_{1-\alpha}$$

- $F_k >> F_{n_2}$
- Assuming the i.i.d. degrees, by the Extreme Value Theory, w.h.p., $\log(F_k) = \gamma^{-1} \log(N)(1 + o(log(N)))$

Sublinear complexity

- $1, \ldots, k$ – top-k nodes in W; $F_1, \ldots, F_k$ – their degrees
- $S_j \sim Binomial(n_1, F_j/N)$
- With normal approximation, and error pr-ty α we need that

$$\sqrt{\frac{n_1}{N}} \frac{F_k - F_{n_2}}{\sqrt{F_k + F_{n_2}}} > z_{1-\alpha}$$

- $F_k >> F_{n_2}$
- Assuming the i.i.d. degrees, by the Extreme Value Theory, w.h.p., $\log(F_k) = \gamma^{-1} \log(N)(1 + o(log(N)))$
- Roughly, $n_1 = O(N^{1-1/\gamma})$

Sublinear complexity

- $1, \ldots, k$ – top-k nodes in W; $F_1, \ldots, F_k$ – their degrees
- $S_j \sim Binomial(n_1, F_j/N)$
- With normal approximation, and error pr-ty α we need that

$$\sqrt{\frac{n_1}{N}} \frac{F_k - F_{n_2}}{\sqrt{F_k + F_{n_2}}} > z_{1-\alpha}$$

- $F_k >> F_{n_2}$
- Assuming the i.i.d. degrees, by the Extreme Value Theory, w.h.p., $\log(F_k) = \gamma^{-1} \log(N)(1 + o(log(N)))$
- Roughly, $n_1 = O(N^{1-1/\gamma})$
- Since $\sum_w S_w = O(n_1)$ w.h.p., n_2 is at most $O(n_1)$

Sublinear complexity

- $1, \ldots, k$ – top-k nodes in W; $F_1, \ldots, F_k$ – their degrees
- $S_j \sim Binomial(n_1, F_j/N)$
- With normal approximation, and error pr-ty α we need that

$$\sqrt{\frac{n_1}{N}} \frac{F_k - F_{n_2}}{\sqrt{F_k + F_{n_2}}} > z_{1-\alpha}$$

- $F_k >> F_{n_2}$
- Assuming the i.i.d. degrees, by the Extreme Value Theory, w.h.p., $\log(F_k) = \gamma^{-1} \log(N)(1 + o(log(N)))$
- Roughly, $n_1 = O(N^{1-1/\gamma})$
- Since $\sum_w S_w = O(n_1)$ w.h.p., n_2 is at most $O(n_1)$
- We conclude that roughly $n = n_1 + n_2 = O(N^{1-1/\gamma})$

Sublinear complexity

- $1, \ldots, k$ – top-k nodes in W; $F_1, \ldots, F_k$ – their degrees
- $S_j \sim Binomial(n_1, F_j/N)$
- With normal approximation, and error pr-ty α we need that

$$\sqrt{\frac{n_1}{N}} \frac{F_k - F_{n_2}}{\sqrt{F_k + F_{n_2}}} > z_{1-\alpha}$$

- $F_k >> F_{n_2}$
- Assuming the i.i.d. degrees, by the Extreme Value Theory, w.h.p., $\log(F_k) = \gamma^{-1} \log(N)(1 + o(log(N)))$
- Roughly, $n_1 = O(N^{1-1/\gamma})$
- Since $\sum_w S_w = O(n_1)$ w.h.p., n_2 is at most $O(n_1)$
- We conclude that roughly $n = n_1 + n_2 = O(N^{1-1/\gamma})$
- Note that the complexity is in terms of $|W| = N$

Sublinear complexity

- $1, \ldots, k$ – top-k nodes in W; $F_1, \ldots, F_k$ – their degrees
- $S_j \sim Binomial(n_1, F_j/N)$
- With normal approximation, and error pr-ty α we need that

$$\sqrt{\frac{n_1}{N}} \frac{F_k - F_{n_2}}{\sqrt{F_k + F_{n_2}}} > z_{1-\alpha}$$

- $F_k >> F_{n_2}$
- Assuming the i.i.d. degrees, by the Extreme Value Theory, w.h.p., $\log(F_k) = \gamma^{-1}\log(N)(1 + o(log(N)))$
- Roughly, $n_1 = O(N^{1-1/\gamma})$
- Since $\sum_w S_w = O(n_1)$ w.h.p., n_2 is at most $O(n_1)$
- We conclude that roughly $n = n_1 + n_2 = O(N^{1-1/\gamma})$
- Note that the complexity is in terms of $|W| = N$
- Popular groups are easier to find than popular users!

Alpha current flow betweenness centrality

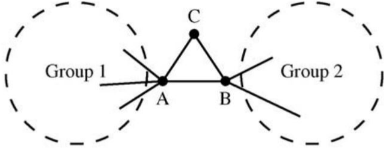
- $G = (V, E)$, $|V| = n$, $|E| = m$

Alpha current flow betweenness centrality

- $G = (V, E)$, $|V| = n$, $|E| = m$
- Betweenness centrality: the fraction of *shortest* paths, andveraged over all source-destination pairs

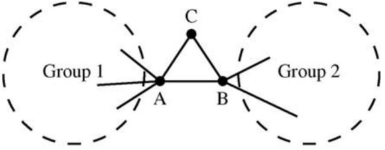
Alpha current flow betweenness centrality

- $G = (V, E)$, $|V| = n$, $|E| = m$
- Betweenness centrality: the fraction of *shortest* paths, andveraged over all source-destination pairs



Alpha current flow betweenness centrality

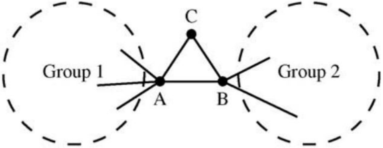
- $G = (V, E)$, $|V| = n$, $|E| = m$
- Betweenness centrality: the fraction of *shortest* paths, andveraged over all source-destination pairs



- Newman (2005), Brandes and Fleischer (2005): current flow (CF) betweenness centrality

Alpha current flow betweenness centrality

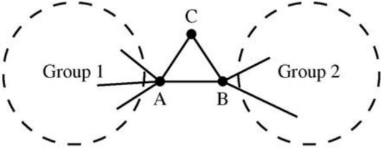
- $G = (V, E)$, $|V| = n$, $|E| = m$
- Betweenness centrality: the fraction of *shortest* paths, andveraged over all source-destination pairs



- Newman (2005), Brandes and Fleischer (2005): current flow (CF) betweenness centrality
- Graph is an electrical network, edges are unit resistances, current is induced to s, t is connected to the ground

Alpha current flow betweenness centrality

- $G = (V, E)$, $|V| = n$, $|E| = m$
- Betweenness centrality: the fraction of *shortest* paths, andveraged over all source-destination pairs



- Newman (2005), Brandes and Fleischer (2005): current flow (CF) betweenness centrality
- Graph is an electrical network, edges are unit resistances, current is induced to s, t is connected to the ground
- The CF-betweenness of edge $e \in E$ is the amount of current through e, averaged over source-destination pairs (s, t)

Alpha current flow betweenness centrality

- The CF centrality has a high computational complexity $(I(n-1) + O(nm\log(n)))$, where $I(n-1)$ is the complexity of the matrix inversion of a $(n-1) \times (n-1)$ matrix

Alpha current flow betweenness centrality

- The CF centrality has a high computational complexity $(I(n-1)+O(nm\log(n)))$, where $I(n-1)$ is the complexity of the matrix inversion of a $(n-1)\times(n-1)$ matrix
- Idea: α-CF betweenness centrality

Alpha current flow betweenness centrality

- The CF centrality has a high computational complexity $(I(n-1)+O(nm\log(n)))$, where $I(n-1)$ is the complexity of the matrix inversion of a $(n-1)\times(n-1)$ matrix
- Idea: α-CF betweenness centrality
  - Each edge has resistence α^{-1}

Alpha current flow betweenness centrality

- The CF centrality has a high computational complexity $(I(n-1)+O(nm\log(n)))$, where $I(n-1)$ is the complexity of the matrix inversion of a $(n-1)\times(n-1)$ matrix
- Idea: α-CF betweenness centrality
  - Each edge has resistence α^{-1}
  - Each node v is connected to the ground node $n+1$ by an edge with resistance $(1-\alpha)^{-1}d_v^{-1}$, where d_v is the degree of v.

Alpha current flow betweenness centrality

- The CF centrality has a high computational complexity $(I(n-1) + O(nm\log(n)))$, where $I(n-1)$ is the complexity of the matrix inversion of a $(n-1) \times (n-1)$ matrix
- Idea: α-CF betweenness centrality
  - Each edge has resistence α^{-1}
  - Each node v is connected to the ground node $n+1$ by an edge with resistance $(1-\alpha)^{-1} d_v^{-1}$, where d_v is the degree of v.
  - In the spirit of PageRank

Alpha current flow betweenness centrality

- The CF centrality has a high computational complexity $(I(n-1)+O(nm\log(n)))$, where $I(n-1)$ is the complexity of the matrix inversion of a $(n-1)\times(n-1)$ matrix
- Idea: α-CF betweenness centrality
  - Each edge has resistence α^{-1}
  - Each node v is connected to the ground node $n+1$ by an edge with resistance $(1-\alpha)^{-1}d_v^{-1}$, where d_v is the degree of v.
  - In the spirit of PageRank
  - Easy to compute

Formal definition

- A unit of current is supplied to a source node $s \in V$

Formal definition

- A unit of current is supplied to a source node $s \in V$
- A destination node $t \in V$ connected to the ground

Formal definition

- A unit of current is supplied to a source node $s \in V$
- A destination node $t \in V$ connected to the ground
- $\varphi_v^{(s,t)}$ is the absolute potential of node $v \in V$

Formal definition

- A unit of current is supplied to a source node $s \in V$
- A destination node $t \in V$ connected to the ground
- $\varphi_v^{(s,t)}$ is the absolute potential of node $v \in V$
- $\varphi_t^{(s,t)} = \varphi_{n+1}^{(s,t)} = 0$

Formal definition

- A unit of current is supplied to a source node $s \in V$
- A destination node $t \in V$ connected to the ground
- $\varphi_v^{(s,t)}$ is the absolute potential of node $v \in V$
- $\varphi_t^{(s,t)} = \varphi_{n+1}^{(s,t)} = 0$
- $\varphi^{(s,t)} = [\varphi_1^{(s,t)}, ..., \varphi_{n-1}^{(s,t)}]^T$

Formal definition

- A unit of current is supplied to a source node $s \in V$
- A destination node $t \in V$ connected to the ground
- $\varphi_v^{(s,t)}$ is the absolute potential of node $v \in V$
- $\varphi_t^{(s,t)} = \varphi_{n+1}^{(s,t)} = 0$
- $\varphi^{(s,t)} = [\varphi_1^{(s,t)}, \dots, \varphi_{n-1}^{(s,t)}]^T$
- Kirchhoff's current law:

$$[\tilde{D}_t - \alpha \tilde{A}_t]\varphi^{(s,t)} = e_s,$$

$\tilde{D}_t$ and $\tilde{A}_t$ are the degree and the adjacency matrices of $G \setminus \{t\}$, e_s is the sth basis vector (Brandes and Fleischer 2005)

Formal definition

- A unit of current is supplied to a source node $s \in V$
- A destination node $t \in V$ connected to the ground
- $\varphi_v^{(s,t)}$ is the absolute potential of node $v \in V$
- $\varphi_t^{(s,t)} = \varphi_{n+1}^{(s,t)} = 0$
- $\varphi^{(s,t)} = [\varphi_1^{(s,t)}, \ldots, \varphi_{n-1}^{(s,t)}]^T$
- Kirchhoff's current law:

  $$[\tilde{D}_t - \alpha \tilde{A}_t]\varphi^{(s,t)} = e_s,$$

  $\tilde{D}_t$ and $\tilde{A}_t$ are the degree and the adjacency matrices of $G \setminus \{t\}$, e_s is the sth basis vector (Brandes and Fleischer 2005)
- $x_e^{(s,t)} = |\varphi_v^{(s,t)} - \varphi_w^{(s,t)}|, \quad (v, w) \in E$ is the difference of potentials

Formal definition

- A unit of current is supplied to a source node $s \in V$
- A destination node $t \in V$ connected to the ground
- $\varphi_v^{(s,t)}$ is the absolute potential of node $v \in V$
- $\varphi_t^{(s,t)} = \varphi_{n+1}^{(s,t)} = 0$
- $\varphi^{(s,t)} = [\varphi_1^{(s,t)}, \ldots, \varphi_{n-1}^{(s,t)}]^T$
- Kirchhoff's current law:

$$[\tilde{D}_t - \alpha \tilde{A}_t]\varphi^{(s,t)} = e_s,$$

  $\tilde{D}_t$ and $\tilde{A}_t$ are the degree and the adjacency matrices of $G \setminus \{t\}$, e_s is the sth basis vector (Brandes and Fleischer 2005)
- $x_e^{(s,t)} = |\varphi_v^{(s,t)} - \varphi_w^{(s,t)}|, \quad (v, w) \in E$ is the difference of potentials
- α-CF betweenness: $x_e^{\alpha} = \dfrac{1}{n(n-1)} \displaystyle\sum_{s,t \in V, s \neq t} x_e^{(s,t)}, \quad e \in E.$

Analysis and computation

Theorem

The voltage drop along the edge (v, w) is given by

$$\varphi_v^{(s,t)} - \varphi_w^{(s,t)} = (c_{s,v} - c_{s,w}) + \frac{c_{s,t}}{c_{t,t}}(c_{t,w} - c_{t,v}),$$

where $C = (c_{v,w}) = [D - \alpha A]^{-1}$.

- It is sufficient to invert the matrix $[D - \alpha A]$ only once. This can be done efficiently

Analysis and computation

Theorem

The voltage drop along the edge (v, w) is given by

$$\varphi_v^{(s,t)} - \varphi_w^{(s,t)} = (c_{s,v} - c_{s,w}) + \frac{c_{s,t}}{c_{t,t}}(c_{t,w} - c_{t,v}),$$

where $C = (c_{v,w}) = [D - \alpha A]^{-1}$.

- It is sufficient to invert the matrix $[D - \alpha A]$ only once. This can be done efficiently
- $\tilde{P}_t$ transition probability matrix of a random walk on $G \setminus \{t\}$

Analysis and computation

Theorem

The voltage drop along the edge (v, w) is given by

$$\varphi_v^{(s,t)} - \varphi_w^{(s,t)} = (c_{s,v} - c_{s,w}) + \frac{c_{s,t}}{c_{t,t}}(c_{t,w} - c_{t,v}),$$

where $C = (c_{v,w}) = [D - \alpha A]^{-1}$.

- It is sufficient to invert the matrix $[D - \alpha A]$ only once. This can be done efficiently
- $\tilde{P}_t$ transition probability matrix of a random walk on $G \setminus \{t\}$
- $\tilde{\pi}_{\cdot,t}(v) = (1-\alpha)\mathbf{e}_v^T [I - \alpha \tilde{P}_t]^{-1}$ is close to Personalized PageRank with teleportation to v. Then we derive:

Analysis and computation

Theorem

The voltage drop along the edge (v, w) is given by

$$\varphi_v^{(s,t)} - \varphi_w^{(s,t)} = (c_{s,v} - c_{s,w}) + \frac{c_{s,t}}{c_{t,t}}(c_{t,w} - c_{t,v}),$$

where $C = (c_{v,w}) = [D - \alpha A]^{-1}$.

- It is sufficient to invert the matrix $[D - \alpha A]$ only once. This can be done efficiently
- $\tilde{P}_t$ transition probability matrix of a random walk on $G \setminus \{t\}$
- $\tilde{\pi}_{\cdot,t}(v) = (1-\alpha)\mathbf{e}_v^T[I - \alpha\tilde{P}_t]^{-1}$ is close to Personalized PageRank with teleportation to v. Then we derive:

$$\varphi_v^{(s,t)} = (1-\alpha)^{-1}\tilde{\pi}_{s,t}(v)d_s^{-1}$$

Datasets

	$\|V\|$	$\|E\|$	$\langle deg(v) \rangle$	$diam(G)$	$C_{\text{clustering}}$	$\langle d(u,v) \rangle$
Dolphins network	62	159	5.13	8	0.259	3.357
VKontakte AMCP	2092	14816	14.16	14	0.338	4.598
Watts-Strogatz	1000	6000	12.00	6	0.422	3.713
Enron	36692	183831	10.02	11	0.4970	$\approx$ 4.8

- The small graphs are used to compare CF and α-CF betweenness
- On the Enron graph, only α-CF betweenness can be computed

Correlations between centrality measures

Kendall tau for centrality measures in the social graph VKontakte AMCP:

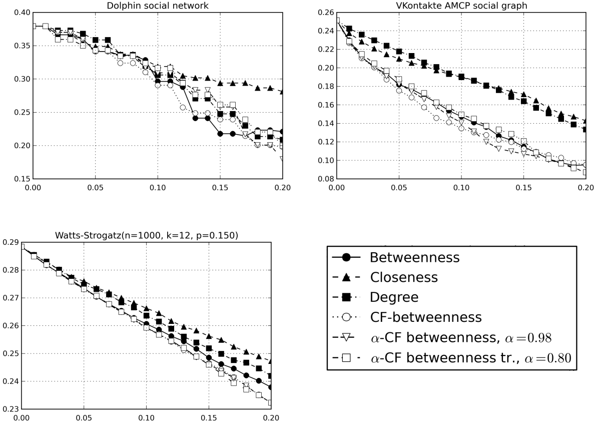
	D	PR	Cl	B/w	CF	(0.8)	tr(0.8)	(0.98)
Degree	1.000	0.655	0.679	0.521	0.545	0.659	0.668	0.599
PageRank	0.655	1.000	0.375	0.662	0.717	0.833	0.811	0.766
Closeness	0.679	0.375	1.000	0.382	0.356	0.424	0.445	0.395
Between.	0.521	0.662	0.382	1.000	0.761	0.760	0.749	0.778
CF	0.545	0.717	0.356	0.761	1.000	0.812	0.833	0.917
αCF(0.8)	0.659	0.833	0.424	0.760	0.812	1.000	0.938	0.878
αCF-tr(0.8)	0.668	0.811	0.445	0.749	0.833	0.938	1.000	0.903
αCF(0.98)	0.599	0.766	0.395	0.778	0.917	0.878	0.903	1.000

Influence on the network connectivity

Inverse average distance: $< d^{-1} >= \frac{1}{n(n-1)} \sum_{u,v \in V, u \neq v} \frac{1}{d(u,v)}$

Influence on the network connectivity

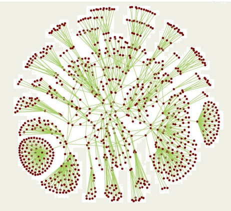
Inverse average distance: $< d^{-1} >= \dfrac{1}{n(n-1)} \displaystyle\sum_{u,v \in V, u \neq v} \dfrac{1}{d(u,v)}$

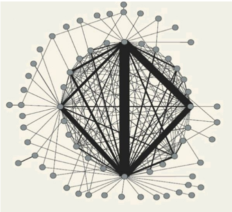


Correlations in power law networks

- **We study the dependencies between degrees of neighboring nodes in graphs with power law degree distribution**

Example: Internet and network of bank transactions





Assortativity coefficient

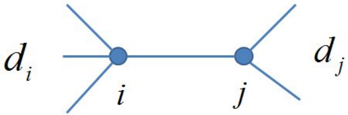
- $G = (V, E)$ undirected graph of n nodes, E'– directed edges
- D_v degree of node $v \in V$

Assortativity coefficient

- $G = (V, E)$ undirected graph of n nodes, E'– directed edges
- D_v degree of node $v \in V$
- Newman (2002): assortativity measure $\rho(G)$

$$\rho(G) = \frac{\frac{1}{|E'|}\sum_{(v,w)\in E'} D_v D_w - \left(\frac{1}{|E'|}\sum_{(v,w)\in E'} \frac{1}{2}(D_v + D_w)\right)^2}{\frac{1}{|E'|}\sum_{(v,w)\in E'} \frac{1}{2}(D_v^2 + D_w^2) - \left(\frac{1}{|E'|}\sum_{(v,w)\in E'} \frac{1}{2}(D_v + D_w)\right)^2}$$

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



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, $\rho(G) < 0$
- Social networks are assortative, $\rho(G) > 0$

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, $\rho(G) < 0$
- Social networks are assortative, $\rho(G) > 0$
- Note: large networks are never strongly disassortative... Dorogovtsev et al. (2010), Raschke et al. (2010)

Convergence of $\rho(G)$ to a non-negtive value

Theorem

Let $(G_n)_{n\geqslant 1}$ be a sequence of graphs of size n satisfying that there exist $\gamma \in (1,3)$ and $0 < c < C < \infty$ such that $cn \leqslant |E| \leqslant Cn$, $cn^{1/\gamma} \leqslant \max_{v\in V_n} D_v \leqslant Cn^{1/\gamma}$ and $cn^{(2/\gamma)\vee 1} \leqslant \sum_{v\in V_n} D_v^2 \leqslant Cn^{(2/\gamma)\vee 1}$. Then, any limit point of the Pearson's correlation coefficient $\rho(G_n)$ is non-negative.

Convergence of $\rho(G)$ to a non-negtive value

Theorem

Let $(G_n)_{n \geqslant 1}$ be a sequence of graphs of size n satisfying that there exist $\gamma \in (1,3)$ and $0 < c < C < \infty$ such that $cn \leqslant |E| \leqslant Cn$, $cn^{1/\gamma} \leqslant \max_{v \in V_n} D_v \leqslant Cn^{1/\gamma}$ and $cn^{(2/\gamma)\vee 1} \leqslant \sum_{v \in V_n} D_v^2 \leqslant Cn^{(2/\gamma)\vee 1}$. Then, any limit point of the Pearson's correlation coefficient $\rho(G_n)$ is non-negative.

Alternative: rank correlations

- $G = (V, E)$, E – set of edges, E' – set of directed edges
- (R_v, R_w) – ranks of (D_v, D_w), where (v, w) is a uniformly chosen directed edge

Alternative: rank correlations

- $G = (V, E)$, E – set of edges, E' – set of directed edges
- (R_v, R_w) – ranks of (D_v, D_w), where (v, w) is a uniformly chosen directed edge
- Ties are resolved at random by adding independent *Uniform*$(0, 1)$ random variables (Mesfioui and Tajar, 2005)

Spearman's rho

- $G = (V, E)$, E – set of edges, E' – set of directed edges
- (R_v, R_w) – ranks of $(D_v + U_e, D_w + U_e')$, where (v, w) is a uniformly chosen directed edge

Spearman's rho

- $G = (V, E)$, E – set of edges, E' – set of directed edges
- (R_v, R_w) – ranks of $(D_v + U_e, D_w + U_e')$, where (v, w) is a uniformly chosen directed edge
- The Spearman's rho (Spearman 1904, H. Hotelling and M.R. Pabst 1936):

$$\rho^{\mathrm{rank}}(G) = \frac{\frac{1}{|E'|}\sum_{(v,w)\in E'} R_v R_w - (|E'|+1)^2/4}{(|E'|^2-1)/12}.$$

Spearman's rho

- $G = (V, E)$, E – set of edges, E' – set of directed edges
- (R_v, R_w) – ranks of $(D_v + U_e, D_w + U_e')$, where (v, w) is a uniformly chosen directed edge
- The Spearman's rho (Spearman 1904, H. Hotelling and M.R. Pabst 1936):

$$\rho^{\text{rank}}(G) = \frac{\frac{1}{|E'|}\sum_{(v,w)\in E'} R_v R_w - (|E'|+1)^2/4}{(|E'|^2-1)/12}.$$

- Pearson's coefficient for (R_v, R_w)
- R_v and R_w are from uniform distribution: $|E'| \cdot \mathit{Uniform}(0, 1)$

Spearman's rho

- $G = (V, E)$, E – set of edges, E' – set of directed edges
- (R_v, R_w) – ranks of $(D_v + U_e, D_w + U'_e)$, where (v, w) is a uniformly chosen directed edge
- The Spearman's rho (Spearman 1904, H. Hotelling and M.R. Pabst 1936):

$$\rho^{\mathrm{rank}}(G) = \frac{\frac{1}{|E'|}\sum_{(v,w)\in E'} R_v R_w - (|E'|+1)^2/4}{(|E'|^2-1)/12}.$$

- Pearson's coefficient for (R_v, R_w)
- R_v and R_w are from uniform distribution: $|E'| \cdot \mathit{Uniform}(0, 1)$
- Factor $|E'|$ cancels, no influence of high dispersion

Convergence criteria in random graphs

$(G_n)_{n\geqslant 1}$ be a sequence of random graphs of size n, $G_n = (V_n, E_n)$.
(X_n, Y_n) degrees on both sides of a uniform directed edge $e \in E_n'$.

Theorem

If every bounded continuous $h\colon \mathbb{R}^2 \to \mathbb{R}$

$$\mathbb{E}_n[h(X_n, Y_n)] \xrightarrow{\mathbb{P}} \mathbb{E}[h(X, Y)],$$

where the r.h.s. is non-random, then

$$\rho^{\mathrm{rank}}(G_n) \xrightarrow{\mathbb{P}} \rho^{\mathrm{rank}} = 12 \cdot \mathrm{Cov}(F_X(X), F_X(Y)),$$

If, in addition, $\mathbb{E}_n[X_n^2] \xrightarrow{\mathbb{P}} \mathbb{E}[X^2] < \infty$, and $\mathrm{Var}(X) > 0$, then

$$\rho(G_n) \xrightarrow{\mathbb{P}} \rho = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}.$$

Preferential Attachment (PA) graph

- Vertex arriving at time $t+1$ attaches to a vertex $v \in [t]$ with probability $(D_v(t)+\delta)/((2+\delta)t+1+\delta)$

Preferential Attachment (PA) graph

- Vertex arriving at time $t+1$ attaches to a vertex $v \in [t]$ with probability $(D_v(t)+\delta)/((2+\delta)t+1+\delta)$
- Dorogovtsev et al. (2010), Grechnikov (2012).

Preferential Attachment (PA) graph

- Vertex arriving at time $t+1$ attaches to a vertex $v \in [t]$ with probability $(D_v(t)+\delta)/((2+\delta)t+1+\delta)$
- DOROGOVTSEV ET AL. (2010), GRECHNIKOV (2012).

Theorem

Let $(G_t^{(m)})_{t\geqslant 1}$ be the PAM. Then

$$\rho^{\mathrm{rank}}(G_t^{(m)}) \xrightarrow{\mathbb{P}} \rho^{\mathrm{rank}},$$

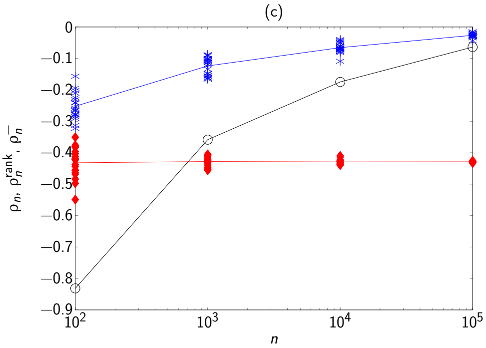
$$\rho(G_t^{(m)}) \xrightarrow{\mathbb{P}} \begin{cases} 0 & \text{if } \delta \leqslant m, \\ \rho & \text{if } \delta > m, \end{cases}$$

where, abbreviating $a = \delta/m$,

$$\rho = \frac{(m-1)(a-1)[2(1+m)+a(1+3m)]}{(1+m)[2(1+m)+a(5+7m)+a^2(1+7m)]}.$$

Preferential Attachment (PA) graph

$\rho(G_n)$ (blue), $\rho^{rank}(G_n)$ (red), and mean $\rho^{-}(G_n)$ (black) in 20 simulations for different n



Web and social networks

Dataset	Description	# nodes	max d	$\rho(G_n)$	$\rho(G_n)^{\mathrm{rank}}$	$\rho^{-}(G_n)$
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	co-starring	1,139,905	11,468	0.3446	0.4689	-0.6737

Web and social networks

Dataset	Description	# nodes	max d	$\rho(G_n)$	$\rho(G_n)^{\mathrm{rank}}$	$\rho^{-}(G_n)$
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	co-starring	1,139,905	11,468	0.3446	0.4689	-0.6737

- Spearman's rho is able to reveal strong negative correlations in large networks

Web and social networks

Dataset	Description	# nodes	max d	$\rho(G_n)$	$\rho(G_n)^{\mathrm{rank}}$	$\rho^-(G_n)$
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	co-starring	1,139,905	11,468	0.3446	0.4689	-0.6737

- Spearman's rho is able to reveal strong negative correlations in large networks
- Still largely open problem: statistical significance of degree-degree correlations

Web and social networks

Dataset	Description	# nodes	max d	$\rho(G_n)$	$\rho(G_n)^{\mathrm{rank}}$	$\rho^{-}(G_n)$
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	co-starring	1,139,905	11,468	0.3446	0.4689	-0.6737

- Spearman's rho is able to reveal strong negative correlations in large networks
- Still largely open problem: statistical significance of degree-degree correlations
- More on correlations in directed networks: talk of Pim

Further research

- Monte Carlo methods for fast evaluation of centrality measures and correlation measures
  - Goal: sublinear complexity

Further research

- Monte Carlo methods for fast evaluation of centrality measures and correlation measures
  - Goal: sublinear complexity
  - Hot topic
- Statistical significance of correlations in networks

Further research

- Monte Carlo methods for fast evaluation of centrality measures and correlation measures
  - Goal: sublinear complexity
  - Hot topic
- Statistical significance of correlations in networks
- Spectral analysis, second-order characteristics of centrality scores (jointly with Toulouse)

Further research

- Monte Carlo methods for fast evaluation of centrality measures and correlation measures
  - Goal: sublinear complexity
  - Hot topic
- Statistical significance of correlations in networks
- Spectral analysis, second-order characteristics of centrality scores (jointly with Toulouse)
- Optimization of the web crawler BUbiNG (jointly with Milano)