UNIVERSITY OF TWENTE.

Monte Carlo methods and mathematical analysis of directed networks

Nelly Litvak
P2: University of Twente, The Netherlands

NADINE Review 2
Brussels, 02-06-2015

P2: University of Twente, The Netherlands

Nelly Litvak, Pim van der Hoorn

P2: University of Twente, The Netherlands

Nelly Litvak, Pim van der Hoorn

Overview:

- Monte Carlo algorithms for networks
- Statistical methods for graphs
- Local and global centralities in directed random graphs

Finding top-k most popular nodes

- Problem: Find top-k network nodes with largest degrees

Finding top-k most popular nodes

- Problem: Find top-k network nodes with largest degrees
- Some applications:
  - Routing via large degree nodes
  - Proxy for various centrality measures
  - Node clustering and classification
  - Epidemic processes on networks
  - Finding most popular entities (e.g. interest groups)

Finding top-k most popular nodes

- Problem: Find top-k network nodes with largest degrees
- Some applications:
  - Routing via large degree nodes
  - Proxy for various centrality measures
  - Node clustering and classification
  - Epidemic processes on networks
  - Finding most popular entities (e.g. interest groups)
  - Many companies maintain network statistics (*twittercounter.com*, *followerwonk.com*, *twitaholic.com*, *www.insidefacebook.com*, *yavkontakte.ru*)

Top-k most popular entities in directed networks

- If the adjacency list of the network is known the top-k list of nodes can be found by the HeapSort with complexity $O(N)$, where N is the total number of nodes.

Top-k most popular entities in directed networks

- If the adjacency list of the network is known the top-k list of nodes can be found by the HeapSort with complexity $O(N)$, where N is the total number of nodes.
- Too high complexity for large networks

Top-k most popular entities in directed networks

- If the adjacency list of the network is known the top-k list of nodes can be found by the HeapSort with complexity $O(N)$, where N is the total number of nodes.
- Too high complexity for large networks
- The network can be accessed only via API, with limited access.
- Randomized algorithms: Find a 'good enough' answer with a small answer of API requests.

Top-k most popular entities in directed networks

- If the adjacency list of the network is known the top-k list of nodes can be found by the HeapSort with complexity $O(N)$, where N is the total number of nodes.
- Too high complexity for large networks
- The network can be accessed only via API, with limited access.
- Randomized algorithms: Find a 'good enough' answer with a small answer of API requests.
- A lot of attention in the literature.

Two-stage algorithm

Two-stage algorithm

- **Stage 1:** Use n_1 API requests to retrieve id's of the followees of n_1 random users
- **Stage 2:** Use n_2 API requests to check *real* degrees of the n_2 users with largest number of followers among the n_1 random users from Stage 1.
- **Result:** Return the identified top-k list of most popular users.

In total, we use $n = n_1 + n_2$ requests to API

Results on Twitter

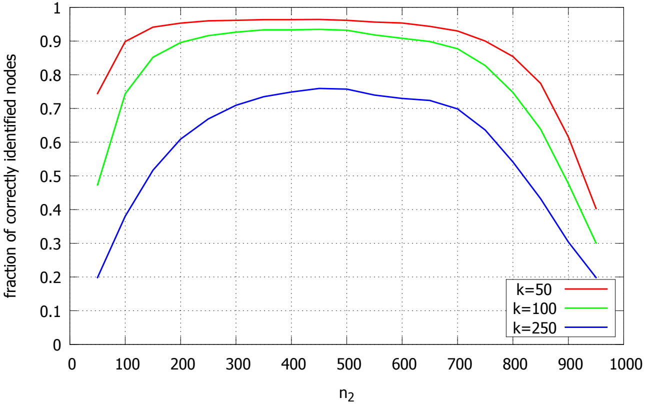


Figure : The fraction of correctly identified top-k most followed Twitter users as a function of n_2, with $n = 1000$.

Known algorithms

- Random-walk based. Cooper, Radzik, Siantos (2012)
  Transitions probabilities along undirected edges (x, y) are proportional to $(d(x)d(y))^b$, where $d(x)$ is the degree of a vertex x and $b > 0$ is some parameter.

Known algorithms

- Random-walk based. Cooper, Radzik, Siantos (2012)
  Transitions probabilities along undirected edges (x, y) are proportional to $(d(x)d(y))^b$, where $d(x)$ is the degree of a vertex x and $b > 0$ is some parameter.
- Random Walk Avrachenkov, L, Sokol, Towsley (2012)
  Random walk with uniform jumps. In an undirected graphs the stationary distribution is a linear function of degrees.
- Crawl-AI and Crawl-GAI. Kumar, Lang, Marlow, Tomkins (2008) At every step all nodes have their *apparent in-degrees* S_j, $j = 1, \dots, N$: the number of discovered edges pointing to this node. Designed for WWW crawl.

Known algorithms

- Random-walk based. Cooper, Radzik, Siantos (2012) Transitions probabilities along undirected edges (x, y) are proportional to $(d(x)d(y))^b$, where $d(x)$ is the degree of a vertex x and $b > 0$ is some parameter.
- Random Walk Avrachenkov, L, Sokol, Towsley (2012) Random walk with uniform jumps. In an undirected graphs the stationary distribution is a linear function of degrees.
- Crawl-AI and Crawl-GAI. Kumar, Lang, Marlow, Tomkins (2008) At every step all nodes have their *apparent in-degrees* S_j, $j = 1, \ldots, N$: the number of discovered edges pointing to this node. Designed for WWW crawl.
- HighestDegree. Borgs, Brautbar, Chayes, Khanna, Lucier (2012) Retrieve a random node, check in-degrees of its out-neighbors. Proceed while resources are available.

Comparison of the algorithms

Table : Percentage of correctly identified nodes from top-100 in Twitter averaged over 30 experiments, $n = 1000$

Algorithm	mean	standard deviation
Two-stage algorithm	92.6	4.7
Random walk (strict)	0.43	0.63
Random walk (relaxed)	8.7	2.4
Crawl-GAI	4.1	5.9
Crawl-AI	23.9	20.2
HighestDegree	24.7	11.8

Comparison of the algorithms

Table : Percentage of correctly identified nodes from top-100 in Twitter averaged over 30 experiments, $n = 1000$

Algorithm	mean	standard deviation
Two-stage algorithm	92.6	4.7
Random walk (strict)	0.43	0.63
Random walk (relaxed)	8.7	2.4
Crawl-GAI	4.1	5.9
Crawl-AI	23.9	20.2
HighestDegree	24.7	11.8

Advantages of the two-stage algorithm:

- does not waste resources
- obtains *exact* degrees of the n_2 'most promising' nodes

Comparison of the algorithms

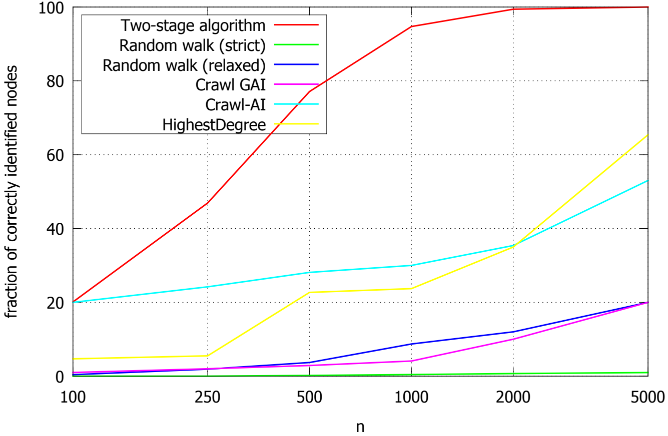


Figure : The fraction of correctly identified top-100 most followed Twitter users as a function of n averaged over 10 experiments.

Performance prediction

$G = (V, E)$ – directed graph, $|V| = N$

- Number the vertices in the decreasing order of their degrees: $F_1 \geqslant F_2 \geqslant \cdots \geqslant F_N$.

Performance prediction

$G = (V, E)$ – directed graph, $|V| = N$

- Number the vertices in the decreasing order of their degrees: $F_1 \geqslant F_2 \geqslant \cdots \geqslant F_N$.
- S_j is the number of followers of node $j = 1, 2, \ldots, N$ among the n_1 randomly chosen vertices in V
- $S_j \sim Binomial(n_1, F_j/N)$

Performance prediction

$G = (V, E)$ – directed graph, $|V| = N$

- Number the vertices in the decreasing order of their degrees: $F_1 \geqslant F_2 \geqslant \cdots \geqslant F_N$.
- S_j is the number of followers of node $j = 1, 2, \ldots, N$ among the n_1 randomly chosen vertices in V
- $S_j \sim Binomial(n_1, F_j/N)$
- $S_{i_1} \geqslant S_{i_2} \geqslant \ldots \geqslant S_{i_N}$ be the order statistics of $S_1, \ldots, S_N$.
- Performance measure:

$$E[\text{fraction of correctly identified top-}k\text{ entities}]$$

$$= \frac{1}{k}\sum_{j=1}^{k} P(j \in \{i_1, \ldots, i_{n_2}\}). \qquad (1)$$

Performance prediction

$G = (V, E)$ – directed graph, $|V| = N$

- Number the vertices in the decreasing order of their degrees: $F_1 \geqslant F_2 \geqslant \cdots \geqslant F_N$.
- S_j is the number of followers of node $j = 1, 2, \ldots, N$ among the n_1 randomly chosen vertices in V
- $S_j \sim Binomial(n_1, F_j/N)$
- $S_{i_1} \geqslant S_{i_2} \geqslant \ldots \geqslant S_{i_N}$ be the order statistics of $S_1, \ldots, S_N$.
- Performance measure:

$$E[\text{fraction of correctly identified top-}k\text{ entities}]$$

$$= \frac{1}{k} \sum_{j=1}^{k} P(j \in \{i_1, \ldots, i_{n_2}\}). \quad (1)$$

- Computation of $P(j \in \{i_1, \ldots, i_{n_2}\})$ is not feasible even if degrees are known

Poisson prediction

- $P(j \in \{i_1, \ldots, i_{n_2}\})$
  $= P(S_j > S_{i_{n_2}}) + P(S_j = S_{i_{n_2}}, j \in \{i_1, \ldots, i_{n_2}\})$
- **Example.** Twitter graph, take $n_1 = n_2 = 500$. Then the average number of nodes i with $S_i = 1$ among the top-l nodes is

$$\sum_{i=1}^{l} P(S_i = 1) = \sum_{i=1}^{l} 500 \, \frac{F_i}{5 \cdot 10^8} \left(1 - \frac{F_i}{5 \cdot 10^8}\right)^{499},$$

  which is 2540.6 for $l = 10,000$ and it is 57.4 for $l = n_2 = 500$. Hence, typically, $[S_{i_{500}} = 1]$. The event $[i \in \{i_1, \ldots, i_{n_2}\}]$ occurs only for a small fraction of nodes i with $[S_i = 1]$.

Poisson prediction

- $P(j \in \{i_1, \ldots, i_{n_2}\})$
  $= P(S_j > S_{i_{n_2}}) + P(S_j = S_{i_{n_2}}, j \in \{i_1, \ldots, i_{n_2}\})$
- **Example.** Twitter graph, take $n_1 = n_2 = 500$. Then the average number of nodes i with $S_i = 1$ among the top-l nodes is

$$\sum_{i=1}^{l} P(S_i = 1) = \sum_{i=1}^{l} 500\,\frac{F_i}{5 \cdot 10^8}\left(1 - \frac{F_i}{5 \cdot 10^8}\right)^{499},$$

  which is 2540.6 for $l = 10,000$ and it is 57.4 for $l = n_2 = 500$. Hence, typically, $[S_{i_{500}} = 1]$. The event $[i \in \{i_1, \ldots, i_{n_2}\}]$ occurs only for a small fraction of nodes i with $[S_i = 1]$.
- Approximation:
  $P(j \in \{i_1, \ldots, i_{n_2}\}) \approx P(S_j > S_{i_{n_2}}) \approx P(S_j > \max\{S_{n_2}, 1\})$

Poisson prediction

- $P(j \in \{i_1, \ldots, i_{n_2}\})$
  $= P(S_j > S_{i_{n_2}}) + P(S_j = S_{i_{n_2}}, j \in \{i_1, \ldots, i_{n_2}\})$
- **Example.** Twitter graph, take $n_1 = n_2 = 500$. Then the average number of nodes i with $S_i = 1$ among the top-l nodes is

$$\sum_{i=1}^{l} P(S_i = 1) = \sum_{i=1}^{l} 500 \, \frac{F_i}{5 \cdot 10^8} \left(1 - \frac{F_i}{5 \cdot 10^8}\right)^{499},$$

  which is 2540.6 for $l = 10,000$ and it is 57.4 for $l = n_2 = 500$. Hence, typically, $[S_{i_{500}} = 1]$. The event $[i \in \{i_1, \ldots, i_{n_2}\}]$ occurs only for a small fraction of nodes i with $[S_i = 1]$.
- Approximation:
  $P(j \in \{i_1, \ldots, i_{n_2}\}) \approx P(S_j > S_{i_{n_2}}) \approx P(S_j > \max\{S_{n_2}, 1\})$
- Assume F_j and F_{n_2} are known, then approximate $S_j \sim Poisson(n_1 F_j / N)$

EVT predictions

- Poisson approximation is not realistic: degrees are unknown

EVT predictions

- Poisson approximation is not realistic: degrees are unknown
- The algorithm finds a few highest degrees with accuracy almost 100%
- Let $\hat{F}_1 \geqslant \hat{F}_2 \geqslant \cdots \geqslant \hat{F}_m$ be the top-m degrees found by the algorithm, $m < k$

EVT predictions

- Poisson approximation is not realistic: degrees are unknown
- The algorithm finds a few highest degrees with accuracy almost 100%
- Let $\hat{F}_1 \geqslant \hat{F}_2 \geqslant \cdots \geqslant \hat{F}_m$ be the top-m degrees found by the algorithm, $m < k$
- The degrees follow a power law distribution with exponent γ

EVT predictions

- Poisson approximation is not realistic: degrees are unknown
- The algorithm finds a few highest degrees with accuracy almost 100%
- Let $\hat{F}_1 \geqslant \hat{F}_2 \geqslant \cdots \geqslant \hat{F}_m$ be the top-m degrees found by the algorithm, $m < k$
- The degrees follow a power law distribution with exponent γ
- Hill's estimator:

$$\hat{\gamma} = \left(\frac{1}{m-1} \sum_{i=1}^{m-1} \log(\hat{F}_i) - \log(\hat{F}_m) \right)^{-1}. \qquad (2)$$

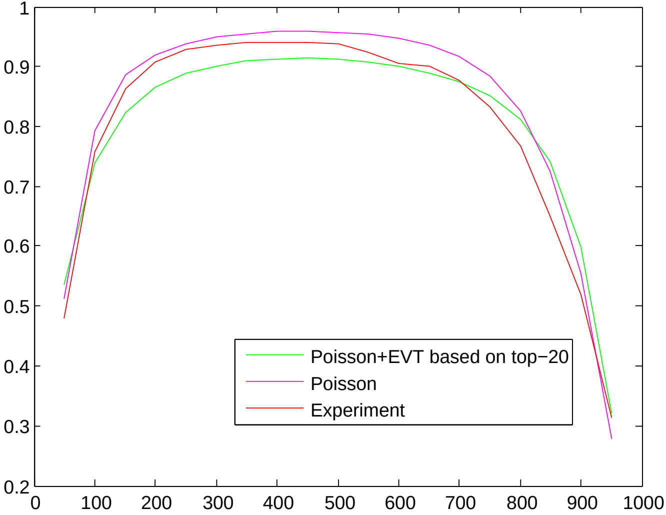
EVT predictions

- Poisson approximation is not realistic: degrees are unknown
- The algorithm finds a few highest degrees with accuracy almost 100%
- Let $\hat{F}_1 \geqslant \hat{F}_2 \geqslant \cdots \geqslant \hat{F}_m$ be the top-m degrees found by the algorithm, $m < k$
- The degrees follow a power law distribution with exponent γ
- Hill's estimator:

$$\hat{\gamma} = \left(\frac{1}{m-1} \sum_{i=1}^{m-1} \log(\hat{F}_i) - \log(\hat{F}_m) \right)^{-1}. \qquad (2)$$

- Estimator for high degrees: Dekkers et al. (1989)
$\hat{f}_j = \hat{F}_m \left(\frac{m}{j-1} \right)^{1/\hat{\gamma}}, \qquad j > 1, j << N.$
- Use $S_j \sim Poisson(n_1 \hat{f}_j / N)$

Performance predictions on the Twitter graph



Optimal parameters

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

Optimal parameters

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

Optimal parameters

- $1, \ldots, k$ – 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

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

Optimal parameters

- $1, \ldots, k$ – 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

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

- $n = O(n_1)$ (SLLN)
- Assume that $k = o(n)$ as $n \to \infty$, then the maximizer of the probability $P(k \in \{i_1, \ldots, i_{n_2}\})$ is

$$n_2 = (3\gamma k^{\gamma} n)^{\frac{1}{\gamma+1}} (1 + o(1)).$$

Sublinear complexity

$|V| = N$

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

- For any fixed $\varepsilon, \delta > 0$, our algorithm finds the fraction $1 - \varepsilon$ of top-k nodes with probability $1 - \delta$ in

  $$n = O(N/a(N))$$

  API requests, as $N \to \infty$, where $a(N) = l(N)N^{\gamma}$ and $l(\cdot)$ is some slowly varying function.

Sublinear complexity

$|V| = N$

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

- For any fixed $\varepsilon, \delta > 0$, our algorithm finds the fraction $1 - \varepsilon$ of top-k nodes with probability $1 - \delta$ in
$$n = \mathrm{O}(N/a(N))$$
API requests, as $N \to \infty$, where $a(N) = l(N)N^{\gamma}$ and $l(\cdot)$ is some slowly varying function.
- For Twitter top-k, $n = O(N^{1-1/\gamma})$

Sublinear complexity

$|V| = N$

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

- For any fixed $\varepsilon, \delta > 0$, our algorithm finds the fraction $1 - \varepsilon$ of top-k nodes with probability $1 - \delta$ in

  $$n = \mathrm{O}(N/a(N))$$

  API requests, as $N \to \infty$, where $a(N) = l(N)N^{\gamma}$ and $l(\cdot)$ is some slowly varying function.
- For Twitter top-k, $n = O(N^{1-1/\gamma})$
- High variability helps a lot!

Sublinear complexity

$|V| = N$

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

- For any fixed $\varepsilon, \delta > 0$, our algorithm finds the fraction $1 - \varepsilon$ of top-k nodes with probability $1 - \delta$ in

  $$n = \mathrm{O}(N/a(N))$$

  API requests, as $N \to \infty$, where $a(N) = l(N)N^{\gamma}$ and $l(\cdot)$ is some slowly varying function.
- For Twitter top-k, $n = O(N^{1-1/\gamma})$
- High variability helps a lot!
- K.Avrachenkov, N.Litvak, L.Ostroumova-Prokhorenkova and E.Suyargulova, **Quick detection of high-degree entities in large directed networks**, IEEE International Conference on Data Mining (ICDM 2014), (arXiv:1410.0571v2[cs.SI]) [M10-WP1.4]

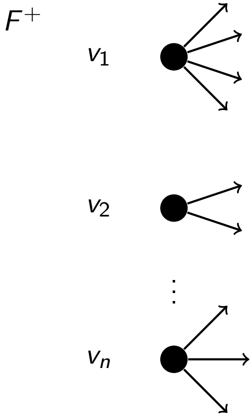
Directed random graphs

- Null-models for statistical analysis of real networks
- Theoretical characterization of centralities in networks
- In the literature, attention is mainly on undirected networks and their geometric properties (degree distributions, distances, component sizes etc.)
- We analyze centralities and statistical estimators in directed random graphs

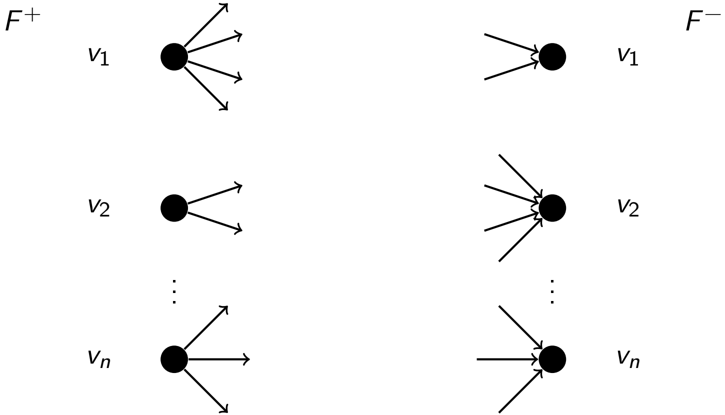
Directed Configuration Model

Directed Configuration Model

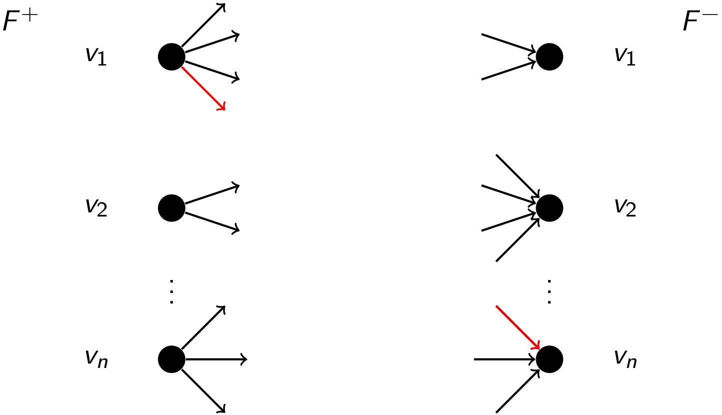
Directed Configuration Model



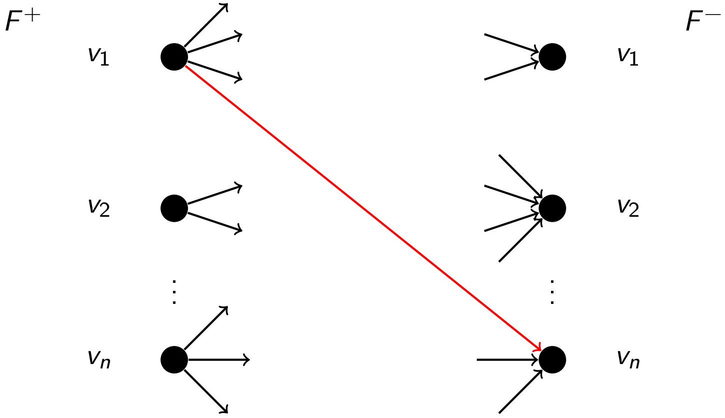
Directed Configuration Model



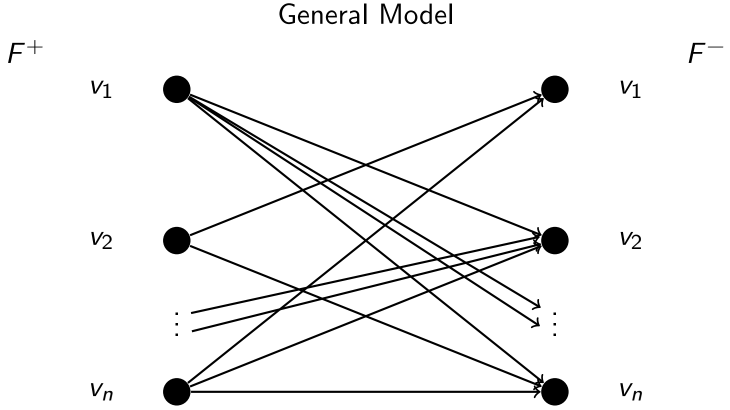
Directed Configuration Model



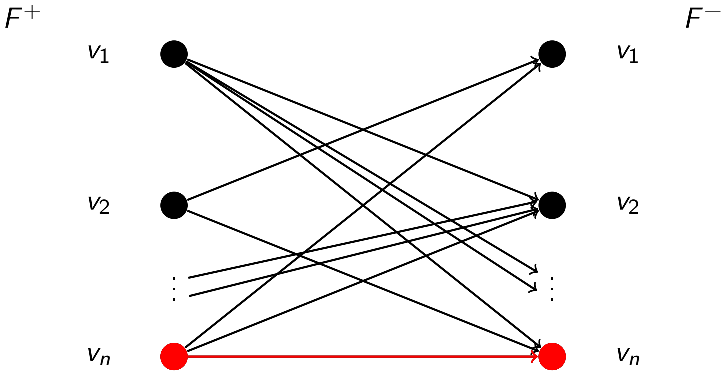
Directed Configuration Model



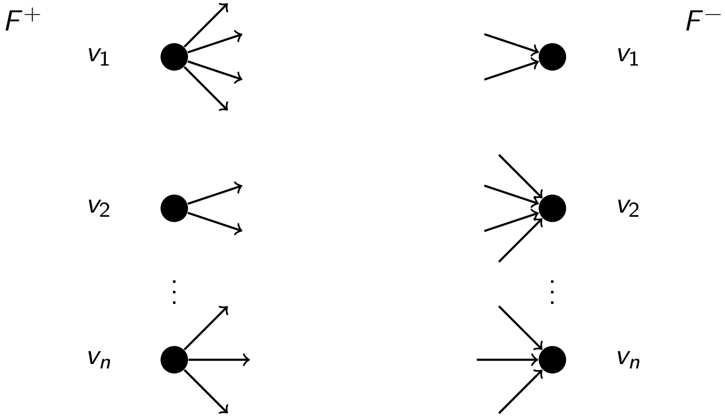
Directed Configuration Model



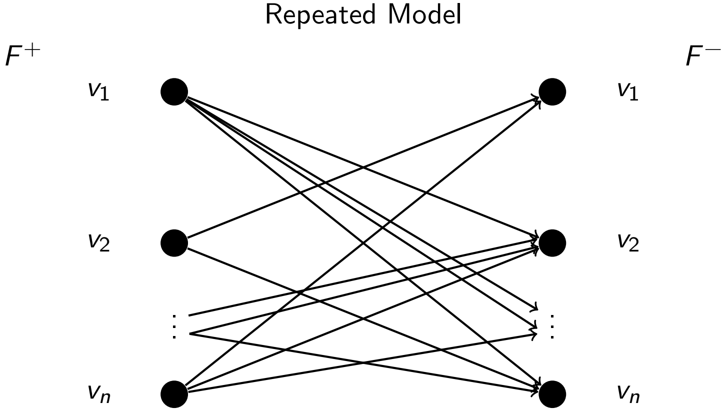
Directed Configuration Model



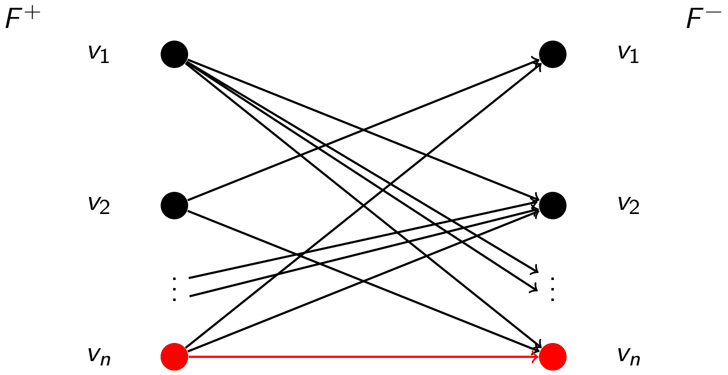
Directed Configuration Model



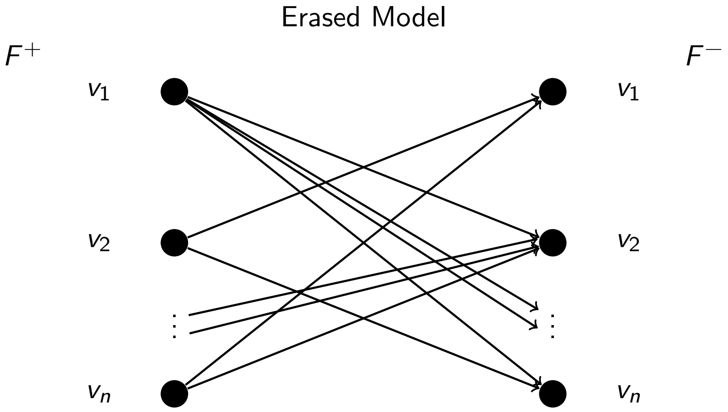
Directed Configuration Model



Directed Configuration Model

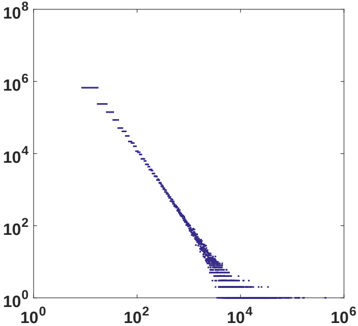


Directed Configuration Model



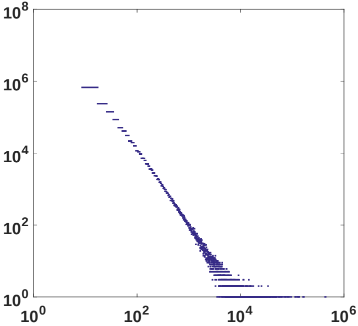
Heavy-tailed degree distributions

Heavy-tailed degree distributions



Loglog plot distribution in-degrees of English Wikipedia (data from U.Milan)

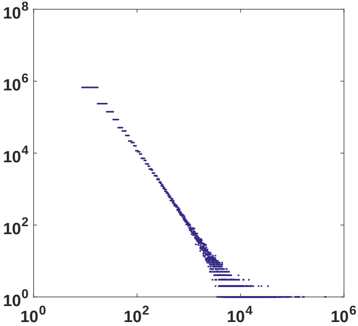
Heavy-tailed degree distributions



Loglog plot distribution in-degrees of English Wikipedia (data from U.Milan)

$$p(k) \approx k^{-\gamma-1}$$

Heavy-tailed degree distributions

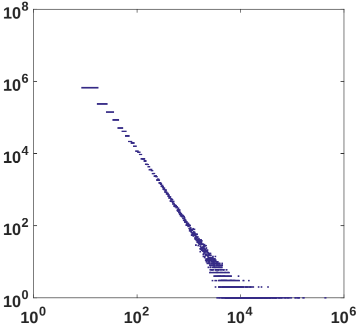


Loglog plot distribution in-degrees of English Wikipedia (data from U.Milan)

$$p(k) \approx k^{-\gamma-1}$$

$$1 < \gamma \leqslant 3$$

Heavy-tailed degree distributions

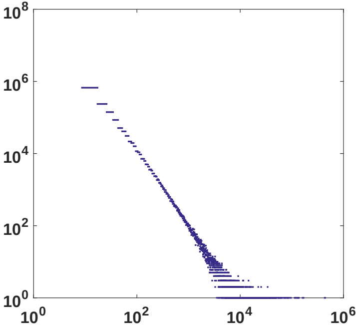


Loglog plot distribution in-degrees of English Wikipedia (data from U.Milan)

$$p(k) \approx k^{-\gamma-1}$$

$$1 < \gamma \leqslant 2$$

Heavy-tailed degree distributions

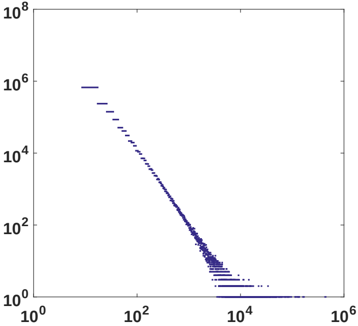


Loglog plot distribution in-degrees of English Wikipedia (data from U.Milan)

$$p(k) \approx k^{-\gamma-1}$$

$$1 < \gamma \leqslant 2 \quad \Rightarrow \quad \mathbb{E}[D] < \infty$$

Heavy-tailed degree distributions



Loglog plot distribution in-degrees of English Wikipedia (data from U.Milan)

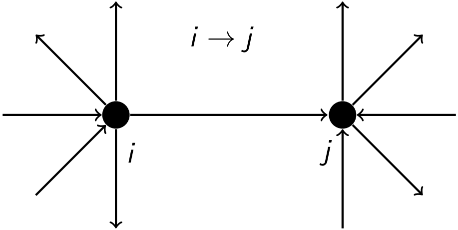
$$p(k) \approx k^{-\gamma-1}$$

$$1 < \gamma \leqslant 2 \quad \Rightarrow \quad \mathbb{E}[D] < \infty \quad \mathbb{E}\left[D^2\right] = \infty$$

Degree-degree correlations

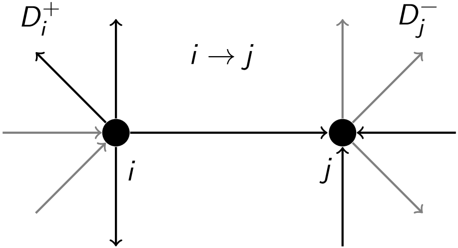
Degree-degree correlations

Given a directed graph $G = (V, E)$.



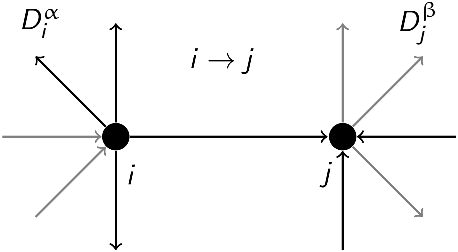
Degree-degree correlations

Given a directed graph $G = (V, E)$.



Degree-degree correlations

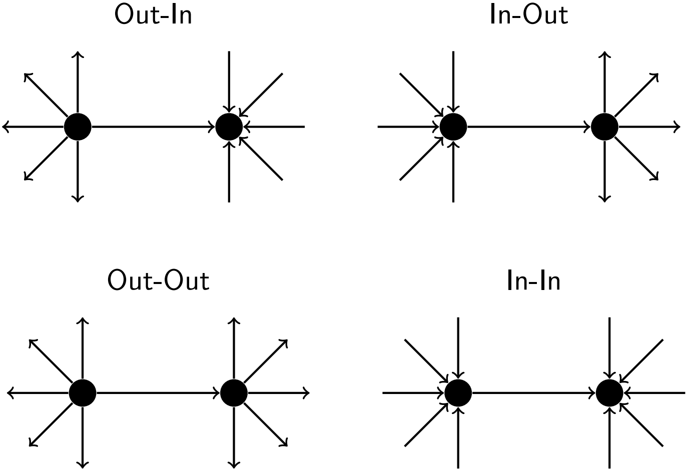
Given a directed graph $G = (V, E)$.



Index degree type by $\alpha, \beta \in \{+, -\}$.

Four types of degree-degree correlation

Four types of degree-degree correlation



Degree-degree correlations in practice

Degree-degree correlations in practice

- Information flow neural networks.
- Stability of P2P networks under attack.
- Epidemics on networks.
- Network Observability.
- Opinion dynamics based on social influence.
- Collaboration in social networks.

Degree-degree correlations in practice

- Information flow neural networks.
- Stability of P2P networks under attack.
- Epidemics on networks.
- Network Observability.
- Opinion dynamics based on social influence.
- Collaboration in social networks.
- ...

Pearson's correlation coefficients

Pearson's correlation coefficients

Given a set of m joint measurements $\{X_i, Y_i\}_{1 \leqslant i \leqslant m}$

Pearson's correlation coefficients

Given a set of m joint measurements $\{X_i, Y_i\}_{1 \leqslant i \leqslant m}$

$$r(X, Y) = \frac{\frac{1}{m}\sum_{i=1}^{m} X_i Y_i - \frac{1}{m^2}\sum_{i=1}^{m} X_i \sum_{i=1}^{m} Y_i}{\sqrt{\mathrm{Var}(X)}\,\sqrt{\mathrm{Var}(Y)}}$$

Pearson's correlation coefficients

Given a set of m joint measurements $\{X_i, Y_i\}_{1 \leqslant i \leqslant m}$

$$r(X,Y) = \frac{\frac{1}{m}\sum_{i=1}^{m} X_i Y_i - \frac{1}{m^2}\sum_{i=1}^{m} X_i \sum_{i=1}^{m} Y_i}{\sqrt{\mathrm{Var}(X)}\,\sqrt{\mathrm{Var}(Y)}}$$

$$\mathrm{Var}(X) = \frac{1}{m}\sum_{i=1}^{m} X_i^2 - \frac{1}{m^2}\left(\sum_{i=1}^{m} X_i\right)^2$$

Pearson's correlation coefficients

Given a set of m joint measurements $\{X_i, Y_i\}_{1 \leqslant i \leqslant m}$

$$r(X, Y) = \frac{\frac{1}{m}\sum_{i=1}^{m} X_i Y_i - \frac{1}{m^2}\sum_{i=1}^{m} X_i \sum_{i=1}^{m} Y_i}{\sqrt{\mathrm{Var}(X)}\,\sqrt{\mathrm{Var}(Y)}}$$

$$\mathrm{Var}(X) = \frac{1}{m}\sum_{i=1}^{m} X_i^2 - \frac{1}{m^2}\left(\sum_{i=1}^{m} X_i\right)^2$$

Given a graph G_n of size n, pick $\alpha, \beta \in \{+, -\}$.
We have E joint measurements $\{D_i^{\alpha}, D_j^{\beta}\}_{i \to j}$

Pearson's correlation coefficients

Given a set of m joint measurements $\{X_i, Y_i\}_{1 \leqslant i \leqslant m}$

$$r(X, Y) = \frac{\frac{1}{m}\sum_{i=1}^{m} X_i Y_i - \frac{1}{m^2}\sum_{i=1}^{m} X_i \sum_{i=1}^{m} Y_i}{\sqrt{\text{Var}(X)}\,\sqrt{\text{Var}(Y)}}$$

$$\text{Var}(X) = \frac{1}{m}\sum_{i=1}^{m} X_i^2 - \frac{1}{m^2}\left(\sum_{i=1}^{m} X_i\right)^2$$

Given a graph G_n of size n, pick $\alpha, \beta \in \{+, -\}$.

We have E joint measurements $\{D_i^{\alpha}, D_j^{\beta}\}_{i \to j}$

$$r_{\alpha}^{\beta}(G_n) := r(D^{\alpha}, D^{\beta})$$

Pearson's correlation coefficients

Given a set of m joint measurements $\{X_i, Y_i\}_{1 \leqslant i \leqslant m}$

$$r(X, Y) = \frac{\frac{1}{m}\sum_{i=1}^{m} X_i Y_i - \frac{1}{m^2}\sum_{i=1}^{m} X_i \sum_{i=1}^{m} Y_i}{\sqrt{\mathrm{Var}(X)}\,\sqrt{\mathrm{Var}(Y)}}$$

$$\mathrm{Var}(X) = \frac{1}{m}\sum_{i=1}^{m} X_i^2 - \frac{1}{m^2}\left(\sum_{i=1}^{m} X_i\right)^2$$

Given a graph G_n of size n, pick $\alpha, \beta \in \{+, -\}$.

We have E joint measurements $\{D_i^{\alpha}, D_j^{\beta}\}_{i \to j}$

$$r_{\alpha}^{\beta}(G_n) := r(D^{\alpha}, D^{\beta})$$

Newman 2003

Convergence of Pearson's correlation coefficients

Convergence of Pearson's correlation coefficients

Convergence of Pearson's correlation coefficients

Theorem 1 (vdHoorn and L 2014)

Let $\alpha, \beta \in \{+, -\}$. Then there exists an area $A_\alpha^\beta \subset \mathbb{R}^2$ such that if $\{G_n\}_{n\in\mathbb{N}}$ is a sequence of graphs with scale-free degree distributions where the tail-exponents $(\gamma_+, \gamma_-) \in A_\alpha^\beta$,

$$\lim_{n\to\infty} r_\alpha^\beta(G_n) \geqslant 0.$$

Convergence of Pearson's correlation coefficients

Theorem 1 (vdHoorn and L 2014)

Let $\alpha, \beta \in \{+, -\}$. Then there exists an area $A_\alpha^\beta \subset \mathbb{R}^2$ such that if $\{G_n\}_{n\in\mathbb{N}}$ is a sequence of graphs with scale-free degree distributions where the tail-exponents $(\gamma_+, \gamma_-) \in A_\alpha^\beta$,

$$\lim_{n\to\infty} r_\alpha^\beta(G_n) \geqslant 0.$$

$$1 < \gamma_\pm \leqslant 2 \in A_\alpha^\beta, \text{ for all } \alpha, \beta \in \{+, -\}$$

Rank correlations: Spearman's rho

Rank correlations: Spearman's rho

Given a graph G_n of size n, $\alpha, \beta \in \{+, -\}$

Rank correlations: Spearman's rho

Given a graph G_n of size n, $\alpha, \beta \in \{+, -\}$

We have E joint measurements $\{D_i^\alpha, D_j^\beta\}_{i \to j}$

Rank correlations: Spearman's rho

Given a graph G_n of size n, $\alpha, \beta \in \{+, -\}$

We have E joint measurements $\{D_i^{\alpha}, D_j^{\beta}\}_{i \to j}$

Compute Pearsons correlation coefficient on $\{D_i^{\alpha}, D_j^{\beta}\}_{i \to j}$

Rank correlations: Spearman's rho

Given a graph G_n of size n, $\alpha, \beta \in \{+, -\}$

Rank the degrees in descending order

We have E joint measurements $\{D_i^{\alpha}, D_j^{\beta}\}_{i \to j} \Rightarrow \{R_i^{\alpha}, R_j^{\beta}\}_{i \to j}$

Compute Pearsons correlation coefficient on $\{R_i^{\alpha}, R_j^{\beta}\}_{i \to j}$

Rank correlations: Spearman's rho

Given a graph G_n of size n, $\alpha, \beta \in \{+, -\}$

Rank the degrees in descending order

We have E joint measurements $\{D_i^\alpha, D_j^\beta\}_{i \to j} \Rightarrow \{R_i^\alpha, R_j^\beta\}_{i \to j}$

Compute Pearsons correlation coefficient on $\{R_i^\alpha, R_j^\beta\}_{i \to j}$

$$\rho_\alpha^\beta(G_n) := r(R^\alpha, R^\beta)$$

Statistical consistency Spearman's rho

Theorem 2 (vdHoorn and L 2014)

Let $\{G_n\}_{n\in\mathbb{N}}$ be a sequence of random graphs, $\alpha, \beta \in \{+,-\}$ and suppose there exist integer valued random variables $\mathcal{D}^\alpha$ and $\mathcal{D}^\beta$ such that

$$p_\alpha^\beta(k,\ell) \xrightarrow{\mathbb{P}} \mathbb{P}\left(\mathcal{D}^\alpha = k, \mathcal{D}^\beta = \ell\right) \quad \text{as } n \to \infty.$$

Then, as $n \to \infty$,

$$\rho_\alpha^\beta(G_n) \xrightarrow{\mathbb{P}} \rho\left(\mathcal{D}^\alpha, \mathcal{D}^\beta\right)$$

Spearman's rho in the Erased Configuration Model

- Simple graph: multiple edges and loops are removed
- Wiring is not entirely neutral

Spearman's rho in the Erased Configuration Model

- Simple graph: multiple edges and loops are removed
- Wiring is not entirely neutral

Theorem 3 (vdHoorn and L 2014)

Let $\{G_n\}_{n\in\mathbb{N}}$ be a sequence of graphs of size n, generated by either the Repeated or Erased Configuration Model and $\alpha, \beta \in \{+, -\}$. Then, as $n \to \infty$,

$$\rho_\alpha^\beta(G_n) \xrightarrow{\mathbb{P}} 0.$$

Spearman's rho in the Erased Configuration Model

- Simple graph: multiple edges and loops are removed
- Wiring is not entirely neutral

Theorem 3 (vdHoorn and L 2014)

Let $\{G_n\}_{n\in\mathbb{N}}$ be a sequence of graphs of size n, generated by either the Repeated or Erased Configuration Model and $\alpha, \beta \in \{+, -\}$. Then, as $n \to \infty$,

$$\rho_\alpha^\beta(G_n) \xrightarrow{\mathbb{P}} 0.$$

- Use Theorem 2

$$p_\alpha^\beta(k, \ell) \xrightarrow{\mathbb{P}} \mathbb{P}\left(\mathcal{D}^\alpha = k, \mathcal{D}^\beta = \ell\right) = \mathbb{P}\left(\mathcal{D}^\alpha = k\right)\mathbb{P}\left(\mathcal{D}^\beta = \ell\right)$$

- ECM is a null-model for degree-degree correlations

Erased model in practice

Erased model in practice

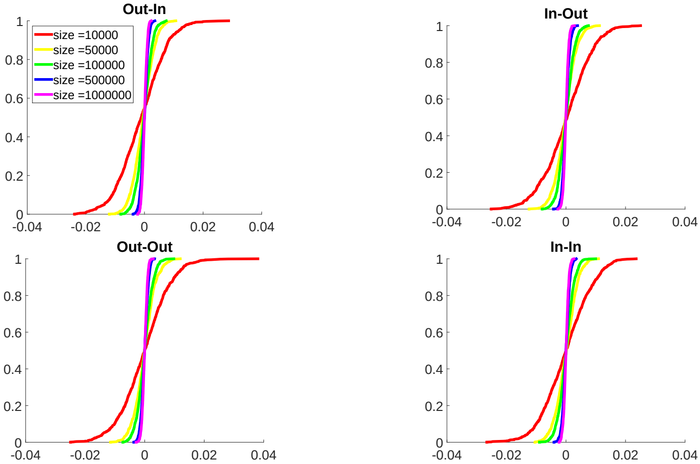


Figure : Empirical cdf of $\rho_{\alpha}^{\beta}(G_n)$ for ECM graphs with $\gamma_{\pm} = 2.1$

Erased model in practice

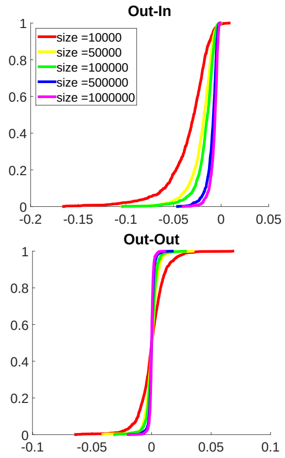
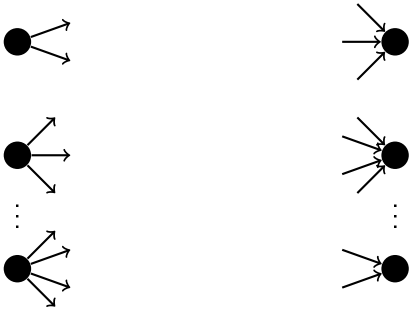


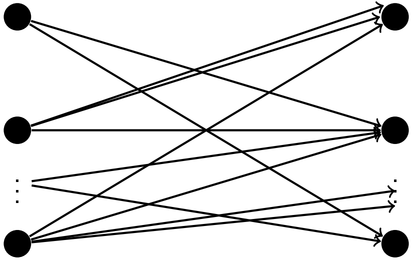
Figure : Empirical cdf of $\rho_\alpha^\beta(G_n)$ for ECM graphs with $\gamma_\pm = 1.5$

Why is Out-In different?

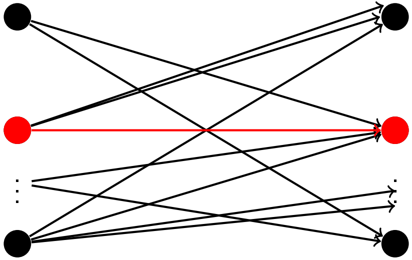
Why is Out-In different?



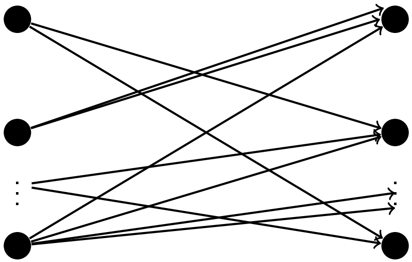
Why is Out-In different?



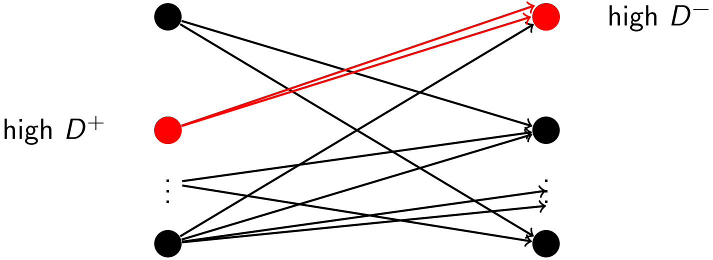
Why is Out-In different?



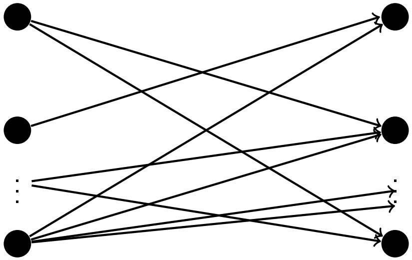
Why is Out-In different?



Why is Out-In different?

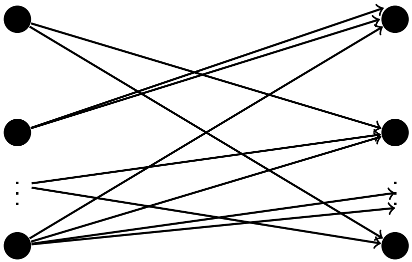


Why is Out-In different?

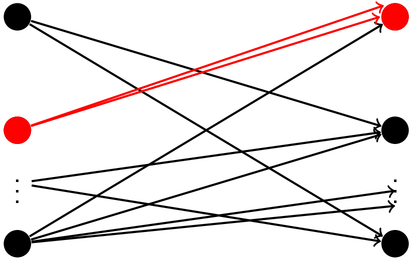


What about In-Out?

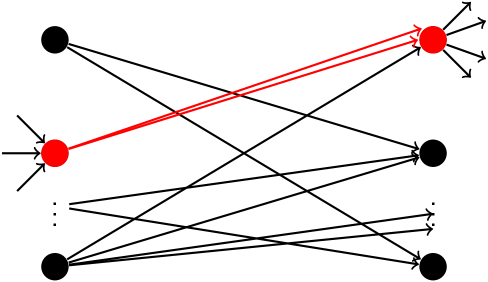
What about In-Out?



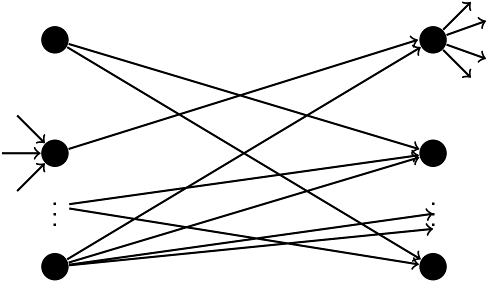
What about In-Out?



What about In-Out?



What about In-Out?



Scaling of ρ_{α}^{β}

Scaling of ρ_α^β

Let G_n be a graph of size n, generated by the ECM and denote by G_n^* the graph before the removal of edges.

Scaling of ρ_α^β

Let G_n be a graph of size n, generated by the ECM and denote by G_n^* the graph before the removal of edges.

Let E_{ij}^c denote the number of erased edges between i and j in ECM.

Scaling of ρ_α^β

Let G_n be a graph of size n, generated by the ECM and denote by G_n^* the graph before the removal of edges.

Let E_{ij}^c denote the number of erased edges between i and j in ECM.

$$D_i^{+\prime} = D_i^+ - \sum_{j=1}^{n} E_{ij}^c.$$

Scaling of ρ_α^β

Let G_n be a graph of size n, generated by the ECM and denote by G_n^* the graph before the removal of edges.

Let E_{ij}^c denote the number of erased edges between i and j in ECM.

$$D_i^{+\,\prime} = D_i^+ - \sum_{j=1}^{n} E_{ij}^c.$$

$$\left|\rho_+^-(G_n) - \rho_+^-(G_n^*)\right| = O\left(\frac{1}{E}\sum_{i,j=1}^{n} \mathbb{E}_n\left[E_{ij}^c\right]\right)$$

A first upper bound

A first upper bound

$$\sum_{i,j=1}^{n} E_{ij}^{c}$$

A first upper bound

$$\sum_{i,j=1}^{n} E_{ij}^{c} = \sum_{i,j=1}^{n} M_{ij} + \sum_{i=1}^{n} S_{ii}$$

A first upper bound

$$\sum_{i,j=1}^{n} E_{ij}^{c} = \sum_{i,j=1}^{n} M_{ij} + \sum_{i=1}^{n} S_{ii}$$

$$\mathbb{E}_n\left[S_{ii}\right] = \frac{D_i^{+} D_i^{-}}{E}$$

A first upper bound

$$\sum_{i,j=1}^{n} E_{ij}^{c} = \sum_{i,j=1}^{n} M_{ij} + \sum_{i=1}^{n} S_{ii}$$

$$\mathbb{E}_n\left[S_{ii}\right] = \frac{D_i^+ D_i^-}{E} \quad \mathbb{E}_n\left[M_{ij}\right] \leqslant \frac{(D_i^+)^2 (D_j^-)^2}{E^2}$$

A first upper bound

$$\sum_{i,j=1}^{n} E_{ij}^{c} = \sum_{i,j=1}^{n} M_{ij} + \sum_{i=1}^{n} S_{ii}$$

$$\mathbb{E}_n\left[S_{ii}\right] = \frac{D_i^+ D_i^-}{E} \quad \mathbb{E}_n\left[M_{ij}\right] \leqslant \frac{(D_i^+)^2 (D_j^-)^2}{E^2}$$

$$\frac{1}{E} \sum_{i,j=1}^{n} \mathbb{E}_n\left[E_{ij}^{c}\right] \leqslant \sum_{i,j=1}^{n} \frac{(D_i^+)^2 (D_j^-)^2}{E^3} + \sum_{i=1}^{n} \frac{D_i^+ D_i^-}{E^2}$$

A first upper bound

$$\sum_{i,j=1}^{n} E_{ij}^{c} = \sum_{i,j=1}^{n} M_{ij} + \sum_{i=1}^{n} S_{ii}$$

$$\mathbb{E}_n\left[S_{ii}\right] = \frac{D_i^+ D_i^-}{E} \quad \mathbb{E}_n\left[M_{ij}\right] \leqslant \frac{(D_i^+)^2 (D_j^-)^2}{E^2}$$

$$\frac{1}{E} \sum_{i,j=1}^{n} \mathbb{E}_n\left[E_{ij}^{c}\right] \leqslant \sum_{i,j=1}^{n} \frac{(D_i^+)^2 (D_j^-)^2}{E^3} + O\left(n^{-1}\right)$$

A first upper bound

$$\sum_{i,j=1}^{n} E_{ij}^{c} = \sum_{i,j=1}^{n} M_{ij} + \sum_{i=1}^{n} S_{ii}$$

$$\mathbb{E}_n\left[S_{ii}\right] = \frac{D_i^+ D_i^-}{E} \quad \mathbb{E}_n\left[M_{ij}\right] \leqslant \frac{(D_i^+)^2 (D_j^-)^2}{E^2}$$

$$\frac{1}{E} \sum_{i,j=1}^{n} \mathbb{E}_n\left[E_{ij}^{c}\right] \leqslant \; O\left(n^{\frac{2}{\gamma_+} + \frac{2}{\gamma_-} - 3}\right) \; + O\left(n^{-1}\right)$$

A first upper bound

$$\sum_{i,j=1}^{n} E^c_{ij} = \sum_{i,j=1}^{n} M_{ij} + \sum_{i=1}^{n} S_{ii}$$

$$\mathbb{E}_n\left[S_{ii}\right] = \frac{D_i^+ D_i^-}{E} \quad \mathbb{E}_n\left[M_{ij}\right] \leqslant \frac{(D_i^+)^2 (D_j^-)^2}{E^2}$$

$$\frac{1}{E} \sum_{i,j=1}^{n} \mathbb{E}_n\left[E^c_{ij}\right] \leqslant \; O\left(n^{\frac{2}{\gamma_+} + \frac{2}{\gamma_-} - 3}\right)$$

A second upper bound

A second upper bound

$$\frac{1}{E}\sum_{i,j=1}^{n}\mathbb{E}_n\left[E_{ij}^c\right] \leqslant 1-\frac{n^2}{E}+\frac{1}{E}\sum_{i,j=1}^{n}\exp\left\{\frac{D_i^+ D_j^-}{E}\right\}$$

CLT for heavy-tailed distributions and Tauberian theorem

A second upper bound

$$\frac{1}{E}\sum_{i,j=1}^{n}\mathbb{E}_n\left[E_{ij}^c\right] \leqslant \frac{n^2}{E}\left(\frac{1}{n^2}\sum_{i,j=1}^{n}\frac{D_i^+ D_j^-}{E} - 1 + \frac{1}{n^2}\sum_{i,j=1}^{n}\exp\left\{\frac{D_i^+ D_j^-}{E}\right\}\right)$$

CLT for heavy-tailed distributions and Tauberian theorem

A second upper bound

$$\frac{1}{E}\sum_{i,j=1}^{n}\mathbb{E}_n\left[E_{ij}^c\right] \leqslant \frac{n^2}{E}\left(\frac{1}{n^2}\sum_{i,j=1}^{n}\frac{D_i^+D_j^-}{E} - 1 + \frac{1}{n^2}\sum_{i,j=1}^{n}\exp\left\{\frac{D_i^+D_j^-}{E}\right\}\right)$$

CLT for heavy-tailed distributions and Tauberian theorem

$$\frac{1}{E}\sum_{i,j=1}^{n}\mathbb{E}_n\left[E_{ij}^c\right] \leqslant O\left(n^{\frac{1}{\gamma_+\wedge\gamma_-}-1}\right) + O\left(n^{1-(\gamma_+\wedge\gamma_-)}\right)$$

A second upper bound

$$\frac{1}{E}\sum_{i,j=1}^{n}\mathbb{E}_n\left[E_{ij}^c\right]\leqslant\frac{n^2}{E}\left(\frac{1}{n^2}\sum_{i,j=1}^{n}\frac{D_i^+D_j^-}{E}-1+\frac{1}{n^2}\sum_{i,j=1}^{n}\exp\left\{\frac{D_i^+D_j^-}{E}\right\}\right)$$

CLT for heavy-tailed distributions and Tauberian theorem

$$\frac{1}{E}\sum_{i,j=1}^{n}\mathbb{E}_n\left[E_{ij}^c\right]\leqslant O\left(n^{\frac{1}{\gamma_+\wedge\gamma_-}-1}\right)+O\left(n^{1-(\gamma_+\wedge\gamma_-)}\right)$$

$$1<\gamma_\pm\leqslant 2$$

A second upper bound

$$\frac{1}{E}\sum_{i,j=1}^{n}\mathbb{E}_n\left[E_{ij}^{c}\right]\leqslant\frac{n^2}{E}\left(\frac{1}{n^2}\sum_{i,j=1}^{n}\frac{D_i^{+}D_j^{-}}{E}-1+\frac{1}{n^2}\sum_{i,j=1}^{n}\exp\left\{\frac{D_i^{+}D_j^{-}}{E}\right\}\right)$$

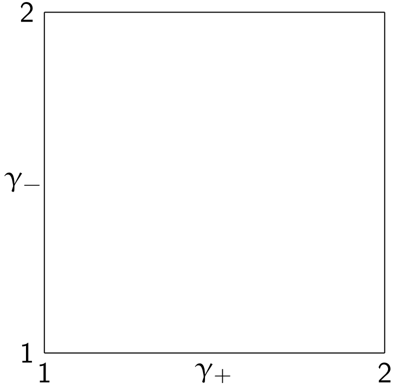
CLT for heavy-tailed distributions and Tauberian theorem

$$\frac{1}{E}\sum_{i,j=1}^{n}\mathbb{E}_n\left[E_{ij}^{c}\right]\leqslant O\left(n^{\frac{1}{\gamma_+\wedge\gamma_-}-1}\right)$$

$$1<\gamma_{\pm}\leqslant 2$$

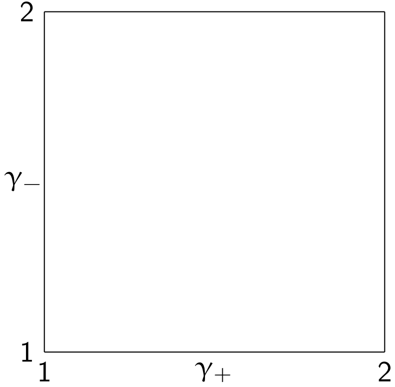
Phase transitions for $\rho^-_+(G_n)$

Phase transitions for $\rho_{+}^{-}(G_n)$



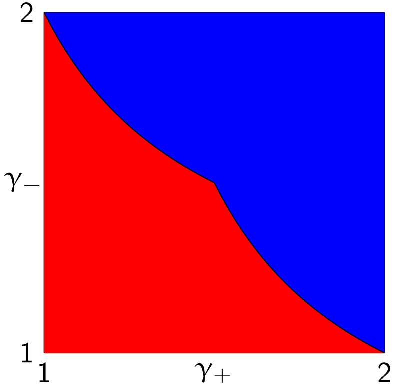
$$\rho_{+}^{-}(G_n) = O\left(\frac{1}{E}\sum_{i,j=1}^{n} \mathbb{E}_n\left[E_{ij}^{c}\right]\right)$$

Phase transitions for $\rho_+^-(G_n)$



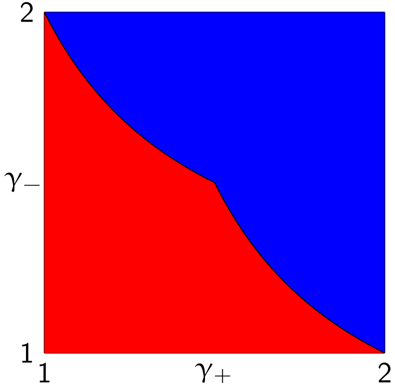
$$\rho_+^-(G_n) = O\left(\frac{1}{E}\sum_{i,j=1}^{n} \mathbb{E}_n\left[E_{ij}^c\right]\right) + O\left(\rho_+^-(G_n^*)\right)$$

Phase transitions for $\rho_{+}^{-}(G_n)$



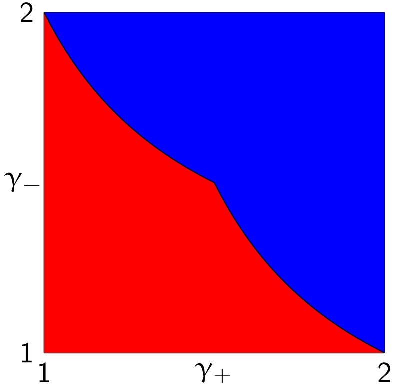
$$\rho_{+}^{-}(G_n) = O\left(\frac{1}{E}\sum_{i,j=1}^{n} \mathbb{E}_n\left[E_{ij}^{c}\right]\right) + O\left(\rho_{+}^{-}(G_n^*)\right)$$

Phase transitions for $\rho_+^-(G_n)$



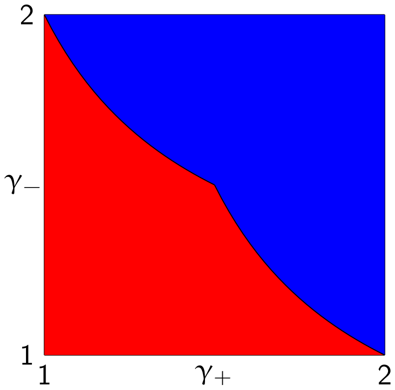
$$\rho_+^-(G_n) = O\left(\frac{1}{E}\sum_{i,j=1}^{n} \mathbb{E}_n\left[E_{ij}^c\right]\right) + O\left(\rho_+^-(G_n^*)\right)$$

Phase transitions for $\rho_+^-(G_n)$



$$\rho_+^-(G_n) = O\left(\frac{1}{E}\sum_{i,j=1}^{n} \mathbb{E}_n\left[E_{ij}^c\right]\right) + O\left(n^{-1/2}\right)$$

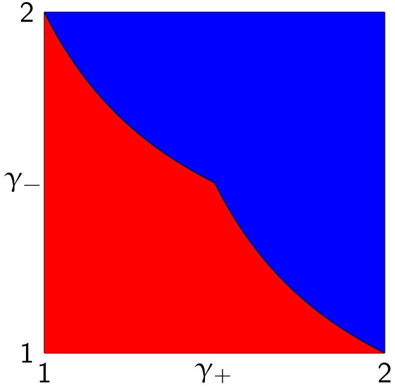
Phase transitions for $\rho_{+}^{-}(G_n)$



$$\frac{1}{\gamma_+ \wedge \gamma_-} - 1 > -\frac{1}{2}$$

$$\rho_{+}^{-}(G_n) = O\left(\frac{1}{E}\sum_{i,j=1}^{n} \mathbb{E}_n\left[E_{ij}^{c}\right]\right) + O\left(n^{-1/2}\right)$$

Phase transitions for $\rho_+^-(G_n)$

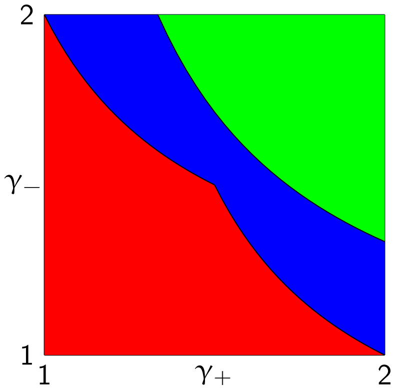


$$\frac{1}{\gamma_+ \wedge \gamma_-} - 1 > -\frac{1}{2}$$

$$\frac{2}{\gamma_+} + \frac{2}{\gamma_-} - 3 < -\frac{1}{2}$$

$$\rho_+^-(G_n) = O\left(\frac{1}{E}\sum_{i,j=1}^{n} \mathbb{E}_n\left[E_{ij}^c\right]\right) + O\left(n^{-1/2}\right)$$

Phase transitions for $\rho_{+}^{-}(G_n)$



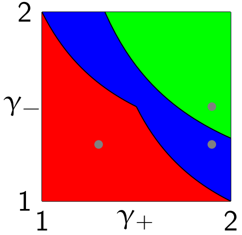
$$\frac{1}{\gamma_+ \wedge \gamma_-} - 1 > -\frac{1}{2}$$

$$\frac{2}{\gamma_+} + \frac{2}{\gamma_-} - 3 < -\frac{1}{2}$$

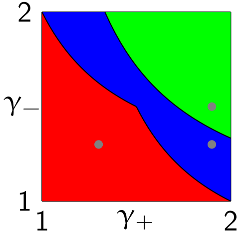
$$\rho_{+}^{-}(G_n) = O\left(\frac{1}{E}\sum_{i,j=1}^{n} \mathbb{E}_n\left[E_{ij}^c\right]\right) + O\left(n^{-1/2}\right)$$

Scaling of $\rho_{+}^{-}(G_n)$ in practice

Scaling of $\rho_+^-(G_n)$ in practice

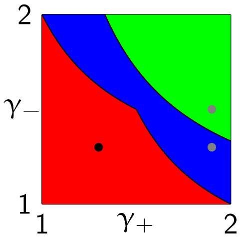


Scaling of $\rho_+^-(G_n)$ in practice

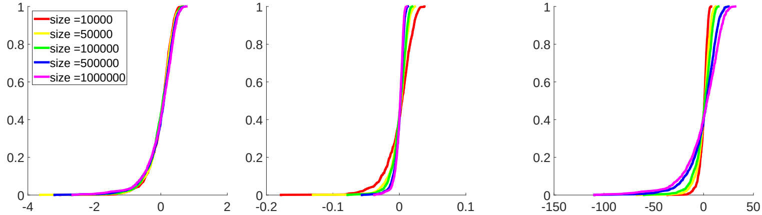


$$\frac{\rho_+^-(G_n) - \mathbb{E}\left[\rho_+^-(G_n)\right]}{N^{f(\gamma_+,\gamma_-)}}$$

Scaling of $\rho_+^-(G_n)$ in practice

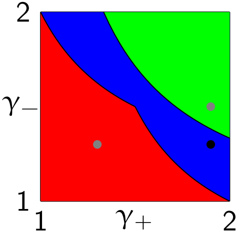


$$\frac{\rho_+^-(G_n) - \mathbb{E}\left[\rho_+^-(G_n)\right]}{N^{f(\gamma_+,\gamma_-)}}$$

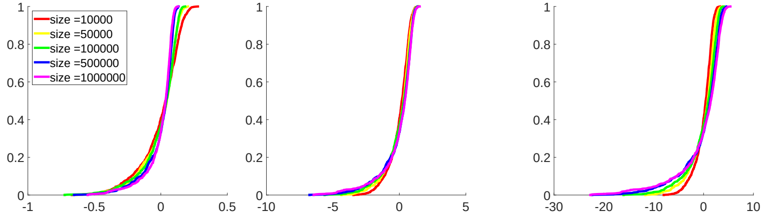


(a) $N^{-1+1/(\gamma_+\wedge\gamma_-)}$ (b) $N^{(2/\gamma_+)+(2/\gamma_-)-3}$ (c) $N^{-1/2}$

Scaling of $\rho_+^-(G_n)$ in practice

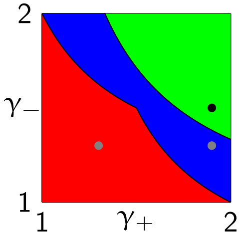


$$\frac{\rho_+^-(G_n) - \mathbb{E}\left[\rho_+^-(G_n)\right]}{N^{f(\gamma_+,\gamma_-)}}$$

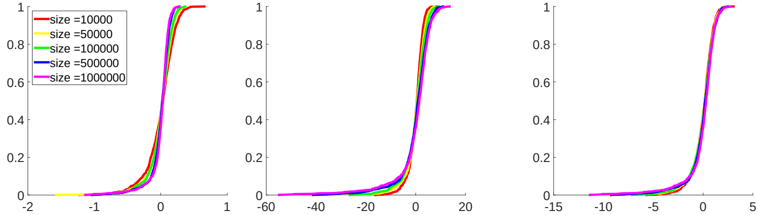


(a) $N^{-1+1/(\gamma_+ \wedge \gamma_-)}$ (b) $N^{(2/\gamma_+)+(2/\gamma_-)-3}$ (c) $N^{-1/2}$

Scaling of $\rho_+^-(G_n)$ in practice



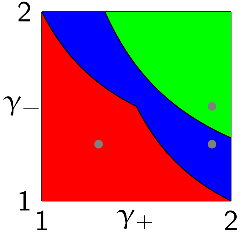
$$\frac{\rho_+^-(G_n) - \mathbb{E}\left[\rho_+^-(G_n)\right]}{N^{f(\gamma_+,\gamma_-)}}$$



(a) $N^{-1+1/(\gamma_+ \wedge \gamma_-)}$ (b) $N^{(2/\gamma_+)+(2/\gamma_-)-3}$ (c) $N^{-1/2}$

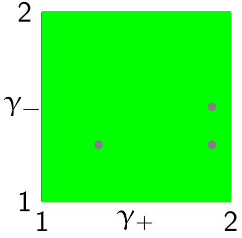
Scaling of $\rho^{+}_{-}(G_n)$ in practice

Scaling of $\rho^{+}_{-}(G_n)$ in practice



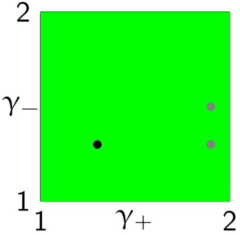
$$\frac{\rho^{+}_{-}(G_n) - \mathbb{E}\left[\rho^{+}_{-}(G_n)\right]}{N^{f(\gamma_+,\gamma_-)}}$$

Scaling of $\rho^+_-(G_n)$ in practice

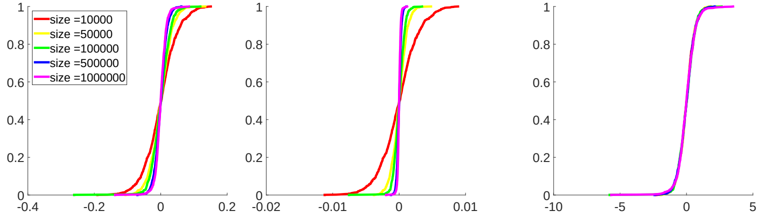


$$\frac{\rho^+_-(G_n) - \mathbb{E}\left[\rho^+_-(G_n)\right]}{N^{f(\gamma_+,\gamma_-)}}$$

Scaling of $\rho_-^+(G_n)$ in practice

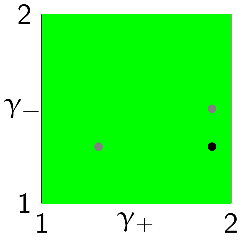


$$\frac{\rho_-^+(G_n) - \mathbb{E}\left[\rho_-^+(G_n)\right]}{N^{f(\gamma_+,\gamma_-)}}$$

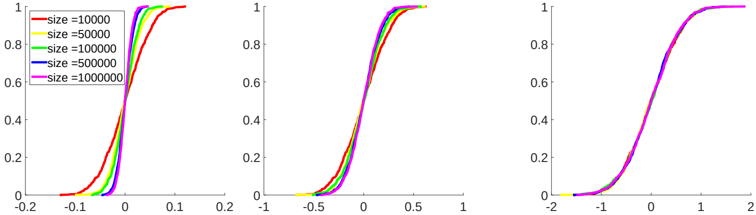


(a) $N^{-1+1/(\gamma_+ \wedge \gamma_-)}$ (b) $N^{(2/\gamma_+)+(2/\gamma_-)-3}$ (c) $N^{-1/2}$

Scaling of $\rho_-^+(G_n)$ in practice

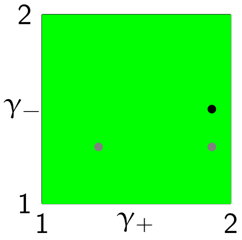


$$\frac{\rho_-^+(G_n) - \mathbb{E}\left[\rho_-^+(G_n)\right]}{N^{f(\gamma_+,\gamma_-)}}$$

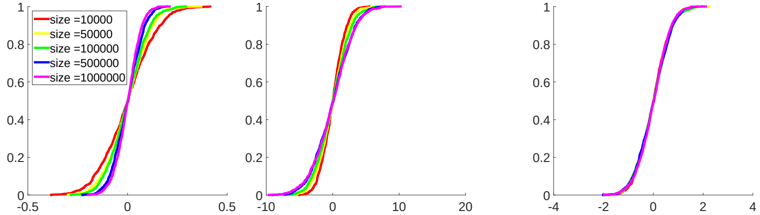


(a) $N^{-1+1/(\gamma_+ \wedge \gamma_-)}$ (b) $N^{(2/\gamma_+)+(2/\gamma_-)-3}$ (c) $N^{-1/2}$

Scaling of $\rho_-^+(G_n)$ in practice



$$\frac{\rho_-^+(G_n) - \mathbb{E}\left[\rho_-^+(G_n)\right]}{N^{f(\gamma_+,\gamma_-)}}$$



(a) $N^{-1+1/(\gamma_+ \wedge \gamma_-)}$ (b) $N^{(2/\gamma_+)+(2/\gamma_-)-3}$ (c) $N^{-1/2}$

Statistical analysis of directed networks

- ECM is easy to construct, and it is a simple graph

Statistical analysis of directed networks

- ECM is easy to construct, and it is a simple graph
- Asymptotically neutrally wired

Statistical analysis of directed networks

- ECM is easy to construct, and it is a simple graph
- Asymptotically neutrally wired
- Finite-size effects result in structural out-in correlations
- We have proved that rank correlations are consistent estimators and characterized their behavior in ECM

Statistical analysis of directed networks

- ECM is easy to construct, and it is a simple graph
- Asymptotically neutrally wired
- Finite-size effects result in structural out-in correlations
- We have proved that rank correlations are consistent estimators and characterized their behavior in ECM
- Our results lay the basis for rigorous statistical analysis of wiring preferences in directed networks of any size

Statistical analysis of directed networks

- ECM is easy to construct, and it is a simple graph
- Asymptotically neutrally wired
- Finite-size effects result in structural out-in correlations
- We have proved that rank correlations are consistent estimators and characterized their behavior in ECM
- Our results lay the basis for rigorous statistical analysis of wiring preferences in directed networks of any size
- P. van der Hoorn and N. Litvak, **Convergence of rank based degree-degree correlations in random directed networks**, Moscow Journal of Combinatorics and Number Theory (2015) (arXiv:1407.7662[math.PR], 2014) [M13-WP4.3]
- P. van der Hoorn and N. Litvak, **Phase transitions for scaling of structural correlations in directed networks**, (arXiv:1504.01535[physics.soc-ph], 2015 [M13- WP4.3]

PageRank in Directed Configuration Model (DCM)

- PageRank R_i of page $i = 1, \dots, n$ is defined as a stationary distribution of a random walk with jumps:

$$R_i = \sum_{j \to i} \frac{c}{d_j} R_j + (1 - c) q_i, \quad i = 1, \dots, n$$

- $d_j = \#$ out-links of page j
- $c \in (0, 1)$, originally 0.85, probability of a random jump
- q_i probability to jump to page i, originally, $q_i = 1/n$

PageRank in Directed Configuration Model (DCM)

- PageRank R_i of page $i = 1, \ldots, n$ is defined as a stationary distribution of a random walk with jumps:

$$R_i = \sum_{j \to i} \frac{c}{d_j} R_j + (1 - c) q_i, \quad i = 1, \ldots, n$$

- $d_j = \#$ out-links of page j
- $c \in (0, 1)$, originally 0.85, probability of a random jump
- q_i probability to jump to page i, originally, $q_i = 1/n$

- Problem: What is the distribution of the PageRank in DCM?

PageRank in Directed Configuration Model (DCM)

- PageRank R_i of page $i = 1, \ldots, n$ is defined as a stationary distribution of a random walk with jumps:

$$R_i = \sum_{j \to i} \frac{c}{d_j} R_j + (1-c)q_i, \quad i = 1, \ldots, n$$

- $d_j = \#$ out-links of page j
- $c \in (0, 1)$, originally 0.85, probability of a random jump
- q_i probability to jump to page i, originally, $q_i = 1/n$

- Problem: What is the distribution of the PageRank in DCM?
- N.Chen, N.Litvak and M.Olvera-Cravioto, **Ranking algorithms on directed configuration networks**, (arXiv:1409.7443v2[math.PR], 2014) [M7-WP5.2]

Bi-directed degree sequence

- Directed graph on n nodes $V = \{v_1, \ldots, v_n\}$.
- Extended bi-degree sequence $(\mathbf{N}_n, \mathbf{D}_n, \mathbf{C}_n, \mathbf{Q}_n) = \{(N_i, D_i, C_i, Q_i) : 1 \leqslant i \leqslant n\}$

$$L_n = \sum_{i=1}^{n} N_i = \sum_{i=1}^{n} D_i$$

Bi-directed degree sequence

- Directed graph on n nodes $V = \{v_1, \dots, v_n\}$.
- Extended bi-degree sequence
  $(\mathbf{N}_n, \mathbf{D}_n, \mathbf{C}_n, \mathbf{Q}_n) = \{(N_i, D_i, C_i, Q_i) : 1 \leqslant i \leqslant n\}$

$$L_n = \sum_{i=1}^{n} N_i = \sum_{i=1}^{n} D_i$$

- **Assumption 1.** Existence of certain limits in the spirit of the weak law of large numbers, including $\frac{1}{n}\sum_{i=1}^{n} D_i^2$ to be bounded in probability (finite variance of the out-degrees).
- **Assumption 2.** In a sequence of random graphs of growing size, the empirical probabilities $P(D_i = k)$ converge to certain distributions.

PageRank in the DCM

- $M = M(n) \in \mathbb{R}^{n \times n}$ is related to the adjacency matrix of the graph:

$$M_{i,j} = \begin{cases} s_{ij} C_i, & \text{if there are } s_{ij} \text{ edges from } i \text{ to } j, \\ 0, & \text{otherwise.} \end{cases}$$

- $Q \in \mathbb{R}^n$ is a personalization vector
- We are interested in the distribution of one coordinate, $R_1^{(n)}$, of the vector $\mathbf{R}^{(n)} \in \mathbb{R}^n$ defined by

$$\mathbf{R}^{(n)} = \mathbf{R}^{(n)} M + Q$$

Original and size-biased distribution

- Given the extended bi-degree sequence $(\mathbf{N}_n, \mathbf{D}_n, \mathbf{C}_n, \mathbf{Q}_n)$:
- Empirical distribution for the root node's parameters:

$$F_n^*(m, q) := \frac{1}{n} \sum_{k=1}^{n} 1(N_k \leqslant m, Q_k \leqslant q),$$

converges to $F^*(m, q) := P(\mathcal{N}_0 \leqslant m, \mathcal{Q}_0 \leqslant q)$

Original and size-biased distribution

- Given the extended bi-degree sequence $(\mathbf{N}_n, \mathbf{D}_n, \mathbf{C}_n, \mathbf{Q}_n)$:
- Empirical distribution for the root node's parameters:

$$F_n^*(m, q) := \frac{1}{n} \sum_{k=1}^{n} 1(N_k \leqslant m, Q_k \leqslant q),$$

converges to $F^*(m, q) := P(\mathcal{N}_0 \leqslant m, \mathcal{Q}_0 \leqslant q)$

- Empirical distribution for a node that has a out-link to any arbitrary node (size-biased by out-degree)

$$F_n(m, q, x) := \sum_{k=1}^{n} 1(N_k \leqslant m, Q_k \leqslant q, C_k \leqslant x) \frac{D_k}{L_n}$$

converges to $F(m, q, x) := P(\mathcal{N} \leqslant m, \mathcal{Q} \leqslant q) P(\mathcal{C} \leqslant x)$.

Main result

$$\mathcal{R} \stackrel{\mathcal{D}}{=} \sum_{j=1}^{\mathcal{N}} \mathcal{C}_j \mathcal{R}_j + \mathcal{Q},$$

- Let $\mathcal{R}$ denote the *endogenous* solution to the SFPE above.
- The *endogenous* solution is the limit of iterations of the recursion starting, say, from $R_0 = \mathbf{1}$.
- **Main result:**

$$R_1^{(n)} \Rightarrow \mathcal{R}^*, \qquad n \to \infty,$$

where $\Rightarrow$ denotes weak convergence and $\mathcal{R}^*$ is given by

$$\mathcal{R}^* := \sum_{j=1}^{\mathcal{N}_0} \mathcal{C}_j \mathcal{R}_j + \mathcal{Q}_0,$$

Methodology

- Three steps, three entirely different techniques.

Methodology

- Three steps, three entirely different techniques.
- **1. Finite approximation.** PageRank is accurately approximated by a finite number of matrix iterations.

Methodology

- Three steps, three entirely different techniques.
- **1. Finite approximation.** PageRank is accurately approximated by a finite number of matrix iterations.
- **2. Coupling with a tree.** Construct a coupling of the DCM graph and a "thorny branching tree" (TBT). The coupling between the graph and the TBT will hold for a number of generations in the tree that is logarithmic in n.

Methodology

- Three steps, three entirely different techniques.
- **1. Finite approximation.** PageRank is accurately approximated by a finite number of matrix iterations.
- **2. Coupling with a tree.** Construct a coupling of the DCM graph and a "thorny branching tree" (TBT). The coupling between the graph and the TBT will hold for a number of generations in the tree that is logarithmic in n.
- **3. Convergence to a weighted branching process.** Show that the rank of the root node of the TBT converges weakly to the stated limit. Chen and Olvera-Cravioto (2014)

Matrix iterations

$$\mathbf{R}^{(n,0)} = B,$$
$$\mathbf{R}^{(n,1)} = \mathbf{R}^{(n,0)}M + Q = BM + Q,$$
$$\dots$$
$$\mathbf{R}^{(n,k)} = \sum_{i=0}^{k-1} QM^i + BM^k, \quad k \geqslant 1.$$

Under event $B_n = \left\{\max_{1\leqslant i\leqslant n}|C_i|D_i \leqslant c,\ \frac{1}{n}\sum_{i=1}^{n}|Q_i| \leqslant H\right\}$

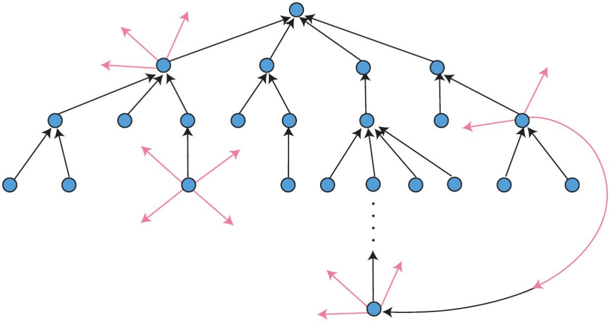
$$\left|\left|\mathbf{R}^{(n,k)} - \mathbf{R}^{(n,\infty)}\right|\right|_1 \leqslant \|\mathbf{r}_0\|_1 c^k + \sum_{i=0}^{\infty}\|\mathbf{Q}\|_1 c^{k+i} = |r_0|nc^k + \|\mathbf{Q}\|_1 \frac{c^k}{1-c}.$$

All nodes are symmetric! Markov inequality:

$$P\left(\left|R_1^{(n,\infty)} - R_1^{(n,k)}\right| > x_n^{-1}\right| B_n\Big) = O\left(x_n c^k\right)$$

Coupling with branching tree

- We start with random node (node 1) and explore its neighbours, labeling the stubs that we have already seen
- τ – the number of generations of WBP completed before coupling breaks



Coupling with branching tree

Lemma (Chen, L, Olvera-Cravioto 2014)

Suppose $(\mathbf{N}_n, \mathbf{D}_n, \mathbf{C}_n, \mathbf{Q}_n)$ *satisfies WLLN,* $\mu = E(\mathcal{N}\mathcal{D})/E(\mathcal{D})$*. Then,*

- *for any* $1 \leqslant k \leqslant h \log n$ *with* $0 < h < 1/(2\log\mu)$*, if* $\mu > 1$*,*
- *for any* $1 \leqslant k \leqslant n^b$ *with* $b < 1/2$*, if* $\mu \leqslant 1$*,*

we have

$$P\left(\tau \leqslant k \middle| \Omega_n\right) = \begin{cases} O\left((n/\mu^{2k})^{-1/2}\right), & \mu > 1, \\ O\left((n/k^2)^{-1/2}\right), & \mu = 1, \\ O\left(n^{-1/2}\right), & \mu < 1, \end{cases}$$

as $n \to \infty$*.*

Remark: μ corresponds to the average number of offspring of a node in TBT.

Numerical results-1

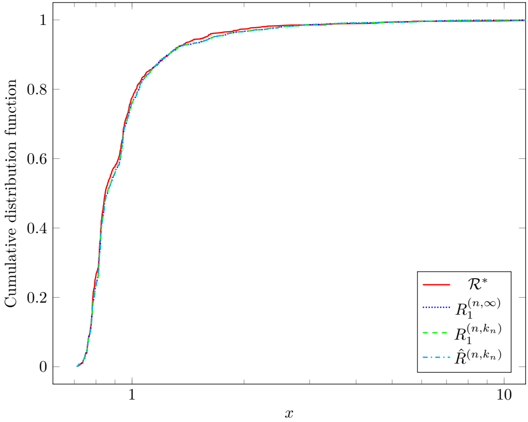


Figure : The empirical CDFs of 1000 samples of $\mathcal{R}^*$, $R_1^{(n,\infty)}$, $R_1^{(n,k_n)}$ and $\hat{R}^{(n,k_n)}$ for $n = 10000$ and $k_n = 9$.

Numerical results-2

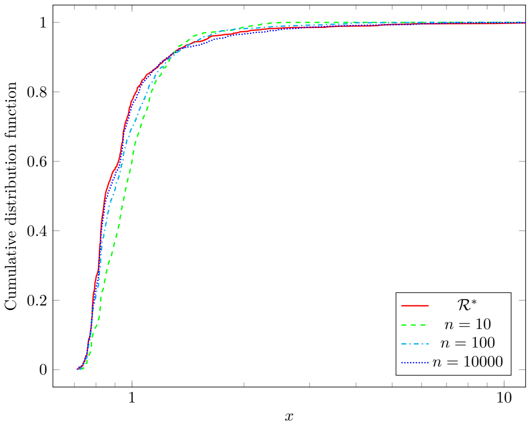


Figure : The empirical CDFs of 1000 samples of $\mathcal{R}^*$ and $R_1^{(n,\infty)}$ for $n = 10$, 100 and 10000.

Wiki graph

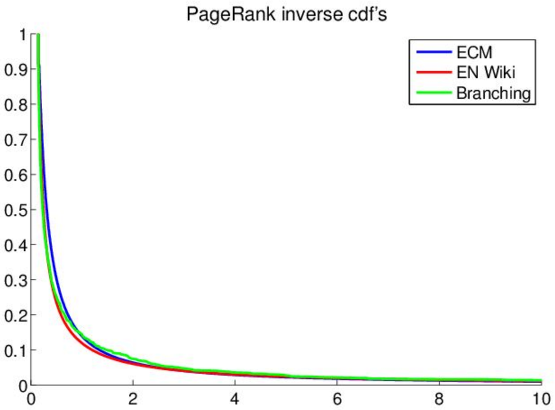


Figure : The empirical distribution of PageRank in English Wikipedia graph and its theoretical prediction. Dataset from U.Milan

Wiki graph

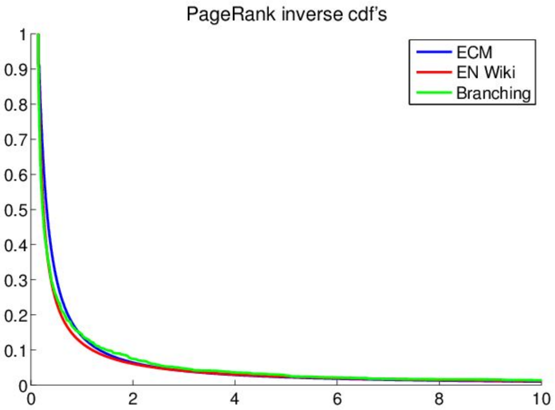


Figure : The empirical distribution of PageRank in English Wikipedia graph and its theoretical prediction. Dataset from U.Milan

Conclusions and ongoing research

- Breakthrough in probabilistic analysis of centralities and relations between them

Conclusions and ongoing research

- Breakthrough in probabilistic analysis of centralities and relations between them
- The methodology developed for analysis of PageRank in DCM can be applied for many other problems (distances, other centralities, other random graphs)

Current work:

- Distances in DCM

Conclusions and ongoing research

- Breakthrough in probabilistic analysis of centralities and relations between them
- The methodology developed for analysis of PageRank in DCM can be applied for many other problems (distances, other centralities, other random graphs)

Current work:

- Distances in DCM
- Analysis of voting models (jointly with U. Milan)

Conclusions and ongoing research

- Breakthrough in probabilistic analysis of centralities and relations between them
- The methodology developed for analysis of PageRank in DCM can be applied for many other problems (distances, other centralities, other random graphs)

Current work:

- Distances in DCM
- Analysis of voting models (jointly with U. Milan)
- Extension to dynamic centralities (jointly with MTA SZTAKI)