UNIVERSITY OF TWENTE.

Finding central nodes in large networks

Nelly Litvak

University of Twente
Eindhoven University of Technology,
The Netherlands

Woudschoten Conference 2017

Google PageRank

- PageRank r_i of page $i = 1, \ldots, n$ is defined as:

$$r_i = \sum_{j\,:\,j \to i} \frac{\alpha}{d_j} r_j + (1-\alpha) q_i, \quad i = 1, \ldots, n$$

- $d_j = \#$ out-links of page j
- $\alpha \in (0,1)$, *damping factor* originally 0.85
- $q_i \geqslant 0$, $\sum_i q_i = 1$, originally, $q_i = 1/n$.

Easily bored surfer model

$$r_i = \sum_{j\,:\,j \to i} \frac{\alpha}{d_j} r_j + (1-\alpha) q_i, \quad i = 1, \ldots, n$$

- r_i is a stationary distribution of a Markov chain

Easily bored surfer model

$$r_i = \sum_{j\,:\,j \to i} \frac{\alpha}{d_j} r_j + (1-\alpha) q_i, \quad i = 1, \ldots, n$$

- r_i is a stationary distribution of a Markov chain
- with probability α follow a randomly chosen outgoing link
- with probability $1-\alpha$ random jump (to page i w.p. q_i)

Easily bored surfer model

$$r_i = \sum_{j : j \to i} \frac{\alpha}{d_j} r_j + (1-\alpha) q_i, \quad i = 1, \ldots, n$$

- r_i is a stationary distribution of a Markov chain
- with probability α follow a randomly chosen outgoing link
- with probability $1-\alpha$ random jump (to page i w.p. q_i)
- Dangling nodes, $d_j = 0$:
  - Random jump from dangling nodes

Easily bored surfer model

$$r_i = \sum_{j\,:\,j \to i} \frac{\alpha}{d_j} r_j + (1-\alpha) q_i, \quad i = 1, \ldots, n$$

- r_i is a stationary distribution of a Markov chain
- with probability α follow a randomly chosen outgoing link
- with probability $1-\alpha$ random jump (to page i w.p. q_i)
- Dangling nodes, $d_j = 0$:
  - Random jump from dangling nodes
  - Stationary distribution $\pi = \mathbf{r}/\|\mathbf{r}\|_1$

Easily bored surfer model

$$r_i = \sum_{j:j \to i} \frac{\alpha}{d_j} r_j + (1-\alpha) q_i, \quad i = 1, \ldots, n$$

- r_i is a stationary distribution of a Markov chain
- with probability α follow a randomly chosen outgoing link
- with probability $1-\alpha$ random jump (to page i w.p. q_i)
- Dangling nodes, $d_j = 0$:
  - Random jump from dangling nodes
  - Stationary distribution $\pi = \mathbf{r}/\|\mathbf{r}\|_1$
- The page is important if many important pages link to it!

Linear equation and eigenvector problem

$$\mathbf{r} = \alpha \mathbf{r} P + (1-\alpha)\mathbf{q}$$

$$\mathbf{r} = \mathbf{r}\left[\alpha P + \frac{1-\alpha}{n}\mathbf{1}^t \mathbf{q}\right] \qquad \text{eigenvector problem}$$

$$\mathbf{r} = (1-\alpha)\mathbf{q}[I - \alpha P]^{-1} = (1-\alpha)\mathbf{q}\sum_{t=0}^{\infty} \alpha^t P^t.$$

Matrix expansion

$$\mathbf{r} = \alpha \mathbf{r} P + (1-\alpha)\mathbf{q}$$

$$\mathbf{r} = (1-\alpha)\mathbf{q}[I - \alpha P]^{-1} = (1-\alpha)\mathbf{q}\sum_{t=0}^{\infty} \alpha^t P^t.$$

- Computation by matrix iterations:

$$\begin{aligned}\mathbf{r}^{(0)} &= (1/n, \ldots, 1/n)\\ \mathbf{r}^{(k)} &= \alpha \mathbf{r}^{(k-1)} P + (1-\alpha)\mathbf{q}\\ &= \mathbf{r}^{(0)}\alpha^k P^k + (1-\alpha)\mathbf{q}\sum_{t=0}^{k-1}\alpha^t P^t\end{aligned}$$

Matrix expansion

$$\mathbf{r} = \alpha \mathbf{r} P + (1-\alpha)\mathbf{q}$$

$$\mathbf{r} = (1-\alpha)\mathbf{q}[I - \alpha P]^{-1} = (1-\alpha)\mathbf{q} \sum_{t=0}^{\infty} \alpha^t P^t.$$

- Computation by matrix iterations:

$$\begin{aligned} \mathbf{r}^{(0)} &= (1/n, \ldots, 1/n) \\ \mathbf{r}^{(k)} &= \alpha \mathbf{r}^{(k-1)} P + (1-\alpha)\mathbf{q} \\ &= \mathbf{r}^{(0)} \alpha^k P^k + (1-\alpha)\mathbf{q} \sum_{t=0}^{k-1} \alpha^t P^t \end{aligned}$$

- Exponentially fast convergence due to $\alpha \in (0,1)$

Matrix expansion

$$\mathbf{r} = \alpha\mathbf{r}P + (1-\alpha)\mathbf{q}$$

$$\mathbf{r} = (1-\alpha)\mathbf{q}[I-\alpha P]^{-1} = (1-\alpha)\mathbf{q}\sum_{t=0}^{\infty}\alpha^t P^t.$$

- Computation by matrix iterations:

$$\begin{aligned}\mathbf{r}^{(0)} &= (1/n,\ldots,1/n)\\ \mathbf{r}^{(k)} &= \alpha\mathbf{r}^{(k-1)}P + (1-\alpha)\mathbf{q}\\ &= \mathbf{r}^{(0)}\alpha^k P^k + (1-\alpha)\mathbf{q}\sum_{t=0}^{k-1}\alpha^t P^t\end{aligned}$$

- Exponentially fast convergence due to $\alpha \in (0,1)$
- Matrix iterations are used to compute PageRank in practice Langville&Meyer 2004, Berkhin 2005

Plan

- Part I: Centrality & computaitonal aspects
- Part II: PageRank

Effect of adding or removing links

$$r_i = (1-\alpha)q_i + (1-\alpha)\sum_{j=1}^{n} q_j \sum_{t=1}^{\infty} \alpha^t (P^t)_{ji}$$

The influence of the nodes on the PageRank of node i decreases *exponentially* with the distance from i.

(X_t) – Markov chain with transition matrix P.

$$\sum_{t=1}^{\infty} \alpha^t (P^t)_{ji} = \sum_{t=1}^{\infty} \alpha^t E_j \mathbf{1}[X_t = i] = E_j[\#\text{ visits to } i \text{ before a jump}]$$

$$= P_j(\text{reach } i \text{ before a jump}) E_i[1 + (\#\text{ returns to } i \text{ before a jump})]$$

- Influence of out-degrees is very limited (Avrachenkov&L 2006)

Effect of adding or removing links

$$r_i = (1-\alpha)q_i + (1-\alpha)\sum_{j=1}^{n} q_j \sum_{t=1}^{\infty} \alpha^t (P^t)_{ji}$$

The influence of the nodes on the PageRank of node i decreases *exponentially* with the distance from i.

(X_t) – Markov chain with transition matrix P.

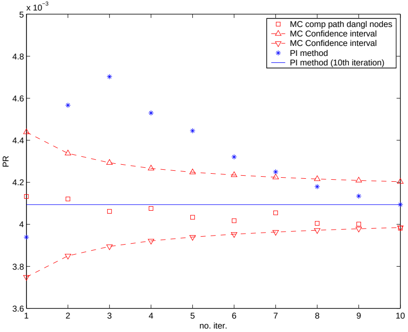
$$\sum_{t=1}^{\infty} \alpha^t (P^t)_{ji} = \sum_{t=1}^{\infty} \alpha^t E_j \mathbf{1}[X_t = i] = E_j[\#\text{ visits to } i \text{ before a jump}]$$

$$= P_j(\text{reach } i \text{ before a jump}) E_i[1 + (\#\text{ returns to } i \text{ before a jump})]$$

- Influence of out-degrees is very limited (Avrachenkov&L 2006)
- Your best PageRank boosting strategy?

Monte-Carlo computations

- Random walk from each node, length *Geometric*$(1-\alpha)$
- Compute the average number of visits to i



Avrachenkov, L, Nemirovsky, Osipova 2007

The influence of α

$$\mathbf{r} = (1-\alpha)\mathbf{q}\sum_{t=0}^{\infty}\alpha^t P^t$$

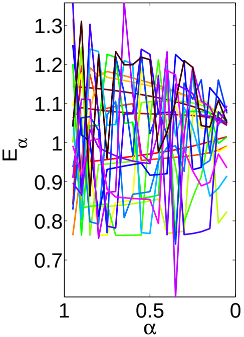


Figure: $-\log$(PageRank) for top-20 Dutch Wiki pages

The influence of α

$$\mathbf{r} = (1-\alpha)\mathbf{q}[I-\alpha P]^{-1} = (1-\alpha)\mathbf{q}\sum_{t=0}^{\infty}\alpha^t P^t$$

- PageRank greately depends on α:
- $\alpha \approx 1$

The influence of α

$$\mathbf{r} = (1-\alpha)\mathbf{q}[I - \alpha P]^{-1} = (1-\alpha)\mathbf{q}\sum_{t=0}^{\infty}\alpha^t P^t$$

- PageRank greately depends on α:
- $\alpha \approx 1$ – ranking by random walk

The influence of α

$$\mathbf{r} = (1-\alpha)\mathbf{q}[I - \alpha P]^{-1} = (1-\alpha)\mathbf{q}\sum_{t=0}^{\infty} \alpha^t P^t$$

- PageRank greately depends on α:
- $\alpha \approx 1$ – ranking by random walk
- $\alpha \approx 0$

The influence of α

$$\mathbf{r} = (1-\alpha)\mathbf{q}[I - \alpha P]^{-1} = (1-\alpha)\mathbf{q}\sum_{t=0}^{\infty} \alpha^t P^t$$

- PageRank greately depends on α:
- $\alpha \approx 1$ – ranking by random walk
- $\alpha \approx 0$ – ranking by in-degree

The influence of α

$$\mathbf{r} = (1-\alpha)\mathbf{q}[I-\alpha P]^{-1} = (1-\alpha)\mathbf{q}\sum_{t=0}^{\infty}\alpha^t P^t$$

- PageRank greately depends on α:
- $\alpha \approx 1$ – ranking by random walk
- $\alpha \approx 0$ – ranking by in-degree
- Langville & Meyer 2004, Baeza-Yates, Boldi & Castillo 2006

How large α should be?

- Initial value 0.85 was found by trial and error

How large α should be?

- Initial value 0.85 was found by trial and error
- Convergence rate of power method is α (Haveliwala & Kamvar 2003)

How large α should be?

- Initial value 0.85 was found by trial and error
- Convergence rate of power method is α (Haveliwala & Kamvar 2003)
- PageRank is more stable with smaller α

How large α should be?

- Initial value 0.85 was found by trial and error
- Convergence rate of power method is α (Haveliwala & Kamvar 2003)
- PageRank is more stable with smaller α
- Ranking by a random-walk is not necessarily a good thing

How large α should be?

- Initial value 0.85 was found by trial and error
- Convergence rate of power method is α (Haveliwala & Kamvar 2003)
- PageRank is more stable with smaller α
- Ranking by a random-walk is not necessarily a good thing

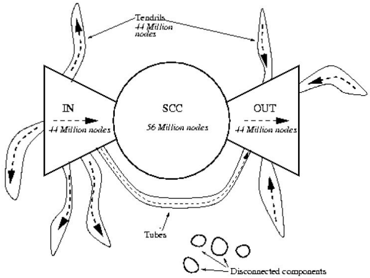


Figure: Broder et al. 2000

How large α should be?

- Initial value 0.85 was found by trial and error
- Convergence rate of power method is α (Haveliwala & Kamvar 2003)
- PageRank is more stable with smaller α
- Ranking by a random-walk is not necessarily a good thing

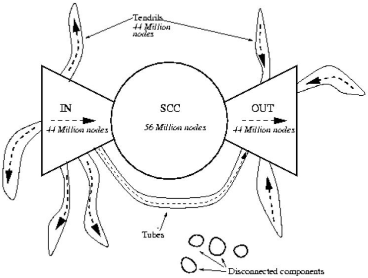


Figure: Broder et al. 2000

- Choose $\alpha = 1/2$ to balance the components (Avrachenkov, L & Kim 2006)

Power laws in complex networks

- Power laws: Internet, WWW, social networks, biological networks, etc...

Power laws in complex networks

- Power laws: Internet, WWW, social networks, biological networks, etc...
- degree of the node $=\#$ (in-/out-) links
- [fraction nodes degree at least k] $= p_k$,
- Power law: $p_k \approx const \cdot k^{-\gamma-1}$, $\alpha > 0$.
- Power law is the model for high variability: some nodes (hubs) have extremely many connections

Power laws in complex networks

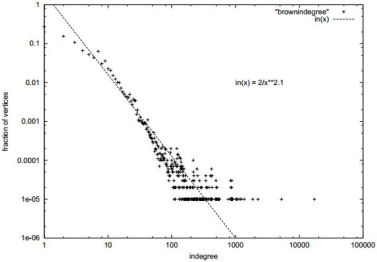
- Power laws: Internet, WWW, social networks, biological networks, etc...
- degree of the node = # (in-/out-) links
- [fraction nodes degree at least k] $= p_k$,
- Power law: $p_k \approx const \cdot k^{-\gamma-1}$, $\alpha > 0$.
- Power law is the model for high variability: some nodes (hubs) have extremely many connections
- $\log p_k = log(const) - (\gamma + 1) \log k$

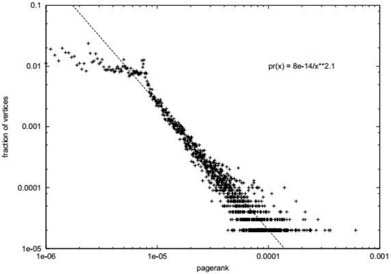
Power laws in complex networks

- Power laws: Internet, WWW, social networks, biological networks, etc...
- degree of the node = # (in-/out-) links
- [fraction nodes degree at least k] $= p_k$,
- Power law: $p_k \approx const \cdot k^{-\gamma-1}$, $\alpha > 0$.
- Power law is the model for high variability: some nodes (hubs) have extremely many connections
- $\log p_k = log(const) - (\gamma + 1) \log k$
- Straight line on the log-log scale

Power law of PageRank

Pandurangan, Raghavan, Upfal 2002.





Stochastic model for PageRank

$$r_i = \sum_{j\,:\,j \to i} \frac{\alpha}{d_j} r_j + (1-\alpha) q_i, \quad i = 1, \ldots, n$$

Stochastic model for PageRank

$$r_i = \sum_{j\,:\,j \to i} \frac{\alpha}{d_j} r_j + (1-\alpha) q_i, \quad i = 1, \ldots, n$$

- Rescale: $R_i = n r_i$, $Q_i \to n(1-\alpha)q_i$ so that $E(R) = 1$
- **Idea:** model R as a solution of stochastic equation (Volkovich&L 2010):

$$R \stackrel{d}{=} \alpha \sum_{j=1}^{N} \frac{1}{D_j} R_j + Q$$

- N: in-degree of the randomly chosen page
- D: out-degree of page that links to the randomly chosen page
- R_j is distributed as R; N, D, R_j are independent
- Denote $C_j = \alpha / D_j$.

Stochastic model for PageRank

$$r_i = \sum_{j\,:\,j \to i} \frac{\alpha}{d_j} r_j + (1-\alpha) q_i, \quad i = 1, \ldots, n$$

- Rescale: $R_i = n r_i$, $Q_i \to n(1-\alpha) q_i$ so that $E(R) = 1$
- **Idea:** model R as a solution of stochastic equation (Volkovich&L 2010):

$$R \stackrel{d}{=} \sum_{j=1}^{N} C_j R_j + Q$$

- N: in-degree of the randomly chosen page
- D: out-degree of page that links to the randomly chosen page
- R_j is distributed as R; N, D, R_j are independent
- Denote $C_j = \alpha / D_j$.

Results for stochastic recursion

$$R \stackrel{d}{=} \sum_{j=1}^{N} C_j R_j + Q$$

Theorem (Volkovich&L 2010)

If $P(Q > x) = o(P(N > x))$, *then the following are equivalent:*

- $P(N > x) \sim L(x)x^{-\gamma}$ *as* $x \to \infty$,
- $P(R > x) \sim aL(x)x^{-\gamma}$ *as* $x \to \infty$, *where* $a = (E[C])^{\gamma}(1 - \mathbb{E}[N]\mathbb{E}[C^{\gamma}])^{-1}$

- Here $a \sim b$ means $a/b \to 1$
- Note that $E[C] = 1/E(N)$, the role of out-degrees is minimal!

Results for stochastic recursion

$$R \stackrel{d}{=} \sum_{j=1}^{N} C_j R_j + Q$$

- Olvera-Cravioto, Jelenkovic 2010, 2012 analyzed the recursion under most general assumptions on C_j's. For example, R can be heavy-tailed even when N is light-tailed.

Results for stochastic recursion

$$R \stackrel{d}{=} \sum_{j=1}^{N} C_j R_j + Q$$

- Olvera-Cravioto, Jelenkovic 2010, 2012 analyzed the recursion under most general assumptions on C_j's. For example, R can be heavy-tailed even when N is light-tailed.
- However, this does not completely explain the behavior of PageRank in networks because the recursion implicitly assumes an underlying *tree* structure.

Results for stochastic recursion

$$R \stackrel{d}{=} \sum_{j=1}^{N} C_j R_j + Q$$

- Olvera-Cravioto, Jelenkovic 2010, 2012 analyzed the recursion under most general assumptions on C_j's. For example, R can be heavy-tailed even when N is light-tailed.
- However, this does not completely explain the behavior of PageRank in networks because the recursion implicitly assumes an underlying *tree* structure.
- We now want to extend the result to random graphs!

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

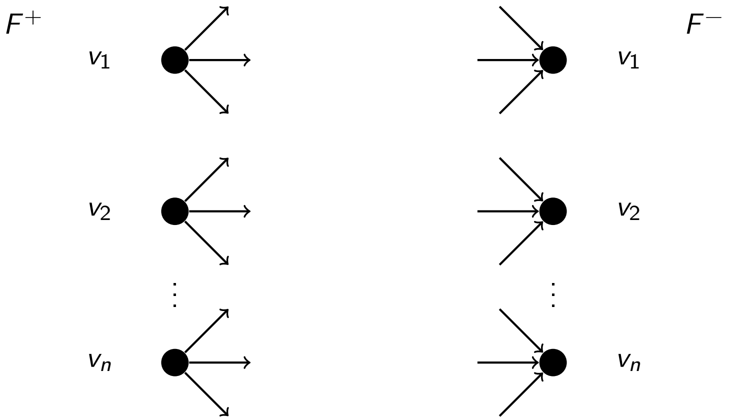
$$L_n = \sum_{i=1}^{n} N_i = \sum_{i=1}^{n} D_i$$

- 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.
- Example: Chen&Olvera-Cravioto 2013

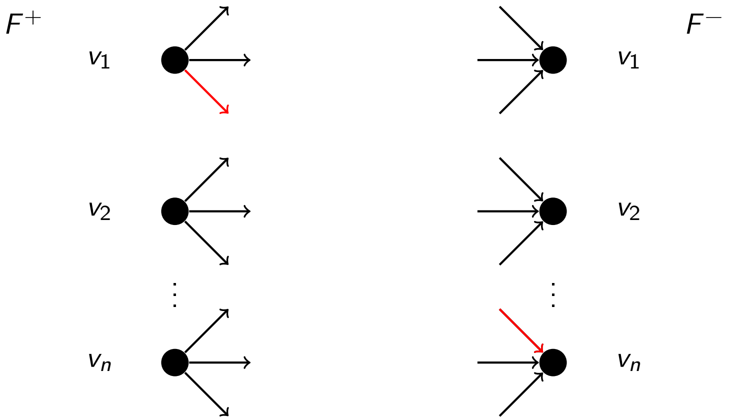
Directed Configuration Model (DCM)

Directed Configuration Model (DCM)

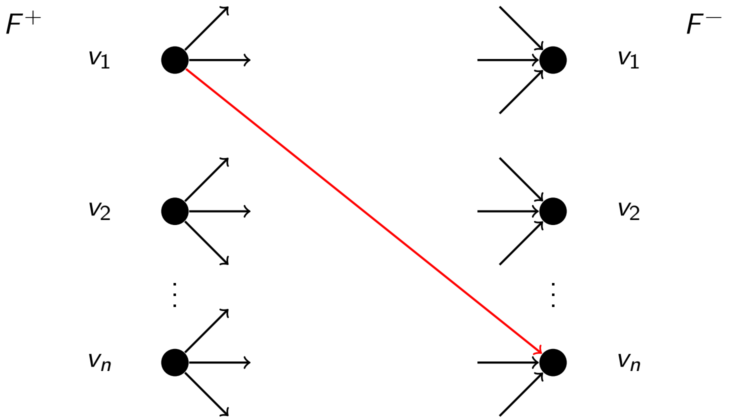
Directed Configuration Model (DCM)



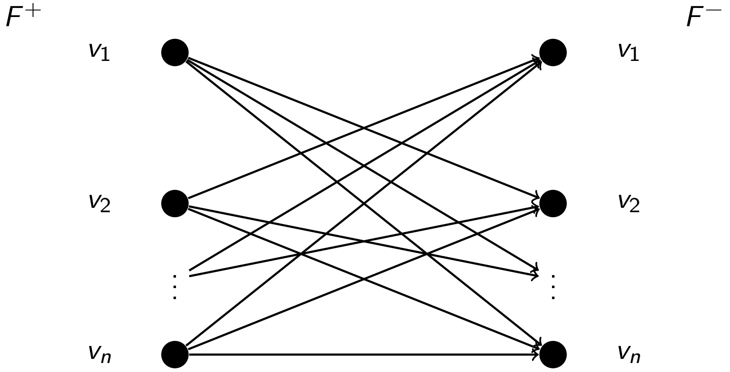
Directed Configuration Model (DCM)



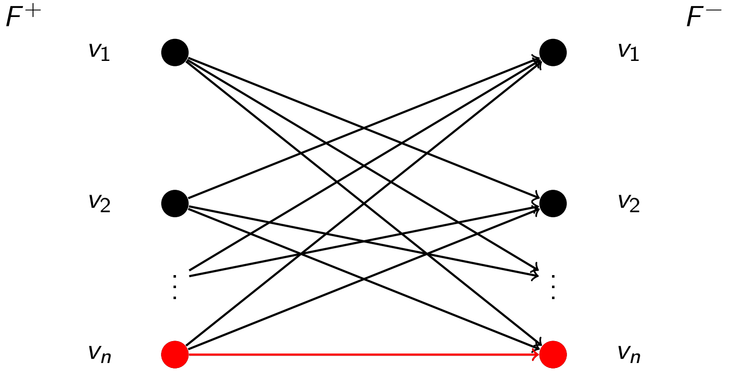
Directed Configuration Model (DCM)



Directed Configuration Model (DCM)



Directed Configuration Model (DCM)



We keep self-loops and double edges.
The result is a multi-graph

PageRank in the DCM

Chen, L, Olvera-Cravioto 2014

- $M = M(n) \in \mathbb{R}^{n \times n}$ 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

Chen, L, Olvera-Cravioto 2016

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

- 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

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

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

- We want to bound $|R_1^{(n,\infty)} - R_1^{(n,k)}|$
- The standard results on mixing times do not help to get rid of the factor n

Convergence for matrix iterations

- All nodes are symmetric!
- $E_n(|R_1^{(n,\infty)} - R_1^{(n,k)}|) = \frac{1}{n} E_n \left|\left|\mathbf{R}^{(n,k)} - \mathbf{R}^{(n,\infty)}\right|\right|_1$

Convergence for matrix iterations

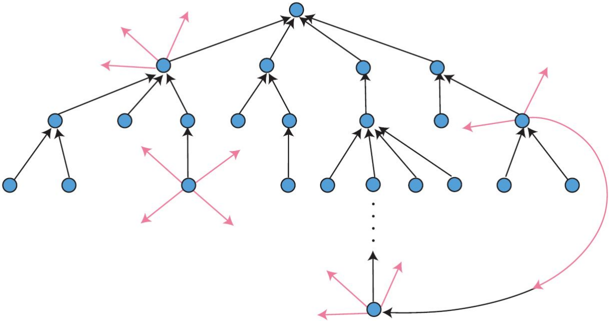
- All nodes are symmetric!
- $E_n(|R_1^{(n,\infty)} - R_1^{(n,k)}|) = \frac{1}{n} E_n \left|\left| \mathbf{R}^{(n,k)} - \mathbf{R}^{(n,\infty)} \right|\right|_1$
- $E_n(|R_1^{(n,\infty)} - R_1^{(n,k)}|) \leqslant |r_0|\alpha^k + \frac{\alpha^k}{n(1-\alpha)} \sum_i |Q_i|$
- Markov inequality:

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

- It is a weaker result than bounding $|R_1^{(n,\infty)} - R_1^{(n,k)}|$, but it is good enough
- Approximation of $R_1^{(n,\infty)}$ by $R_1^{(n,k)}$
- Next, approximate $R_1^{(n,k)}$ by the PageRank of a root of a tree with depth k

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 (The Coupling Lemma)

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.

The Coupling Lemma: idea of the proof

- $\hat{Z}_s$ # individuals in generation s of the tree
- $\hat{V}_s$ # outbound stubs of all nodes in generation s

The Coupling Lemma: idea of the proof

- $\hat{Z}_s$ # individuals in generation s of the tree
- $\hat{V}_s$ # outbound stubs of all nodes in generation s
- $\hat{Z}_s$, $\hat{V}_s$ are not much larger than their means:

$$E_n\left[\hat{Z}_s\right] \approx \mu^{s+1}, \quad E_n\left[\hat{V}_s\right] \approx \lambda\mu^s, \quad \lambda = E[D^2]/\mu.$$

The Coupling Lemma: idea of the proof

- $\hat{Z}_s$ # individuals in generation s of the tree
- $\hat{V}_s$ # outbound stubs of all nodes in generation s
- $\hat{Z}_s$, $\hat{V}_s$ are not much larger than their means:

$$E_n\left[\hat{Z}_s\right] \approx \mu^{s+1}, \quad E_n\left[\hat{V}_s\right] \approx \lambda\mu^s, \quad \lambda = E[D^2]/\mu.$$

- An inbound stub of a node in the rth generation will be the first one to be paired with a labeled outbound stub with a probability not larger than

$$P_r := \frac{1}{L_n}\sum_{s=0}^{r} \hat{V}_s \approx \frac{\lambda\mu^r}{n(\mu-1)}.$$

The Coupling Lemma: idea of the proof

- $\hat{Z}_s$ # individuals in generation s of the tree
- $\hat{V}_s$ # outbound stubs of all nodes in generation s
- $\hat{Z}_s$, $\hat{V}_s$ are not much larger than their means:

$$E_n\left[\hat{Z}_s\right] \approx \mu^{s+1}, \quad E_n\left[\hat{V}_s\right] \approx \lambda\mu^s, \quad \lambda = E[D^2]/\mu.$$

- An inbound stub of a node in the rth generation will be the first one to be paired with a labeled outbound stub with a probability not larger than

$$P_r := \frac{1}{L_n}\sum_{s=0}^{r} \hat{V}_s \approx \frac{\lambda\mu^r}{n(\mu-1)}.$$

- $\{\tau = r\}$ is equivalent to the event that $Binomial(\hat{Z}_r, P_r)$ r.v. is greater or equal than 1.

The Coupling Lemma: idea of the proof

- $\hat{Z}_s$ # individuals in generation s of the tree
- $\hat{V}_s$ # outbound stubs of all nodes in generation s
- $\hat{Z}_s$, $\hat{V}_s$ are not much larger than their means:

$$E_n\left[\hat{Z}_s\right] \approx \mu^{s+1}, \quad E_n\left[\hat{V}_s\right] \approx \lambda\mu^s, \quad \lambda = E[D^2]/\mu.$$

- An inbound stub of a node in the rth generation will be the first one to be paired with a labeled outbound stub with a probability not larger than

$$P_r := \frac{1}{L_n}\sum_{s=0}^{r}\hat{V}_s \approx \frac{\lambda\mu^r}{n(\mu-1)}.$$

- $\{\tau = r\}$ is equivalent to the event that $Binomial(\hat{Z}_r, P_r)$ r.v. is greater or equal than 1.
- Markov's inequality: $P(\tau = r) \leqslant \hat{Z}_r P_r = O\left(\mu^{2r} n^{-1}\right)$, $r \leqslant k$.

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

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

Numerical results-1

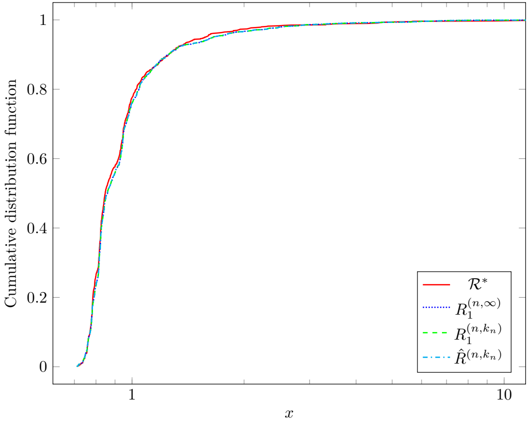


Figure: The empirical CDFs of 1000 samples of $\mathcal{R}^*$, $R_1^{(n,\infty)}$, $R_1^{(n,k_n)}$ and $\hat{R}^{(n,k_n)}$ for $n = 10000$ and $k_n = 9$.

Numerical results-2

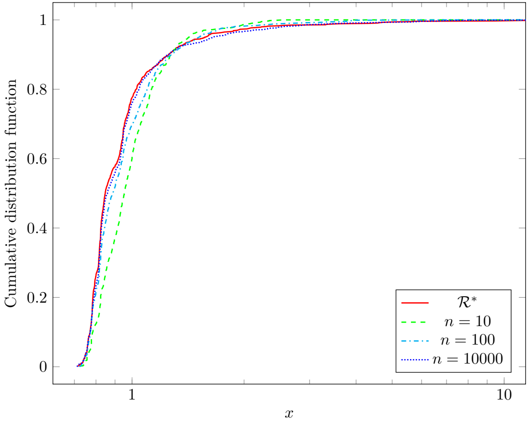
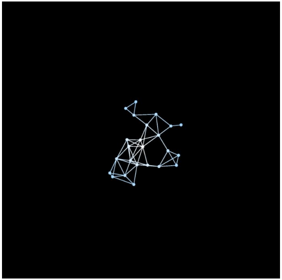
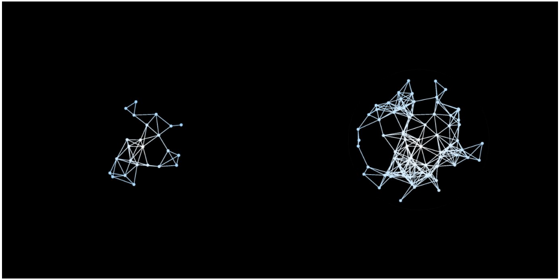


Figure: The empirical CDFs of 1000 samples of $\mathcal{R}^*$ and $R_1^{(n,\infty)}$ for $n = 10$, 100 and 10000.

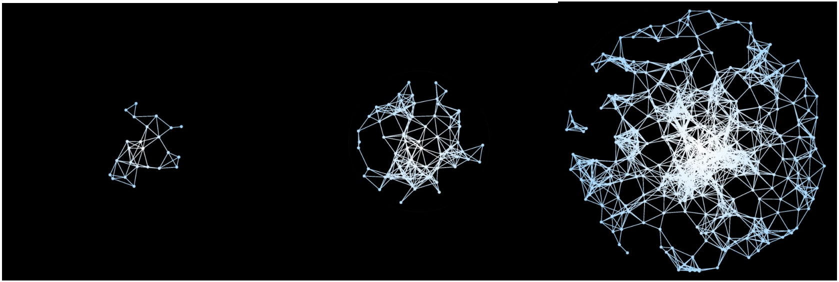
Graph limits



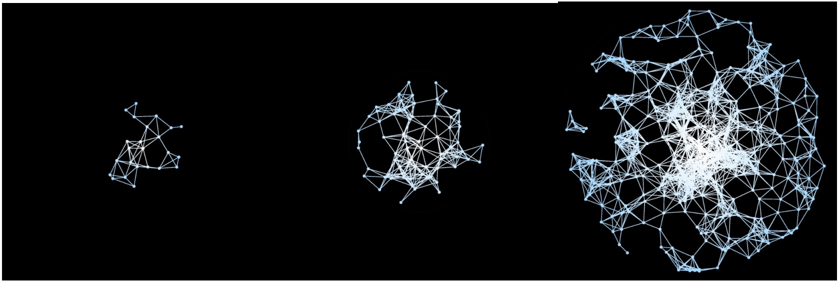
Graph limits



Graph limits

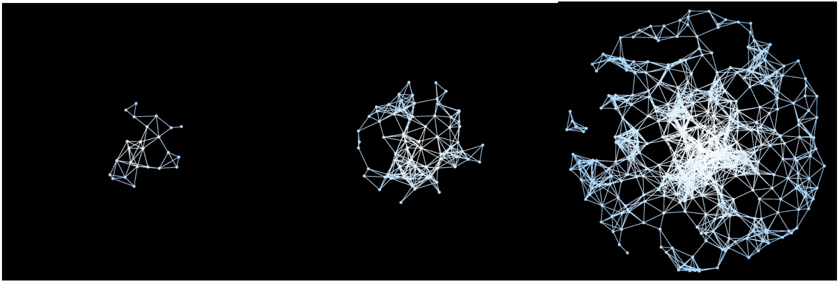


Graph limits



- Local weak convergence (Benjamini& Schramm 2001)

Graph limits



- Local weak convergence (Benjamini& Schramm 2001)
- Ongoing work vdHofstad, Garavaglia, L (2017)

Thank you!