PageRank on directed complex networks

Mariana Olvera-Cravioto

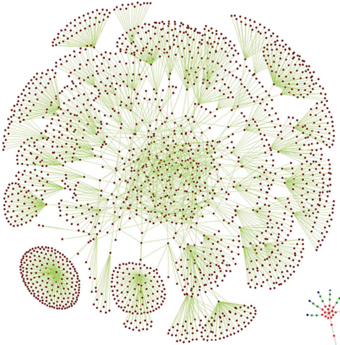
UC Berkeley

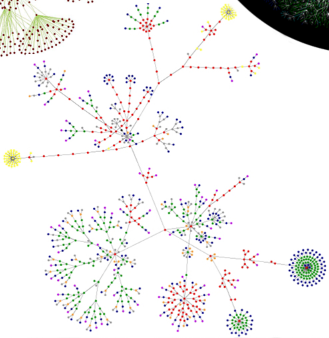
molvera@berkeley.edu

January 25th, 2018

An abundance of information

- An era of massive amounts of information that need to be **organized**.
- **A famous example:** The World Wide Web.
  - Organize webpages based on their "popularity", "relevance", etc.
  - Search engines based on ranking algorithms.
- **Other important examples:** Twitter, healthcare networks, scientific citations, customer reviews, etc.
- Information represented by graphs: nodes, edges, and node attributes.
- Different types of graphs require different ranking schemes.





The problem to solve

- We want to analyze the "typical" behavior of ranking algorithms on large directed graphs.

The problem to solve

- We want to analyze the “typical” behavior of ranking algorithms on large directed graphs.
  - Can we characterize nodes with very high ranks?

The problem to solve

- We want to analyze the “typical” behavior of ranking algorithms on large directed graphs.
  - Can we characterize nodes with very high ranks?
  - Can we determine the distribution of the ranks?

The problem to solve

- We want to analyze the "typical" behavior of ranking algorithms on large directed graphs.
  - Can we characterize nodes with very high ranks?
  - Can we determine the distribution of the ranks?
- Our approach:

  STEP 1: Start with an appropriate random graph model.
  STEP 2: Show that we can analyze the rank via a fixed-point equation.
  STEP 3: Characterize the solutions to this fixed-point equation.

The WWW graph

- WWW seen as a directed graph (webpages = nodes, links = edges).
- For ranking purposes we can think of it as being a *simple* graph.
- Empirical observations:

$$\text{fraction pages} > k \text{ in-links} \propto k^{-\alpha}, \qquad \alpha = 1.1$$

$$\text{fraction pages} > k \text{ out-links} \propto k^{-\beta}, \qquad \beta = 1.72$$

- We want a directed random graph model that matches the degree distributions.

The WWW graph

- WWW seen as a directed graph (webpages = nodes, links = edges).
- For ranking purposes we can think of it as being a *simple* graph.
- Empirical observations:

$$\text{fraction pages} > k \text{ in-links} \propto k^{-\alpha}, \qquad \alpha = 1.1$$

$$\text{fraction pages} > k \text{ out-links} \propto k^{-\beta}, \qquad \beta = 1.72$$

- We want a directed random graph model that matches the degree distributions.
- The power-law hypothesis for PageRank:

$$\text{fraction pages with rank} > k \propto k^{-\alpha}$$

Model 1: The directed configuration model

- Directed graph on n nodes $V_n = \{1, 2, \ldots, n\}$.
- In-degree and out-degree:
  - $d_i^+ =$ in-degree of node $i =$ number of edges pointing to i.
  - $d_i^- =$ out-degree of node $i =$ number of edges pointing out from i.
- We call $(\mathbf{d}^+, \mathbf{d}^-) = (\{d_i^+\}, \{d_i^-\})$ a bi-degree-sequence if

$$\sum_{i=1}^{n} d_i^+ = \sum_{i=1}^{n} d_i^-$$

- **Target joint degree distribution:**

$$F(x, y) = P(\mathscr{D}^+ \le x, \mathscr{D}^- \le y)$$

with $(\mathscr{D}^+, \mathscr{D}^-) \in \mathbb{N}^2$.

Model 1: The directed configuration model

- By adding some randomness into the bi-degree sequence we can obtain $(\mathbf{D}^+, \mathbf{D}^-)$ such that

$$\frac{1}{n}\sum_{i=1}^{n} 1(D_i^+ \le x, D_i^- \le y) \xrightarrow{P} F(x,y), \qquad n \to \infty,$$

  see, e.g., the algorithm proposed in (Chen-OC '12).
- Given the bi-degree sequence, assign to each node i a number of inbound and outbound half edges according to the sequence.
- We obtain a graph by randomly pairing the inbound half edges with the outbound ones.
- The result is a *multigraph* (e.g., with self-loops and multiple edges in the same direction) on the nodes V_n.
- Conditionally on the resulting graph being simple, it is uniformly chosen among all graphs having $(\mathbf{D}^+, \mathbf{D}^-)$ as their bi-degree sequence.

Model 2: The inhomogeneous random digraph

- Consider a directed graph on the set of vertices $V_n = \{1, 2, \ldots, n\}$ having edges in E_n.
- Each vertex i is assigned a *type* $\mathbf{x}_i \in \mathcal{S}$.
- Types are distributed according to some measure μ.
- Let $\kappa(\mathbf{x}, \mathbf{y}) : \mathcal{S}^2 \to \mathbb{R}_+$ and construct the graph by independently drawing an edge from i to j with probability

$$p_{ij}^{(n)} = \mathbb{P}_n((i,j) \in E_n) = 1 \wedge \frac{\kappa(\mathbf{x}_i, \mathbf{x}_j)(1 + \varphi_n(\mathbf{x}_i, \mathbf{x}_j))}{n}, \quad 1 \leq i \neq j \leq n,$$

  where $\mathbb{P}_n(\cdot) = P(\cdot | \{\mathbf{x}_i : 1 \leq i \leq n\})$ and $|\varphi_n(\mathbf{x}_i, \mathbf{x}_j)| \to 0$.
- The limiting degree joint distribution is given by mixed Poisson r.v.s with mixing distributions determined by the types.

Model 2: The inhomogeneous random digraph

- Examples with $\mathbf{x} = (x^+, x^-)$ and $\kappa(\mathbf{x}, \mathbf{y}) = \theta^{-1} x^- y^+$:
  - Directed Erdős-Rényi model:
  $$p_{ij}^{(n)} = \frac{\lambda}{n}$$
  - Directed Chung-Lu model:
  $$p_{ij}^{(n)} = \frac{x_i^- x_j^+}{l_n} \wedge 1, \qquad l_n = \sum_{i=1}^{n} (x_i^+ + x_i^-)$$
  - Directed generalized random graph:
  $$p_{ij}^{(n)} = \frac{x_i^- x_j^+}{l_n + x_i^- x_j^+}$$
  - Directed Poissonian random graph or Norros-Reittu model:
  $$p_{ij}^{(n)} = 1 - e^{-x_i^- x_j^+ / l_n}$$

Google's PageRank

- PageRank computes the rank of a vertex as:

$$r_i = (1-c)q_i + c\sum_{j\to i}\frac{r_j}{D_j^-},$$

where the sum is taken over all vertices pointing to vertex i, D_j^- is the number of outbound links of page j, $\mathbf{q} = (q_1, \ldots, q_n)$ is a probability vector, known as personalization, n is the total number of vertices in the graph, and c is a damping factor, usually $c = 0.85$.

- Multiply both sides by n to obtain a "scale free" rank.
- In matrix notation,

$$\mathbf{R} = (1-c)\mathbf{q} + \mathbf{R}\mathbf{M}, \qquad \mathbf{M} = \text{matrix of weights.}$$

Matrix iterations

- Since $\mathbf{M}^k \to 0$ as $k \to \infty$, $\mathbf{R}$ admits the representation

$$\mathbf{R} = (1-c)\mathbf{q}\sum_{i=0}^{\infty}\mathbf{M}^i.$$

- Hence, we can approximate $\mathbf{R}$ with finitely many matrix iterations

$$\mathbf{R}^{(k)} = (1-c)\mathbf{q}\sum_{i=0}^{k}\mathbf{M}^i.$$

Matrix iterations

- Since $\mathbf{M}^k \to 0$ as $k \to \infty$, $\mathbf{R}$ admits the representation

$$\mathbf{R} = (1-c)\mathbf{q}\sum_{i=0}^{\infty}\mathbf{M}^i.$$

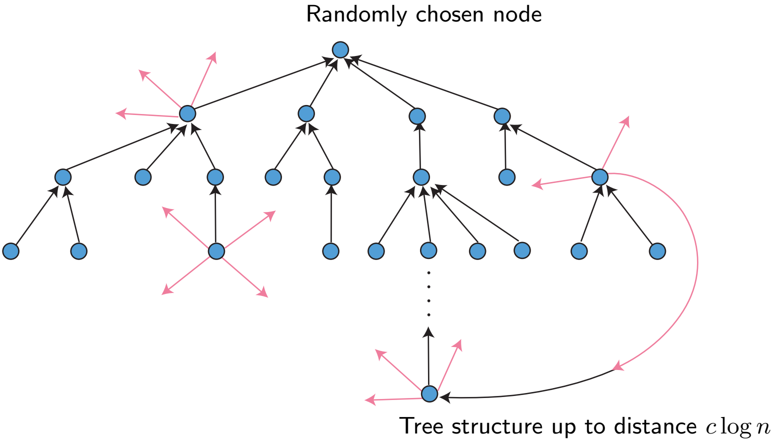
- Hence, we can approximate $\mathbf{R}$ with finitely many matrix iterations

$$\mathbf{R}^{(k)} = (1-c)\mathbf{q}\sum_{i=0}^{k}\mathbf{M}^i.$$

- **Remark:** $\mathbf{R}^{(k)}$ contains only the "local" behavior of the graph.

Random graph approximation

- Many random graph models have a *local tree-like* behavior.
- Both the DCM and the IRD do.



Connection to the fixed point equation

- Consider the more general setting where

$$R_i = Q_i + \sum_{j \to i} \frac{\zeta_j}{D_j} \cdot R_j.$$

- PageRank corresponds to $Q_i = q_i(1-c)n$ and $\zeta_j = c$.

Connection to the fixed point equation

- Consider the more general setting where

$$R_i = Q_i + \sum_{j \to i} \frac{\zeta_j}{D_j} \cdot R_j.$$

- PageRank corresponds to $Q_i = q_i(1-c)n$ and $\zeta_j = c$.
- A stochastic approximation (independent in-degree and out-degree):

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

  where $C_j = \zeta_j / \mathcal{D}_j$, $|\zeta_j| \le c < 1$ for all j, $N \in \mathbb{N}$, and $\{R_j\}$ are i.i.d. copies of R independent of $(Q, N, \{C_j\})$.
- $R =$ rank, $N =$ in-degree, $\mathcal{D}_i =$ neighbors' *size-biased* out-degree, $R_i =$ neighbors' ranks.

In the presence of degree-degree correlations

- If the in-degree and out-degree of the same vertex are dependent, then $C_j = \zeta_j / \mathcal{D}_j$ and R_j are too.
- By setting $X_j = C_j R_j$ we obtain a new fixed-point equation:

$$R^* = Q_0 + \sum_{j=1}^{N_0} X_j, \qquad X \stackrel{\mathcal{D}}{=} CQ + \sum_{j=1}^{N} C X_j,$$

where (Q, N, C) are arbitrarily dependent, with $C = \zeta / \mathcal{D}$ as before, (Q_0, N_0) are the attributes of a vertex chosen uniformly at random, and R^* is the rank of this randomly chosen vertex.

Assumptions for the DCM

- Let ξ be uniformly chosen from $\{1, 2, \ldots, n\}$, and set

$$F_n(m, k, q, x) = \mathbb{P}_n(D_\xi^+ \leq m, D_\xi^- \leq k, Q_\xi \leq q, \zeta_\xi \leq x)$$

  and

$$F(m, k, q, x) = P(\mathscr{D}^+ \leq x, \mathscr{D}^- \leq k, Q \leq q, \zeta \leq x),$$

- Let d_1 denote the Wasserstein metric of order 1.
- Assume:
  - $d_1(F_n, F) \xrightarrow{P} 0$, as $n \to \infty$.
  - $E[\mathscr{D}^+] = E[\mathscr{D}^-]$.
  - $E[(\mathscr{D}^+)^{1+\delta} + (\mathscr{D}^-)^2 + \mathscr{D}^+\mathscr{D}^- + |Q| + |Q|\mathscr{D}^-] < \infty$ for some $\delta > 0$ and $|\zeta| \leq c < 1$ a.s.
  - Some other technical conditions.

Assumptions for the IRD

- Suppose types are of the form $\mathbf{X}_i = (W_i^+, W_i^-, Q_i, \zeta_i)$ and the kernel $\kappa(\mathbf{X}_i, \mathbf{X}_j) = W_i^- W_j^+/\theta$.
- Let ξ be uniformly chosen from $\{1, 2, \ldots, n\}$, and set

$$F_n(u, v, q, x) = \mathbb{P}_n(W_\xi^+ \leq u, W_\xi^- \leq v, Q_\xi \leq q, \zeta_\xi \leq x)$$

  and

$$F(u, v, q, x) = P(W^+ \leq u, W^- \leq v, Q \leq q, \zeta \leq x),$$

- Assume:
  - $d_1(F_n, F) \xrightarrow{P} 0$, as $n \to \infty$.
  - $|\varphi_n(\mathbf{X}_i, \mathbf{X}_j)| \xrightarrow{P} 0$ for each i, j.
  - $E[(W^+)^{1+\delta} + (W^-)^2 + W^+W^- + |Q| + |Q|W^-] < \infty$ for some $\delta > 0$ and $|\zeta| \leq c < 1$ a.s.
  - Some other technical conditions.

The limiting distribution for PageRank

- **Theorem:** (OC '18) Let R_ξ denote the rank of a uniformly chosen vertex in either the DCM or the IRD. Then, under the assumptions for each model, there exists a r.v. $\mathcal{R}^*$ such that

$$R_\xi \Rightarrow \mathcal{R}^* \qquad \mathbb{E}_n[R_\xi] \xrightarrow{P} E[\mathcal{R}^*], \qquad n \to \infty,$$

with

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

where the $\{X_j\}$ are i.i.d. copies of the attracting endogenous solution to

$$X \stackrel{\mathcal{D}}{=} \mathcal{C}\mathcal{Q} + \sum_{j=1}^{\mathcal{N}} \mathcal{C} X_j, \tag{1}$$

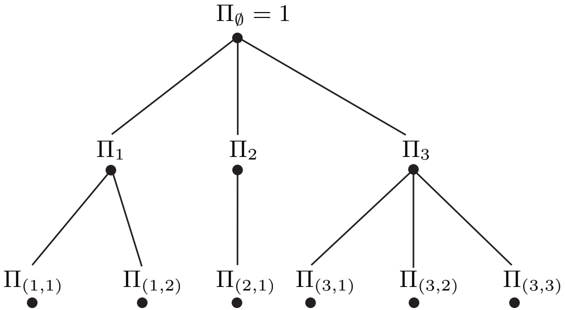
and are independent of $(\mathcal{Q}_0, \mathcal{N}_0)$.

Some remarks

- The convergence in distribution was first proved in (Chen-Litvak-OC '17) for the DCM and in (Lee-OC '17) for the IRD, both under independence between the in-degree and out-degree.
- $(\mathcal{Q}_0, \mathcal{N}_0)$ correspond to the limiting personalization and in-degree, respectively, of a randomly chosen vertex.
- The distribution of $(\mathcal{Q}_0, \mathcal{N}_0)$ is directly related to distribution F, whereas the distribution of $(\mathcal{Q}, \mathcal{N}, \mathcal{C})$ is size-biased.
- The endogenous solution to (1) can be constructed on a *weighted branching process*.

The weighted branching process

- Number of offspring N, mark $(Q, C_1, C_2, \dots)$.



- Each node in the tree has a weight $\Pi_{(i_1,\dots,i_n)}$ defined via the recursion

$$\Pi_{i_1} = C_{i_1}, \qquad \Pi_{(i_1,\dots,i_n)} = C_{(i_1,\dots,i_n)}\Pi_{(i_1,\dots,i_{n-1})}, \qquad n \geq 2,$$

and $\Pi = 1$ is the weight of the root node.

The attracting endogenous solution

- Consider the SFPE

$$R \stackrel{\mathcal{D}}{=} \sum_{i=1}^{N} C_i R_i + Q$$

where $\{R_i\}$ are i.i.d., independent of $(Q, N, C_1, C_2, \dots)$, having the same distribution as R, $Q, \{C_i\}$ real-valued random variables, $N \in \mathbb{N} \cup \{\infty\}$.

The attracting endogenous solution

- Consider the SFPE
$$R \stackrel{\mathcal{D}}{=} \sum_{i=1}^{N} C_i R_i + Q$$
where $\{R_i\}$ are i.i.d., independent of $(Q, N, C_1, C_2, \ldots)$, having the same distribution as R, $Q, \{C_i\}$ real-valued random variables, $N \in \mathbb{N} \cup \{\infty\}$.
- The attracting endogenous solution is given by
$$R = \sum_{\mathbf{i} \in \mathcal{T}} Q_{\mathbf{i}} \Pi_{\mathbf{i}}.$$
- It is well defined provided $E\left[\sum_{i=1}^{N} |C_i|^\beta\right] < 1$ for some $0 < \beta \leq 1$, or if $E\left[\sum_{i=1}^{N} C_i^2\right] < 1$ and $E[Q] = 0$.

Our particular SFPE

- For the stochastic fixed point equation

$$X \stackrel{\mathcal{D}}{=} \sum_{i=1}^{\mathcal{N}} \mathcal{C} X_i + \mathcal{C}\mathcal{Q}$$

describing PageRank, we have $(Q, N, C_1, C_2, \dots) = (\mathcal{C}\mathcal{Q}, \mathcal{N}, \mathcal{C}, \mathcal{C}, \dots)$, and the stability condition is satisfied since $E[\mathcal{N}|\mathcal{C}|] \leq c < 1$.

Asymptotic behavior of the solution

- The results in (OC '12) for linear SFPEs give:
- **Theorem:** Suppose $\mathcal{C} \geq 0$ and define $\rho_\alpha = E[\mathcal{N}\mathcal{C}^\alpha]$ for $\alpha > 0$. Then,
  - If $P(\mathcal{N}\mathcal{C} > x) \in \mathcal{R}_{-\alpha}$ with $\alpha > 1$, $E[|\mathcal{Q}\mathcal{C}|^{\alpha+\epsilon}] < \infty$ and $\rho_{\alpha+\epsilon} < \infty$ for some $\epsilon > 0$, $E[\mathcal{Q}\mathcal{C}] > 0$, and $\rho_1 \vee \rho_\alpha < 1$, then
  $$P(X > x) \sim \frac{(E[\mathcal{Q}\mathcal{C}])^\alpha}{(1-\rho_1)(1-\rho_\alpha)} P(\mathcal{N}\mathcal{C} > x), \qquad x \to \infty.$$
  - If $P(\mathcal{Q}\mathcal{C} > x) \in \mathcal{R}_{-\alpha}$ with $\alpha > 1$, $E[|\mathcal{Q}\mathcal{C}|^\beta] < \infty$ for all $0 < \beta < \alpha$, $\rho_1 \vee \rho_\alpha < 1$ and $E[(\mathcal{N}\mathcal{C})^{\alpha+\epsilon}] < \infty$ for some $\epsilon > 0$, then
  $$P(X > x) \sim (1-\rho_\alpha)^{-1} P(\mathcal{Q}\mathcal{C} > x), \qquad x \to \infty.$$
- **Note:** $\mathcal{N}\mathcal{C} = \mathcal{N}|\zeta|/\mathcal{D}$, where $(\mathcal{N}, \mathcal{D}, \zeta)$ are the size-biased in-degree, out-degree, and weight, respectively.

Asymptotic behavior of PageRank

- Recall the limiting PageRank:

$$\mathcal{R}^* = \sum_{i=1}^{\mathcal{N}_0} X_i + \mathcal{Q}_0$$

- Suppose that $P(\mathcal{N}_0 > x) \in \mathcal{R}_{-\alpha}$ for some $\alpha > 1$ and $E[|\mathcal{Q}_0|^{\alpha+\epsilon}] < \infty$ for some $\epsilon > 0$. Then,
  - If $P(X > x) \in \mathcal{R}_{-\alpha}$ and $E[X] > 0$,

$$P(\mathcal{R}^* > x) \sim P\left(\max_{1 \leq i \leq \mathcal{N}_0} X_i > x\right) + P(\mathcal{N}_0 > x/E[X]), \qquad x \to \infty.$$

  - If $E[|X|^{\alpha+\epsilon}] < \infty$ for some $\alpha > 0$,

$$P(\mathcal{R}^* > x) \sim P(\mathcal{N}_0 > x/E[X]), \qquad x \to \infty.$$

The power-law hypothesis for PageRank holds!

The impact of degree-degree correlations

- When the in-degree and out-degree are independent $\mathcal{N}_0 \stackrel{\mathcal{D}}{=} \mathcal{N}$ and

$$P(\mathcal{N}\mathcal{C} > x) \sim E[\mathcal{C}^\alpha] P(\mathcal{N} > x),$$

which leads to a heavy-tailed X.

- When the in-degree and out-degree are positively correlated we may have $E[(\mathcal{N}\mathcal{C})^{\alpha+\epsilon}] < \infty$, which in turn may lead to $E[|X|^{\alpha+\epsilon}] < \infty$ (provided $\mathcal{Q}$ is light enough).
- In other words,

The contribution of the neighbors to the rank disappears!

Thank you for your attention.