Branching Distributional Equations and their Applications

Mariana Olvera-Cravioto

UNC Chapel Hill

molvera@unc.edu

August 22nd, 2018

Google's PageRank

- PageRank computes the rank of a webpage as:

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

  where, $\{1, 2, \ldots, n\}$ are the pages under consideration, the sum is taken over all pages pointing to i, D_j is the number of outbound links of page j, $\mathbf{q} = (q_1, \ldots, q_n)$ is a personalization vector, and c is a damping factor, usually $c = 0.85$.
- Multiply both sides by n to obtain a "scale free" rank.
- In matrix notation,

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

The problem to solve

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

The problem to solve

- We want to analyze the "typical" behavior of a large class 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 a large class 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 a large class of ranking algorithms on large directed graphs.
  - Can we characterize nodes with very high ranks?
  - Can we determine the distribution of the ranks?
  - Can we propose new algorithms that will have a pre specified typical behavior?

The problem to solve

- We want to analyze the "typical" behavior of a large class of ranking algorithms on large directed graphs.
  - Can we characterize nodes with very high ranks?
  - Can we determine the distribution of the ranks?
  - Can we propose new algorithms that will have a pre specified typical behavior?
- 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.
- Interestingly,

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

- **The power law hypothesis:** This is true in any scale-free graph.

A first random graph model

- We consider a directed version of the "configuration model".
- Directed graph on n nodes $V = \{1, 2, \ldots, n\}$.
- 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 from i.
- $(\mathbf{D}^+, \mathbf{D}^-) = (\{d_i^+\}, \{d_i^-\})$ is called a bi-degree-sequence.
- **Target distributions:**

$$\begin{aligned} \textit{In-degree:} \quad & F = (f_k : k = 0, 1, 2, \ldots), \quad \text{and} \\ \textit{Out-degree:} \quad & G = (g_k : k = 0, 1, 2, \ldots). \end{aligned}$$

- Assume F has finite $1 + \epsilon$ moments and G has finite $2 + \epsilon$ moments.

The directed configuration model

- Assume we have a bi-degree sequence $(\mathbf{D}^+, \mathbf{D}^-)$ that is graphical with high probability and whose in-degree and out-degree distributions are F and G, respectively.
- A method based on i.i.d. samples was given 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 nodes $\{1, 2, \ldots, n\}$.
- **Theorem:** (Chen-OC, '12) By erasing the self loops and multiple edges in the same direction we obtain a simple graph with asymptotic degree distributions F and G as $n \to \infty$.

Pros and cons of the configuration model

- **Pros:**
  - It can be used to "fit" any degree distribution.
  - Conditionally on the pairing process resulting in a simple graph, the graph is uniformly chosen among all simple graphs having that bi-degree sequence. (Exercise)
  - A uniformly chosen graph is a good "null" model.
- **Cons:**
  - The model is somewhat "artificial".
  - It does not provide any explanation as to why some nodes have high degrees whereas others do not.
  - The probability of the pairing process resulting in a simple graph is asymptotically zero if the degrees have infinite variance, and erasing self-loops and multiple-edges destroy the *uniformity*.

Inhomogeneous random digraphs

- A beautifully simple model: the Erdős-Rényi graph.

Inhomogeneous random digraphs

- A beautifully simple model: the Erdős-Rényi graph.
- The (undirected) Erdős-Rényi graph on n vertices is constructed by deciding whether each of the $\binom{n}{2}$ possible edges is present according to an independent coin-flip with probability $p = \lambda/n$.
- The resulting graph is simple by construction.
- Problem.... it produces degrees that are too homogeneous (Poisson to be precise... Exercise).
- **Goal:** generalize the Erdős-Rényi graph to produce directed inhomogeneous graphs with heavy-tailed degrees.

A large family of models

- Consider a digraph $\mathcal{G}(V_n, E_n)$ on the set of vertices $V_n = \{1, 2, \ldots, n\}$ having edges in E_n.
- Each vertex $i \geq 1$ 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

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

where $\mathbb{P}_n(\cdot) = P(\cdot | \{\mathbf{x}_i\}_{i=1}^n)$ and $|\varphi_n(\mathbf{x}_i, \mathbf{x}_j)| \to 0$ sufficiently fast.

Some examples included in the family

- 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:
$$p_{ij}^{(n)} = 1 - e^{-x_i^- x_j^+ / l_n}$$

Degree distributions

- Let (D_i^+, D_i^-) denote the in-degree and out-degree of vertex i.
- Define

$$\lambda^+(\mathbf{x}) = \int_{\mathcal{S}} \kappa(\mathbf{x}, \mathbf{y}) \mu(d\mathbf{y}) \quad \text{and} \quad \lambda^-(\mathbf{x}) = \int_{\mathcal{S}} \kappa(\mathbf{y}, \mathbf{x}) \mu(d\mathbf{y})$$

- **Theorem:** (Cao-OC, '17) Let ξ be the index of a uniformly chosen vertex in V_n. Under some regularity conditions on the kernel κ, we have

$$(D_\xi^+, D_\xi^-) \Rightarrow (Z^+, Z^-), \qquad n \to \infty,$$

where (Z^+, Z^-) is a pair of mixed Poisson r.v.'s with mixing parameters $\lambda^+(\mathbf{X})$ and $\lambda^-(\mathbf{X})$, respectively, conditionally independent given $\mathbf{X}$, and $\mathbf{X}$ distributed according to μ.

Joint degree distribution

- Does the model produce *scale-free graphs*?

Joint degree distribution

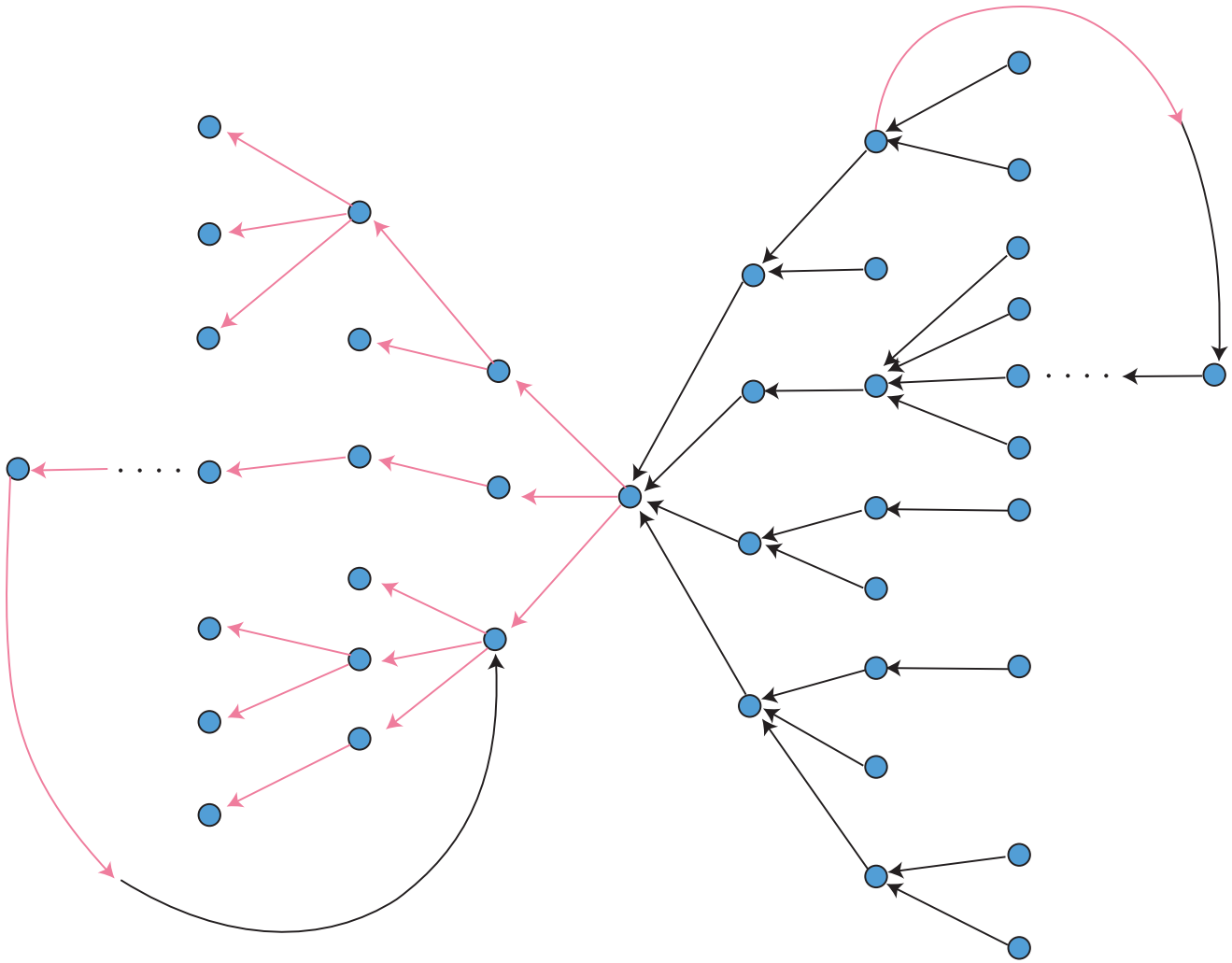
- Does the model produce *scale-free graphs*?
- **Theorem:** (Cao-OC, '17) Suppose that μ is such that if $\mathbf{X}$ is distributed according to μ, then $(\lambda^+(\mathbf{X}), \lambda^-(\mathbf{X}))$ has a (non-standard) multivariate regularly varying distribution with scaling functions $a(t) \in \mathcal{RV}(1/\alpha)$ and $b(t) \in \mathcal{RV}(1/\beta)$. Then (Z^+, Z^-) is also multivariate regularly varying with the same scaling functions.
- **Remark:** in the examples, $(\lambda^+(\mathbf{x}), \lambda^-(\mathbf{x})) = (cx^+, (1-c)x^-)$.

Pros and cons of the inhomogeneous random digraph

- **Pros:**
  - It always produces simple graphs.
  - It provides an explanation for the inhomogeneity of the degrees based on that of the "types".
  - Arcs are independent of each other.
- **Cons:**
  - We can only obtain degree distributions that are mixed Poisson (with arbitrary mixing distributions).
  - The graphs produced, although inhomogeneous, are not necessarily realistic.

The local tree-like structure

- Many random graph models have a *local tree-like* behavior.
- Both the configuration model and the inhomogeneous random digraph do.

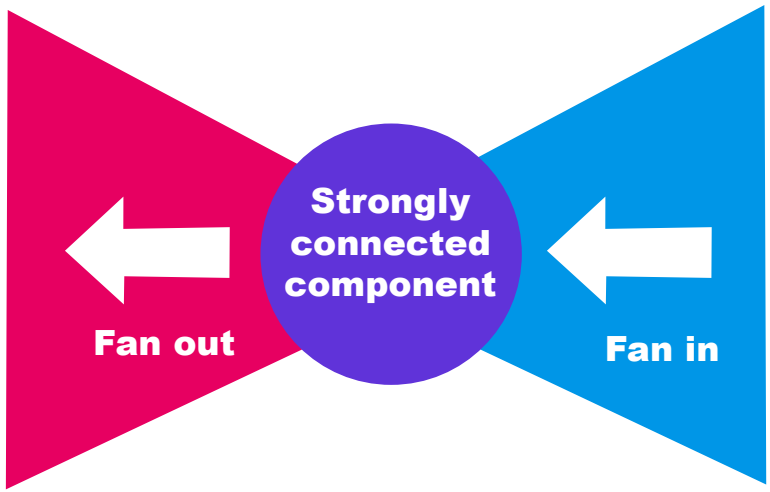


Analysis through branching processes

- The analysis of random graphs in general is based in many cases on their local tree-like structure.
- The key idea is that the exploration of the graph starting from a randomly chosen node can be coupled with a suitable branching process.
- **Examples:** the Erdős-Rényi graph can be coupled with a Galton-Watson tree with a Poisson number of offspring, and the inhomogeneous random graph can be coupled with a multi-type branching process.
- Directed graphs such as the ones described earlier can be coupled with "marked" branching processes.

The bow-tie structure

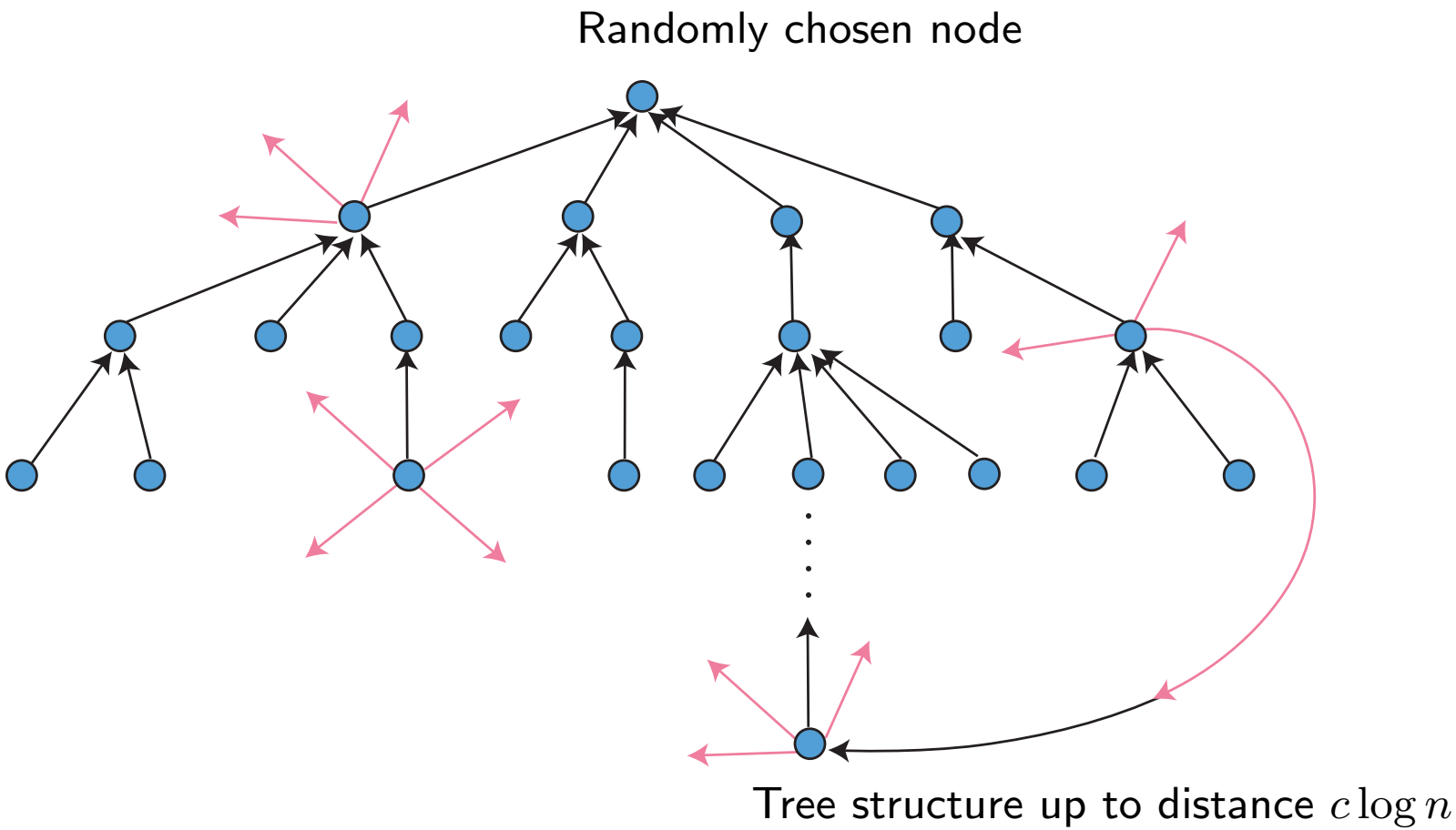
- We say that vertices i and j belong to a strongly connected component if there is a path from i to j and a path from j to i.
- The largest strongly connected component is said to be *giant* if it has at least ϵn vertices for some $\epsilon > 0$.
- Both the DCM and the IRD have a **phase transition** for the existence of a giant strongly connected component.



- The phase transition is determined by the survival probabilities of the coupled branching processes.

Coupling for analyzing PageRank

- We only need to analyze the fan-in of a randomly chosen vertex.
- We need to keep track of both the in-degrees and out-degrees.



More on coupling

- In general, the in-degree becomes the offspring and the out-degree becomes a "mark".
- Tree is rooted at the randomly chosen vertex.
- All other nodes have a **size bias**.
- For the DCM:
  - The coupling tree is a Galton-Watson process.
  - The root has offspring according to F, all other nodes according to:

$$h(m) = P(\mathcal{N} = m) = \frac{E[1(D^+ = m)D^-]}{E[D^-]}$$

- For the IRD:
  - Coupling tree is a multi-type BP with types distributed according to μ .
  - For rank-1 kernels, i.e. $\mathbf{x} = (x^+, x^-)$, $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \theta^{-1} x_i^- x_j^+$, it can be reduced to a single-type.
  - The root has offspring according to Z^+, a mixed Poisson with mixing distribution $\lambda^+(\mathbf{X})$, $\mathbf{X} \sim \mu$, all other nodes according to:

$$h(m) = P(\mathcal{N} = m) = \frac{E[1(Z^+ = m)X^-]}{E[X^-]}$$