PageRank on directed preferential attachment graphs

Mariana Olvera-Cravioto

Joint work with Sayan Banerjee and Prabhanka Deka

UNC Chapel Hill

molvera@email.unc.edu

November 23rd, 2021

General attachment graphs

- Let $G(V_n, E_n)$ denote a directed multigraph on the vertices $V_n = \{1, 2, \ldots, n\}$ with edges in the set E_n.
- We will construct a sequence of multigraphs $\{G(V_n, E_n) : n \geq 1\}$ by adding one vertex at a time.
- Each vertex n will be assigned from the start a number $d_n^+ \geq 1$ of outbound edges.
- Upon arrival, vertex n connects its d_n^+ outbound edges to the existing graph according to some random rule.
- Let $D_i(n-1, k-1)$ denote the **total degree** of vertex i after $k-1$ edges of vertex n have been attached to the graph.
- **Note:** $D_n(n-1, 0) = d_n^+$.

Preferential and uniform attachment

- Let $f(x) = ax + b$, with $\inf_{x \geq 1} f(x) > 0$.
- Attachment probability:

$$P\left(k^{th} \text{ edge of vertex } n \text{ attaches to vertex } i\right) = \frac{f(D_i(n-1,k-1))}{\sum_{j=1}^{n} f(D_j(n-1,k-1))}, \qquad i = 1, 2, \ldots, n$$

- **Preferential attachment:** $f(x) = x + b$
- **Uniform attachment:** $f(x) = b$
- The usual case studied in the literature has $d_n^+ \equiv m$ for all $n \geq 1$.
- The resulting graph $G(V_n, E_n)$ has no directed cycles.

Graph exploration on marked directed graphs

- Let $\mathcal{G}_i^{(k)}$ denote the subgraph of $G(V_n, E_n)$ obtained from exploring the in-component of depth k of vertex i.
- When encountering a vertex j we include as a mark its out degree d_j^+.
- In general, vertices can have marks of the form $\mathbf{X}_i \in \mathcal{S}$, with $\mathcal{S}$ a separable metric space with metric ρ.
- Let $\mathcal{G}_i^{(k)}(\mathbf{X})$ denote the graph $\mathcal{G}_i^{(k)}$ including its vertex marks.

Graph isomorphism and probability space

- **Definition:** We say that two multigraphs $G(V,E)$ and $G'(V',E')$ are **isomorphic** if there exists a bijection $\sigma : V \to V'$ such that

$$l(i) = l(\sigma(i)) \text{ and } e(i,j) = e(\sigma(i),\sigma(j)), \qquad i \in V,\ (i,j) \in E$$

  where $l(i)$ is the number of self-loops of vertex i and $e(i,j)$ is the number of edges from vertex i to vertex j; we write $G \simeq G'$.

- Let $\mathbb{P}_n(\cdot) = P(\cdot\,|\mathbf{X}_i, 1 \leq i \leq n)$ denote the conditional probability space given the vertex marks.

Local weak limits

- **Definition:** We say that the sequence of graphs $\{G(V_n, E_n) : n \geq 1\}$ admits a **strong coupling** with a marked rooted graph $\mathcal{G}_*(\mathbf{X}^*)$ if for I_n uniformly chosen from V_n, and any fixed $k \geq 1$,

$$\mathbb{P}_n\left(\mathcal{G}_{I_n}^{(k)} \not\simeq \mathcal{G}_*^{(k)}\right) \xrightarrow{P} 0, \qquad n \to \infty,$$

  and if σ is the bijection between $\mathcal{G}_*^{(k)}$ and $\mathcal{G}_{I_n}^{(k)}$, and $V_*^{(k)}$ is the vertex set of $\mathcal{G}_*^{(k)}$, then for any $\epsilon > 0$

$$\mathbb{P}_n\left(\bigcap_{i \in V_*^{(k)}} \{\rho(\mathbf{X}_{\sigma(i)}, \mathbf{X}_i^*) \leq \epsilon\}, \mathcal{G}_{I_n}^{(k)} \simeq \mathcal{G}_*^{(k)}\right) \xrightarrow{P} 1, \qquad n \to \infty.$$

- **Note:** $\mathcal{G}_*^{(k)}$ denotes the neighborhood of depth k of $\mathcal{G}_*$.
- If the marks are discrete, we can take $\epsilon = 0$.

Local weak limits... cont.

- **Definition:** We say that the sequence of graphs $\{G(V_n, E_n) : n \geq 1\}$ converges in the **local weak sense in probability** to a marked rooted graph $\mathcal{G}_*(\mathbf{X}^*)$ if:
  - for any fixed graph $G = G(V, E)$ and
  - any $\{B_i : i \in V\} \subseteq \mathcal{S}$ satisfying $P(\mathbf{X}^* \in \partial B_i) = 0$,

  we have for any fixed $k \geq 1$

$$\frac{1}{n}\sum_{i=1}^{n} 1\left(\mathcal{G}_i^{(k)} \simeq G, \bigcap_{j \in V} \{\mathbf{X}_{\sigma(j)} \in B_j\}\right) \xrightarrow{P} P\left(\mathcal{G}_*^{(k)} \simeq G, \bigcap_{j \in V} \{\mathbf{X}^*_{\sigma'(j)} \in B_j\}\right)$$

  as $n \to \infty$, where σ, σ' denote the bijections defining the isomorphisms in each side.

Strong couplings: Conditions

- Let $\{G(V_n, E_n) : n \geq 1\}$ be the sequence of directed general attachment graphs with attachment function $f(x) = ax + b$.
- Suppose $\inf_{x \geq 1} f(x) > 0$.
- Define

$$\nu_n(\cdot) = \frac{1}{n} \sum_{i=1}^{n} 1(\mathbf{X}_i \in \cdot)$$

- Suppose

$$d_1(\nu_n, \nu) \xrightarrow{P} 0, \qquad n \to \infty,$$

  where d_1 is the Wasserstein metric of order one.
- For this talk, we only need $\mathbf{X}_i = d_i^+$.

Strong couplings: Describing the limit

- Let $\{\xi(t) : t \geq 0\}$ be a Markovian pure birth process with $\xi(0) = 0$ and birth rates

$$P(\xi(t + dt) = k + 1 | \xi(t) = k) = f(k)dt + o(dt)$$

- Let $\lambda > 0$ be the Malthusian rate of the process, i.e., $E\left[\int_0^\infty e^{-\lambda s}\xi(ds)\right] = 1$.
- Let $\{\xi^{(n,i)} : i \geq 1, n \geq 0\}$ be i.i.d. copies of ξ, and let $\{\mathcal{D}_n^+ : n \geq 0\}$ be an i.i.d. sequence distributed according to ν, and independent of everything else.
- Define

$$\bar{\xi}^{(n)} = \sum_{i=1}^{\mathcal{D}_n^+} \xi^{(n,i)}$$

- Let $\{\mathcal{B}(t) : t \geq 0\}$ be a CTBP driven by $\{\bar{\xi}^{(n)} : n \geq 0\}$, where $\bar{\xi}^{(n)}$ is the birth process associated to the nth node to be born.

Strong couplings: Main theorem

- Let $\mathcal{T}_t$ denote the discrete skeleton of $\mathcal{B}(t)$.
- Let $\mathcal{T}_t(\mathcal{D}^+)$ denote the tree $\mathcal{T}_t$ where the kth birth is assigned as its mark

$$\mathcal{D}_k^+ = \sum_i d_i^+ 1(S_{i-1} < k \leq S_i),$$

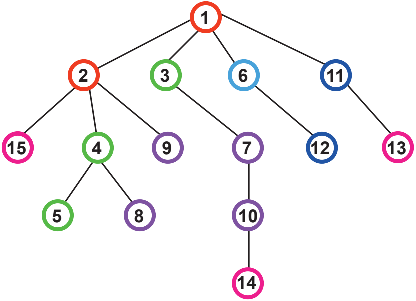
  where $S_n = d_1^+ + \cdots + d_n^+$, $S_0 = 0$.
- Let $\tau \sim$Exponential(λ), independent of $\{\mathcal{B}(t) : t \geq 0\}$.
- **Theorem:** [Banerjee-Deka-OC '21] $\{G(V_n, E_n) : n \geq 1\}$ converges in the local weak sense in probability to $\mathcal{T}_\tau(\mathcal{D}^+)$, and it admits a strong coupling with $\mathcal{T}_\tau(\mathcal{D}^+)$.

Related results

- The local limit for the preferential attachment case with $d_n^+ \equiv m$ was established by [Berger-Borgs-Chayes-Saberi '14] in terms of the Pólya point graph.
- Main result is given in terms of local weak convergence in probability.
- The local limit for the general f case and $d_n^+ \equiv 1$ was established by [Rudas-Tóth-Valkó '06].
- The uniform attachment graph with $d_n^+ \equiv m$ was described in [Garavaglia-van der Hofstad '17], without the local weak limit.

Collapsed branching processes

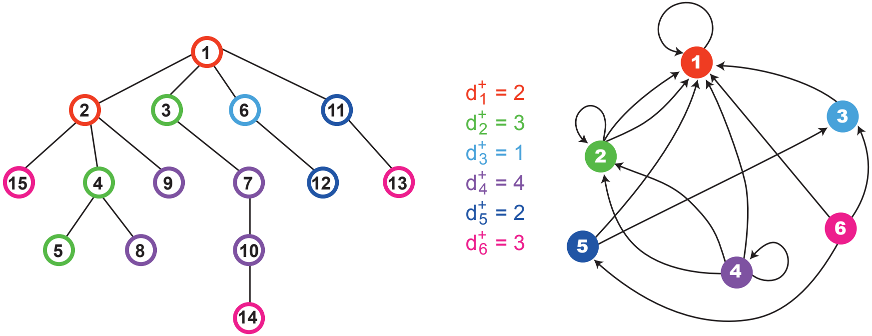
- The proof of the theorem is obtained by collapsing the branching process $\{\mathcal{B}(t) : t \geq 0\}$.
- The procedure works with general functions f satisfying $\inf_{x\geq 1} f(x) > 0$.



$d_1^+ = 2$
$d_2^+ = 3$
$d_3^+ = 1$
$d_4^+ = 4$
$d_5^+ = 2$
$d_6^+ = 3$

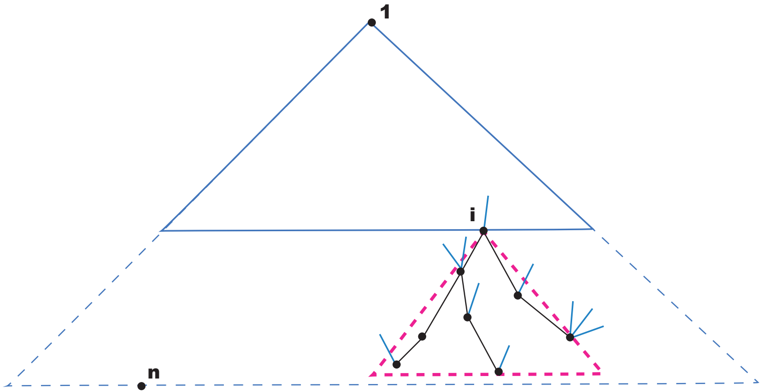
Collapsed branching processes

- The proof of the theorem is obtained by collapsing the branching process $\{\mathcal{B}(t) : t \geq 0\}$.
- The procedure works with general functions f satisfying $\inf_{x\geq 1} f(x) > 0$.



Local limit of collapsed branching processes

- Suppose f also satisfies $f(x) \leq Cx$ for some constant $C < \infty$.
- The local limit is obtained by showing the collapsing procedure results in a tree w.h.p.



Degree distributions: preferential attachment

- Let $\mathcal{D}^-$ denote the degree of the root of $\mathcal{T}_\tau$.
- Let $\overline{F}(x) = P(\mathcal{D}^+ > x)$ and let $\mu = E[\mathcal{D}^+]$.
- Suppose $\overline{F}$ is either light-tailed or regularly varying.
- For the preferential attachment case $f(x) = x + b/\mu$, with $b > -\mu$, then

$$\begin{aligned} P(\mathcal{D}^- > x) &\sim \mu P\left(\xi(\tau) > x\right) + \overline{F}\left(x/E[\xi(\tau)]\right) \\ &\sim C_{\mu,b}\, x^{-2-b/\mu} + \overline{F}(x), \qquad x \to \infty \end{aligned}$$

- In other words, the graph $G(V_n, E_n)$ is asymptotically **scale-free**.

Degree distributions: uniform attachment

- Let $\mathcal{D}^-$ denote the degree of the root of $\mathcal{T}_\tau$.
- Let $\overline{F}(x) = P(\mathcal{D}^+ > x)$ and let $\mu = E[\mathcal{D}^+]$.
- For the uniform attachment case $f(x) = b$, with $b > 0$, then

$$\mathcal{D}^- \stackrel{\mathcal{D}}{=} \text{Poisson}(b\mathcal{D}^+\tau)$$

- In particular, if $\mathcal{D}^+ \equiv m$, then

$$\mathcal{D}^- \stackrel{\mathcal{D}}{=} \text{Geometric}(1/(m+1))$$

  and if $\overline{F}$ is regularly varying with index $\alpha \geq 1$, then

$$P(\mathcal{D}^- > x) \sim E[(b\tau)^\alpha]\overline{F}(x), \qquad x \to \infty$$

- $G(V_n, E_n)$ can be asymptotically **scale-free** or have **light-tailed** degrees, depending on $\overline{F}$.

Google's PageRank

- Arguably, one of the most important notions of node centrality in directed complex networks.
- PageRank assigns a *universal* rank to each vertex in a directed graph by solving the system of linear equations:

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

  where r_i is the rank of vertex i, d_i^+ is its out-degree, q_i its personalization value, and $0 < c < 1$ is the damping factor.
- Provided $\mathbf{q} = (q_1, \ldots, q_n)$ is a probability vector, PageRank can be interpreted as the stationary distribution of the "lazy surfer" random walk on the graph.

A linear algebra representation

- **Scale-free PageRank:** $R_i = nr_i$, $Q_i = (1-c)q_i$

  $$R_i = (1-c)Q_i + \sum_{j \to i} \frac{c}{D_j^+} R_j$$

  where $R_i = nr_i$, $Q_i = q_i$.
- In matrix form:

  $$\mathbf{R} = \mathbf{Q} + \mathbf{R}M, \quad \text{equiv.} \quad \mathbf{R} = \mathbf{Q} \sum_{r=0}^{\infty} M^r,$$

  where $\mathbf{R} = (R_1, \ldots, R_n)$, $\mathbf{Q} = (Q_1, \ldots, Q_n)$, and $M = CA$, with A the adjacency matrix of the graph and C the diagonal matrix whose ith element is $C_{ii} = c/(D_i^+ \vee 1)$.
- **Note:** If A has a zero row, we replace the corresponding row of M with $c(q_1, \ldots, q_n)$.

Locality of PageRank

- Note that the matrix M satisfies $\|M\|_\infty = c < 1$.
- $M^k \to 0$ as $k \to \infty$ geometrically fast.
- We can approximate $\mathbf{R}$ with finitely many iterations:

$$\mathbf{R} \approx \mathbf{Q} \sum_{r=0}^{k} M^r =: \mathbf{R}^{(k)}$$

- **Observation:** $\mathbf{R}^{(k)}$ contains only local information about the in-neighborhoods of depth k of each vertex.

PageRank is a local computation!

The power-law hypothesis

- Let R_{I_n} denote the PageRank of a typical vertex in a graph $G(V_n, E_n)$:

$$R_{I_n} = \sum_{i=1}^{n} R_i 1(I_n = i), \qquad I_n \text{ uniform in } V_n$$

- Suppose there exists $\mathcal{R}^*$ such that

$$R_{I_n} \Rightarrow \mathcal{R}^*, \qquad n \to \infty$$

- Folklore says that on **scale-free** graphs where the in-degree distribution follows a power-law with index $\alpha > 0$, i.e.,

$$P(\mathcal{D}^- > x) \sim C x^{-\alpha}, \qquad x \to \infty,$$

  the PageRank distribution will also follow a power-law with the same index, i.e.,

$$P(\mathcal{R}^* > x) \sim H x^{-\alpha}, \qquad x \to \infty$$

Static graphs

- **Static directed graphs:** Erdős-Rényi, Chung-Lu, Norros-Reittu, generalized random graph, configuration model.
- All these random graphs have as their local weak limit a **marked (delayed) Galton-Watson process**.
- The offspring distribution for the root is given by the limiting in-degree of the graph; all other nodes have a size-biased distribution.
- It is known that the power-law hypothesis holds for these models, i.e., if

$$(D^-_{I_n}, Q_{I_n}) \xrightarrow{d_1} (\mathcal{D}^-, \mathcal{Q}), \qquad \mathcal{D}^- \in RV(\alpha),$$

for some $\alpha > 1$, then

$$R_{I_n} \xrightarrow{d_1} \mathcal{R}^* \in RV(\alpha)$$

[Chen-Litvak-OC '17, OC '21].

Limiting PageRank on static graphs

- Moreover, $\mathcal{R}^*$ can be represented as:

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

where the $\{X_i\} \in RV(\alpha)$ are i.i.d., independent of $(\mathcal{D}^-, \mathcal{Q})$, and are distributed as the special endogenous solution to a **stochastic fixed-point equation**.

- $X \stackrel{D}{=} c\mathcal{R}/\mathcal{D}^+$, where $\mathcal{R}$ and $\mathcal{D}^+$ are the limiting PageRank and out-degree of an inbound neighbor of vertex I_n (*size-biased*).

- Heavy-tailed analysis gives the most likely way to achieve a high rank:

$$P(\mathcal{R}^* > x) \sim P\left(\max_{1 \le i \le \mathcal{D}^-} c\mathcal{R}_i/\mathcal{D}_i^+ > x \right) + P(\mathcal{D}^- > x/E[c\mathcal{R}/\mathcal{D}^+])$$

Peer review **Popularity**

PageRank on general attachment graphs

- Consider a general attachment graph $G(V_n, E_n)$ with attachment function $f(x) = ax + b$, $\inf_{x \geq 1} f(x) > 0$.
- Let $d_n^+ \equiv m \geq 1$ and $q_n \equiv 1$ for all $n \geq 1$.
- Let $\mathcal{T}_\tau(\mathcal{D}^+)$ be the local weak limit of $G(V_n, E_n)$.
- Let $\mathcal{R}^*$ denote the PageRank of the root of $\mathcal{T}_\tau(\mathcal{D}^+)$.
- **Theorem:** [Banerjee-OC '21] Let R_{I_n} be the PageRank of a uniformly chosen vertex I_n. Then,

$$R_{I_n} \Rightarrow \mathcal{R}^* \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n} 1(R_i \in \cdot) \xrightarrow{P} P(\mathcal{R}^* \in \cdot)$$

as $n \to \infty$.

Tail behavior of $\mathcal{R}^*$

- Moreover, there exist constants $0 < C_1, C_2 < \infty$ such that
  - **Preferential attachment:** $f(x) = x + b/m$, $b \geq 0$

$$C_1 x^{-(2+b/m)/(1+(m+b)c/m)} \leq P(\mathcal{R}^* > x) \leq C_2 x^{-(2+b/m)/(1+(m+b)c/m)}$$

  - **Uniform attachment:** $f(x) = b$, $b > 0$

$$C_1 x^{-1/c} \leq P(\mathcal{R}^* > x) \leq C_2 x^{-1/c}$$

- **Observations:**
  a. $\mathcal{R}^*$ has heavy tails in both cases.
  b. In uniform attachment graphs $\mathcal{D}^-$ is light-tailed, but $\mathcal{R}^*$ is heavy-tailed.
  c. In preferential attachment graphs the tail index of $\mathcal{D}^-$ and $\mathcal{R}^*$ do not coincide (PageRank is heavier), i.e.,

The power-law hypothesis fails!

Remarks

- For static graphs, the ranks of sibling nodes are independent of each other.
- Large in-degree vertices are uniformly spread out throughout the graph.
- For general attachment graphs this is no longer true.
- Large in-degree vertices will tend to have highly ranked offspring.
- Dependence among sibling nodes persists even when the in-degree is light-tailed, as in uniform attachment graphs.

Thank you for your attention.