Processes on random graphs: Modeling the web, social networks and opinion dynamics

Lecture 3

Mariana Olvera-Cravioto

UNC Chapel Hill

molvera@email.unc.edu

February 7th, 2024

Modeling large graphs using random graph theory

- So far, we have thought of the graph G representing the social network as fixed.
- **Idea:** think of G as a realization from some random graph model.
- **Question:** can we find a random graph model that could have produced the specific graph G?

Modeling large graphs using random graph theory

- So far, we have thought of the graph G representing the social network as fixed.
- **Idea:** think of G as a realization from some random graph model.
- **Question:** can we find a random graph model that could have produced the specific graph G?
- **Answer:** depends on how many properties of G we need to model....

Modeling large graphs using random graph theory

- So far, we have thought of the graph G representing the social network as fixed.
- **Idea:** think of G as a realization from some random graph model.
- **Question:** can we find a random graph model that could have produced the specific graph G?
- **Answer:** depends on how many properties of G we need to model....
- "First order" properties:
  - Degree distribution(s) (scale free property)
  - Connectivity
  - Typical distances (small world phenomenon)
  - Community structure

Random graph models

- "First order" properties are easy to model.
- **Static** models describe a "snapshot" of a graph.
- **Dynamic** models describe the evolution of a graph as it grows are called .
- Static models that can model first order properties include:
  - Erdős-Rényi model
  - Chung-Lu or expected given degree model
  - Norros-Reittu or Poissonian random graph
  - Generalized random graph
  - Configuration model
  - Stochastic block model
- Dynamic models include the Albert-Barabási or preferential attachment model and its generalizations.
- Our focus from now on will be on static models.

Matrix form of PageRank

- Recall that our goal is to analyze the distribution of a **typical** vertex in both the PageRank vector and the opinion model.
- Both problems define linear recursions on a fixed $G = (V, E; \mathscr{A})$.
- **Scale-free PageRank:**

$$R_i = Q_i + \sum_{j \to i} \frac{c}{D_j^+} R_j, \qquad i \in V,$$

where $R_i = |V| r_i$, $Q_i = (1-c)|V| q_i$, D_j^+ the out-degree of vertex j.

- In matrix form:

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

where $\mathbf{R} = (R_1, \dots, R_{|V|})$, $\mathbf{Q} = (Q_1, \dots, Q_{|V|})$, and $M = \Gamma A$, with Γ a diagonal matrix of "weights" and A the adjacency matrix of the graph.

Matrix form of the opinion model

- **Opinion model:**

$$R_i^{(k+1)} = \sum_{j=1}^{n} c(i,j) R_j^{(k)} + W_i^{(k)} + (1-c-d) R_i^{(k)}, \qquad i \in V,$$

  where $R_i^{(k)}$ denotes the opinion of vertex i at time k.
- Let $\mathbf{R}^{(k)} = (R_1^{(k)}, \ldots, R_{|V|}^{(k)})'$.
- Explicit computation gives that if we let $\mathbf{W}^{(k)} = (W_1^{(k)}, \ldots, W_{|V|}^{(k)})'$, then

$$\mathbf{R}^{(k)} = \sum_{t=0}^{k-1} \sum_{s=0}^{t} a_{s,t} C^s \mathbf{W}^{(k-t)} + \sum_{s=0}^{k} a_{s,k} C^s \mathbf{R}^{(0)}$$

  for some matrix $C \in [0,1]^{|V| \times |V|}$ and coefficients $\{a_{s,t}\}$.
- The matrix C contains the weights each vertex assigns to its neighbors.

Strict contractions

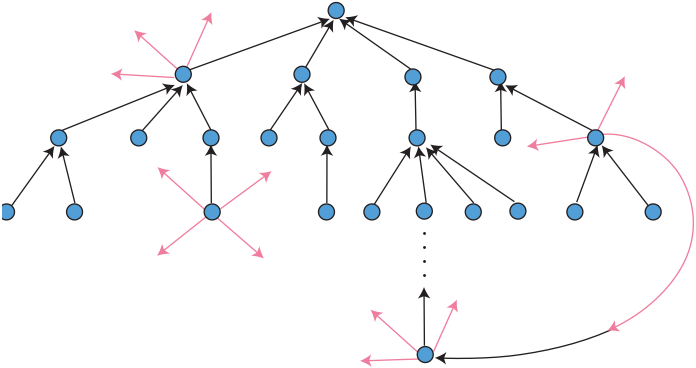
- Both problems lead to linear recursions on a directed graph.
- Moreover, the matrices M and C are strictly **substochastic**.
- The limits $\lim_{k\to\infty} M^k = \lim_{k\to\infty} C^k = 0$ hold.
- Under a suitable Wasserstein metrics, both recursions define **strict contractions**.
- **Consequence:** we can accurately approximate the PageRank vector and the stationary opinion vector with finitely many iterations, i.e., with

$$\sum_{r=0}^{k} M^r \qquad \text{and} \qquad \mathbf{R}^{(k)}, \quad \text{respectively.}$$

- **Key observation:** the processes are **local**, since every vertex is only influenced by its inbound neighborhood of depth k!

Locally tree-like graphs

- Most random graph models are *locally tree-like*.
- Sample $G_n = (V_n, E_n; \mathscr{A}_n)$ from any of the models mentioned today.
- Choose I_n uniformly in V_n and **explore** its in-component.



Graph exploration on marked directed graphs

- Let $\mathcal{G}_i^{(k)}$ denote the subgraph of $G_n = (V_n, E_n; \mathscr{A}_n)$ 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^+, as well as any other vertex attributes that we may need.
- In general, vertices can have marks of the form $\mathbf{X}_i \in \mathcal{S}$, with $\mathcal{S}$ a Polish 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 $\mathbb{P}_n(\cdot) = P(\cdot \, | \mathbf{a}_i, 1 \leq i \leq n)$ denote the conditional probability space given the latent variables needed to generate the graph.

Local weak limits

- **Definition:** We say that the sequence of graphs $\{G_n : n \geq 1\}$ admits a **strong coupling** with a rooted tree $\mathcal{T}(\boldsymbol{\mathcal{X}})$ if for any finite set of uniformly chosen vertices $\{1, \ldots, \ell\}$, there exists a collection of independent copies of $\mathcal{T}(\boldsymbol{\mathcal{X}})$, denoted $\{\mathcal{T}_{\emptyset(i)}(\boldsymbol{\mathcal{X}})\}_{i=1}^{\ell}$, such that for any $k \geq 0$ and $\epsilon > 0$,

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

- If the marks are discrete, we can take $\epsilon = 0$.
- The existence of a strong coupling implies **local weak convergence in probability** *(Aldous, Benjamini-Schramm)*.

Strong couplings

- Strong couplings exist for all the random graph models mentioned earlier.
- All the **static** random graph models have as their local weak limits a (delayed) marked, single or multi type, Galton-Watson tree.
- Strong couplings also exist for **dynamic** random graphs (e.g., preferential attachment), but their local weak limits are continuous time branching processes stopped at a random time.
- Strong couplings also exist for **semi-sparse** random graphs, however, the coupled trees have distributions that depend on n, and are not locally finite as $n \to \infty$.

General maps on directed graphs

- Consider maps on directed graphs of the form:

$$R_i^{(k+1)} = \Phi\left(\mathbf{X}_i, R_i^{(k)}, \eta_i^{(k+1)}, \{(\mathbf{X}_j, \xi_j^{(k+1)}, R_j^{(k)}) : j \to i\}\right), \quad i \in V_n$$

  where $(\eta_i^{(k)}, \{\xi_j^{(k)} : j \in V\})$ are random noises, and the $\{\mathbf{X}_i\}$ are vertex attributes.
- Let $\mathbf{R}^{(k)} = (R_1^{(k)}, \ldots, R_n^{(k)})$.
- If the map Φ defines a strict contraction under a suitable metric, then

$$\mathbf{R}^{(k)} \Rightarrow \mathbf{R} \qquad k \to \infty$$

- Consider the behavior of the typical vertex $R_{I_n}^{(k)}$ and R_{I_n}, where I_n is uniformly chosen from V_n.

Exchange of limits

- If $G_n = (V_n, E_n; \mathscr{A}_n)$ is locally tree-like and the map Φ is nice enough, we can **exchange the limits**.

$$\begin{tikzcd}[column sep=huge, row sep=huge]
R_{I_n}^{(k)} \arrow[r, "k \to \infty"] \arrow[d, "n \to \infty"] & R_{I_n} \arrow[d, "n \to \infty"] \\
\mathcal{R}_{\emptyset}^{(k)} \arrow[r, "k \to \infty"] & \mathcal{R}_{\emptyset}
\end{tikzcd}$$

- $\mathcal{R}_{\emptyset}^{(k)}$ and $\mathcal{R}_{\emptyset}$ correspond to the finite time and stationary version, respectively, of the map Φ on the local weak limit of $\{G_n : n \geq 1\}$.

Back to PageRank

- To understand the PageRank algorithm, we analyze the distribution of the Page rank of a typical vertex, I_n, on a random graph having as its local weak limit a (delayed) marked Galton-Watson process.
- E.g. the directed configuration model or any of the rank-1 inhomogeneous random digraph models.
- The **delay** refers to the fact that the root has a different distribution than all other nodes in the tree, due to the **size-bias** produced by the exploration process.

The limiting PageRank

- The limiting random variable $\mathcal{R}_\emptyset$ is the personalized PageRank of the root of the coupled Galton-Watson tree.
- When the in-degree and out-degree are *asymptotically independent*, $\mathcal{R}_\emptyset$ admits the representation $\mathcal{R}_\emptyset \stackrel{\mathcal{D}}{=} \mathcal{R}^*$, where $\mathcal{R}^*$ is the solution to the distributional fixed-point equation:

$$\mathcal{R}^* \stackrel{\mathcal{D}}{=} \sum_{i=1}^{\mathcal{N}} \mathcal{C}_i \mathcal{R}_i^* + \mathcal{Q},$$

where the $\{\mathcal{R}_i^*\}$ are i.i.d. copies of $\mathcal{R}^*$, independent of $(\mathcal{Q}, \mathcal{N}, \{\mathcal{C}_i\})$, $\mathcal{N}$ has the in-degree distribution and $\mathcal{C}_i = c/\mathcal{D}_i^+$, with $\mathcal{D}_i^+$ the size-biased out-degree.

What can $\mathcal{R}^*$ tell us?

- We analyze the large deviations of $\mathcal{R}^*$, since they correspond to vertices with very high ranks.
- Since most real-world graphs are scale-free in their in-degree, we focus on graphs where $\mathcal{N}$ has a regularly varying distribution.

What can $\mathcal{R}^*$ tell us?

- We analyze the large deviations of $\mathcal{R}^*$, since they correspond to vertices with very high ranks.
- Since most real-world graphs are scale-free in their in-degree, we focus on graphs where $\mathcal{N}$ has a regularly varying distribution.
- A **heavy tail** analysis leads to an interesting insight *(Jelenković-OC '12)*.
- The most likely path to achieving a high rank is:

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

Peer review **Popularity**

- This characterizes the webpages with very high PageRanks.
- It also explains why PageRank captures better the "relevance" of a page.

PageRank under degree correlations

- When the in-degree and the out-degree are not *asymptotically independent*, $\mathcal{R}_\emptyset$ admits the representation

$$\mathcal{R}_\emptyset = \sum_{i=1}^{\mathcal{N}} Y_i + \mathcal{Q},$$

where $\{Y_i\}$ are i.i.d. copies of the solution to the distributional fixed-point equation

$$Y \stackrel{\mathcal{D}}{=} \mathcal{C}^* \mathcal{Q}^* + \sum_{j=1}^{\mathcal{N}^*} \mathcal{C}^* Y_j$$

with $\{Y_j\}$ i.i.d. and independent of $(\mathcal{Q}^*, \mathcal{N}^*, \mathcal{C}^*)$ (size-biased versions of $(\mathcal{Q}, \mathcal{N}, \mathcal{C})$).
- The asymptotic behavior of $\mathcal{R}_\emptyset$ changes, and the **peer review** effect disappears.
- PageRank and **degree centrality** do essentially the same.

Opinion dynamics on a dSBM

- Since in this model the community structure is important, we use a directed stochastic block model (dSBM) for our analysis.
- A dSBM with K communities has edge probabilities of the form:

$$p_{ij}^{(n)} = P((i,j) \in E_n) = \frac{\kappa(J_i, J_j)\theta_n}{n}, \qquad i \neq j,$$

  where $\kappa : \{1, \ldots, K\} \times \{1, \ldots, K\} \to [0, \infty)$, and $J_i \in \{1, 2, \ldots, K\}$ is the community label of vertex i.
- The parameter θ_n can be used to create dense graphs.
- We can also use a **degree corrected** dSBM to obtain a scale-free graph.

Local weak limit of a dSBM

- The local weak limit of the dSBM is a multi-type Galton-Watson process with a type for each community.
- When the in-degree and out-degree are *asymptotically independent*, there is no size-bias on the in-degree (the out-degree plays no role in this model).
- For each $i \in V_n$ and each $k \geq 1$, let $\mathcal{T}_{\emptyset(i)}^{(k)}(\boldsymbol{\mathcal{X}})$ denote the coupled depth-k marked branching tree rooted at vertex i and having the distribution of the local weak limit of $G = (V_n, E_n; \mathscr{A}_n)$.
- **Note:** It is possible to couple all n graph explorations with their local weak limits simultaneously.

Trajectories and stationary behavior

- For each $i \in V_n$ and each $k \geq 1$ let $\mathcal{R}^{(k)}_{\emptyset(i)}$ denote the opinion at time k of the root $\emptyset(i)$ of $\mathcal{T}^{(k)}_{\emptyset(i)}(\boldsymbol{\mathcal{X}})$, computed according to our model.
- Let $\mathcal{J}_{\mathbf{i}}$ denote the community label of node $\mathbf{i}$.
- The vector $\boldsymbol{\mathcal{R}}^{(k)} = (\mathcal{R}^{(k)}_{\emptyset(1)}, \ldots, \mathcal{R}^{(k)}_{\emptyset(n)})'$ does **NOT** have **independent components**.
- Consider the **trajectories** $(R_i^{(0)}, R_i^{(1)}, \ldots, R_i^{(k)})$, as well as the **stationary** version R_i, $i \in V_n$.
- The stationary behavior of the process $\{\mathbf{R}^{(k)} : k \geq 0\}$ is determined by a limiting vector $(\mathcal{J}_\emptyset, \mathcal{R}_\emptyset)$ satisfying:

$$\mathcal{R}^{(k)}_\emptyset \Rightarrow \mathcal{R}_\emptyset, \qquad k \to \infty.$$

Sparse approximation... cont.

- Suppose G_n is a dSBM and θ_n is a constant.
- **Theorem:** (Lin-OC '24+) For any fixed $k \geq 1$,

$$\lim_{n\to\infty} \max_{0\leq r\leq k} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_n\left[\left|R_i^{(k)} - \mathcal{R}_{\emptyset(i)}^{(k)}\right|\right] = 0,$$

and for any bounded and continuous function $f : \mathbb{R}^{k+1} \to \mathbb{R}$,

$$\frac{1}{n}\sum_{i=1}^{n} f(R_i^{(0)}, \ldots, R_i^{(k)}) \xrightarrow{P} E\left[f(\mathcal{R}_\emptyset^{(0)}, \ldots, \mathcal{R}_\emptyset^{(k)})\right], \qquad n\to\infty.$$

Moreover, if $\mathbf{R} = (R_1, \ldots, R_n)'$ is distributed according to the stationary distribution of $\{\mathbf{R}^{(k)} : k \geq 0\}$, then, for any continuous and bounded function $f : \mathbb{R} \to \mathbb{R}$,

$$\frac{1}{n}\sum_{i=1}^{n} f(R_i) \xrightarrow{P} E\left[f(\mathcal{R}_\emptyset)\right], \qquad n\to\infty.$$

Computing means and variances

- Since the local weak limit is a K-type marked Galton-Watson process, the random variables
$$\mathcal{Y}^{(j)} \stackrel{\mathcal{D}}{=} (\mathcal{R}_\emptyset | \mathcal{J}_\emptyset = j),$$
where $\mathcal{J}_\emptyset \in \{1, \ldots, K\}$ is the community label of the root $\emptyset$, are tractable.
- **Note:** when $c + d = 1$, they satisfy a system of distributional fixed-point equations.
- These equations allow us to compute
$$E[\mathcal{Y}^{(j)}] \qquad \text{and} \qquad \mathrm{Var}(\mathcal{Y}^{(j)})$$
for each $j \in \{1, \ldots, K\}$.
- **Observation:** these are enough to characterize consensus and polarization, as well as to study the effects of cognitive biases.

Semi-sparse and dense graphs

- Although most real-world social networks are sparse, we may want to also analyze denser graphs, e.g., whose degrees grow as $\log n$ or faster.
- In this setting, a **mean-field** analysis is more appropriate.
- In the PageRank and opinion model examples, whenever the mean degree grows to infinity, we can approximate the matrices

  $$M^s \qquad \text{and} \qquad C^s,$$

  for $s \geq 1$, with their expected values.
- The resulting approximations are provably accurate and also very tractable.

Final remarks

- PageRank was studied in (Avrachenkov-Kadavankandy-Litvak '18) for an SBM with average degree growing faster than $(\log n)^b$, with $b > 1$.
- The opinion model was studied in (Andreou-OC '24) for a dSBM with average degree growing to infinity arbitrarily slowly.
- The semi-sparse range, i.e., with average degree growing at most as $(\log n)^b$ for $b \geq 1$, can be analyzed using **local approximations**.
- **Observation:** in semi-sparse to dense graphs, the interactions among the vertices do not matter.

References

- **Local weak convergence:**

[1] D. Aldous and J.M. Steele. The objective method: probabilistic combinatorial optimization and local weak convergence. In Probability on discrete structures, pages 1-72. Springer, 2004.

[2] I. Benjamini and O. Schramm. Recurrence of distributional limits of finite planar graphs. In Selected Works of Oded Schramm, pages 533-545. Springer, 2011.

[3] A. Garavaglia, R. van der Hofstad, and N. Litvak. *Local weak convergence for PageRank*. Annals of Applied Probability, 30(1):40-79, 2020.

[4] M. Olvera-Cravioto. *Strong couplings for static locally tree-like random graphs.* Journal of Applied Probability, 59(4):1261-1285.

[5] R. van der Hofstad. *Random Graphs and Complex Networks, Vol. II*. Cambridge University Press, 2016.

References

- **PageRank:**

[6] K. Avrachenkov, A. Kadavankandy, and N. Litvak. *Mean field analysis of personalized PageRank with implications for local graph clustering*. Journal of statistical physics, 173:895-916, 2018.

[7] N. Chen, N. Litvak, and M. Olvera-Cravioto. *Generalized PageRank on directed configuration networks*. Random Structures & Algorithms, 56(61):722-774, 2020.

[8] P.R. Jelenković and M. Olvera-Cravioto. *Information ranking and power laws on trees*. Adv. in Appl. Probab., 42:1057-1093, 2010.

[9] M. Olvera-Cravioto. *PageRank's behavior under degree correlations*. Annals of Applied Probability, 31(3):1403-1442, 2021.

[10] Y. Volkovich and N. Litvak. (2010). *Asymptotic analysis for personalized web search*. Adv. in Appl. Probab., 42 577-604, 2010.

- **Opinion dynamics:**

[11] P. Andreou and M. Olvera-Cravioto. *Opinion dynamics on non-sparse networks with community structure*, ArXiv:2401.04598, 2024.

[12] N. Fraiman, T. Lin, and M. Olvera-Cravioto. *Opinion dynamics on directed complex networks*. Mathematics of Operations Research, to appear. 2024.

Thank you for your attention.