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

Lecture 2

Mariana Olvera-Cravioto

UNC Chapel Hill

molvera@email.unc.edu

February 6th, 2024

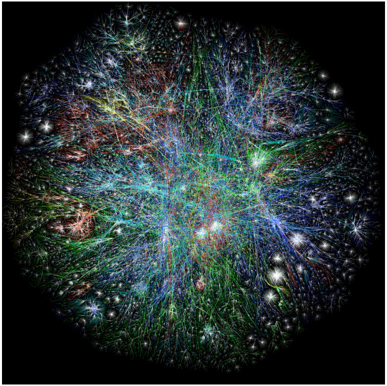
Problem I

Google's PageRank algorithm

The Internet

- The Internet is a giant network of computers around the world connected through "wires".
- Think of the Internet as a giant graph consisting of vertices and edges:

vertices = servers/computers
edges = a wired connection between them



The World Wide Web

- The WWW is a "virtual" network connecting webpages through links.
- It defines a **directed** graph where:

vertices = webpages

edges = directed links from one webpage to another

- The WWW was officially born in 1991 with the creation of the first *browser*, a software interface that allowed users to access many different types of files stored in many different computers.

Modeling complex networks

- Many real-world graphs are extraordinarily big, e.g., millions or billions of vertices.
- Most of them are fairly sparse, i.e., the ratio

$$\frac{\#\text{ edges}}{\#\text{ vertices}}$$

is not too big.
- Many share two key properties:
  - **Small world:** the typical distance between vertices is small compared to the total number of vertices.
  - **Scale-free:** the proportion of vertices with k (inbound/outbound) neighbors decays as a power of k, e.g.,

$$\frac{\#\text{ vertices with } k \text{ neighbors}}{\text{total } \#\text{ vertices}} \approx Ck^{-\alpha}$$

Other properties of complex networks

- Many interesting graphs are disconnected, but may have a **giant** connected component.
- Some graphs have many "clusters" (groups of vertices that have more connections among themselves than with the rest of the graph).
- Some directed graphs exhibit high levels of correlation between the number of inbound neighbors and the number of outbound neighbors of a given vertex.

Other properties of complex networks

- Many interesting graphs are disconnected, but may have a **giant** connected component.
- Some graphs have many "clusters" (groups of vertices that have more connections among themselves than with the rest of the graph).
- Some directed graphs exhibit high levels of correlation between the number of inbound neighbors and the number of outbound neighbors of a given vertex.
- These properties influence how fast a message can spread through a network and/or how many vertices it can reach.
- **They also influence which vertices are more "central" to the network.**

Relevance and centrality

- Intuitively, a vertex in a graph is **central** if many paths go through it.
- One of the most popular measures of centrality is the one computed by Google's PageRank algorithm.
- The idea behind Google's search engine is that **relevant** webpages should be those that are **central** to the network.
- **Why?**

Relevance and centrality

- Intuitively, a vertex in a graph is **central** if many paths go through it.
- One of the most popular measures of centrality is the one computed by Google's PageRank algorithm.
- The idea behind Google's search engine is that **relevant** webpages should be those that are **central** to the network.
- **Why?** Links are created by people, and people will tend to create links to webpages that have relevant/interesting content.
- How does PageRank find "central" vertices?

The PageRank algorithm

- Let n denote the number of vertices in the WWW.
- The PageRank of webpage i, denoted r_i, is a number in $[0, 1]$ that measures its "centrality" within the network.
- r_i is a "universal" rank, i.e., it does not change from one search to another, and it has nothing to do with the content of webpage i.
- r_i depends only on the topology of the graph, i.e., on the structure determined by the edges connecting the vertices.
- **Relevance is contagious:** If a relevant webpage has a link pointing to another webpage, it makes it relevant too, but if it points to too many webpages this effect is reduced.

Computing the PageRank vector

- To compute the PageRank vector $\mathbf{r} = (r_1, \ldots, r_{|V|})$ we solve the system of linear equations:
$$r_i = \frac{1-c}{|V|} + c \sum_{j \to i} \frac{r_j}{d_j^+},$$
where the sum is taken over all the inbound neighbors to webpage i, d_j^+ is the number of outbound neighbors of webpage j, and $c \in (0,1)$ is a constant known as the *damping factor*, usually $c = 0.85$.

Computing the PageRank vector

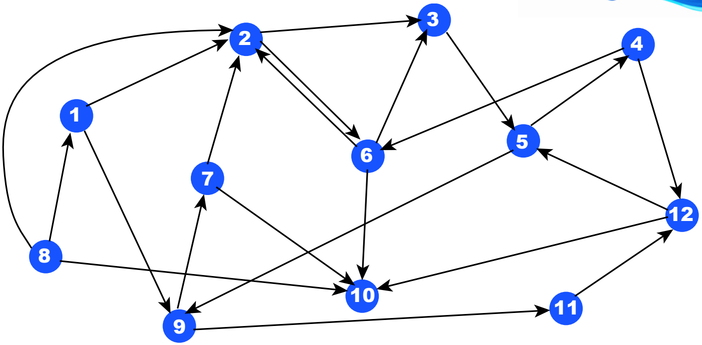
- To compute the PageRank vector $\mathbf{r} = (r_1, \ldots, r_{|V|})$ we solve the system of linear equations:

$$r_i = \frac{1-c}{|V|} + c \sum_{j \to i} \frac{r_j}{d_j^+},$$

  where the sum is taken over all the inbound neighbors to webpage i, d_j^+ is the number of outbound neighbors of webpage j, and $c \in (0,1)$ is a constant known as the *damping factor*, usually $c = 0.85$.
- **Why does this work?**

The random surfer interpretation

- Recall that the goal is to rank vertices according to their "centrality" within the network.
- Imagine you had a web surfer who navigates the WWW by choosing which links to follow at random.
- Specifically, when the surfer visits webpage i, she will choose where to go next with equal probability among all the outbound links of webpage i.
- In other words, this is a **random walk** on the graph.


1
2
3
4
5
6
7
8
9
10
11
12

Random walks on connected graphs

- Let $\{X_k : k \geq 0\}$ denote the stochastic process that tells us the identity of the vertex our surfer visits on the kth step.
- $\{X_k : k \geq 0\}$ is a Markov chain on the set of vertices of the graph.
- If the underlying graph is connected, and we let $k \to \infty$, the proportion of visits to vertex i converges, i.e.,

$$\lim_{k \to \infty} \frac{\text{Number of visits to vertex } i \text{ in the first } k \text{ steps}}{k} = \pi_i$$

  exists, and corresponds to the **stationary probability** of vertex i.
- The **stationary probability** of vertex i has the interpretation of being the long-run proportion of time that our random surfer spends in vertex i.
- When the damping factor $c = 0$, we have $r_i = \pi_i$!

Random walks on disconnected graphs

- The problem with the WWW is that it is a disconnected graph.
- On a disconnected graph the random walk can get "stuck".
- To fix this imagine the surfer has a coin that lands *heads* with probability c and *tails* with probability $1 - c$.
- At each step, before choosing which link to follow next, she flips the coin:
  - If it lands *heads* she chooses with equal probability any of the outbound links if there is one, or chooses from all n webpages if there are no outbound links.
  - If it lands *tails* she chooses with equal probability any of the $|V|$ webpages in the WWW.
- The **stationary probability** of vertex i is equal to its PageRank, i.e.,

$$\pi_i = r_i!$$

PageRank today

- The algorithm that Google uses today has greatly evolved since the original PageRank.
- Each website in the WWW still has a "universal" rank, although the way it is computed has become more sophisticated.
- The order in which the results of a search are displayed depends also on the user's computer, i.e., results are **personalized**.
- Personalized PageRank:

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

where $\mathbf{q} = (q_1, \ldots, q_{|V|})$ is a probability vector that determines where to go after a *tail*.

Understanding Personalized PageRank

- **Question:** Which pages are getting highly ranked?

Understanding Personalized PageRank

- **Question:** Which pages are getting highly ranked?
- To answer this question, consider the empirical distribution of the PageRank vector on a fixed graph $G = (V, E)$, i.e.,

$$\frac{1}{|V|} \sum_{i \in V} 1(r_i \in A),$$

where r_i is the PageRank of vertex i.
- **The power-law hypothesis:** *On scale-free graphs, the distribution of PageRank follows a power-law with the same exponent as the in-degree.*

Understanding Personalized PageRank

- **Question:** Which pages are getting highly ranked?
- To answer this question, consider the empirical distribution of the PageRank vector on a fixed graph $G = (V, E)$, i.e.,

$$\frac{1}{|V|} \sum_{i \in V} 1(r_i \in A),$$

  where r_i is the PageRank of vertex i.
- **The power-law hypothesis:** *On scale-free graphs, the distribution of PageRank follows a power-law with the same exponent as the in-degree.*

  $\longrightarrow$ in-degree is involved

Understanding Personalized PageRank

- **Question:** Which pages are getting highly ranked?
- To answer this question, consider the empirical distribution of the PageRank vector on a fixed graph $G = (V, E)$, i.e.,

$$\frac{1}{|V|} \sum_{i \in V} 1(r_i \in A),$$

  where r_i is the PageRank of vertex i.
- **The power-law hypothesis:** *On scale-free graphs, the distribution of PageRank follows a power-law with the same exponent as the in-degree.*

  $\longrightarrow$ in-degree is involved
- However, in general,

  Set of high in-degree nodes $\neq$ Set of high PageRank nodes

Understanding Personalized PageRank

- **Question:** Which pages are getting highly ranked?
- To answer this question, consider the empirical distribution of the PageRank vector on a fixed graph $G = (V, E)$, i.e.,

$$\frac{1}{|V|}\sum_{i\in V} 1(r_i \in A),$$

  where r_i is the PageRank of vertex i.

- **The power-law hypothesis:** *On scale-free graphs, the distribution of PageRank follows a power-law with the same exponent as the in-degree.*

  $\longrightarrow$ in-degree is involved

- However, in general,

  Set of high in-degree nodes $\neq$ Set of high PageRank nodes

  $\longrightarrow$ more than just the in-degree

Empirical distributions and the typical vertex

- We will explain which vertices get highly ranked by analyzing the PageRank distribution on a random graph.
- In particular, our approach will focus on its large deviations.
- Note that if I is a uniformly chosen vertex in V, then

$$P(r_I \in A|G) = \frac{1}{|V|} \sum_{i \in V} 1(r_i \in A)$$

- We call I a **typical vertex**.
- **Key idea:** on large random graphs, r_I converges to a tractable random variable.
- **Note:** the components of the PageRank vector are usually $O(1/n)$, so we will rescale them first.

Other network centrality measures

- **Degree centrality:** for vertex i,

$$C_D(i) = D_i = \sum_{j \neq i} A_{ij}$$

On directed graphs we define the in-degree and out-degree separately.

- **Closeness centrality:** let $d(i,j)$ denote the hop distance from vertex i to vertex j, and define

$$C_C(i) = \frac{n-1}{\sum_j d(i,j)}$$

where n is the number of vertices in the graph.

- **Betweeness centrality:** let g_{jk} denote the number of paths connecting vertices j and k, and let $g_{jk}(i)$ denote the number of those paths that go through vertex i,

$$C_B(i) = \sum_j \sum_{k \neq j} \frac{g_{jk}(i)}{g_{jk}}$$

Problem 2

Modeling opinions on social networks

Modeling opinions on social networks

- **Motivation:** model the evolution of opinions on a large social network.
- As for the PageRank problem, we focus on a **typical** vertex.
- **Goal:** obtain a characterization of the typical stationary behavior of the process being studied, that is:
  - Tractable
  - Easy to estimate from simple network statistics
  - Valid with high probability on almost any real-world complex network

Modeling opinions on social networks

- We model individuals as vertices on a marked directed graph $G = (V, E; \mathscr{A})$.
- An edge from vertex i to vertex j, (i, j), is interpreted as:

  "individual j **listens** to individual i".
- Individuals hold **opinions** about a given topic.
- Opinions take values on the interval $[-1, 1]$.
- There may be an **external media** that broadcasts a variety of opinions.
- At each time step $k = 1, 2, \ldots$, each individual **listens** to the opinions of all its inbound neighbors and those in the media, and then updates her own opinion.
- Individuals weigh the opinions they listen to in a personalized way, and may also control what media they listen to.

Model parameters: vertex attributes

- Let $(c(i,1), c(i,2), \dots c(i,n)) \geq 0$ be the vector of weights for her neighbors' opinions; $c(i,k) \equiv 0$ if $(k,i) \notin E$.
- Weights are assumed to satisfy:

$$\sum_{j=1}^{n} c(i,j) = c < 1 \qquad \text{if } d_i^- = \sum_{j=1}^{n} 1(j \to i) > 0.$$

- Individuals have an **internal opinion** $q_i \in [-1, 1]$.
- The internal opinion remains static throughout the process, and may influence its dynamics.
- We call a vertex i with $d_i^- = 0$ a **stubborn agent**.

Model parameters: vertex attributes

- Each vertex $i \in V$ in the graph has a **mark** $\mathbf{x}_i$.
- Vertex marks usually include their in-degree and out-degree, but they can also include many other vertex attributes.
- In our model, marks include:
  - Internal opinion
  - Community label
  - Amount of trust given to each inbound neighbor
- Vertex marks are assumed to take values on a Polish space $\mathcal{S}$.
- We equip $\mathcal{S}$ with a metric ρ.

Model parameters: external media

- Let $W_i^{(k)}$ denote the external media signal received by individual i at time k, $k = 0, 1, 2, \ldots$.
- The media signals $\{W_i^{(k)} : k \geq 0\}$ are i.i.d. given $\mathbf{x}_i$ and the $\{W_i^{(k)} : i \in V, k \geq 0\}$ are conditionally independent given $\{\mathbf{x}_i : i \in V\}$.
- Media signals satisfy

$$|W_i^{(k)}| \leq d + c - \sum_{j \in V} c(i, j),$$

  for some $d \in (0, 1)$.
- Let $\nu(\mathbf{x}_i)$ denote the distribution of $W_i^{(0)}$.
- Let $R_i^{(k)}$ denote the **opinion** of individual i at time k.

The DeGroot-Friedkin-Johnsen model

- The **DeGroot-Friedkin-Johnsen** model is widely used in the social sciences for modeling opinions.
- All individuals in the graph $G = (V, E; \mathscr{A})$ update their opinions simultaneously at step $k+1$ according to the recursion:

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

- **Special cases:**
  - $d_i^- \geq 1$ for all $i \in V \longrightarrow$ no stubborn agents
  - $c + d = 1 \longrightarrow$ no memory
  - $\{W_i^{(k)} : k \geq 0\}$ independent of $\mathbf{x}_i \longrightarrow$ pure noise
  - $\{W_i^{(k)} : k \geq 0\} \sim \nu(\mathbf{x}_i) \longrightarrow$ media signal that depends on individual's attributes

Goals for the model

- We want a model for the evolution of opinions on a social network that can predict complex behavior.
- The type of graphs covered in the analysis should be able to model real-world social networks.
- The opinions of individuals should be allowed to depend on their particular attributes (e.g., political inclinations).
- We want to model phenomena known as *confirmation bias* and *selective exposure*.
- The model should exhibit polarization under strong biases.
- **Goal:** explain when consensus is possible and quantify the potential of various depolarizing interventions.

Markov chain on a fixed graph

- The opinion model

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

on a marked directed graph $G = (V, E; \mathscr{A})$ defines a **Markov chain** on $[-1,1]^{|V|}$.

- Let $\mathbf{R}^{(k)} = (R_1^{(k)}, \ldots, R_{|V|}^{(k)})$.
- **Theorem:** (Fraiman-Lin-OC '22) Suppose G is locally finite and $d > 0$. Then, there exists a random vector $\mathbf{R}$ such that

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

Typical behavior

- Let $\mathbf{R} = (R_1, \ldots, R_{|V|})$ be the vector of stationary opinions.
- **Goal:** describe the distribution of R_I, where I is uniformly chosen in V.
- R_I represents the *typical* opinion of an individual in the network.
- The distribution of R_I also describes the proportion of individuals in the graph G having opinions in $A \subseteq [-1, 1]$, i.e.,

$$P(R_I \in A | G) = \frac{1}{|V|} \sum_{i \in V} 1(R_i \in A).$$

- In small graphs the distribution of $\mathbf{R}$ will greatly depend on G.
- On large graphs, the dependence on the detailed structure of G decreases, and only its main statistical properties matter.

Using the model to understand opinion formation

- Models like the DeGroot-Friedkin-Johnsen model have been widely used to study the question:

  Is there consensus as $k \to \infty$?

- By **consensus** we mean: does the process $\{\mathbf{R}^{(k)} : k \geq 0\}$ converge to a stationary distribution concentrated around **one** point?
- **Question:** Under what conditions can we expect consensus to exist?
- Other questions we explore are the effects of: **confirmation bias**, **selective exposure** and the presence of **bots**.
- Assume the media signals take the form:

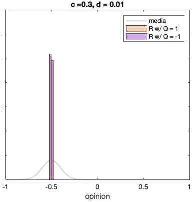
$$W_i^{(k)} = dZ_i^{(k)} + q_i \left(c - \sum_{j \in V} c(i,j) \right)$$

Parameters in the simulations

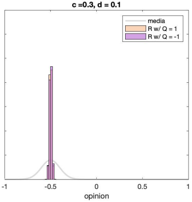
- **Trust on the media:** parameterized by d.
- **Community influence:** parameterized by c.
- **Assortativity:** individuals are more likely to connect to individuals from the same community than to individuals from a different community.
- **Selective exposure:** individuals listen to media signals that are different depending on their community.
- **Confirmation bias:** individuals put more weigh on the opinions of neighbors from their own communities.
- **Bots:** artificial accounts that send extreme signals and are stubborn (modeled as separate communities).
- **Influencers:** individuals who are central to the network and who can reach many people.

The role of the media and the trust put on it

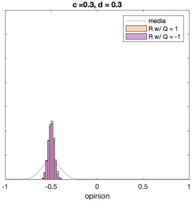
- Assortative dSBM, 2 communities having internal opinions in $\{-1, 1\}$.
- Media signals follow a truncated normal $N(-0.5, 0.01)$ distribution.
- Everybody listens to the same media.
- μ_i is the mean opinion in community i (standard deviation).
- Community influence is $c = 0.3$ in all simulations.



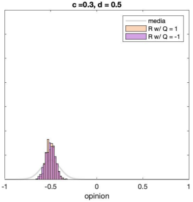
(a) Lack of Trust
$\mu_1 = 0.5$ (0.0000)
$\mu_2 = -0.5$ (0.0000)



(b) Little Trust
$\mu_1 = 0.5$ (0.0141)
$\mu_2 = -0.5$ (0.0141)



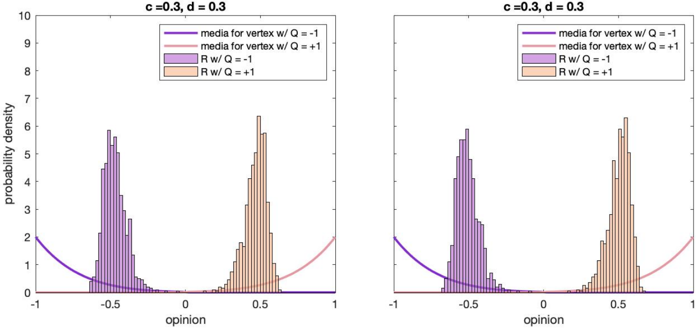
(c) Moderate Trust
$\mu_1 = 0.5$ (0.0332)
$\mu_2 = -0.5$ (0.0332)



(d) High Trust
$\mu_1 = 0.5$ (0.0511)
$\mu_2 = -0.5$ (0.0511)

Polarization due to selective exposure

- Neutral dSBM, 2 communities having internal opinions in $\{-1, 1\}$.
- Individuals choose the media signals they want to listen to, given by the translated Beta distributions shown in the figures.
- Individuals can choose to trust neighbors from their own community more.

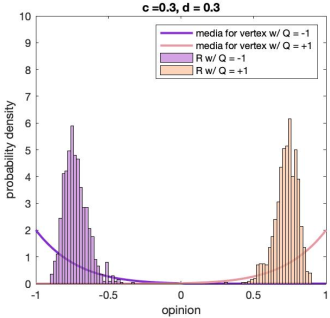


(a) Uniform trust to all neighbors.
$\mu_1 = 0.47$ (0.0748),
$\mu_2 = -0.47$ (0.0755)

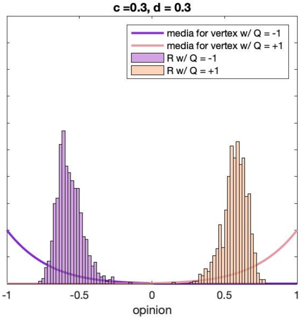
(b) Higher trust to same community neighbors.
$\mu_1 = 0.50$ (0.0742),
$\mu_2 = -0.51$ (0.0748)

Polarization due to bots

- Assortative dSBM, 2 communities having internal opinions in $\{-1, 1\}$.
- Selective exposure with biased media.
- There are bots in the network that target people according to their community label, and push them towards the extremes.



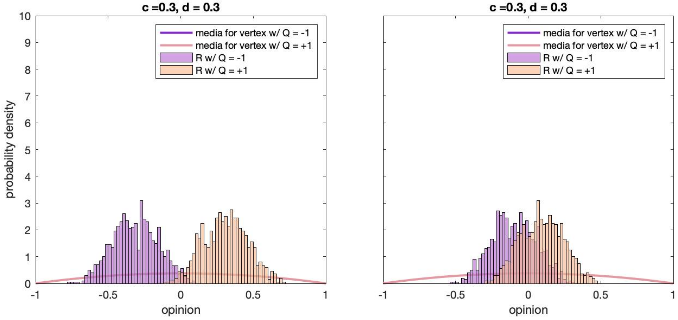
(a) Uniform trust to neighbors.
$\mu_1 = 0.73\ (0.0787)$,
$\mu_2 = -0.72\ (0.0837)$



(b) Discount trust on artificial accounts.
$\mu_1 = 0.57\ (0.0843)$,
$\mu_2 = -0.57\ (0.0877)$

Depolarizing with a balanced media

- Assortative dSBM, 2 communities having internal opinions in $\{-1, 1\}$.
- Targeted polarizing bots sending extreme signals.
- Media is neutral and the same for everyone.

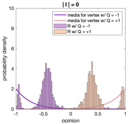


(a) Uniform trust to neighbors
$\mu_1 = 0.31\ (0.1520)$,
$\mu_2 = -0.31\ (0.1572)$

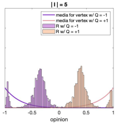
(b) Little trust in artificial accounts
$\mu_1 = 0.09\ (0.1487)$,
$\mu_2 = -0.10\ (0.1536)$

Depolarizing with influencers

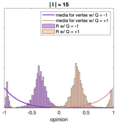
- Degree-corrected, assortative dSBM, selective exposure, no bots.
- Top influencers in the network **countermessage** their followers.
- As number of influencers increases ($d = 0.1$ top, $d = 0.3$ bottom).



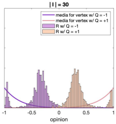
(a) $|\mu_1 - \mu_2| = 0.9469$



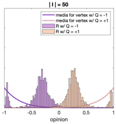
(b) $|\mu_1 - \mu_2| = 0.8727$



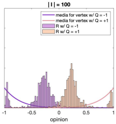
(c) $|\mu_1 - \mu_2| = 0.8168$



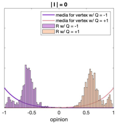
(d) $|\mu_1 - \mu_2| = 0.7709$



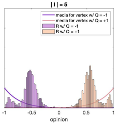
(e) $|\mu_1 - \mu_2| = 0.7148$



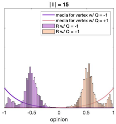
(f) $|\mu_1 - \mu_2| = 0.6139$



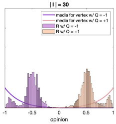
(g) $|\mu_1 - \mu_2| = 1.2106$



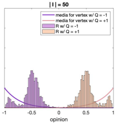
(h) $|\mu_1 - \mu_2| = 1.1405$



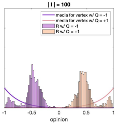
(i) $|\mu_1 - \mu_2| = 1.1140$



(j) $|\mu_1 - \mu_2| = 1.0828$



(k) $|\mu_1 - \mu_2| = 1.0519$



(l) $|\mu_1 - \mu_2| = 0.9973$

References

- **Google's PageRank:**

[1] S. Brin and L. Page. *The anatomy of a large-scale hypertextual Web search engine.*. Comput. Networks ISDN Systems, 30(1-7):107-117, 1998.

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

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

- **Opinion dynamics:**

[5] M. H. DeGroot. *Reaching a consensus*. Journal of the American Statistical Association, 69(345):118-121, 1974.

[6] N.E. Friedkin and E.C. Johnsen. *Social influence and opinions*. Journal of Mathematical Sociology, 15(3-4):193-206, 1990.

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

Next lecture

- We will talk about how random graph theory, heavy-tailed asymptotics, and mean-field approximations can help answer today's questions.
- Our analysis will be based on static random graph models.
- The key technique is something known as **local weak convergence**.
- Explicit formulas for distributions, means and variances are obtained through **distributional fixed-point equations**.
- For denser graphs, the approach uses **mean-field** analysis.