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

Lecture 1

Mariana Olvera-Cravioto

UNC Chapel Hill

molvera@email.unc.edu

February 5th, 2024

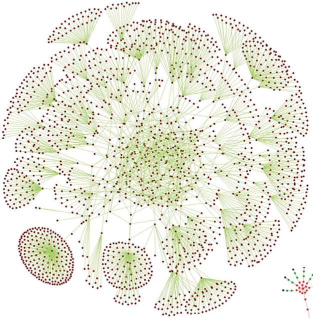
Social networks and graphs

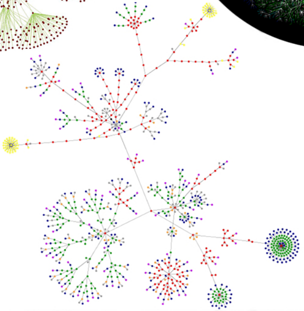
- The internet, the web, Facebook, X (Twitter), LinkedIn, Instagram, WhatsApp, WeChat, Snapchat, Pinterest, Reddit, etc. are all examples of **networks**.
- In **social networks**, connections occur among people.
- A connection between two people can mean many different things depending on the network, e.g., friendship, hyperlinks, follower-followed relations, etc.
- There are also many networks that do not involve people at all, e.g., the internet, neural connections in the brain, interactions between proteins in biology, articles in a citation network, etc.
- When analyzing networks, it is often convenient to think of them as **graphs**.

Graphs

- A graph consists on a set of **vertices**, V, and a set of **edges** E.
- Graphs can be **undirected** or **directed**.
- In an undirected graph, the relation between the vertices is symmetric, while in a directed graph it is not.
- We will call the vertices $V = \{1, 2, \ldots, n\}$, and write $i \to j$ to mean there is an edge (perhaps undirected) from vertex i to vertex j.
- In an undirected graph, the **degree** of a vertex is the number of edges incident to it.
- In a directed graph, the **in-degree** is the number of inbound edges and the **out-degree** is the number of outbound edges.

Different types of graphs





Types of graphs

- **Simple graphs:** a graph that has no self-loops nor multiple edges between any two vertices.
- **Multigraphs:** a graph that may have self-loops or multiple edges between two vertices.
- **Connected graphs:** (undirected) graphs where every pair of vertices is connected through a path.
- **Strongly connected graphs:** (directed) graphs where for any pair of vertices i and j, there exists a directed path from i to j and one from j to i, not necessarily the same.
- **Complete graphs:** there is an edge between every pair of vertices in the graph.
- **Sparse graphs:** the average number of edges is of the same order as the number of vertices.

Structures and properties

- Some structures that can be of interest when studying graphs are:
  - **Cycles:** paths that start and end with the same vertex without repeating vertices.
  - **Cliques:** complete subgraphs.
  - **Distance between two vertices:** length of the minimum path connecting two vertices; in directed graphs the path must be directed.
  - **Component of a vertex:** the set of vertices that can be reached through (directed) paths from a given vertex.
- Some properties of interest:
  - **Diameter:** the maximum distance between two points in the graph.
  - **Components:** sizes of the largest, second largest, etc.
  - **Cycle lengths:** the typical length of cycles in the graph.
  - **Clustering:** the proportion of triangles (3-cliques) vs. open wedges.
  - **Communities:** subsets of vertices that have more edges among their vertices than with vertices outside the set.

Some questions of interest

- Is the graph (strongly) connected?
  - If not, does there exist a **giant** (strongly) connected component? (In a graph with n vertices, a giant has βn vertices for some $\beta > 0$)
  - What is the size of the smaller components?
- What is the diameter of the graph?
- What is the typical distance between vertices in the graph?
- What is the degree distribution, e.g.,

$$p_n(k) = \frac{1}{n} \sum_{i=1}^{n} 1(d_i = k), \qquad d_i = \text{degree of vertex } i,$$

  in the graph?
- Does the graph have clusters/communities?
- Are there vertices that are more "influential" or "central" to the network?

The small world phenomenon

- In the late 60's, a social psychologist named Stanley Milgram conducted a set of experiments to try to determine the typical length of paths connecting two individuals in the United States.
- A letter addressed to somebody in Boston would be given to a set of randomly chosen people in different states in the Midwest, strangers to the recipient, with the instruction to help it reach its destination by sending it to an acquaintance.
- **Result:** it took an average of 6 people to connect the first sender and the final recipient, something that became known as the

  small world or **six degrees of separation**

  phenomenon.
- Interestingly, the small world property is very common in large real-world networks.

Scale-free networks

- Recall that the degree of a vertex $i \in V = \{1, 2, \dots, n\}$ in an undirected graph, denoted d_i, is the number of edges incident to it.
- The proportion of vertices having degree $k = 0, 1, 2, \dots,$ is given by

$$p_n(k) = \frac{1}{n} \sum_{i=1}^{n} 1(d_i = k)$$

- We call $\{p_n(k) : k \geq 0\}$ the **degree distribution**.
- If the degree distribution of a graph satisfies

$$p_n(k) \propto k^{-\gamma}$$

  for some $\gamma > 0$ (usually $\gamma \in (2, 3)$), we say that the graph is **scale-free**.
- In a scale-free graph there are vertices that have really large degrees, even if the average degree is small.

Random graph models

- Some real networks are too big to be analyzed exactly.
- Some may even be constantly changing.
- **Idea:** we can think of our specific real-world graph as just one "typical" element of a larger class.
- If we can show that a property holds for a large class of graphs, it is likely it will hold for our specific graph.
- **Random graphs** are mathematical models that can help us understand large real-world graphs.
- No random graph model can mimic all the properties of a specific real-world graph, so we focus on choosing models that share certain properties that are important to the problem we want to analyze.

Large graph limit

- Random graph models consist of a vertex set $V_n = \{1, 2, \ldots, n\}$ and a set of rules for determining whether a given edge is present or not based on some random events.
- Their mathematical analysis is usually done under the **large graph limit** $n \to \infty$ on a sequence of graphs $\{G_n = (V_n, E_n) : n \geq 1\}$.
- Taking the limit $n \to \infty$ simplifies computations in order for us to identify general properties.
- In practice, establishing results in the large graph limit means that our findings are likely to be true for sufficiently large graphs.

Static vs. evolving models

- Random graph models can be broadly classified into two categories: **static models** and **evolving** or **growing models**.
- Static models are meant to represent a "snapshot" of a large network.
- In static models G_n and G_{n+1} can be totally different.
- Evolving models are meant to describe the growth of a graph as vertices get added to the graph (usually one at a time), so G_n and G_{n+1} share most edges.
- In many evolving models edges and vertices never disappear, so G_n is a subgraph of G_{n+1}.

The Erdős-Rényi random graph

- The simplest model for a random graph is the **Erdős-Rényi model**.
- Consider a graph with vertex set $V_n = \{1, 2, \ldots, n\}$.
- There are a total of $\binom{n}{2}$ possible edges in the graph, and each of them will be chosen to be present or not with a coin flip.
- Suppose you have a coin that lands heads with probability $p \in (0, 1)$.
- For each pair of vertices i and j, toss the coin; if it lands heads, draw an edge between i and j, otherwise do nothing.
- Equivalently, if A denotes the adjacency matrix of the graph, let

$$A_{ij} = A_{ji} = 1(\text{coin-flip is a head}), \qquad i \neq j,$$

and set $A_{ii} = 0$.

Properties of the Erdős-Rényi model

- This is the most studied random graph model there is.
- Some of its connectivity properties are:
  - If $np < 1$ the graph will consists of only small components of size $O(\log n)$.
  - If $np \to c > 1$ the graph will contain a unique **giant** connected component, with all other components of size $O(\log n)$.
  - If $np = 1$ the largest component will have size $O(n^{2/3})$.
  - If $p < (1-\epsilon)n^{-1} \log n$ the graph will most likely be **disconnected**.
  - If $p > (1+\epsilon)n^{-1} \log n$ the graph will most likely be **connected**.
- When the graph is connected, it exhibits the **small-world** property, with typical distance of order $O(\log n)$.

Degree distribution

- To compute the degree distribution we can use binomial probabilities.
- Fix a vertex $i \in V_n$, then its degree is given by

$$D_i = \sum_{j=1}^{n} \chi_{i,j}, \qquad \chi_{i,j} = 1((i,j) \in E_n)$$

- Note that the $\chi_{i,j}$ are independent Bernoulli r.v.s with parameter p.
- Therefore, since all vertices have the same distribution, for all $i \in V_n$,

$$P(D_i = k) = P(D_1 = k) = P(\mathsf{Bin}(n,p) = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Moreover, if $np \to c$ as $n \to \infty$, we have that

$$\lim_{n\to\infty} P(D_1 = k) = \frac{e^{-c} c^k}{k!}, \qquad k \geq 0,$$

i.e., a Poisson distribution with mean c.... not **scale-free**.

Poisson vs. scale-free

- The Poisson distribution is **light-tailed**, i.e., its tail decreases exponentially fast.
- Poisson random variables tend to take values close to their mean.
- A scale-free distribution is **heavy-tailed**, i.e.,

$$\sum_{k=0}^{\infty} e^{\epsilon k} P(D = k) = \infty$$

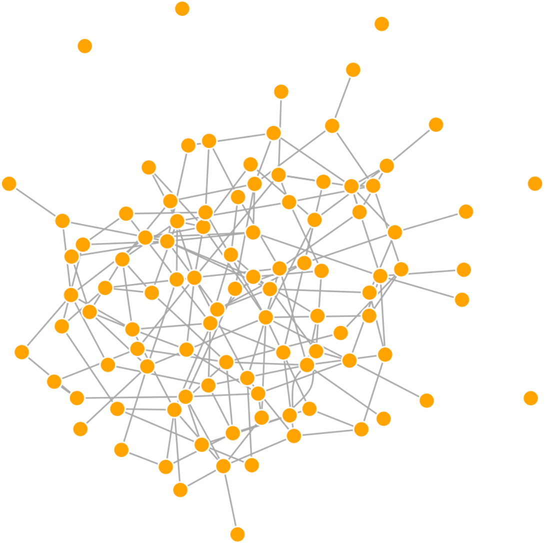
  for all $\epsilon > 0$.
- Heavy-tailed random variables can take extremely large values.
- In particular, for any $k \geq 1$,

$$\lim_{m\to\infty} P(D > k + m | D > m) = 1$$

  which can be interpreted as:

"Given that D is large, most likely it is huge."

An Erdős-Rényi graph



Inhomogeneous random graphs

- Erdős-Rényi graphs are quite **homogeneous**, i.e., all the vertices have degrees close to their common mean.
- Real-world networks are often scale-free.
- We can create random graphs that have inhomogeneous degrees by allowing the edge probabilities to vary from vertex to vertex.
- To each vertex $i \in V_n$ assign a value $w_i \geq 0$, and define the edge probability

$$p_{ij}^{(n)} := P((i,j) \in E_n) = \frac{w_i w_j}{l_n} \wedge 1, \qquad i \neq j,$$

  where $l_n = w_1 + \cdots + w_n$.
- The adjacency matrix of the graph is given by:

$$A_{ij} = \begin{cases} 1, & \text{with probability } p_{ij}^{(n)}, \\ 0 & \text{with probability } 1 - p_{ij}^{(n)}. \end{cases}$$

Inhomogeneous random graphs... cont.

- Each edge is determined independently of other edges.
- This choice of edge probabilities corresponds to the **Chung-Lu model**.
- The expected degree of vertex $i \in V_n$ is:

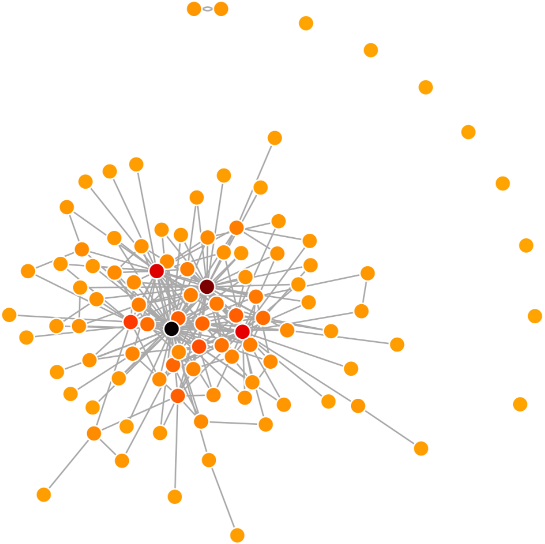
$$E[D_i] = \sum_{j=1}^{n} p_{ij}^{(n)} \approx w_i$$

- If we let

$$F(x) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} 1(w_i \leq x),$$

  then the degree distribution "looks" like F (in fact, p_n converges to a mixed Poisson with mixing distribution F).
- If we set $w_i = p$ for all $i \in V_n$ we get an Erdős-Rényi model.
- Scale-free graphs can be obtained by choosing F to be a power-law distribution.

An inhomogeneous random graph



Graphs with communities

- Inhomogeneous random graphs can be **scale-free** and will have the **small-world** property.
- However, they do not have community structure.
- Suppose we want to generate a graph with K communities.
- To each vertex $i \in V_n$ assign a community label $J_i \in \{1, 2, \ldots K\}$.
- Now sample edges independently using 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 $\kappa : \{1, \ldots, K\} \times \{1, \ldots, K\} \to [0, \infty)$.
- The parameter θ_n can be used to create dense graphs.
- The size of community $k \in \{1, \ldots, K\}$ is $n\pi_k^{(n)} = \sum_{i=1}^{n} 1(J_i = k)$.

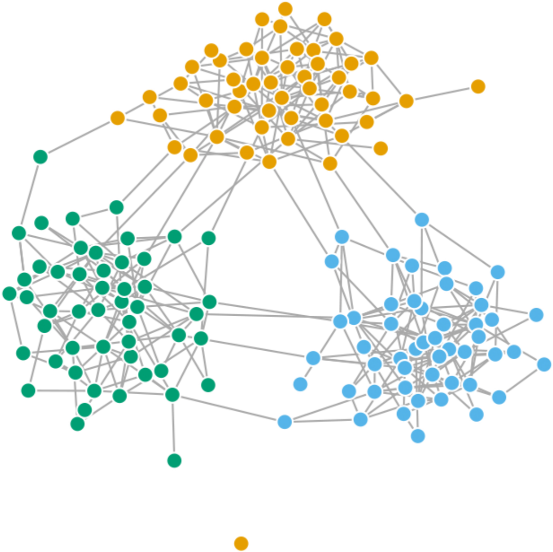
Graphs with communities... cont.

- This construction is known as a **stochastic block model**.
- In order to create communities we choose $\kappa(a,b)$ be "large" for $a=b$, and "small" for $a \neq b$.
- The expected degree of a vertex in community $m \in \{1,\dots,K\}$ is:

$$E[D_i|J_i=m] = \sum_{j=1}^{n} \frac{\kappa(m,J_j)}{n} = \sum_{r=1}^{K} \kappa(m,r)\pi_r^{(n)}$$

- Stochastic block models are homogeneous within each community, but can have different expected degree from one community to another.
- **Degree corrected** versions of the stochastic block model can create inhomogeneity while preserving the community structure.

A stochastic block model



Graphs with clustering

- The **global clustering coefficient** of a graph is

$$\frac{\text{number of triangles}}{\text{number of open wedges}}$$

- Inhomogeneous random graphs do not have significant clustering.
- In fact, inhomogeneous random graphs are **locally tree-like**.
- They have "long" cycles of length $O(\log n)$.
- The clustering coefficient in the models we have seen converges to zero as $n \to \infty$.
- Real-world graphs often have positive clustering coefficients, especially social networks.

Graphs with clustering... cont.

- To construct a graph with non-negligible clustering, we start by generating a **bipartite graph** with vertex sets $V_n = \{1, \ldots, n\}$ and $\mathcal{A}_m = \{a_1, \ldots, a_m\}$, $n, m \geq 1$.
- To each vertex $i \in V_n$ assign a value $w_i \geq 0$ and define

$$p_i = \frac{\gamma w_i}{n} \wedge 1,$$

  where $\gamma > 0$ is a fixed parameter.
- Next, for each $i \in V_n$ toss a coin that lands heads with probability p_i with each of the vertices in $\mathcal{A}_m$, and draw an edge if it is a head.
- Let $N(i) \subseteq \mathcal{A}_m$ be the set of neighbors of i.
- We will now construct a new graph $G_n = (V_n, E_n)$, with adjacency matrix A by setting:

$$A_{ij} = 1(N(i) \cap N(j) \neq \varnothing)$$

Graphs with clustering... cont.

- This model is called a **random intersection graph**.
- Let $F(x) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} 1(w_i \le x)$ be the weight distribution, and assume it has finite mean.
- If we choose $m = \lfloor \beta n \rfloor$, the degree of vertex $i \in V_n$ in G_n will have (approximately) the distribution of

$$\mathsf{Poisson}(\beta\gamma w_i) + \mathsf{Poisson}(\gamma),$$

  with the two Poisson r.v.s independent of each other.
- As with inhomogeneous random graphs, we can obtain the scale-free property by choosing F to be a power-law distribution.
- The parameters β, γ can be used to tune the clustering coefficient to cover the entire range $(0, 1)$, with small values of $\beta\gamma$ producing higher clustering.

An intersection graph

The Albert-Barabási model

- All the random graph models we have seen so far are **static**.
- Static models do not explain how graphs grow.
- **Evolving** models propose a mechanism for choosing how a new vertex will connect to the existing graph.
- Vertices are labeled in the order in which they arrive to the graph.
- One of the most famous evolving random graph models is the **Albert-Barabási graph** or **preferential attachment model**.
- This model assumes that an incoming vertex will choose a vertex to connect to with probability proportional to its degree.
- In other words, newcomers “prefer” to attach to high degree vertices.

The Albert-Barabási model... cont.

- The model starts with one vertex that has a self-loop.
- At each time step, a new vertex arrives and connects by drawing one edge either to itself, or to an existing vertex.
- Let $D_i(k)$ be the degree of vertex i after k vertices have arrived.
- When vertex $k+1$ arrives it attaches to vertex i with probability:

$$p_i(k) = \begin{cases} \frac{D_i(k)}{2k+1}, & i = 1, \ldots, k, \\ \frac{1}{2k+1}, & i = k+1. \end{cases}$$

- This model produces **scale-free** graphs with degree distribution:

$$P_k(n) = \frac{1}{n} \sum_{i=1}^{n} 1(D_i(n) = k) \approx 4k^{-3}$$

for large n.

Preferential attachment models

- A generalization of the model allows each new vertex to attach using $m \geq 1$ edges, and attaches the jth edge of vertex $k+1$ to vertex i with probability:

$$p_i(k) = \frac{D_i(k, j-1) + \delta}{\sum_{v=1}^{k} (D_v(k, j-1) + \delta)}, \qquad i = 1, \ldots, k, k+1,$$

  where $\delta > -m$ and $D_i(t, j)$ is the degree of vertex i after t vertices have arrived and j edges of vertex $t+1$ have been attached.

- This model generates **scale-free** graphs with degree distribution

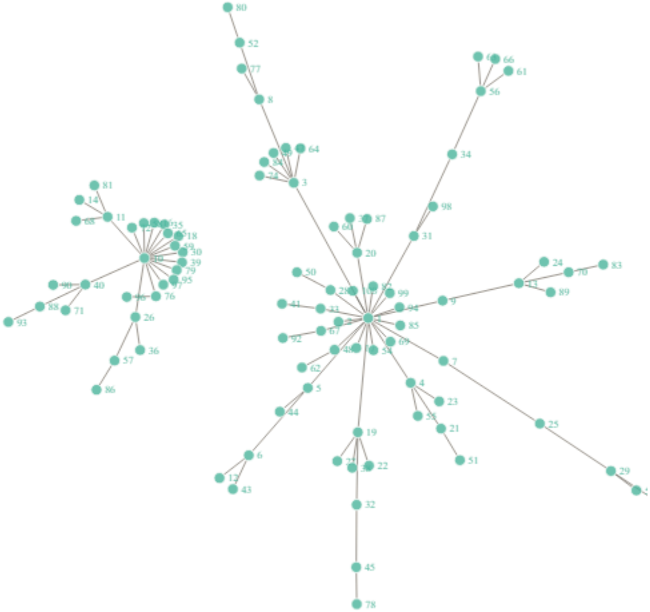
$$P_k(n) = \frac{1}{n} \sum_{i=1}^{n} 1(D_i(n, m) = k) \approx C_{m,\delta}\, k^{-\tau}$$

  for large n, where $\tau = 3 + \delta/m$.

Preferential attachment models... cont.

- In preferential attachment models, the degrees of **older** vertices are very different from those of **younger** ones.
- In contrast, all the static models we discussed have **exchangeable** vertices.
- The "time-stamp" of a vertex, i.e., its time of arrival, gives us a lot of information about its properties.
- Older vertices tend to have larger degrees.
- The **largest degree** grows as $O(n^{-1/(2+\delta/m)})$ as $n \to \infty$.

An Albert-Barabási graph



References and next lecture

- The topics covered in today's lecture are now **classic**.
- Textbooks:

[1] Remco van der Hofstad. *Random Graphs and Complex Networks, Vol. I*. Cambridge University Press, 2016.

[2] Béla Bollobas. *Random Graphs*. 2nd Edition, Cambridge University Press, 2001.

- **Next lecture:**
  - We will talk about two problems: Google's PageRank algorithm and an opinion dynamics model.
  - Both problems can be stated as (stochastic) processes on a fixed large directed graph.
  - When we model the underlying graph as a realization from a suitable random graph model, we can obtain interesting insights and tractable formulas.