Introduction to Network Models

Amir Ajourlou

Laboratory for Information and Decision Systems
Institute for Data, Systems, and Society
Massachusetts Institute of Technology

Lecture 10
Recall the diameter of a graph: let $d_{ij}$ be the distance between nodes $i$ and $j$ (i.e., length of the shortest path between $i$ and $j$).

$$\text{diameter} = \max_{i,j} d_{ij}.$$ 

We will show that the diameter of the ER graph varies as $\ln n$.

**Heuristic Argument:**
- Let $c$ denote the average degree of a node, $c = (n - 1)p$.
- The average number of nodes $s$ steps away from a randomly chosen node is $c^s$.
- The number of nodes reached is equal to the total number of nodes when $c^s \approx n$, or $s \approx \frac{\ln n}{\ln c}$.
- Every node is within $s$ steps of the starting point, implying that the diameter is approximately $\frac{\ln n}{\ln c}$.
- This argument works when $s$ is small (breaks down when $c^s$ become comparable with $n$ since number of nodes within distance $s$ cannot exceed number of nodes in the whole graph).
Consider two different starting nodes $i$ and $j$. The average number of nodes $s$ and $t$ steps away from them will be equal to $c^s$ and $c^t$ (assume both remain smaller than order $n$).

We have $d_{ij} > s + t + 1$ if and only if there is no edge between the surfaces. Since there are on average $c^s \times c^t$ pairs of nodes between surfaces, this implies $P(d_{ij} > s + t + 1) = (1 - p)^{c^{s+t}}$. Denoting $l = s + t + 1$, we have

$$P(d_{ij} > l) = (1 - p)^{c^{l-1}} \approx \left(1 - \frac{c}{n}\right)^{c^{l-1}}.$$
Taking logs of both sides, we find
\[ \ln P(\text{dist} > l) = c^{l-1} \ln \left(1 - \frac{c}{n}\right) \approx -\frac{c^l}{n}, \]
where we used \( \ln(1 + x) \approx x \) (which holds for large \( n \)). Therefore,
\[ P(\text{dist} > l) = \exp\left(-\frac{c^l}{n}\right). \]

The diameter is the smallest \( l \) such that \( P(\text{dist} > l) \) is zero. The preceding will tend to zero only if \( c^l \) grows faster than \( n \), i.e., \( c^l = an^{1+\epsilon} \) for some constant \( a \) and \( \epsilon \to 0 \) (note that this can be achieved while keeping both \( c^s \) and \( c^t \) smaller than \( n \)).

Rearranging for \( l \), we obtain the diameter as
\[
l = \frac{\ln a}{\ln c} + \lim_{\epsilon \to 0} \frac{1 + \epsilon}{\ln c} \ln n = A + \frac{\ln n}{\ln c},
\]

Example: Let \( n = 7 \times 10^9 \) and \( c = 1000 \). Then, \( l = \frac{\ln n}{\ln c} = 3.3 \).
Stochastic block model (SBM)

- ER graphs are too homogeneous
  - No community structure arises
- What if probabilities \( p \) are not the same for all edges?
  - Divide the nodes into blocks
  - Edge probability \( p \) is larger within blocks
  - Edge probability \( q \) is smaller between blocks
- If \( p = q \), we recover the traditional ER graph
SBM with two symmetric communities

- Also called the *planted bisection* model \( \Rightarrow \) Equal size communities

![Graph with blue and red communities]

© Source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/.

- When can we recover both communities from observing the graph?
  - Detection \( \Rightarrow \) \( \mathbb{P}(\frac{d(\hat{X}, X)}{n} < 0.5 - \epsilon) \rightarrow 1 \) [Mossel, Neeman, Sly, 2012]
    \[ p = \frac{a}{n}, \; q = \frac{b}{n}, \text{ Detection iff } (a - b)^2 > 2(a + b) \]
  - Recovery \( \Rightarrow \) \( \mathbb{P}(\hat{X} = X) \rightarrow 1 \) [Abbe, Bandeira, Hall, 2016]
    \[ p = \frac{a \log n}{n}, \; q = \frac{b \log n}{n}, \text{ Recovery iff } \frac{a + b}{2} \geq 1 + \sqrt{ab} \]
The $G_{n,p}$ model and real-world networks

- For large graphs, $G_{n,p}$ suggests $P[d]$ with an exponential tail
  $\Rightarrow$ Unlikely to see degrees spanning several orders of magnitude

- Concentrated distribution around the mean $\mathbb{E}[D_v] = (n - 1)p$

- Q: Is this in agreement with real-world networks?
Degree distributions of the WWW analyzed in [Broder et al '00]

⇒ Web a digraph, study both in- and out-degree distributions

Majority of vertices naturally have small degrees

⇒ Nontrivial amount with orders of magnitude higher degrees
Seems to be a structural pattern

- More heavy-tailed degree distributions found in [Barabasi-Albert '99]
- Caveat: Their mathematics is not very precise and some of their conclusions are incorrect

These heterogeneous, diffuse degree distributions are not exponential
Power-law degree distributions

Log-log plots show roughly a linear decay, suggesting the power law

\[ P[d] \propto d^{-\alpha} \Rightarrow \log P[d] = C - \alpha \log d \]

- Power-law exponent (negative slope) is typically \( \alpha \in [2, 3] \)
- Normalization constant \( C \) is mostly uninteresting

- Power laws often best followed in the tail, i.e., for \( d \geq d_{\text{min}} \)
The Scale-Free Property

Figure 4.4

(a) Comparing a Poisson function with a power-law function ($\theta = 2.1$) on a linear plot. Both distributions have $\theta_k = 10$.

(b) The same curves as in (a), but shown on a log-log plot, allowing us to inspect the difference between the two functions in the high-$k$ regime.

(c) A random network with $\theta_k = 3$ and $N = 50$, illustrating that most nodes have comparable degree $k_{\theta \theta} k_{\theta \theta}$.

(d) A scale-free network with $\theta = 2.1$ and $\theta_k = 3$, illustrating that numerous small-degree nodes coexist with a few highly connected hubs.

The Largest Hub

All real networks are finite. The size of the WWW is estimated to be $N \approx 10^{12}$ nodes; the size of the social network is the Earth's population, about $N \approx 7 \times 10^9$. These numbers are huge, but finite. Other networks pale in comparison: The genetic network in a human cell has approximately 20,000 genes while the metabolic network of the *E. Coli* bacteria has only about a thousand metabolites. This prompts us to ask: How does the network size affect the size of its hubs? To answer this we calculate the expected maximum degree, $k_{\text{max}}$, called the natural cutoff of the degree distribution $p_k$.

It represents the expected size of the largest hub in a network. It is instructive to perform the calculation first for the exponential distribution $p_k$. For a network with minimum degree $k_{\text{min}}$, the normalization condition provides $C = e^{-k_{\text{min}}}$. To calculate $k_{\text{max}}$ we assume that in a network of $N$ nodes we expect at most one node in the $(k_{\text{max}}, \infty)$ regime (ADVANCED TOPICS 3.B). In other words the probability to observe a node whose degree exceeds $k_{\text{max}}$ is $1/N$:

$$P(d) = d^{-2.1}$$

$P(d)$

$P(d)=d^{-2.1}$

Poisson

$P(d)=d^{-2.1}$

Poisson

Erdős-Renyi’s Poisson degree distribution exhibits a sharp cutoff

⇒ Power laws upper bound exponential tails for large enough $d$

Scale-free network: degree distribution with power-law tail
Popularity as a network phenomenon

- **Popularity** is a phenomenon characterized by extreme imbalances
  - How can we quantify these imbalances? Why do they arise?

  ![Bell Curve vs. Power Law Distribution](image-url)

- Basic models of network behavior can be very insightful
  - Result of coupled decisions, correlated behavior in a population
Barabasi-Albert model

- Network model capturing the notion of preferential attachment
- Initial graph size $M$, connection number $m$, and stopping time $T$
  - 1) Start with $M$ fully connected nodes
  - 2) Add a new node and randomly connect it to $m$ existing nodes
  - 3) Random connections with probability proportional to degrees
  - 4) Repeat $T$ times
  - Turns out this model has existed in literature in one way or another for 50 years.
    - Barabasi and Albert rediscovered and popularized it
    - Click here for a brief history.

- Degree distribution of resulting graph is power law up to a certain degree
- for degrees up to $n^{1/6}$
  - https://www.youtube.com/watch?v=4GDqJVtPEGg
Does the internet have an Achille’s heel?

Barabasi and Albert claimed the network of routers connecting the internet is scale-free

⇒ They claimed degree distribution follows a power law

If true, potentially, by attacking popular nodes we can make the network fail:

⇒ NO (fortunately)

Preferences attachment implies power-law degree distribution

However, the converse is NOT true! [Li, Alderson, Doyle, Willinger 2005]

Power law can arise from constrained optimization of network performance

you need more than random graph models to talk about internet

© A.K. Peters, Ltd.. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/.
1.022 Introduction to Network Models
Fall 2018

For information about citing these materials or our Terms of Use, visit: https://ocw.mit.edu/terms.