1.022 Introduction to Network Models

Amir Ajorlou

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

Lecture 10

Diameter of the ER graph

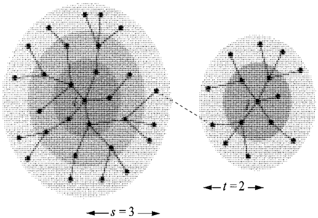
- Recall the diameter of a graph: let d_{ij} be the distance between nodes i and j (i.e., length of the shortest path between i and j).

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

- 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

$$P(d_{ij} > l) = (1 - p)^{c^{l-1}} \approx \left(1 - \frac{c}{n}\right)^{c^{l-1}}.$$

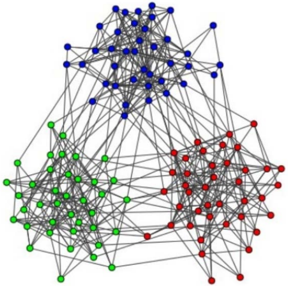


Diameter of the ER graph

- Taking logs of both sides, we find
$$\ln P(d_{ij} > l) = c^{l-1} \ln\left(1 - \frac{c}{n}\right) \approx -\frac{c^l}{n},$$
where we used $\ln(1+x) \approx x$ (which holds for large n). Therefore,
$$P(d_{ij} > l) = exp\left(-\frac{c^l}{n}\right).$$
- The diameter is the smallest l such that $P(d_{ij} > l)$ 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
$$l = \frac{\ln a}{\ln c} + \lim_{\epsilon \to 0} \frac{(1+\epsilon)\ln n}{\ln c} = A + \frac{\ln n}{\ln c},$$
- Example: Let $n = 7 \times 10^9$ and $c = 1000$. Then, $l = \frac{\ln n}{\ln c} = 3.3$.

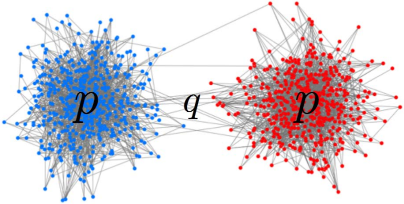
Stochastic block model (SBM)

- ER graphs are too homogeneous
  - $\Rightarrow$ No community structure arises
- What if probabilities p are not the same for all edges?
  - $\Rightarrow$ Divide the nodes into blocks
  - $\Rightarrow$ Edge probability p is larger within blocks
  - $\Rightarrow$ 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



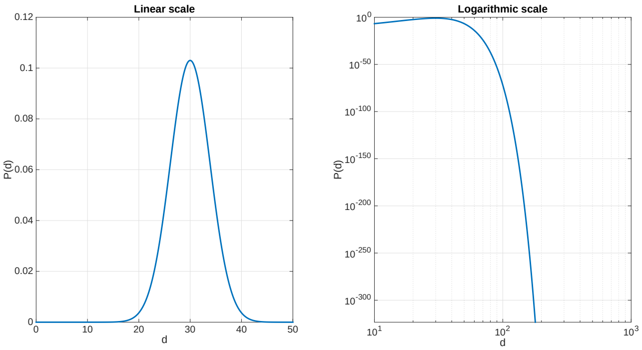
- When can we recover both communities from observing the graph?
- Detection $\Rightarrow \mathbb{P}(\frac{d(\hat{\mathbf{X}},\mathbf{X})}{n} < 0.5 - \epsilon) \to 1$ [Mossel, Neeman, Sly, 2012]

  $\Rightarrow p = a/n,\ q = b/n$, Detection iff $(a-b)^2 > 2(a+b)$
- Recovery $\Rightarrow \mathbb{P}(\hat{\mathbf{X}} = \mathbf{X}) \to 1$ [Abbe, Bandeira, Hall, 2016]

  $\Rightarrow p = \frac{a \log n}{n},\ q = \frac{b \log n}{n}$, Recovery iff $\frac{a+b}{2} \geq 1 + \sqrt{ab}$

The $G_{n,p}$ model and real-world networks

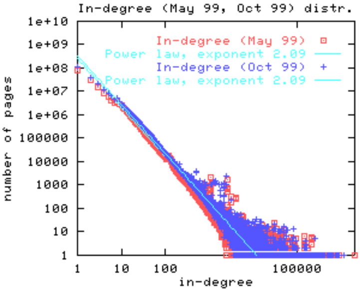
- For large graphs, $G_{n,p}$ suggests $\mathrm{P}[d]$ with an exponential tail
  - $\Rightarrow$ Unlikely to see degrees spanning several orders of magnitude

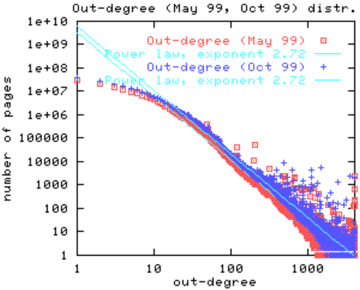


- Concentrated distribution around the mean $\mathbb{E}[D_v] = (n-1)p$
- Q: Is this in agreement with real-world networks?

World Wide Web

- Degree distributions of the WWW analyzed in [Broder et al '00]
  - $\Rightarrow$ Web a digraph, study both in- and out-degree distributions



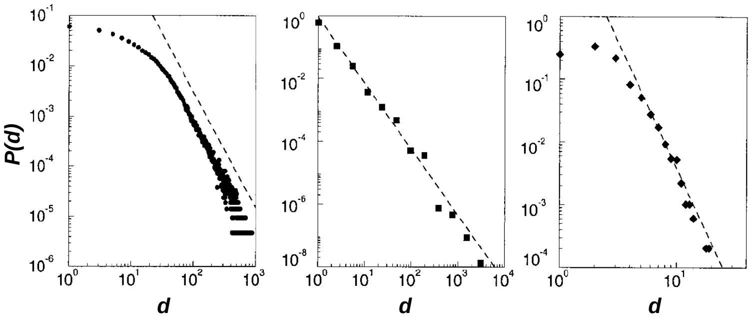


 Source: Broder, Andrei, Ravi Kumar, Farzin Maghoul, et al. "Graph structure in the Web." *Computer Networks* 33 (2000) 309–20.

- Majority of vertices naturally have small degrees
  - $\Rightarrow$ Nontrivial amount with orders of magnitude higher degrees

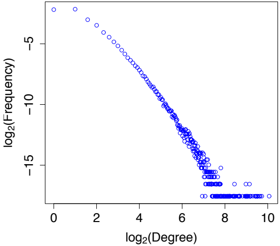
Seems to be a structural pattern

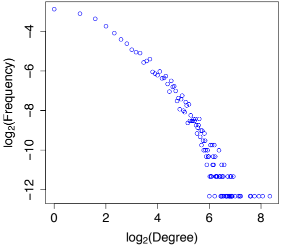
- More heavy-tailed degree distributions found in [Barabasi-Albert '99]
- Caveat: Their mathematics is not very precise and some of their conclusions are incorrect



Author collaboration | Web graph | Power grid

- These heterogeneous, diffuse degree distributions are not exponential



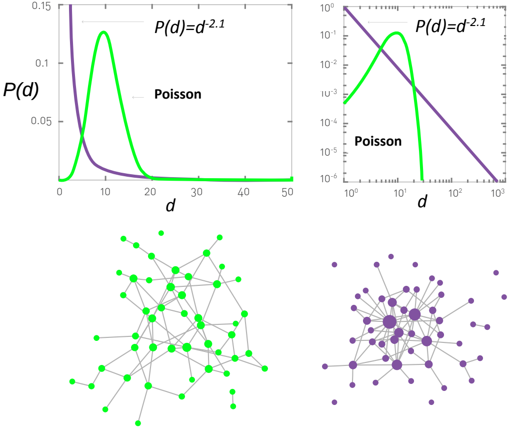


- Log-log plots show roughly a linear decay, suggesting the power law

$$\mathrm{P}\left[d\right] \propto d^{-\alpha} \Rightarrow \log \mathrm{P}\left[d\right] = C - \alpha \log d$$

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

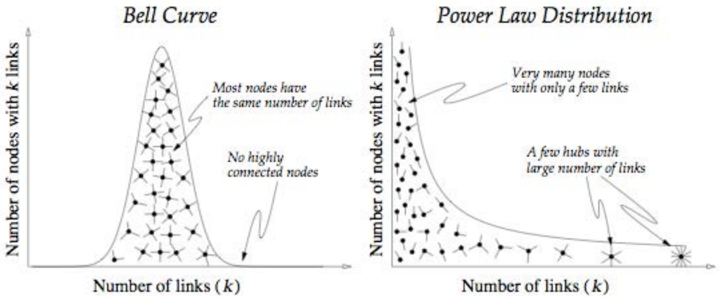
- Power laws often best followed in the tail, i.e., for $d \geq d_{\text{min}}$



- Erdös-Renyi's Poisson degree distribution exhibits a sharp cutoff
  - $\Rightarrow$ 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?



- Basic models of network behavior can be very insightful

  $\Rightarrow$ 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)
- Preferential 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