Network Models

Random, Small World & Preferential Attachment Models

— SL07 —

Philippe Cudré-Mauroux

pcm@unifr.ch

TABLE OF CONTENTS — SL07

Network Models: Why? (1/2)

- Large social networks are extremely difficult to analyze
  - Millions of nodes
  - Billions of edges
  - Constant growth
  - Real-time updates
- Example: Facebook
  - 2011: 721 million nodes (users)
  - 2011: 69 billion edges (friends)
  - 2017: 1.86 billion monthly active users
  - Many, many further node & edge types

Network Models: Why? (2/2)

- *Models* can be used to
  - Generate realistic networks (e.g., for controlled experiments)
  - Understand underlying phenomena of existing networks (e.g., growth)
  - Predict future characteristics of existing networks

Map of Facebook friends ©Paul Butler

Properties of Real-World Networks

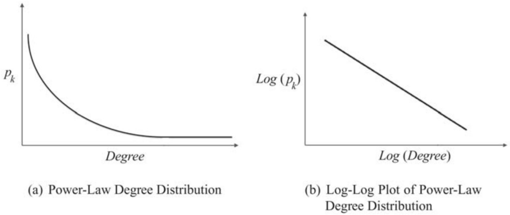
- Real-world networks share common characteristics
- When designing models, we aim to accurately mimic these common characteristics
- For networks in particular, three network attributes exhibit consistent measurements across real-world networks:
  - Degree distribution
  - Clustering coefficient
  - Average path length

Degree Distribution (1/3)

- Real-world networks are typically characterized by a power-law degree distribution
  - Many individuals with a few friends and a handful of users with tens of thousands of friends
  - Many sites are visited less than a thousand times a month, whereas a few are visited more than a million times daily
  - Most social media users are active on a few sites, whereas a few individuals are active on hundreds of sites

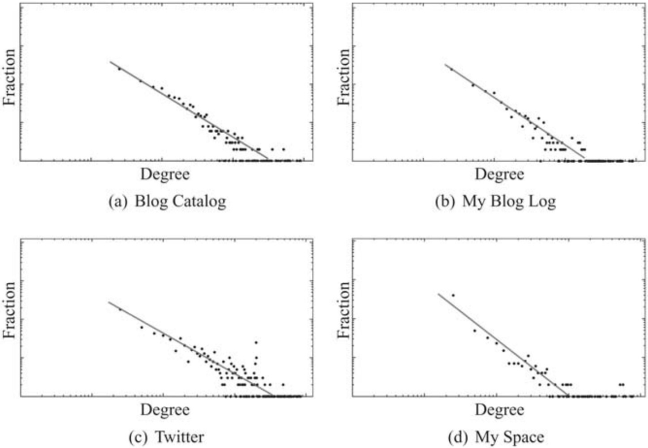
Degree Distribution (2/3)

- Mathematically, if k is the node degree and p_k the fraction of nodes with degree k then
  - $p_k = ak^{-b}$, where b is the power-law exponent and a is the power-law intercept; or, taking the log
  - $\log p_k = -b \log k + \log a$, a straight line with slope $-b$ and intercept $\log a$
  - Networks having a power-law degree distribution are often called scale-free



Power-Law Degree Distribution ©SMM

Degree Distribution (3/3)



(a) Blog Catalog (b) My Blog Log (c) Twitter (d) My Space

Log-Log Plots for Power-Law Degree Distribution in Social Media Networks ©SMM

Clustering Coefficient

- In real-world social networks, friendships are highly transitive
  - Friends of an individual are often friends with one another
  - These friendships form triads of friendships that are frequently observed in social networks
- The clustering coefficient of a node v is defined as follows:
  $C(v) = \frac{\text{\#connected pairs of v's neighbors}}{\text{\#pairs of v's neighbors}}$

Average Clustering Coefficient in Real-World Networks [SMM]

Web	Facebook	Flickr	LiveJournal	YouTube
0.081	0.14 [w/ 100 friends]	0.31	0.33	0.13

Average Path Length

- In real-world networks, any two members of the network are usually connected via short paths
  - This is known as the small-world phenomenon

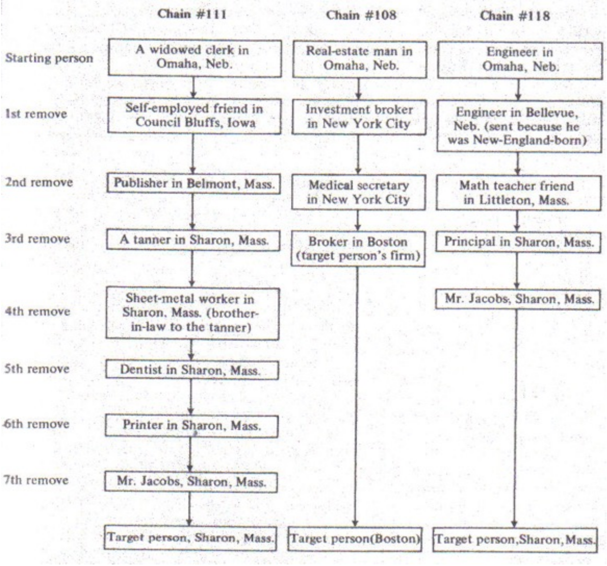
Average Path Length in Real-World Network [from SMM]

Web	Facebook	Flickr	LiveJournal	YouTube
16.1	4.7	5.7	5.9	5.1

Milgram's Small World Experiment (1/3)

- Finding short chains of acquaintances linking pairs of people in USA who didn't know each other (1967)
  - Source person in Nebraska and Kansas
  - Target person in Massachusetts
  - The letter could be only be given to persons one knows on a first name basis (acquaintances)

Milgram's Small World Experiment (2/3)



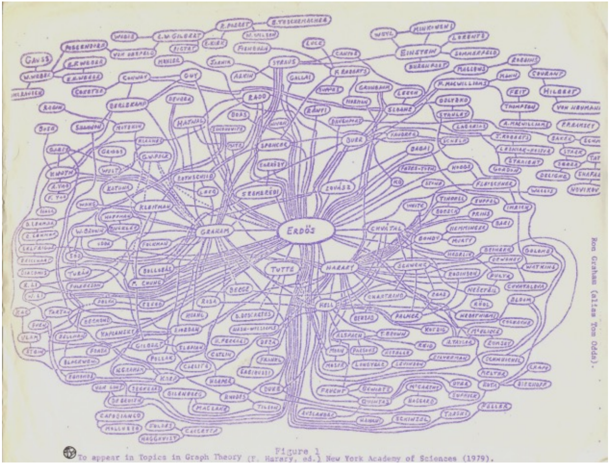
Milgram's Small World Experiment (3/3)

- Average length of the chains that were completed lied between 5 and 6 steps
  - Coined as "Six degrees of separation" principle
  - This was far less than assumed under "grid-like" assumptions!
  - Why are there short chains of acquaintances linking together arbitrary pairs of strangers?
  - Why is the diameter low?

WHAT ABOUT FACEBOOK?

- "Each person in the world (at least among the 1.59 billion people active on Facebook) is connected to every other person by an average of three and a half other people." Smriti Bhagat et al. 2016
  - Average: 3.57
  - Mark Zuckerberg: 3.17
  - Sheryl Sandberg: 2.92
  - Average was 3.74 in 2011

ERDÖS NUMBER (1/2)



Erdös Number (2/2)

- Paul Erdös was a famous mathematician who published more than 1500 papers
  - Many researchers are proud of being his collaborator
  - A person who writes a paper with him has an Erdös Number of 1
  - A person who writes a paper with a person whose Erdös Number is 1 has an Erdös Number of 2. And so on.
  - The median Erdös number is 5
  - The mean is 4.69, and the standard deviation is 1.27

The Kevin Bacon Game

- Invented in 1994 by two students at Albright College
  - The goal is to link any actor to Kevin Bacon through no more than six connections
  - Where two actors are connected if they have appeared in a movie together

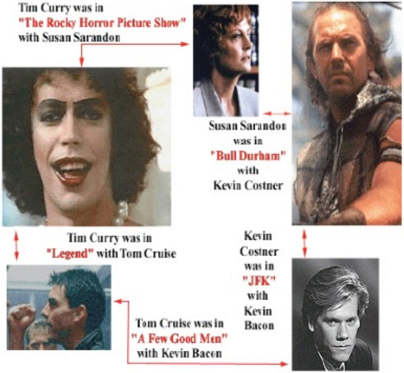


Table of contents — SL07

Random Graphs

- Now let us survey a few models approximating real networks
- We start with random graphs:
  - A random graph has a fixed number of nodes n
  - Any of the $\binom{n}{2}$ edges is created independently, with probability p
  - A random graph model can be unequivocally denoted by $G(n, p)$
  - Note that one random graph model $G(n, p)$ can create many different instances of graphs

Expected Degree

- The expected degree of a random graph is $(n-1)p$
- *Proof*:
  - each node can have $n-1$ links (neighbors)
  - each link exists with probability p
- Similarly, the expected number of edges is $\binom{n}{2}p$

Degree Distribution

- The degree distribution of a random graph follows a binomial distribution
  - $P(d_v = d) = \binom{n-1}{d} p^d (1-p)^{n-1-d}$
- For $n \to \infty$, the degree distribution tends to a Poisson distribution, which differs from the power-law degree distribution observed in real-world networks.

Clustering Coefficient

- The expected, global clustering coefficient of a random graph is p
- *Proof*:
  - The global clustering coefficient of a graph corresponds to the probability of two neighbors of a node being connected
  - In random graphs, for any two nodes, this probability is the same and is equal to p

Average Path Length

- The average path length in a random network is $l \approx \frac{\log n}{\log c}$ where c is the expected node degree
- *Proof [sketch]*:
  - Let D denote the expected diameter of the graph (length of the longest shortest path between any pair of nodes)
  - Starting at one node, one can visit c nodes after 1 step, c^2 nodes after 2 steps, etc.
  - Almost all nodes should be visited after D steps, i.e., $c^D \approx n$
  - In a random graph, the expected diameter size D tends to the average path l in the limit, hence $c^l \approx n$ and taking the log on both sides $l \approx \frac{\log n}{\log c}$

Modeling Real-World Networks with Random Graphs

- Given a real-world network, one can simulate it through a random graph model
  - Measure the average degree c of the network
  - Derive the corresponding p (slide 20): $p = \frac{c}{n-1}$
  - Simulate the network using $G(n, p)$

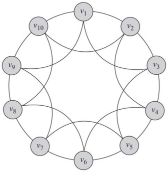
Example of Real VS Simulated Networks [from SMM]

Network	n	Real Network l	Real Network C	Simulated Network l	Simulated Network C
Film Actors	225K	3.65	0.79	2.99	0.0003
Medline co-authors	1.5M	4.6	0.56	4.91	0.0002
E. Coli	282	2.9	0.32	3.04	0.026

Table of contents — SL07

Small-World Model (1/3)

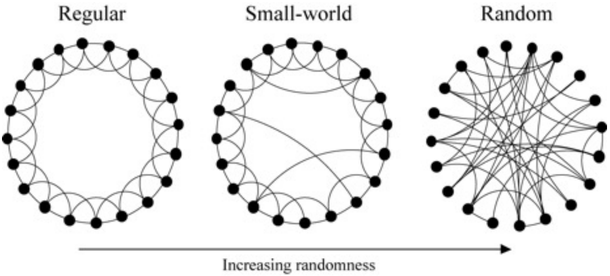
- Random graphs can model the average path length of real-world networks properly, but vastly underestimate their clustering coefficient
- To mitigate this problem, Duncan J. Watts and Steven Strogatz suggested a new model: the small-world model
- They start from an "egalitarian" model where each node is connected to the same number of neighbors through a regular pattern: a regular ring lattice



Regular Ring Lattice of Degree 4 ©SMM

SMALL-WORLD MODEL (2/3)

- The lattice can model transitivity well, however its average path length is too high
- To overcome this problem, the small-world model adds a rewiring step where each edge is rewired to a random destination with a probability β
- By varying β one can tune the degree of randomness of the resulting networks and introduce shortcuts



SMALL-WORLD MODEL (3/3)

Algorithm 4.1 Small-World Generation Algorithm

Require: Number of nodes $|V|$, mean degree c, parameter β

return A small-world graph $G(V, E)$

G = A regular ring lattice with $|V|$ nodes and degree c

for node v_i (starting from v_1), and all edges $e(v_i, v_j)$, $i < j$ **do**

  v_k = Select a node from V uniformly at random.

  if rewiring $e(v_i, v_j)$ to $e(v_i, v_k)$ does not create loops in the graph or multiple edges between v_i and v_k **then**

    rewire $e(v_i, v_j)$ with probability β: $E = E - \{e(v_i, v_j)\}$, $E = E \cup \{e(v_i, v_k)\}$;

  end if

end for

Return $G(V, E)$

Small-World Generation Algorithm ©SMM

Degree Distribution

- The degree distribution of small-world networks is as follows:
  $P(d_v = d) = \sum_{n=0}^{min(d-c/2,c/2)} \binom{c/2}{n}(1-\beta)^n \beta^{c/2-n} \frac{(\beta c/2)^{d-c/2-n}}{(d-c/2-n)} e^{-\beta c/2}$
  where $P(d_v = d)$ is the probability of observing degree d for node v
- Similar to the Poisson degree distribution observed in random graphs
- In practice, most nodes have similar degrees due to the underlying lattice (contrary to real-world networks whose degree distributions follow a power-law rule)

Clustering Coefficient

- The clustering coefficient for a regular lattice of degree d is $C(d) = \frac{3(d-2)}{4(d-1)}$
- The clustering coefficient for a small-world network varies between the one for a regular lattice and the one for a random graph depending on β: $C(\beta) \approx (1-\beta)^3 C(d)$
- *Proof [sketch]*:
    - A triad of friends is *not* rewired with a probability $(1-\beta)^3$
    - C is then simply the original coefficient from the lattice ($C(d)$) multiplied by that probability
    - Note that we neglect the triads created by the rewirings

AVERAGE PATH LENGTH

- The average path length for a small-world network varies between the one for a regular lattice and the one for a random graph depending on β
- No simple analytical formula exists for the average path length of small-world networks
- However, it can be computed empirically for different values of β

Modeling Real-World Networks with Small-World Graphs

- Given a real-world network, one can simulate it through a small-world graph model
  - Measure the average degree c of the network
  - Derive the corresponding β using the formula for $C(\beta)$
  - Simulate the network using the resulting model

Example of Real VS Simulated Networks [from SMM]

	n	**Real Network**		**Simulated Network**	
Network		l	C	l	C
Film Actors	225K	3.65	0.79	4.2	0.73
Medline co-authors	1.5M	4.6	0.56	5.1	0.52
E. Coli	282	2.9	0.32	4.46	0.31

Table of contents — SL07

Preferential Attachment Model (1/2)

- One of the issues with the small-world model is that the edges are rewired at random, resulting in a fairly constant degree distribution
- Barabási and Albert solved that issue by developing the preferential attachment model
  - When new nodes are added to networks, they are more likely to connect to well-connected nodes
  - Rich-get-richer phenomenon
- Nodes are added one at a time
  - The new node v_i connects to another node v_j with a probability $P(v_j) = \frac{d_j}{\sum_k d_k}$ where d_k is the degree of node k
  - New nodes are limited to max m neighbors

PREFERENTIAL ATTACHMENT MODEL (2/2)

Algorithm 4.2 Preferential Attachment

Require: Graph $G(V_0, E_0)$, where $|V_0| = m_0$ and $d_v \geq 1 \; \forall \; v \in V_0$, number of expected connections $m \leq m_0$, time to run the algorithm t

return A scale-free network
`//Initial graph with` m_0 `nodes with degrees at least 1`
$G(V, E) = G(V_0, E_0)$;
for 1 to t **do**
  $V = V \cup \{v_i\}$; `// add new node` v_i
  while $d_i \neq m$ **do**
    Connect v_i to a random node $v_j \in V$, $i \neq j$ (i.e., $E = E \cup \{e(v_i, v_j)\}$) with probability $P(v_j) = \frac{d_j}{\sum_k d_k}$.
  end while
end for
Return $G(V, E)$

Preferential Attachment Generation Algorithm ©SMM

CLUSTERING COEFFICIENT

- In general, not many triangles are formed by the Barábasi-Albert model, since edges are created independently and one at a time
- $C = \frac{m_0 - 1}{8} \frac{(\ln t)^2}{t}$, where m_0 is the number of nodes initially in the network and t is the number of steps for the growth of the network
- As nodes get added, the clustering coefficient gets smaller and fails to model the high clustering coefficient observed in real-world networks

Average Path Length

- The average path length of the preferential attachment model increases logarithmically with the number of nodes present in the network: $l \sim \frac{\log |V|}{\log \log |V|}$
- Hence, on average, preferential attachment models generate shorter path lengths than random graphs

Modeling Real-World Networks with Preferential Attachment Models

- Given a real-world network, one can simulate it through preferential attachment:
  - Measure the average degree of the network
  - Set m to that value
  - Simulate the network using the resulting model

Network	n	Real Network		Simulated Network	
		l	C	l	C
Film Actors	225K	3.65	0.79	4.9	0.005
Medline co-authors	1.5M	4.6	0.56	5.36	0.0002
E. Coli	282	2.9	0.32	2.37	0.03

- Generates realistic degree distributions and small average path lengths; however, it fails to exhibit the high clustering coefficient observed in real-world networks.

Conclusions (1/2)

- Analyzing large social networks is hard
  - But models can help
  - Still a very active research topic today
- Our world is smaller than one might think
  - ... thanks to the "small world" property exhibited by many real-world networks
  - ... and is still shrinking!

Conclusions (2/2)

- Random graphs exhibit a realistic average path length but also Poisson degree distributions and an unrealistically low clustering coefficient
- Small-world networks exhibit high transitivity and short path lengths (both commonly observed in real-world networks), but have a degree distribution similar to the Poisson degree distribution observed in random graphs
- Preferential attachment graphs follow a power-law degree distribution, exhibit realistic average path lengths but unrealistically low clustering coefficients