CPSC 619-600
Networks and Distributed Processing
Spring 2017

Random Graphs

Dmitri Loguinov
Texas A&M University

March 21, 2017

Agenda

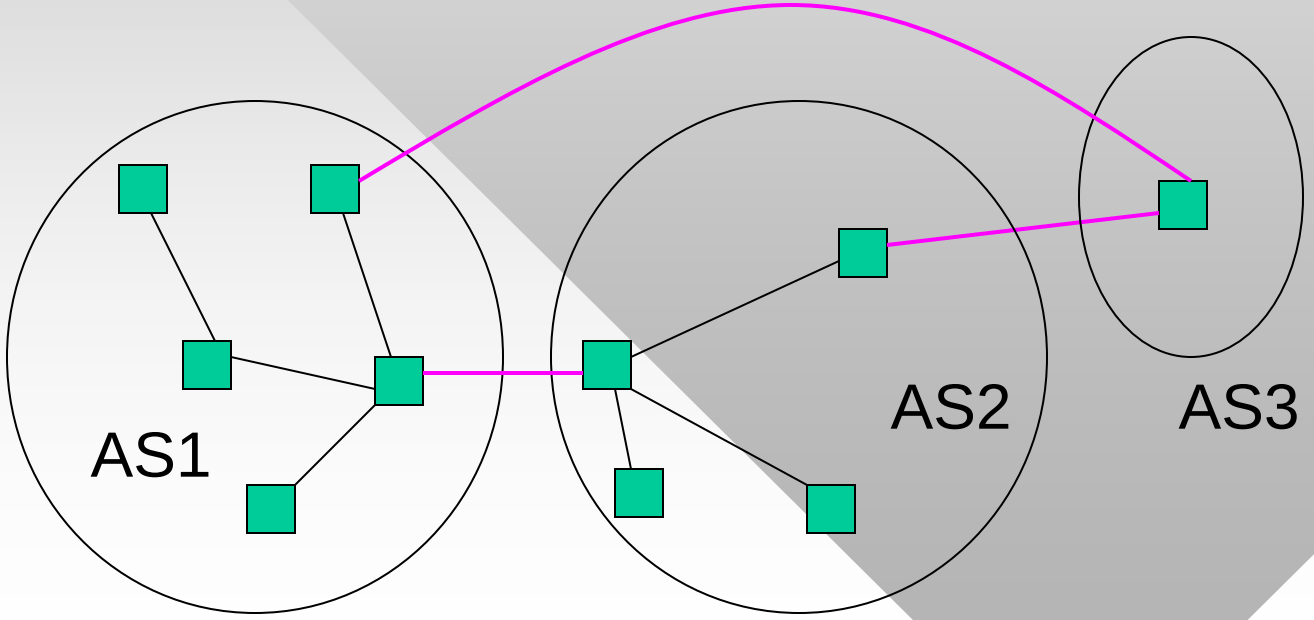
- Topology-modeling goals
- Real-life distributions
- Observed graphs
  - Kevin Bacon graph
  - Collaborator networks
  - Internet AS graph

Introduction

- The first half of this course dealt with microscopic network behavior
  - Packet arrivals, queue size distribution, packet loss, waiting time, etc.
  - We finished with the M/M/1 queue, while the rest of the queues are much more complex in derivations
- We now look at networks from a macroscopic view
  - Topics include topology models, construction of random networks in P2P, and congestion control
  - No concern for how individual packets behave

Introduction 2

- The Internet itself is a large graph
  - Consists of millions of routers connected together
  - Contains autonomous systems (AS), which are ISP-level sub-graphs of the Internet



Introduction 3

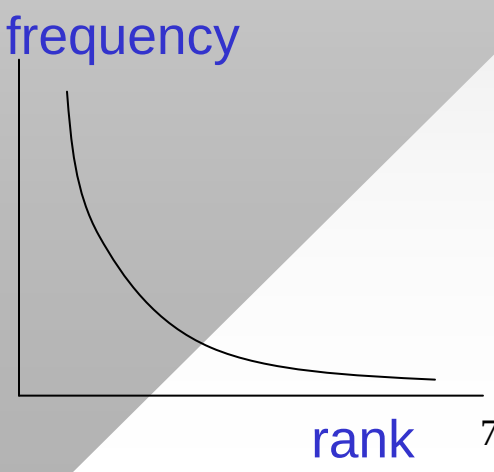
- What would be a good model for the AS graph?
  - We may use such models to create simulated topologies over which we test new protocols
  - We may also be able to better understand how the Internet was formed, its current characteristics, and where its evolution is taking it
- Besides the Internet, random graphs arise in other areas of networking
  - P2P systems constructed dynamically by end users; WWW constructed by hyperlinks; router-level Internet
  - Multicast trees are also random graphs
- Our goal is to understand how to create random graphs and model their properties

Real-Life Distributions

- Many real-life phenomena exhibit distributions that have puzzled researchers for a long time
  - Common math assumptions: Gaussian or exponential
- Example: distribution of word frequencies/popularity in the English language
  - George Zipf (1902-1950), Harvard professor, originally studied this problem and found that word popularity followed what's known today as the Zipf distribution
- Other occurrences of Zipf distributions
  - TCP flow size (number of packets)
  - Web object size (in bytes)
  - File popularity in P2P networks

Real-Life Distributions 2

- The problem addressed by Zipf:
  - Consider frequency of each word
  - Sort all words by their frequency f in descending order (from the most popular to the least)
  - Linearly assign rank r to each word, $r = 1, 2, 3\ldots$
  - Plot frequency as function of rank, $f(r)$, and model the result
- Consider an example based on some article
  - Count word frequencies
  - “the” occurs 10,000 times, rank 1
  - “of” occurs 5,000, rank 2
  - “to” 4,500, rank 3, and so on
  - Now plot frequency vs. rank



Real-Life Distributions 3

- Zipf observed that this curve was a power function with the exponent close to -1
  - In other words, $f(r) \sim r^{-b}$, where b is close to 1
- This shape of the PDF, CCDF, or PMF is usually called heavy-tailed or power-law (more later)
- Zipf studied the population of US cities and also found them to be heavy-tailed
- Vilfredo Pareto applied similar modeling to the income distribution; however, his concern was the probability that the income was above some threshold m
  - In other words, he found that $P(X \geq m) \sim m^{-\alpha}$

Real-Life Distributions 4

- The traditional Pareto distribution used outside of renewal theory is non-shifted:

$$P(X_i \leq x) = 1 - (x/\beta)^{-\alpha}, x \geq \beta$$

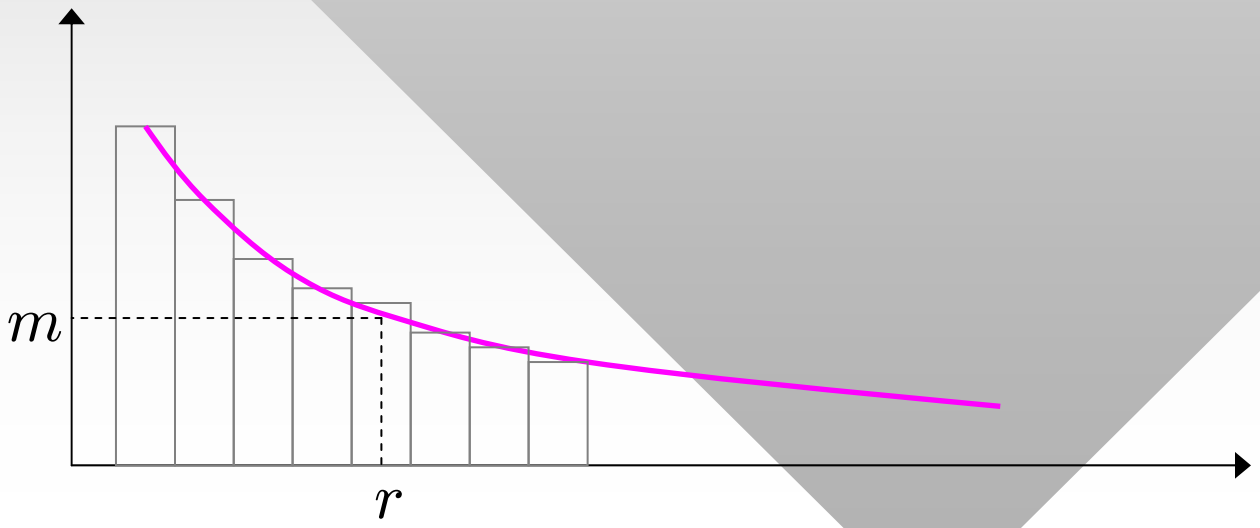
- In renewal theory, we had a slightly different version to allow for arbitrarily small delays X

$$P(X \leq x) = 1 - (1 + x/\beta)^{-\alpha}, x \geq 0$$

- Most papers assume the former version
- Parameter α is called shape and β scale

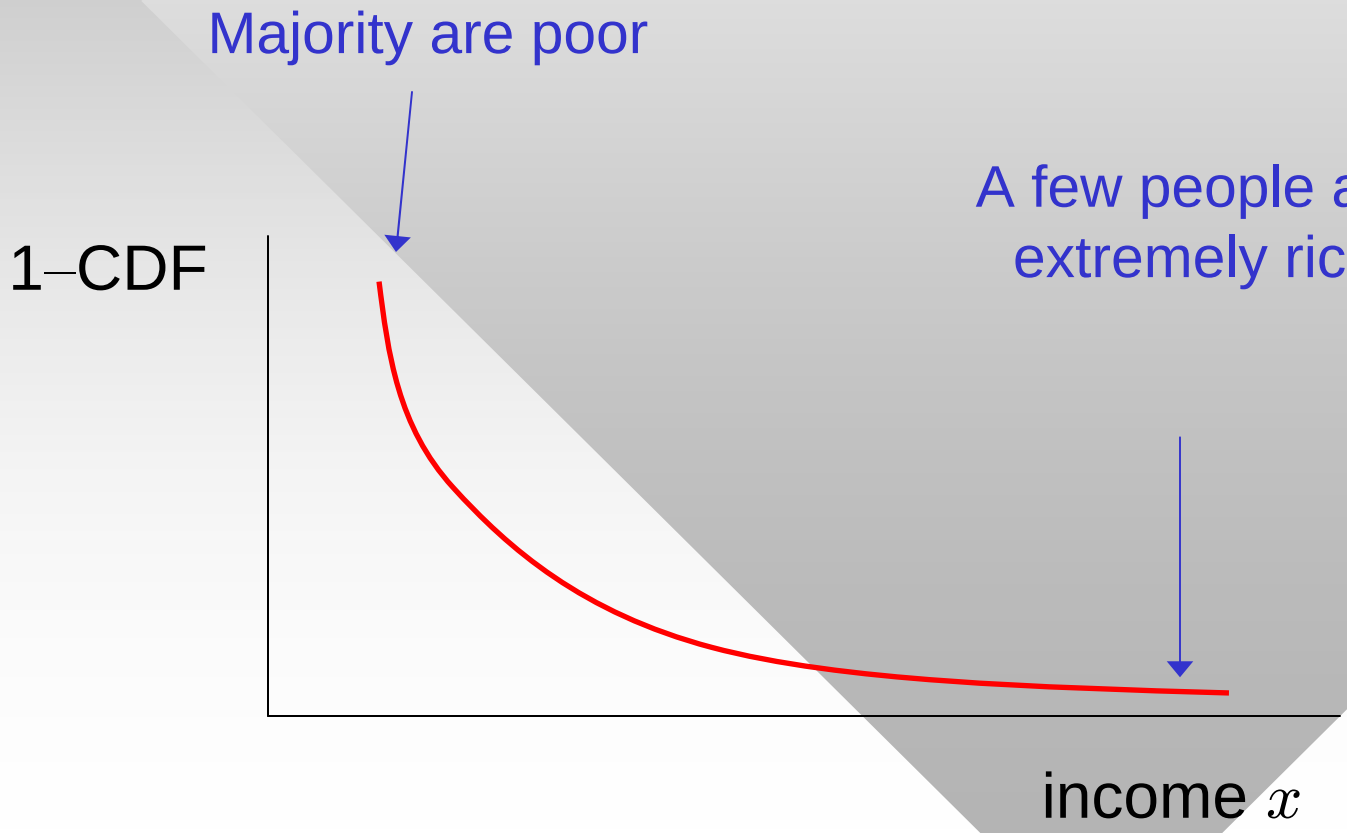
Real-Life Distributions 5

- Theorem: Zipf and Pareto are the same distribution
- Proof: Assume income is Zipf-distributed
  - Now sort people by their income m from highest to lowest; plot function $m(r)$ and observe that $m \sim r^{-b}$ (and $r \sim m^{-1/b}$)
  - Suppose the person with rank r has income m, then we can say that $P(X \geq m) = r \ / \ N$, or $P(X \geq m) \sim m^{-1/b}$, which means it's Pareto with $\alpha = 1/b$ ($N =$ total number of people)



Real-Life Distributions 6

- What does Pareto/Zipf distribution of income mean?
  - A small percentage of people hold a large majority of wealth
  - Extremely large deviation from the mean is possible



Real-Life Distributions 5

- Heavy-tailed distributions are generally characterized by the existence of extremely large values
  - It is not uncommon to observe X_i that is larger than average by a factor of $1{,}000$ or $10{,}000$
  - Most non-heavy-tailed distributions do not exhibit such behavior and the deviation from the mean is small
- Example: what is the probability that an exponential variable exceeds its mean by a factor of 100?

$$P(X > 100E[X]) = e^{-\lambda\frac{100}{\lambda}} = e^{-100} = 3.7\times10^{-44}$$

Real-Life Distributions 6

- Now solve the same for Pareto with $\alpha = 2$
  - Keep in mind that in non-shifted Pareto $E[X] = \alpha\beta/(\alpha–1)$

$$P(X > 100E[X]) = \left(100\frac{\alpha}{\alpha - 1}\right)^{-2} = 2.5\times10^{-5}$$

  - This is 10^{39} times more likely than in the previous model
- Definition: a distribution is called heavy-tailed if its tail decays to zero as a power function (i.e., $x^{-\alpha}$) or slower

$$\exists\alpha > 0 : \frac{x^{-\alpha}}{P(X > x)} = O(1)$$

Real-Life Distributions 7

- For example, exponential distributions are not heavy-tailed:

$$\frac{x^{-\alpha}}{P(X > x)} = x^{-\alpha} e^{\lambda x} \to \infty$$

  - The tail decays faster than any power function $x^{-\alpha}$

- Many distributions in nature are Gaussian (e.g., human weight / height, number of leafs on a tree, fish size within same species, zebra running speed)

- Most socially created metrics are heavy-tailed
  - City population, income, song popularity, the number of in-links per website, word-usage frequency, etc.
  - Why is this the case?

Real-Life Graphs

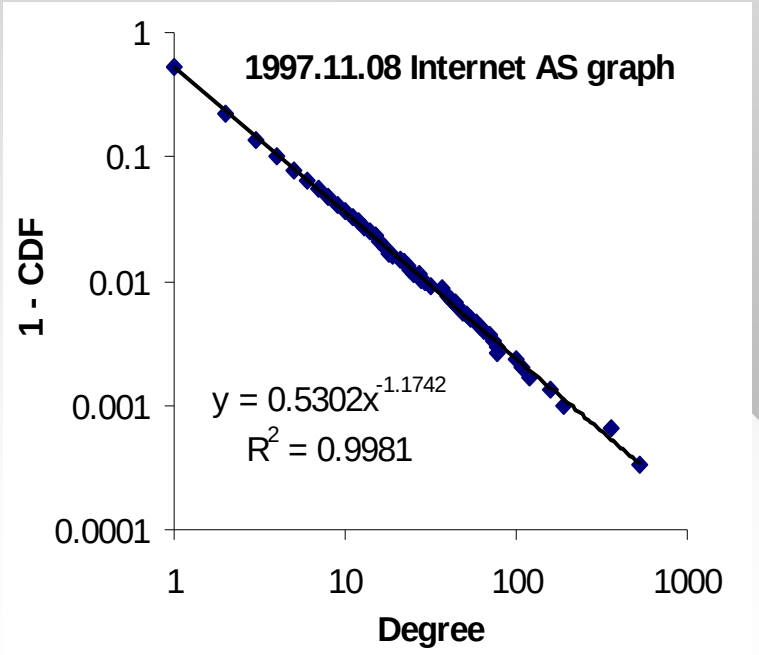
- Property 1: heavy-tailed degree
- Simon in 1955 noticed heavy tails in the distribution of papers published by scientists
  - In other words, if X_i is the number of papers published by a person i, then X_i has a Zipf distribution
  - E.g., 10% of the authors publish 90% of the papers
- Simon later analyzed the co-author graph
  - Think of each person as a node in the graph
  - Each collaboration is a link between two nodes
  - The degree is how many people you collaborated with
  - Turns out this degree is also heavy-tailed

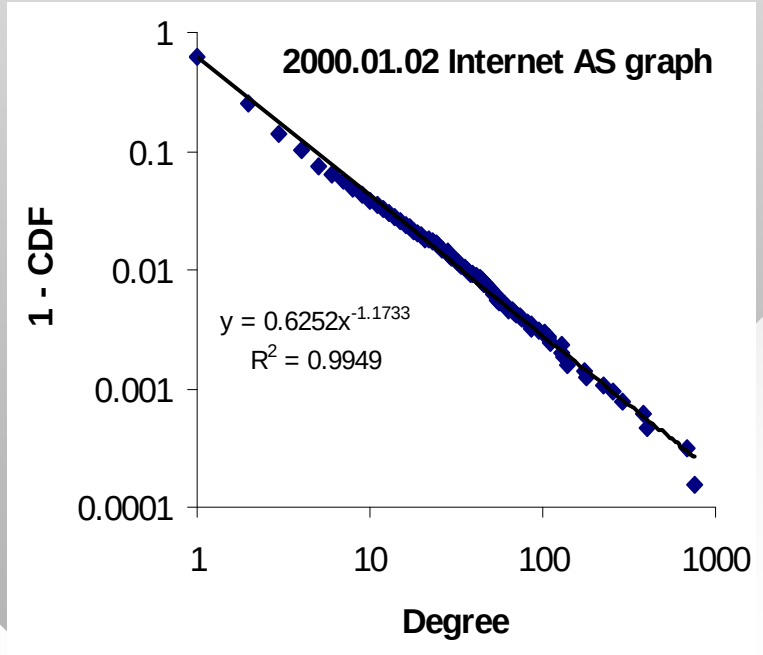
Real-Life Graphs 2

- Faloutsos *et al*. (SIGCOMM 1999)
  - Studied the inter-AS graph derived from several BGP dumps and discovered the power-law degree distribution
  - Used datasets from 11/1997, 4/1998, and 12/1998
- Faloutsos brothers first modeled the degree of each node using a Zipf distribution
  - Found parameter b was close to 0.8
- Then they modeled it using a Pareto distribution
  - Found parameter α was close to 1.2
- In fact, these findings are the same since $b = 1/\alpha$

Real-Life Graphs 3

- Most ISPs have only 1 or 2 peering links to other ISPs
  - At the same time, a handful of ISPs maintain over $1{,}500$ peering connections (80% of connections in 1% of ISPs)





Real-Life Graphs 4

- Property 2: small diameter (or average distance)
- The Kevin Bacon graph
  - Each link is a movie in which two people co-starred
  - The graph starts with Kevin Bacon and contains 355,000+ actors known to the movie database
  - Maximum distance from Bacon is 4 for US actors, 8 internationally (very small diameter)
  - As before, some actors have starred with a ton of other actors, while the majority starred with only a few
- In the 1960s, Milgram performed one of the first experiments to deduce the diameter of the social network in the US

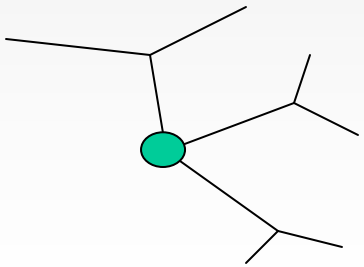
Real-Life Graphs 5

- He asked random people in Nebraska to deliver an envelope to a person in Boston through a chain of acquaintances
  - Each envelope found its destination in no more than 6 hops
  - The phrase “6-degrees of separation” was later used to describe various networks with small diameter
- Average distance of the Internet AS-graph is 4.5 hops
  - This holds regardless of its size (ranging from $3{,}000$ nodes in 1997 to $20{,}000$ in 2006)
  - The graph is self-organizing such that routing is efficient

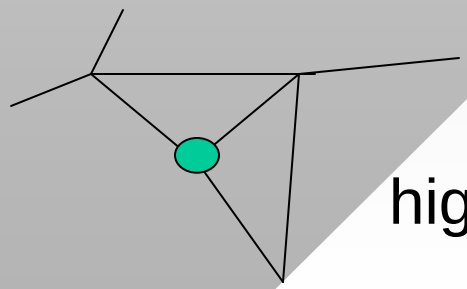
Real-Life Graphs 6

- Property 3: high clustering
  - Determines how well neighbors of a node link to each other
- Assume that Γ_x consists of all edges between the neighbors of x (not including edges to/from x)
- Definition: clustering coefficient γ_x of node x is

$$\gamma_x = \frac{|\Gamma_x|}{d_x(d_x - 1)/2}$$



low γ



high γ

Real-Life Graphs 7

- This is the ratio of the number of actual edges between the neighbors of x to the maximum possible (the graph is assumed to be undirected)
- Definition: clustering of a graph G is:

$$\gamma(G) = E[\gamma_x | d_x \geq 2]$$

- Clustering $\gamma(G)$ is always between 0 and 1
  - Many real graphs demonstrate clustering over 0.3
  - E.g., the Internet stays between 0.35 and 0.45 and the actor graph has $\gamma(G)=0.78$

Real-Life Graphs 8

- Definition: a graph is called small-world if it has properties 2 and 3:
  - Diameter is close to optimal for a given degree
  - Clustering coefficient $\gamma(G)$ stays high as $n \to \infty$
- Typical goal of topology modeling is to achieve all three properties simultaneously and have a connected graph (property 4)
- We perform this in several steps
  - Small diameter and connectivity
  - Heavy-tailed degree
  - Constant clustering