How Do “Real” Networks Look?

Networked Life
NETS 112
Fall 2015
Prof. Michael Kearns
Roadmap

- Next several lectures: “universal” structural properties of networks
- Each large-scale network is unique microscopically, but with appropriate definitions, striking macroscopic commonalities emerge
- Main claim: “typical” large-scale network exhibits:
  - heavy-tailed degree distributions $\rightarrow$ “hubs” or “connectors”
  - existence of giant component: vast majority of vertices in same component
  - small diameter (of giant component) : generalization of the “six degrees of separation”
  - high clustering of connectivity: friends of friends are friends
- For each property:
  - define more precisely; say what “heavy”, “small” and “high” mean
  - look at empirical support for the claims
- First up: heavy-tailed degree distributions
How Do “Real” Networks Look?
I. Heavy-Tailed Degree Distributions
What Do We Mean By Not “Heavy-Tailed”?

- Mathematical model of a typical “bell-shaped” distribution:
  - the Normal or Gaussian distribution over some quantity \( x \)
  - Good for modeling many real-world quantities... but not degree distributions
  - if mean/average is \( \mu \) then probability of value \( x \) is:
    \[
    \text{probability}(x) \propto e^{-(x-\mu)^2}
    \]
  - main point: exponentially fast decay as \( x \) moves away from \( \mu \)
  - if we take the logarithm:
    \[
    \log(\text{probability}(x)) \propto -(x - \mu)^2
    \]
- Claim: if we plot \( \log(x) \) vs \( \log(\text{probability}(x)) \), will get strong curvature
- Let’s look at some (artificial) sample data...
  - (Poisson better than Normal for degrees, but same story holds)
data sampled from a Normal distribution with mean 100.00 and std 1.00

same data on a log-log scale

log(freq(x))

log(x)

data sampled from a Normal distribution with mean 100.00 and std 10.00

same data on a log-log scale

log(freq(x))

log(x)

data sampled from a Normal distribution with mean 100.00 and std 20.00

same data on a log-log scale

log(freq(x))

log(x)
What *Do We Mean By* “Heavy-Tailed”?

- One mathematical model of a typical “heavy-tailed” distribution:
  - the Power Law distribution with exponent $\beta$
  
  \[
  \text{probability}(x) \propto \frac{1}{x^\beta}
  \]

  - main point: inverse polynomial decay as $x$ increases
  - if we take the logarithm:

  \[
  \log(\text{probability}(x)) \propto -\beta \log(x)
  \]

- Claim: if we plot $\log(x)$ vs $\log(\text{probability}(x))$, will get a straight line!
- Let’s look at (artificial) some sample data…
Erdos Number Project Revisited
Figures 1 and 2: In-degree and out-degree distributions subscribe to the power law. The law also holds if only off-site (or "remote-only") edges are considered.

*Degree Distribution of the Web Graph [Broder et al.]*
FIG. 1. The distribution function of connectivities for various large networks. (A) Actor collaboration graph with $N = 212,250$ vertices and average connectivity $\langle k \rangle = 28.78$; (B) World wide web, $N = 325,729$, $\langle k \rangle = 5.46$ (6); (C) Powergrid data, $N = 4,941$, $\langle k \rangle = 2.67$. The dashed lines have slopes (A) $\gamma_{actor} = 2.3$, (B) $\gamma_{ww}$ = 2.1 and (C) $\gamma_{power}$ = 4.

**Actor Collaborations; Web; Power Grid [Barabasi and Albert]**
FIG. 2. Histograms of the number of papers written by scientists in four of the databases. As with Fig. 1, the solid lines are least-squares fits to Eq. (1).

*Scientific Productivity (Newman)*
Zipf’s Law

• Look at the frequency of English words:
  – “the” is the most common, followed by “of”, “to”, etc.
  – claim: frequency of the n-th most common \( \sim \frac{1}{n} \) (power law, \( a \sim 1 \))

• General theme:
  – rank events by their frequency of occurrence
  – resulting distribution often is a power law!

• Other examples:
  – North America city sizes
  – personal income
  – file sizes
  – genus sizes (number of species)
  – the “long tail of search” (on which more later…)
  – let’s look at log-log plots of these

• People seem to dither over exact form of these distributions
  – e.g. value of \( a \)
  – but not over heavy tails
iPhone App Popularity
Summary

• Power law distribution is a good mathematical model for heavy tails; Normal/bell-shaped is not
• Statistical signature of power law and heavy tails: linear on a log-log scale
• Many social and other networks exhibit this signature
• Next “universal”: small diameter
How Do “Real” Networks Look?

II. Small Diameter
What Do We Mean By “Small Diameter”?

• First let’s recall the definition of diameter:
  – assumes network has a single connected component (or examine “giant” component)
  – for every pair of vertices $u$ and $v$, compute shortest-path distance $d(u,v)$
  – then (average-case) diameter of entire network or graph $G$ with $N$ vertices is

$$diameter(G) = \frac{2}{(N(N - 1))} \sum_{u,v} d(u,v)$$

  – equivalent: pick a random pair of vertices $(u,v)$; what do we expect $d(u,v)$ to be?

• What’s the smallest/largest diameter($G$) could be?
  – smallest: 1 (complete network, all $N(N-1)/2$ edges present); independent of $N$
  – largest: linear in $N$ (chain or line network)

• “Small” diameter:
  – no precise definition, but certainly $<< N$
  – Travers and Milgram: $\sim 5$; any fixed network has fixed diameter
  – may want to allow diameter to grow slowly with $N$ (?)
  – e.g. $\log(N)$ or $\log(\log(N))$
Empirical Support

- Travers and Milgram, 1969:
  - diameter ~ 5-6, N ~ 200M

- Columbia Small Worlds, 2003:
  - diameter ~4-7, N ~ web population?

- Lescovec and Horvitz, 2008:
  - Microsoft Messenger network
  - Diameter ~6.5, N ~ 180M

- Backstrom et al., 2012:
  - Facebook social graph
  - diameter ~5, N ~ 721M
Summary

• So far: naturally occurring, large-scale networks exhibit:
  – heavy-tailed degree distributions
  – small diameter

• Next up: clustering of connectivity
How Do “Real” Networks Look?
III. Clustering of Connectivity
The Clustering Coefficient of a Network

• Intuition: a measure of how “bunched up” edges are

• The clustering coefficient of vertex u:
  – let \( k = \text{degree of } u = \text{number of neighbors of } u \)
  – \( k(k-1)/2 = \text{max possible # of edges} \) between neighbors of u
  – \( c(u) = (\text{actual # of edges between neighbors of } u)/(k(k-1)/2) \)
  – fraction of pairs of friends that are also friends
  – \( 0 \leq c(u) \leq 1 \); measure of *cliquishness* of u’s neighborhood

• Clustering coefficient of a graph G:
  – \( CC(G) = \text{average of } c(u) \) over all vertices u in G

\[
\begin{align*}
k &= 4 \\
k(k-1)/2 &= 6 \\
c(u) &= 4/6 = 0.666…
\end{align*}
\]
What Do We Mean By “High” Clustering?

- CC(G) measures how likely vertices with a common neighbor are to be neighbors themselves.
- Should be compared to how likely random pairs of vertices are to be neighbors.
- Let p be the edge density of network/graph G:

\[ p = \frac{E}{N(N - 1)/2} \]

- Here E = total number of edges in G.
- If we picked a pair of vertices at random in G, probability they are connected is exactly p.
- So we will say clustering is high if CC(G) >> p.
Clustering Coefficient Example 1

$\frac{1}{2 \times \frac{1}{2}} = 1$

$\frac{3}{4 \times \frac{3}{2}} = \frac{1}{2}$

$\frac{2}{3 \times \frac{2}{2}} = \frac{2}{3}$

$\frac{2}{3 \times \frac{2}{2}} = \frac{2}{3}$

$\frac{1}{2 \times \frac{1}{2}} = 1$

C.C. = $(1 + \frac{1}{2} + 1 + \frac{2}{3} + \frac{2}{3})/5 = 0.7666…$

$p = \frac{7}{(5 \times 4/2)} = 0.7$

Not highly clustered
Clustering Coefficient Example 2

- Network: simple cycle + edges to vertices 2 hops away on cycle
- By symmetry, all vertices have the same clustering coefficient
- Clustering coefficient of a vertex v:
  - Degree of v is 4, so the number of possible edges between pairs of neighbors of v is $4 \times \frac{3}{2} = 6$
  - How many pairs of v’s neighbors actually are connected? 3 --- the two clockwise neighbors, the two counterclockwise, and the immediate cycle neighbors
  - So the c.c. of v is $\frac{3}{6} = \frac{1}{2}$
- Compare to overall edge density:
  - Total number of edges = $2N$
  - Edge density $p = \frac{2N}{N(N-1)/2} \sim \frac{4}{N}$
  - As N becomes large, $\frac{1}{2} \gg \frac{4}{N}$
  - So this cyclical network is highly clustered
Divide $N$ vertices into $\sqrt{N}$ groups of size $\sqrt{N}$ (here $N = 25$)
Add all connections within each group (cliques), connect “leaders” in a cycle
$N - \sqrt{N}$ non-leaders have C.C. = 1, so network C.C. $\rightarrow 1$ as $N$ becomes large
Edge density is $p \sim 1/\sqrt{N}$
<table>
<thead>
<tr>
<th></th>
<th>$L_{\text{actual}}$</th>
<th>$L_{\text{random}}$</th>
<th>$C_{\text{actual}}$</th>
<th>$C_{\text{random}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOVIE ACTORS</td>
<td>3.65</td>
<td>2.99</td>
<td>0.79</td>
<td>0.00027</td>
</tr>
<tr>
<td>POWER GRID</td>
<td>18.7</td>
<td>12.4</td>
<td>0.080</td>
<td>0.005</td>
</tr>
<tr>
<td>C. ELEGANS</td>
<td>2.65</td>
<td>2.25</td>
<td>0.28</td>
<td>0.05</td>
</tr>
</tbody>
</table>

$L$ = Path Length; $C$ = Clustering Coefficient.