Introduction to Network Analysis

Small-world Phenomena

- Name, address, occupation of the target were known; no sending was allowed
- 18 packages returned back to Boston
- Mean path result was just 5.9 steps
- Small-world effect was confirmed in many other experiments

Bonus observations in the experiment
- Most of the packages were received through 3 target’s friends
- People are good in finding short paths (later was shown that it is hard to find shortest path without knowing full information)

Similar experiments
- Emails: only 384 out of 24K were received/results confirmed, 4 steps
- Microsoft .NET Messenger Service: 6.6 people

Stanley Milgram
1933-1984
Degree Distributions

The probability that randomly chosen node has degree $k$

Classical undirected random graph models $G_{n,p}$

When graphs are small:

$$\binom{n-1}{k} p^k (1-p)^{n-1-k} \quad \text{(Binomial distribution)}$$

When graphs are large:

$$\frac{(np)^k}{k!} e^{-np} \quad \text{(Poisson distribution)}$$
Degree Distributions

$p_0 = \frac{1}{9}, \ p_1 = \frac{2}{9}, \ p_2 = \frac{1}{9}, \ p_3 = \frac{2}{9}, \ p_4 = \frac{3}{9}$

The probability that randomly chosen node has degree $k$

The tail is much longer

$>2K$

Internet at the level of autonomous systems

World Wide Web

From Newman “Networks, an Introduction”
Power Laws (aka scale-free)

Internet at the level of autonomous systems

logarithmic scales; bigger range of bins

\[ \ln p_k = -\alpha \ln k + c \text{ or } p_k = Ck^{-\alpha}, \text{ where } C = e^c \]

typical \( \alpha \in [2, 3] \) (see handout)

Problem of histograms: statistics is poor at the tail of the distribution

Solution I: different sizes of bins
Cumulative Distribution

Probability at a random vertex has degree \( k \) or greater
\[
P_k = \sum_{k'=k}^{\infty} p_{k'}
\]

Let \( p_k \) follows a power law in its tail, i.e.,
\[
p_k = C k^{-\alpha} \quad \text{for} \quad k \geq k_{\text{min}}.
\]

Then
\[
P_k = C \sum_{k'=k}^{\infty} k'^{-\alpha} \\
\approx C \int_{k}^{\infty} k'^{-\alpha} \, dk' = \frac{C}{\alpha - 1} k^{-(\alpha-1)}
\]

\[
\alpha = 1 + N \left( \sum_i \ln \frac{d(i)}{k_{\text{min}} - 1/2} \right)^{-1}
\]

Advantages:
- no bins
- easy calculation
- can be plotted as normal function at log-log scale
- binning loses the information; cumulative distribution preserves everything

Disadvantages
- less easy to interpret than histograms
- successive points are correlated

Figure 8.7: Cumulative distribution function for the degrees of vertices on the Internet. For a distribution with a power-law tail, as is approximately the case for the degree distribution of the Internet, the cumulative distribution function, Eq. (8.4), also follows a power law, but with a slope 1 less than that of the original distribution.

From Newman “Networks, an Introduction”
Cumulative Distribution

Figure 8.8: Cumulative distribution functions for in- and out-degrees in directed networks. (a) The in-degree distribution of the World Wide Web, from the data of Broder et al. [56]. (b) The out-degree distribution for the same Web data set. (c) The in-degree distribution of a citation network, from the data of Redner [280]. The distributions follow approximate power-law forms in each case.

From Newman “Networks, an Introduction”
Power Laws

More examples: city populations, moon craters, solar flares, computer files, words frequencies in human languages, hits on web pages, publications per scientist, book sales, ...

**Normalization:** we have to find $C$ such that $\sum_{k=0}^{\infty} p_k = 1$

After eliminating $k = 0$

$$C = \frac{1}{\sum_{k=1}^{\infty} k^{-\alpha}} = \frac{1}{\zeta(\alpha)}, \text{ i.e., } p_k = \frac{k^{-\alpha}}{\zeta(\alpha)}, \text{ where } p_0 = 0$$

However, pure power-law behavior is not perfect for real-world networks

**Normalization over the tail:**

$$p_k = \frac{k^\alpha}{\sum_{k=k_{\min}}^{\infty} k^{-\alpha}} = \frac{k^{-\alpha}}{\zeta(\alpha, k_{\min})}$$

or if we approximate it then $C \approx 1/ \left( \int_{k_{\min}}^{\infty} k^{-\alpha} \, dk \right) = (\alpha - 1) k_{\min}^{\alpha-1}$
Moments: The $m$th moment of the distribution is defined as

$$\langle k^m \rangle = \sum_{k=0}^{\infty} k^m p_k = \sum_{k=0}^{k_{\min}-1} k^m p_k + C \sum_{k=k_{\min}}^{\infty} k^m k^{-\alpha}$$

if power law begins with some $k_{\min}$

$m$th moment exists (finite) when $\alpha > m + 1$ (integrate the second term)

Remark: This estimate works for arbitrarily large network with the same power law distribution. For finite network $\langle k^m \rangle = \frac{1}{n} \sum_{i \in V} d(i)^m$

**Top-heavy distributions or 80/20 rule:** how many edges are connected to the highest degree vertices?

$$\int_{x_{1/2}}^{\infty} p(x) \, dx = \frac{1}{2} \int_{x_{\text{min}}}^{\infty} p(x) \, dx,$$

Point that divides distribution in two halves

$$x_{1/2} = 2^{1/(\alpha-1)} x_{\text{min}}.$$

Further reading: Newman “Power laws, Pareto distributions and Zipf’s law”
Introduction to Network Analysis

Cumulative distributions for Internet nodes

Noncumulative histogram for Internet nodes

Power law is not observed