Lecture 6: Scale (1/2)

Why are large scale systems skewed? (i.e., Why makes popular Youtube Videos)?

COMS 4995-1: Introduction to Social Networks
Thursday, September 22nd
Why do I have an iPhone with my cell provider?
- It’s technically better?
- It’s cool?
- Because of my social network?

How many $ a cellphone provider spends on retaining users?
Outline

* The ubiquitous power-law

* The three sources of power law
  – reinforcement
  – optimization
  – artefacts of random stopping

* How to handle the long tail?
What is a power-law?

* Distribution: different values of a variable metric
  - Among individuals (size, income), webpage (#links) ...

* Large values
  - $P[X>x]$ is decreasing to 0 as $x$ increases, but how fast?
  - Light-tailed: $\exists \lambda, \text{cst}$, for any $x$, $P[X>x] / \exp(-\lambda x) < \text{cst}$
  - Heavy-tailed: $\forall \lambda$, $\lim_{x \to \infty} P[X>x] / \exp(-\lambda x) = \infty$

* Power-law is a special case of heavy tailed
  - Polynomial decrease: $P[X>x] \sim c \cdot x^{-\alpha}$ for large $x$
What makes power-law special?

* #1: very large values are rare but not impossible
  – Ex.: α=2, c=1, P[X>1000] = 1/1,000,000
* #2: a few large values have a big impact
  – Unfairness: the 20% richest own 80% of wealth
    … and the 5% richest own 75% of wealth
  – Variance, mean can be infinite
* #3: invariance by {conditioning + mult. rescaling}
  – Exponential: P[X>x+1|X>x] = e^{-λ} for any x
  – Power law: P[X>2x|X>x]=2^{-α} for any x
What makes this graph different?

Power laws, Pareto distributions and Zipf’s law.
Power law in comparison
How to spot a power law?

- Plot CCDF or density using log. x and log. y axis
  - Is the shape linear? For how many magnitude orders?
  - What is the slope?
- In contrast, light tailed or lognormal decreases fast
- Care must be taken to identify a power-law
  - Density is sensitive to binning, CCDF preferred
  - Validation: goodness of fit test, alternative
  - Coefficient Maximum likelyhood

\[ \hat{\alpha} = 1 + n \left[ \sum_{i=1}^{n} \ln \frac{x_i}{x_{\text{min}}} \right]^{-1} \]

Power-law distributions in empirical data.
The ubiquitous power law

Power laws, Pareto distributions and Zipf’s law.
Non-power law exist too!

- Length of a relationship
- # birds in species
- # of email addresses kept
- Size of forest fire

Power laws, Pareto distributions and Zipf’s law.
Power law in Internet

On power-law relationships of the internet topology
Faloutsos, Faloutsos & Faloutsos. SIGCOMM (1999)

Graph structure in the web.
Cha, M., et. al. (2009). Analyzing the video popularity characteristics of large-scale user generated content systems.
What about other social networks

Understanding individual human mobility patterns.

The scaling laws of human travel.
Power law in time statistics

The origin of bursts and heavy tails in human dynamics.
A. Barabasi, Nature (2005)
The ubiquitous power law

- Popularity of items: Amazon, YouTube, Flickr
- Degree of nodes in large graphs
  - links to website, connection between Internet Ases
  - Collaboration between actors, scientists
- Across a broad spectrum of natural system
  - # Species in genus, proteine seq. genome, words
- Time and space: characterizes human activities
Outline

* The ubiquitous power-law

* The three sources of power law
  – reinforcement
  – optimization
  – artefacts of random stopping

* How to handle the long tail?
Power law by reinforcement

Species inside genus
1. new species appear?
   Mutation (random)
2. new genus appear?
   Large mutation

Key: More species implies more mutations

A mathematical theory of evolution …
G. Yule, Phil. Trans. Roy. Soc. London (1925)
The Yule process

* N balls (i.e. species) into bins (i.e. genus)
  - #balls in a bin = #species in this genus

* At each iteration t→t+1:
  - Mutation occurs in 1 species uniformly chosen
  - With probability p, this creates a new genus
  - Otherwise, this creates a new species in the genus

* Let $X_i(t) = \# \{\text{genus containing exactly } i \text{ species}\}$
Example of evolution
Example of evolution

\[ t=1 \]

\[ \text{Example of evolution} \]

\[ t=2 \]

\[ \text{Example of evolution} \]

\[ t=3 \]

\[ \text{Example of evolution} \]
Example of evolution (2)
Thm: There exists $c_1, c_2, \ldots$ such that a.s. $X_i(t)/t \to c_i$

- we have $c_1 = p/(2-p)$ and $c_i = c_{i-1}(1-\alpha/i + O(i^{-2}))$
  where $\alpha = (2-p)/(1-p)$
- this implies $c_i \propto i^{-\alpha}$, which explains power law

Proof follows two ingredients:

1. Analysis of the expected value evolution
2. A probabilistic “concentration” result and its consequence
\* Let $N(t)$ be the number of species: $N(t) = N_0 + t$

\* Let us consider number of genus with 1 species

$$X_1(t + 1) = \begin{cases} 
X_1(t) + 1 & \text{with probability } p \\
X_1(t) - 1 & \text{with probability } (1 - p) \frac{X_1(t)}{N(t)} \\
X_1(t) & \text{otherwise}
\end{cases}$$

$$E[X_1(t + 1)] = E[X_1(t)] + p - \frac{E[X_1(t)]}{N(t)} \times (1 - p)$$
We have \[ E[X_1(t + 1)] = E[X_1(t)] + p - \frac{E[X_1(t)]}{N(t)} * (1 - p) \]

- This implies that \( X_1(t)/t \to C_1 = p/(2-p) \) as \( t \) goes large.
A similar analysis of recurrence

Leads to evolution of expectation: