Lecture 5: Structure (2/3)

"Why do super nodes emerge? How does Google find and rank them?"

COMS 6998-2: Social Networks
Monday, October 6th
Outline

* The ubiquitous power-law
* Generative models of power law
* The node ranking problem
Distribution: different values of a variable metric
- Among individuals (size, income), webpage (#links) ...
- Typically interesting because it varies

How to represent a distribution?

CDF (P[V<x] as function x)  PDF (P[V=x] as f. of x)
Empirical distribution

* How they look in reality?
  - Instead of $P[V=x]$ looks at frequency of observations
Lorenz curve

Distribution of wealth
- x-axis: fraction of total income
- Y-axis: fraction of population

What is a power-law?

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

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

* Power-law is a special case of heavy tailed
  - Polynomial decrease: \( P[X>x] \sim c \, x^{-\alpha} \) for large \( x \)
#1: very large values are rare but not impossible
- Ex.: $\alpha=2$, $c=1$, $P[X>1000] = 1/1,000,000$
#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

  - Mathematically speaking:
    * Variance can be infinite
    * Sometimes even the mean is infinite!
#3: Small values can have a big impact collectively
- Long tail effect
What makes power-law special?

* #4: invariance by \{conditioning + mult. rescaling\}

Exponential: \( P[X > x+1 | X > x] = e^{-\lambda} \) for any \( x \)

“memory less: if you have already waited for one hour, then the remaining waiting time is the same!”

Power law: \( P[X > 2x | X > x] = 2^{-\alpha} \) for any \( x \)

“Catastrophe: if you have already waited for two hours, then you will wait twice more than after one hour!”
What makes this graph different?

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

- Plot CCDF \( P[V>x] \) 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 likelihood

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

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


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
* The ubiquitous power-law
* Generative models of power law
* The node ranking problem
Power law by reinforcement

Species inside genus
1. new species appear?  
Mutations (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 $X_i(t) = \# \{\text{genus containing exactly } i \text{ species}\}$
Analysis of the Yule process

* 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 $c_i = c_{i-1} \left(1 - \frac{\alpha}{i} + O(i^{-2})\right)$
    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
Consequences on Graph

We observe that in-degree of webpage is distributed as a power law.

Can this be the result of a reinforcement process?

Graph structure in the web.
Nodes (i.e. webpages) join the graph in sequence, creating edges linking to previous nodes.

Each node $u$ creates $k$ outgoing edges as follows.

1. Choose a node $v$ uniformly
2. With prob. $p$, edge $u \rightarrow v$ is created,
3. With prob. $(1-p)$, edge $u \rightarrow w$ for $w$ a children of $v$,

Similarly (see assignment):

- Fraction of nodes with $i$ incoming links $\sim i^{\alpha}$
- We have again $\alpha = (2-p)/(1-p)$
Some consequences

- A model for graph with skewed degree distribution
  - Different from “semi-edge+random assignment”
- Assuming we remove a fraction 0<f<1 of nodes?
  - If done randomly, a fraction 0<c<1-f still connected
  - If removed in order of decreasing degree ("hubs")
    for some f<1 no positive fraction remains connected
- Thm: Diameter of the graph is O(log(N)).

The diameter of a scale-free random graph.
B. Bollobas, O. Riordan, Combinatorica. (2004)

Practice: Make your own power law

What would it take to modify today’s apple policy
... and obtain a power law?
1. A lot of apple
2. With fixed probability
   – Every m apple, send it to x who already received one
   Ask x to send it to one who received the apple by her
A generalization of our previous construction
- States that the increase/decrease of an entity should be made proportional to its size.

Discrete version:
- Let $X(t)$ be a variable (e.g. your income) discretized $Y(t) = j$ if and only if your income $X(t)$ is in $[m \cdot e^j; m \cdot e^{j+1}]$
- proportional evolution: $P[Y(t+1) = j \mid Y(t) = i] = f(j-i)$
- E.g.: $f(1) = p; f(-1) = 1-p$, then $P[X(\infty) > x] = \frac{x}{m} \ln\left(\frac{p}{1-p}\right)$

A generalization of our previous construction
- States that the increase/decrease of an entity should be made proportional to its size.

Continuous version:
- Let $X(t)$ be a variable (e.g. your income)
- proportional evolution: $X(t+1) = F(t) X(t)$
- For large $t$, $X(t)$ approaches lognormal distribution (i.e. $\log(X(t)) \sim$ normal dist; $X(t)$ heavy tailed, not PL)

**Skewed distribution**
- Can be explained by reinforcement, propor. effect.
- This also applies to degree distribution in graph, that also exhibit small-world and “hub” structure

**Be careful**
- Lognormal and power law may both be obtained
- Graphs obtained do not exhibit clustering coefficient
- There are other sources explaining power-laws!
Other explanation Internet AS graph

Internet routers degrees follow power law.

How was this measured?

Traceroute: pick a source compute shortest path to many destinations.

Reality test: Internet AS graph

What about running Traceroute on uniform random graphs?

After sampling degree resembles power law!

Traceroute does not discover all edges
- As it needs to be on the shortest paths from a destination to the source
- Moreover, this reinforces the presence of edges near the source

Thm: For general degree dist. random graphs,
- the degree distribution changes radically
- It may exhibit heavy tail/power law artificially

On the bias of traceroute sampling.
D Achlioptas, A Clauset, D Kempe, C Moore. J. ACM (2009)
Frequency of items, popularity of objects and distributions of many variables are highly skewed

- This proves that it cannot be the sum of independent effect (Central limit theory would predict Normal).
- It could be reinforcement, or the result of a bias (typically random stopping),
- In some cases (caching, topology) it can relate to optimal efficiency