CS 5614: (Big) Data Management Systems

B. Aditya Prakash

Lecture #20: Graph Mining 1
Networks are everywhere!

Facebook Network [2010]

Gene Regulatory Network [Decourty 2008]

Human Disease Network [Barabasi 2007]

The Internet [2005]
What else do they have in common?
High School Dating Network


Blue: Male
Pink: Female

Interesting observations?
The Internet

Skewed Degrees
Robustness
Karate Club Network
WARM-UP AND BASICS
A Question

- How many of you think your friends have more friends than you? 😊

- A recent Facebook study
  - Examined all of FB’s users: 721 million people with 69 billion friendships.
    - about 10 percent of the world’s population!
  - Found that user’s friend count was less than the average friend count of his or her friends, 93 percent of the time.
  - Users had an average of 190 friends, while their friends averaged 635 friends of their own.
Possible Reasons?

- You are a loner?
- Your friends are extroverts?
- There are more extroverts than introverts in the world?
Example

Average number of friends?

Source: S. Strogatz, NYT 2012
Example

Average number of friends
= ( 1 + 3 + 2 + 2 ) / 4
= 2

Source: S. Strogatz, NYT 2012
Example

Average number of friends
\[= \frac{1 + 3 + 2 + 2}{4}\]
\[= 2\]

Average number of friends of friends

Source: S. Strogatz, NYT 2012
Average number of friends
= \frac{(1 + 3 + 2 + 2)}{4}
= 2

Average number of friends of friends
= \frac{(3 + 1 + 2 + 2 + 3 + 2 + 3 + 2)}{8}
= \frac{(1 \times 1) + (3 \times 3) + (2 \times 2) + (2 \times 2)}{8}

Source: S. Strogatz, NYT 2012
Example

Average number of friends
\[ = \frac{1 + 3 + 2 + 2}{4} \]
\[ = 2 \]

Average number of friends of friends
\[ = \frac{3 + 1 + 2 + 2 + 3 + 2 + 3 + 2}{8} \]
\[ = \frac{(1 \times 1) + (3 \times 3) + (2 \times 2) + (2 \times 2)}{8} \]
\[ = 2.25! \]

Source: S. Strogatz, NYT 2012
Actually it is (almost) always true!

- Proof?
Actually it is (almost) always true!

- Proof?

\[ E[X] = \frac{\sum x_i}{N} \]
Actually it is (almost) always true!

- Proof?

\[
E[X] = \frac{\sum x_i}{N} \\
Var[X] = E[(X - E[X])^2] \\
= E[X^2] - E[X]^2
\]
Actually it is (almost) always true!

- Proof?

\[
E[X] = \sum x_i / N
\]

\[
Var[X] = E[(X - E[X])^2]
\]

\[
= E[X^2] - E[X]^2
\]

\[
\frac{E[X^2]}{E[X]} = E[X] + \frac{Var[X]}{E[X]}
\]
Actually it is (almost) always true!

Proof?

\[ E[X] = \sum x_i / N \]

\[ \text{Var}[X] = E[(X - E[X])^2] \]

\[ = E[X^2] - E[X]^2 \]

\[ \frac{E[X^2]}{E[X]} = E[X] + \frac{\text{Var}[X]}{E[X]} \]

Essentially, it is true if there is any spread in # of friends (non-zero variance)!
Implications

- **Immunization**
  - Acquaintance immunization
    - Immunize friend-of-friend
- **Early warning of outbreaks**
  - Again, monitor friends of friends

---

**Figure 1.** Network Illustrating Structural Parameters.

This real network of 105 students shows variation in structural attributes and topological position. Each circle represents a person and each line represents a friendship tie. Nodes A and B have different "degree," a measure that indicates the number of ties. Nodes with higher degree also tend to exhibit higher "centrality" (node A with six friends is more central than B and C who both only have four friends). If contagions infect people at random at the beginning of an epidemic, central individuals are likely to be infected sooner because they lie a shorter number of steps (on average) from all other individuals in the network. Finally, although nodes B and C have the same degree, they differ in "transitivity" (the probability that any two of one's friends are friends with each other). Node B exhibits high transitivity with many friends that know one another. In contrast, node C's friends are not connected to one another and therefore they offer more independent possibilities for becoming infected earlier in the epidemic.

---

**Figure 2.** Theoretical expectations of differences in contagion between central individuals and the population as a whole.

A contagious process passes through two phases, one in which the number of infected individuals exponentially increases as the contagion spreads, and one in which incidence exponentially decreases as susceptible individuals become increasingly scarce. These dynamics can be modeled by a logistic function. Central individuals lie on more paths in a network compared to the average person in a population and are therefore more likely to be infected early by a contagion that randomly infects some individuals and then spreads from person to person within the network. This shifts the S-shaped logistic cumulative incidence function forward in time for central individuals compared to the population as a whole (left panel). It also shifts the peak infection rate forward (right panel).

---

**Social Network Sensors**

Prakash 2017
Network structure is important!!

- A network is a collections of nodes with relations between some nodes

**Object:** nodes, vertices $N$

**Relations:** links, edges $E$

**System:** graphs, networks $G(N, E)$
Networks and Graphs

- **Networks**: typically a real system
  - Metabolic Network, Social Network, etc.
- **Graphs**: typically the mathematical representation
  - Web graph, Planar graphs etc.

But we use it interchangeably
Networks: Which representation?

- Connect people who work together: professional network
- Connect authors and papers: co-authorship network
- Connect all people whose name is John Smith?
Networks: Which representation?

- Choice is important
  - In some cases there is a unique unambiguous representation
  - In most others, YOU have to choose
    - Depends on what you want to ask
Undirected vs Directed Graphs

- Undirected
  - Links are symmetrical
  
- Examples
  - Friendships (on FB!)
  - Collaborators
Undirected vs Directed Graphs

- Directed
  - Links are directed
  - Examples
    - Following on Twitter
    - Phone calls
Graph connectivity

- Connected (undirected) graphs
  - There is a path between any two vertices

![Graph examples]

- Bridge edge: If we erase it, the graph becomes disconnected.
- Articulation point: If we erase it, the graph becomes disconnected.
- Largest Component: Giant Component
- Isolated node (node F)

![Multiple components]

Prakash 2017
Graph connectivity

- Connected (undirected) graphs
  - There is a path between any two vertices

Largest component:
- GIANT component

Bridge edge:
- If we erase it, the graph becomes disconnected.

Articulation point:
- If we erase it, the graph becomes disconnected.

D
C
A
B
H
F
G

Largest component: GIANT component

D
C
A
B
H
F
G
Graph connectivity

- Extension to directed graphs
  - **Strongly** connected: has a path from every node to every other node
  - **Weakly** connected: is connected if edge directions are disregarded

![Graph](image)

Weakly connected, but not strongly connected

Why?
Next Class

- The Web as a graph
- More network properties
- The $G(n,p)$ model
Networks with attributes

- Edges and Nodes can have attributes
- **Weighted graphs**
  - Numerical Attribute on edges
  - Can be negative! (trust vs distrust, transcription regulatory networks)
  - E.g. frequency of communication on a phone call graph
- Other types of attributes
  - Ranking (bff, second bff, ...)
  - Type (relative, co-worker, ...)
  - More aggregate global properties (centralities etc.)
Adjacency Matrices

- Representing edges (who is adjacent to whom) as a matrix \(\rightarrow [\text{matrix algebra!}]\)

\[
A_{ij} = 1 \text{ if node } i \text{ has an edge to node } j \\
= 0 \text{ otherwise}
\]

A is symmetric for an undirected graph
Other ways

- **Incidence Matrix**
  - $b_{ij} = 1$ if vertex $i$ and edge $j$ are incident

- **Adjacency lists**
  - “edge lists”
  - \{(2,3), (2,5) \ldots\}
  - More efficient if network is sparse and large
Bipartite Networks

- Two-mode networks
  - Edges occur only across groups
  - The red and blue nodes are ‘independent sets’
  - E.g. people-to-places they visit

- Can be folded into ‘one-mode’ networks
  - People who visit the same places
Networks

- DBLP: bi-partite network author-papers folded author-author collaboration ....
- Twitter follower-followee: directed weighted
- Facebook friendship: undirected, unweighted
- Mobile phone calls: directed, weighted
- Protein-protein interactions: undirected, unweighted with self-interactions
NETWORK FEATURES
How to characterize a graph?

- Think of them as network ‘features’
Degree Distribution

- Degree $k_i = \text{the number (or the total weight)}$ of edges incident to node $i$

- For directed networks: in-degree and out-degree

- More terms: source node has in-degree 0, sink node has out-degree 0
Degree Distribution $P[k]$

- $P[k] = \text{probability that a randomly chosen node has degree } k$
- Empirically,

\[ P[k] = \frac{N_k}{N} \]

($N_k = \text{the total number of degree } k \text{ nodes}$)
A path is a sequence of nodes $n_i$ where each pair $n_i, n_{i+1}$ have an edge.

IMPORTANT: a path can intersect itself, pass through same edge multiple times.

- E.g. AB is a path
- ABCEDA is a path
- ABA is NOT a path
Matrix Method: Number of paths between u and v

- **Length 1:** if $A_{uv} = 1$ then 1, otherwise 0
- **Length 2:** if $A_{uk} A_{kv} = 1$, then there is a path from u to v via k. so $P^{(2)}_{uv} = \sum A_{uk} A_{kv} = [A^2]_{uv}$
- ..... 
- ..... 
- **Length l:** $P^{(l)}_{uv} = [A^l]_{uv}$
Distance in a graph

- **Distance** \((u, v)\) = \# nodes on the **shortest path** connecting \(u\) and \(v\)
  - if no path between \(u\) and \(v\), then \(dist(u,v) = \infty\)

Dist(B, C) = 1
Dist(C, B) = 3

Note: in directed graphs, \(dist(.)\) is NOT symmetric
How far are the nodes?

- **Diameter**: Longest *shortest* path between any \( u \) and \( v \)

\[
d = \max_{u,v \neq u} \text{dist}(u,v)
\]

- **Average path length**: (typically done for connected components)

\[
\bar{d} = \frac{1}{\# \text{pairs}_{u,v \neq u}} \sum_{u,v \neq u} \text{dist}(u,v)
\]

\[
\# \text{pairs} = \binom{N}{2} = \frac{N(N-1)}{2}
\]
Finding shortest paths

- **Classic** Computer Science problem
- Breadth-First Search
- Dijkstra’s algorithm
- Bellman-ford algorithm
- Floyd-Warshall algorithm
- ..........
RANDOM GRAPH MODELS
Q: When should there be an edge between A and B?
The Simplest (non-trivial) Graph Model

Q: When should there be an edge between A and B?
A: With probability $p$
The Simplest (non-trivial) Graph Model

- What about other edges?

Do the same!

\textit{Independently}

The Erdos-Renyi Random graph model
The E-R Model [1960]

- **Two variants**
  - $\mathbf{G}(n,p)$: each undirected edge $(u, v)$ between $n$ nodes occurs with probability $p$ (independently)
  - $\mathbf{G}(n,m)$: total $m$ undirected edges, picked at uniformly randomly from $n$ nodes

What are the characteristics of networks from this model?
Characteristics of the E-R Model

- Heavily Analyzed with many results! (see Bollobas: Random Graphs)

- We’ll go over a major few...
First: \( n \) and \( p \) do NOT uniquely determine a graph

\[ n = 15, \; p = 1/2 \]
Reminder: Network Properties

- Key properties:
  - Degree Distribution: \( P[k] \)
  - Path length: \( l \)
  - Clustering Co-efficient: \( C_i \)

- **Useful to characterize a network with**

- Many others: e.g. largest eigenvalue of the adjacency matrix, the laplacian eigen-gap etc. etc. (we will see some of these later!)
Main result: G(n,p)’s degree distribution is Binomial(n-1,p)

\[ P[k] = \binom{n-1}{k} p^k (1 - p)^{n-1-k} \]

Select k nodes out of the n-1 remaining

Probability of having k edges and NOT having the rest

If \( p = \frac{c}{n} \), then

\[ \lim_{n \to \infty} P[k] = \text{Poisson}(c) = \frac{c^k}{k!} e^{-c} \]

The Poisson Limit Theorem
**Expected Degree**

\[ E[d_v] = \sum P[k] \times k = (n - 1)p \]

But how close are the actual degrees to the Expectation?

**Chernoff Bounds:** If \( X \) is Binomial\((n, p)\) and \( \varepsilon \leq 2/3 \), then

\[
P[|X - E[X]| \geq \varepsilon E[X]] \leq 2 \exp\left(-\frac{1}{3} \varepsilon^2 E[X]\right)
\]

The probability \( X \) deviates from \( E[X] \) is exponentially small: or the degree *concentrates* on the mean.
Path Length: diameter \( d \)

- We **skip** the proof!
  - A bit involved
  - Uses the concept of ‘expansion’
  - Random graphs are ‘good’ expanders: sparse graphs with good connectivity: so diameters are not ‘too large’

- **Main result:** The diameter of an E-R graph is \( O(\log n) \)
Clustering Co-efficient $C_i$

- **Recall:** $C_i = \frac{2e_i}{k_i(k_i - 1)}$ \[ e_i \text{ is the number of edges between the neighbors of node } i \]

- **For an E-R graph**
  
  $e_i = \binom{k_i}{2} = \frac{k_i(k_i - 1)}{2}$
  
  $p = \frac{k_i(k_i - 1)}{2}p$

- **So**
  
  $C_i = \frac{2k_i(k_i - 1)}{2k_i(k_i - 1)}p = p$

- **Main result:** CC of an ER graph is small
What happens when $p$ is varied?

$p$

0

Empty Graph

........

1

Complete Graph

Here?
(Size of) The Giant Component

- $r = \text{fraction of vertices NOT in the GCC}$
  
  $= \text{the probability of a randomly chosen node belongs to the GCC}$

- All of the following are **scaling** results, that is as $n \to \infty$
  
  - Control the scaling somehow: a natural scaling is the expected degree is a constant i.e. $p = c/n$
Figure 1: (a) Graphical solutions to Eq. (8), showing the curve \( y = 1 - e^{-cS} \) for three choices of \( c \) along with the curve \( y = S \). The locations of their intersections give the numerical solutions to Eq. (8). Any solution \( S > 0 \) implies a giant component. (b) The solution to Eq. (8) as a function of \( c \), showing the discontinuous emergence of a giant component at the critical point \( c = 1 \). In the "super-critical" regime \( c > 1 \), the lines always intersect at a second point \( S > 0 \), implying the existence of a giant component. The transition between the two "phases" happens at \( c = 1 \), which is called the "critical point".

2.3.1 Branching processes and percolation

Another way to prove that the giant component exists for \( c > 1 \) is to model the exploration of a component as a Galton-Watson branching process (this approach is more mathematical, but has the advantage of allowing us to say much more about both the giant component and the size and shape of the components not part of the giant one). There are three cases to consider: (i) the sub-critical regime of \( c < 1 \), the critical regime of \( c = 1 \) and the super-critical regime of \( c > 1 \). In general, this approach considers a "percolation" or branching process on the exploration of the component containing a vertex \( i \). In the sub-critical regime, the average number of "offspring" or neighbors in the branching process is less than 1 and the branching process tends to die out very quickly (there is a close analogy here with Poisson processes). This leads to a network composed of mostly small trees. In the super-critical regime, the average number of offspring in the branching process is larger than 1 and the process exhibits exponential growth. At the critical point, the average number of offspring is exactly 1, and the size of the branching process is controlled by the variance rather than the average.
What happens when p is varied?

- **p**:
  - 0: Empty Graph
  - 1/n: GCC emerges
  - \( \log \frac{n}{n} \): No isolated nodes remarkably is also the threshold for connectivity!
  - 1: Complete Graph
ARE REAL GRAPHS RANDOM?
## Recap

<table>
<thead>
<tr>
<th></th>
<th>Random Graphs ( G(n,p) )</th>
<th>Real Graphs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree Distribution</td>
<td>Binomial((n,p))</td>
<td></td>
</tr>
<tr>
<td>Clustering Co-efficient</td>
<td>(~ p) (SMALL)</td>
<td></td>
</tr>
<tr>
<td>Diameter (Average Shortest Path)</td>
<td>(O(\log n))</td>
<td></td>
</tr>
</tbody>
</table>

Prakash 2017
Degree Distribution of Web is heavily Skewed

$P[k] \approx k^{-\alpha}, \ 2 < \alpha < 3$

Broder et al 2000
AS routers [Faloutsos\textsuperscript{3} 1999]

\[ P[k] \approx k^{-\alpha}, \quad 2 < \alpha < 3 \]
Other Networks [Barabasi, Albert 1999]

\[ P[k] \approx k^{-\alpha}, \ 2 < \alpha < 3 \]
Clustering Co-efficient is much larger than $G(n,p)$

\[ C \text{ of } G(n,p) \approx p = \langle k \rangle / N \]
Diameter of real graphs are short

\[ \text{diam of } G(n,p) \approx \log N / \log <k> \]
## Recap

<table>
<thead>
<tr>
<th></th>
<th>Random Graphs $G(n, p)$</th>
<th>Real Graphs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Degree Distribution</strong></td>
<td>Binomial($n, p$)</td>
<td>Skewed (Power-Law type)</td>
</tr>
<tr>
<td><strong>Clustering Coefficient</strong></td>
<td>$\approx p$ (SMALL)</td>
<td>Much larger</td>
</tr>
<tr>
<td><strong>Diameter (Average Shortest Path)</strong></td>
<td>$O(\log n)$</td>
<td>About the same</td>
</tr>
</tbody>
</table>
Are real graphs random?

- **NO!**

- **Main problems:**
  - No local structure
  - GCC typically does not arise from phase transitions
  - Degree distributions skewed
Why study $G(n,p)$?

Essentially, all models are wrong, but some are useful.

(George E. P. Box)

izquotes.com
Why study $G(n,p)$?

- Help calculate many quantities, and compare to real data
- Helps to understand to what extent a property is a result of random process
- Techniques are very useful
POWER-LAWS
Definition

- $P(x) = C x^{-a}$ (\(x \geq x_{\text{min}}\))
- E.g. prob( city population between \(x\) and \(dx\))

\[
\log(p(x)) \quad \log(x_{\text{min}}) \quad \log(x)
\]
For discrete variables

- $P(k) = C \ k^{-\alpha} \ (k > 0)$

- Or the Yule distribution

\[
P(k) = CB(k, a)
\]

\[
B(k, a) = \frac{\Gamma(k) \Gamma(a)}{\Gamma(k + a)} \approx k^{-a}
\]
Power laws, Pareto distributions and Zipf's law


FIG. 2 Left: histogram of the populations of all US cities with populations of 1000 or more. Right: another histogram of the same data, but plotted on logarithmic scales. The approximate straight-line form of the histogram in the right panel implies that the distribution follows a power law. Data from the 2000 US Census.

Power-law distributions occur in an extraordinarily diverse range of phenomena. In addition to city populations, the sizes of earthquakes [3], moon craters [4], solar flares [5], computer files [6] and wars [7], the frequency of use of words in any human language [2, 8], the frequency of occurrence of personal names in most cultures [9], the numbers of papers scientists write [10], the number of citations received by papers [11], the number of hits on web pages [12], the sales of books, music recordings and almost every other branded commodity [13, 14], the numbers of species in biological taxa [15], people's annual incomes [16] and a host of other variables all follow power-law distributions.

Power laws also occur in many situations other than the statistical distributions of quantities. For instance, Newton's famous $1/r^2$ law for gravity has a power-law form with exponent $\alpha = 2$. While such laws are certainly interesting in their own way, they are not the topic of this paper. Thus, for instance, there has been some discussion of the "allometric" scaling laws seen in the physiognomy and physiology of biological organisms [17], but since these are not statistical distributions they will not be discussed here.
Power laws, Pareto distributions and Zipf's law

Power-law distributions are the subject of this article. In the following sections, I discuss ways of detecting power-law behaviour, give empirical evidence for power laws in a variety of systems and describe some of the mechanisms by which power-law behaviour can arise.

Readers interested in pursuing the subject further may also wish to consult the reviews by Sornette [18] and Mitzenmacher [19], as well as the bibliography by Li.

Figure 1: Left: histogram of heights in centimetres of American males. Data from the National Health Examination Survey, 1959–1962 (US Department of Health and Human Services). Right: histogram of speeds in miles per hour of cars on UK motorways. Data from Transport Statistics 2003 (UK Department for Transport).

Figure 2: Left: histogram of the populations of all US cities with 10000 or more. Right: another histogram of the same data, but plotted on logarithmic scales. The approximate straight-line form of the histogram in the right panel implies that the distribution follows a power law. Data from the 2000 US Census.

Power-law distributions occur in an extraordinarily diverse range of phenomena. In addition to city populations, the sizes of earthquakes [3], moon craters [4], solar flares [5], computer files [6] and wars [7], the frequency of use of words in any human language [2, 8], the frequency of occurrence of personal names in most cultures [9], the numbers of papers scientists write [10], the number of citations received by papers [11], the number of hits on web pages [12], the sales of books, music recordings and almost every other branded commodity [13, 14], the numbers of species in biological taxa [15], people's annual incomes [16] and a host of other variables all follow power-law distributions.

While such laws are certainly interesting in their own way, they are not the topic of this paper. Thus, for instance, there has in recent years been some discussion of the "allometric" scaling laws seen in the physiognomy and physiology of biological organisms [17], but since these are not statistical distributions they will not be discussed here.

[Newman, 2005]
Exponential vs Power-Laws

\[ p(x) = cx^{-0.5} \]

Exponential decays much faster!

\[ p(x) = cx^{-1} \]

Power Laws have longer ‘tails’

\[ p(x) = c^{-x} \]
Degree Distribution of Web is heavily Skewed

$P[k] \approx k^{-\alpha}, \ 2 < \alpha < 3$

Broder et al 2000
AS routers [Faloutsos 1999]

$$P[k] \approx k^{-\alpha}, \ 2 < \alpha < 3$$
Other Networks [Barabasi, Albert 1999]

\[ P[k] \approx k^{-\alpha}, \quad 2 < \alpha < 3 \]
Power-laws are everywhere

[Clauset et al. 2007]
Exploiting Long Tails

ANATOMY OF THE LONG TAIL

Online services carry far more inventory than traditional retailers. Rhapsody, for example, offers 19 times as many songs as Wal-Mart’s stock of 39,000 tunes. The appetite for Rhapsody’s more obscure tunes (charted below in yellow) makes up the so-called Long Tail. Meanwhile, even as consumers flock to mainstream books, music, and films (right), there is real demand for niche fare found only online.

THE NEW GROWTH MARKET: OBSCURE PRODUCTS YOU CAN’T GET ANYWHERE BUT ONLINE

Sources: Erik Brynjolfsson and Jeffrey Hu, MIT, and Michael Smith, Carnegie Mellon; Barnes & Noble; Netflix; RealNetworks

C. Andersen, WIRED, 2004
Some examples which are NOT P.Ls

- (a) The abundance of North American bird species,
- (b) The number of entries in people's email address books of 16,881 users of a large university computer system [33].
- (c) The size in acres of all wildfires occurring on US federal land between 1986 and 1996 (National Fire Occurrence Database, USDA Forest Service and Department of the Interior). Note that the horizontal axis is logarithmic.

The labels on the left refer to the panels in the figure. Exponents, geology, physics and astronomy, and this on its own is an extraordinary statement.
Binomial vs Power-Law Degree Distribution

Bell Curve
- Most nodes have the same number of links
- No highly connected nodes

Power Law Distribution
- Very many nodes with only a few links
- A few hubs with large number of links

Number of nodes with k links vs Number of links (k)

Network comparisons:
- Left: Bell Curve network
- Right: Power Law Distribution network
Scale Free Networks

- Networks with power-law tails in degree distribution

- Name comes from Physics
  - Scale invariance: no characteristic scale
  - \( f(a \cdot x) = a^c f(x) \) :::: A scale free function
    - Verify it holds in power-laws
Heavy-Tailed Distributions

- Formally, \( P(x) \) is heavy tailed if
  \[
  \lim_{x \to \infty} \frac{P(X > x)}{e^{-\lambda x}} = \infty
  \]

- Exponential:
  \[
  P(X) = \lambda e^{-\lambda x}
  \]
  \[
  P(X > x) = 1 - P(X \leq x) = e^{-\lambda x}
  \]

- Normal:
  \[
  P(X) = \frac{1}{\sqrt{2\pi \sigma}} e^{\frac{(x-\mu)^2}{2\sigma^2}}
  \]
  Both are not heavy-tailed
Heavy Tail Examples: AKA Long Tails, Zipf’s Law, Pareto’s Law etc.

\[ P(x) \propto x^{-\alpha} \]

- Power law
- Power law with cutoff
- Stretched exponential
- Log-normal

\[ x^{-\alpha} e^{-\lambda x} \]

\[ x^{\beta-1} e^{-\lambda x^\beta} \]

\[ \frac{1}{x} \exp \left[ -\frac{(\ln x - \mu)^2}{2\sigma^2} \right] \]
Normalizing Constant for Power-Laws

- **Q:** $p(x) = cx^{-\alpha}$, with $x > x_{\text{min}}$ what is $c$?
- **See that:**
  \[ \int_{x \geq x_{\text{min}}} p(x) \, dx = 1 \]

  \[ c \int_{x \geq x_{\text{min}}} x^{-\alpha} = 1 \]

  \[ \Rightarrow c = (\alpha - 1)x_{\text{min}}^{\alpha - 1} \]

  \[ p(x) = \frac{\alpha - 1}{x_m} \left( \frac{x}{x_m} \right)^{-\alpha} \]
Average value? \( E[X] \) of a power-law

- \( P(x) \) is a power-law

\[
E [ x ] = \int_{x_m}^{\infty} x \ p(x) \, dx = z \int_{x_m}^{\infty} x^{-\alpha+1} \, dx \\
= - \frac{(\alpha-1)x_m^{\alpha-1}}{2-\alpha} \left[ \infty^{2-\alpha} - x_m^{2-\alpha} \right] 
\]

Need: \( \alpha > 2 \)

\[
E [ X ] = \frac{\alpha - 1}{\alpha - 2} x_m 
\]
Infinite Moments

- Average: \( E[X] = \frac{\alpha - 1}{\alpha - 2} \times m \) (if \( \alpha \leq 2 \), \( E[X] \) is \( \infty \))

- Variance: if \( \alpha \leq 3 \), \( \text{Var}[X] = \infty \)

In real networks \( 2 < \alpha < 3 \) so:
\( E[x] = \text{const} \)
\( \text{Var}[x] = \infty \)
Three Versions of P.L.

PDF = frequency-count plot
Zipf plot = rank-frequency plot
NCDF = CCDF (Complementary CDF)

USEFUL!

IF ONE PLOT IS A P.L., SO ARE THE OTHER TWO

PDF = frequency-count plot
Zipf plot = rank-frequency plot
NCDF = CCDF (Complementary CDF)

IF ONE PLOT IS A P.L., SO ARE THE OTHER TWO

Prob (area = x)

area

Prob (area >= x)

Prob (area = x)

area

Prob (area >= x)

Prob (area = x)
Estimating $\alpha$ from data

- One way: Fit a line on the log-log plot using least squares i.e. solve $\arg \min_{\alpha} (\log(y) - \alpha \log(x))^2$
Better way: Use the three versions

- Plot CCDF: $P(X \geq x)$. Then estimated $\alpha = 1 + \alpha'$
  - Try to prove it!
Case-Study: FlickR

Log scale, $\alpha=1.75$

CCDF, Log scale, $\alpha=1.75$

CCDF, Log scale, $\alpha=1.75$, exp. cutoff
Another way: Max. Likelihood Estimation

\[ P(x) = \frac{\alpha - 1}{x_m} \left( \frac{x}{x_m} \right)^{-\alpha} \]

\[ L(\alpha) = \ln(\prod_i^n p(d_i)) = \sum_i^n \ln p(d_i) \]
\[ = \sum_i^n \ln(\alpha - 1) - \ln(x_m) - \alpha \ln \left( \frac{d_i}{x_m} \right) \]

- So we want \( \alpha^* = \arg \max_{\alpha} L(\alpha) \)

\[ \frac{dL(\alpha)}{d\alpha} = 0 \Rightarrow \frac{n}{\alpha-1} - \sum_i \ln \left( \frac{d_i}{x_m} \right) = 0 \]

\[ \alpha^* = 1 + n \left[ \sum_i^n \ln \left( \frac{d_i}{x_m} \right) \right]^{-1} \]
Max. Expected Degree in a Scale-Free network

- Let $K$ be the max. expected degree $\rightarrow$ the expected number of nodes with degree $> K$ should be $< 1$

$$
\int_{K}^{\infty} p(x) \, dx \approx \frac{1}{n}
$$

$$
= \frac{(a-1)x_{m}^{a-1}}{-(a-1)} \left[ 0 - K^{1-a} \right] = x_{m}^{a-1} \cdot K^{-(a-1)} \approx \frac{1}{n}
$$

$$
K = x_{m} n^{1/(\alpha - 1)}
$$
Max. Expected Degree: Consequences

- Q: Why don’t we see networks in real-life with $\alpha = 4, 5, 6 \ldots$?
- In real networks $K \approx 1000$
- How large should $n$ be if $\alpha = 4, 5, 6 \ldots$?
  - If $\alpha = 5$,
    
    $$K = x_m \frac{1}{n^{\alpha-1}}$$

    $$n = \left(\frac{K}{x_m}\right)^{\alpha-1} \approx 10^{12}$$
Generative Processes for P.L.s

- CAN NOT arise from sums of independent event
- Central Limit Theorem

\[ X, X_1, \ldots, X_n : \text{rand. vars with mean } \mu, \text{ variance } \sigma^2 \]
\[ S_n = \sum_i X_i : E[S_n] = n\mu, \text{ var}[S_n] = n\sigma^2, \text{ SD}[S_n] = \sigma\sqrt{n} \]

\[ P(S_n = E[S_n] + x \cdot \text{SD}[S_n]) \sim \frac{1}{2\pi} e^{-\frac{x^2}{2}} \]
Generative processes

- Combination of exponentials
- Forest fire
- Richer gets richer
- Random walks
- ....
Summary: Scale-Free networks

- **Second moment** $\langle k^2 \rangle$ diverges
- **Average** $\langle k \rangle$ diverges
- **Ultra small world behavior**
- **Regime full of anomalies**

*Source: J. Leskovec*