Network Robustness and Preferential Attachment

CS224W: Analysis of Networks
Jure Leskovec, Stanford University
http://cs224w.stanford.edu
Overview

- (1) Power-laws in Networks
- (2) Network Robustness
- (3) Preferential Attachment
Which interesting graph properties do we observe that need explaining?

- **Small-world model:**
  - Diameter
  - Clustering coefficient

- **Node degree distribution**
  - What fraction of nodes has degree $k$ (as a function of $k$)?
  - Prediction from simple random graph models: $p(k) = \text{exponential function of } k$
  - Observation: Often a power-law: $p(k) \propto k^{-\alpha}$
Expected based on $G_{np}$

$$P(k) \propto k^{-\alpha}$$

Found in data
Take a network, plot a histogram of $P(k)$ vs. $k$.
Plot the same data on log-log scale:

How to distinguish:
\[ P(k) \propto \exp(-k) \text{ vs. } P(k) \propto k^{-\alpha} \]

Take logarithms:
if \[ y = f(x) = e^{-x} \] then
\[ \log(y) = -x \]
If \[ y = x^{-\alpha} \] then
\[ \log(y) = -\alpha \log(x) \]
So on log-log axis power-law looks like a straight line of slope \(-\alpha\)!
Node Degrees: Faloutsos

- First observed in Internet Autonomous Systems [Faloutsos, Faloutsos and Faloutsos, 1999]

Internet domain topology
Node Degrees: Web

- The World Wide Web [Broder et al., 2000]
Node Degrees: Barabasi&Albert

- Other Networks [Barabasi-Albert, 1999]
Above a certain $x$ value, the power law is always higher than the exponential!
Exponential vs. Power-Law

- Power-law vs. Exponential on log-log and semi-log (log-lin) scales

\[ p(x) = cx^{-0.5} \]
\[ p(x) = cx^{-1} \]

**log-log**
- x ... logarithmic axis
- y ... logarithmic axis

**semi-log**
- x ... linear axis
- y ... logarithmic axis
Exponential vs. Power-Law

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)

Examples of networks:

-Bell Curve network (top left)
- Power Law network (bottom right)
Power-Law Degree Exponents

- Power-law degree exponent is typically $2 < \alpha < 3$
  - Web graph:
    - $\alpha_{in} = 2.1$, $\alpha_{out} = 2.4$ [Broder et al. 00]
  - Autonomous systems:
    - $\alpha = 2.4$ [Faloutsos, 99]
  - Actor-collaborations:
    - $\alpha = 2.3$ [Barabasi-Albert 00]
  - Citations to papers:
    - $\alpha \approx 3$ [Redner 98]
  - Online social networks:
    - $\alpha \approx 2$ [Leskovec et al. 07]
Definition:
Networks with a power-law tail in their degree distribution are called “scale-free networks”

Where does the name scale-free come from?

- **Scale invariance**: There is no characteristic scale
  - Scale invariance means laws do not change if scales of length, energy, or other variables, are multiplied by a common factor

- **Scale-free function**: $f(ax) = a^\lambda f(x)$
  - Power-law function: $f(ax) = a^\lambda x^\lambda = a^\lambda f(x)$

Log() or Exp() are not scale free!

- $f(ax) = \log(ax) = \log(a) + \log(x) = \log(a) + f(x)$
- $f(ax) = \exp(ax) = \exp(x)^a = f(x)^a$
Power-Laws are Everywhere

Many other quantities follow heavy-tailed distributions

[Cluset-Shalizi-Newman 2007]
Not Everyone Likes Power-Laws 😊

CMU grad-students at the G20 meeting in Pittsburgh in Sept 2009
Heavy Tailed Distributions

- Degrees are heavily skewed:
  Distribution \( P(X > x) \) is **heavy tailed if:**

\[
\lim_{x \to \infty} \frac{P(X > x)}{e^{-\lambda x}} = \infty
\]

- **Note:**
  - Normal PDF: \( p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \)
  - Exponential PDF: \( p(x) = \lambda e^{-\lambda x} \)
    - then \( P(X > x) = 1 - P(X \leq x) = e^{-\lambda x} \)
    are not heavy tailed!
What is the normalizing constant?

\[ p(x) = Z x^{-\alpha} \quad Z = ? \]

- \( p(x) \) is a distribution: \( \int p(x) \, dx = 1 \)

\[ 1 = \int_{x_m}^{\infty} p(x) \, dx = Z \int_{x_m}^{\infty} x^{-\alpha} \, dx \]

\[ = - \frac{Z}{\alpha-1} \left[ x^{-\alpha+1} \right]_{x_m}^{\infty} = - \frac{Z}{\alpha-1} \left[ \infty^{1-\alpha} - x_m^{1-\alpha} \right] \]

\[ \Rightarrow Z = (\alpha - 1)x_m^{\alpha-1} \]

Need: \( \alpha > 1 \) !

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

Integral:

\[ \int (ax)^n = \frac{(ax)^{n+1}}{a(n+1)} \]
What’s the expected value of a power-law random variable $X$?

$$E[X] = \int_{x_m}^{\infty} x \, p(x) \, dx = Z \int_{x_m}^{\infty} x^{-\alpha+1} \, dx$$

$$= \frac{Z}{2-\alpha} \left[ x^{2-\alpha} \right]_{x_m}^{\infty} = \frac{(\alpha-1)x_m^{\alpha-1}}{-(\alpha-2)} \left[ \infty^{2-\alpha} - x_m^{2-\alpha} \right]$$

$$\Rightarrow E[X] = \frac{\alpha - 1}{\alpha - 2} \, x_m$$

Need: $\alpha > 2$!

Power-law density:

$$p(x) = \frac{\alpha - 1}{x_m} \left( \frac{x}{x_m} \right)^{-\alpha}$$

$$Z = \frac{\alpha - 1}{x_m^{1-\alpha}}$$
Mathematics of Power-Laws

- **Power-laws have infinite moments!**

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

- If \( \alpha \leq 2 \): \( E[X] = \infty \)
- If \( \alpha \leq 3 \): \( Var[X] = \infty \)
  
  - Average is meaningless, as the variance is too high!

- **Consequence:** Sample average of \( n \) samples from a power-law with exponent \( \alpha \)

In real networks \( 2 < \alpha < 3 \) so:

- \( E[X] = \text{const} \)
- \( Var[X] = \infty \)
Estimating Power-law Exponent
Estimating Power-Law Exponent $\alpha$

Estimating $\alpha$ from data:

1. Fit a line on log-log axis using least squares:
   
   Solve $\arg\min_{\alpha} (\log(y) - \alpha \log(x) + b)^2$

BAD!
Estimating Power-Law Exponent $\alpha$

Estimating $\alpha$ from data:

- Plot Complementary CDF (CCDF) $P(X \geq x)$. Then the estimated $\alpha = 1 + \alpha'$ where $\alpha'$ is the slope of $P(X \geq x)$.

- **Fact:** If $p(x) = P(X = x) \propto x^{-\alpha}$ then $P(X \geq x) \propto x^{-(\alpha-1)}$

  - $P(X \geq x) = \sum_{j=x}^{\infty} p(j) \approx \int_{x}^{\infty} Z y^{-\alpha} dy =$

  - $= - \frac{Z}{1-\alpha} [y^{1-\alpha}]_{x}^{\infty} = \frac{Z}{1-\alpha} x^{-(\alpha-1)}$
Estimating $\alpha$ from data:

- **Use maximum likelihood approach:**
  - The log-likelihood of observed data $d_i$:
    - $L(\alpha) = \ln(\prod_i^n p(d_i)) = \sum_i^n \ln p(d_i)$
    - $= \sum_i^n \left( \ln(\alpha - 1) - \ln(x_m) - \alpha \ln\left(\frac{d_i}{x_m}\right)\right)$
  - **Want to find $\alpha$ that max $L(\alpha)$:** Set $\frac{dL(\alpha)}{d\alpha} = 0$
    - $\frac{dL(\alpha)}{d\alpha} = 0 \Rightarrow \frac{n}{\alpha - 1} - \sum_i^n \ln\left(\frac{d_i}{x_m}\right) = 0$
    - $\Rightarrow \hat{\alpha} = 1 + n \left[\sum_i^n \ln\left(\frac{d_i}{x_m}\right)\right]^{-1}$

**Power-law density:**

$p(x) = \frac{\alpha - 1}{x_m} \left(\frac{x}{x_m}\right)^{-\alpha}$
Flickr: Fitting Degree Exponent

\[ P_k \sim k^{-1.75} \]

CCDF, Log scale, \( \alpha=1.75 \)

CCDF, Log scale, \( \alpha=1.75 \), exp. cutoff
Random vs. Scale-free network

Random network
(Erdos-Renyi random graph)
Degree distribution is Binomial

Scale-free (power-law) network
Degree distribution is Power-law
Consequence of Power-Law Degrees
The interest in the robustness problem has two origins:

→ Robustness of a system is an important problem in many areas

→ Many real networks are not regular, but have a scale-free topology
How does network connectivity change as nodes get removed? [Albert et al. 00; Palmer et al. 01]

Nodes can be removed:

- **Random failure:**
  - Remove nodes uniformly at random

- **Targeted attack:**
  - Remove nodes in order of decreasing degree

This is important for robustness of the internet as well as epidemiology
Network Robustness

- **Networks with equal number of nodes and edges:**
  - ER random graph
  - Scale-free network

- **Study the properties of the network as an increasing fraction of nodes are removed**
  - Node selection:
    - Random (corresponds to random failures)
    - Nodes with largest degrees (corresponds to targeted attacks)

- **Measures:**
  - Fraction of nodes in the largest connected component
  - Average shortest path length between nodes in the largest component
Network Robustness

- Study the properties of the network as an increasing fraction of nodes are removed
  - Node selection:
    - Random failures: select node to remove at random
    - Targeted attacks: remove nodes by decreasing degree
Graphs are resilient to random failures, but are sensitive to targeted attacks:

Degree exponent $\gamma=3.5$
What proportion of random nodes must be removed in order for the size \((S)\) of the giant component to drop to 0?

- Infinite scale-free networks with \(\gamma < 3\) never break down under random node failures

Source: Cohen et al., Resilience of the Internet to Random Breakdowns
Real networks are resilient to random failures

- \( G_{np} \) has better resilience to targeted attacks
  - E.g., we need to remove all pages of degree >5 to disconnect the Web. But this is a very small fraction of all web pages!
Resilience in Real Networks

- Random failures
- Targeted attack

Source: Error and attack tolerance of complex networks. Réka Albert, Hawoong Jeong and Albert-László Barabási

The first few % of nodes removed:
- $G_{np}$
- SF: Scale-free

Notice how targeted attacks very quickly disconnect the network.
Preferential Attachment Model
Exponential vs. Power-Law Tails

Model: $G_{np}$
New nodes are more likely to link to nodes that already have high degree

Herbert Simon’s result:
- Power-laws arise from “Rich get richer” (cumulative advantage)

Examples
- **Citations** [de Solla Price ‘65]: New citations to a paper are proportional to the number it already has
  - **Herding**: If a lot of people cite a paper, then it must be good, and therefore I should cite it too
- **Sociology**: Matthew effect, [http://en.wikipedia.org/wiki/Matthew_effect](http://en.wikipedia.org/wiki/Matthew_effect)
- “For whoever has will be given more, and they will have an abundance. Whoever does not have, even what they have will be taken from them.”
- Eminent scientists often get more credit than a comparatively unknown researcher, even if their work is similar
Model: Preferential attachment

- Preferential attachment:
  [de Solla Price ‘65, Albert-Barabasi ‘99, Mitzenmacher ‘03]
  - Nodes arrive in order 1, 2, ..., n
  - At step \( j \), let \( d_i \) be the degree of node \( i < j \)
  - A new node \( j \) arrives and creates \( m \) out-links
  - Prob. of \( j \) linking to a previous node \( i \) is proportional to degree \( d_i \) of node \( i \)

\[
P(j \rightarrow i) = \frac{d_i}{\sum_k d_k}
\]
We analyze the following **simple** model:

- Nodes arrive in order $1, 2, 3, \ldots, n$
- When node $j$ is created it makes a **single out-link** to an earlier node $i$ chosen:
  - **1)** With prob. $p$, node $j$ links to $i$ chosen uniformly at random (from among all earlier nodes)
  - **2)** With prob. $1 - p$, node $j$ chooses $i$ uniformly at random & links to a random node $l$ that $i$ points to
    - **This is same as saying:** With prob. $1 - p$, node $j$ links to node $l$ with prob. proportional to $d_l$ (the in-degree of $l$)
- **Our graph is directed:** Every node has out-degree 1
**Claim:** The described model generates networks where the fraction of nodes with **in-degree** \( k \) scales as:

\[
P(d_i = k) \propto k^{-(1 + \frac{1}{q})}
\]

where \( q = 1 - p \)

So we get power-law degree distribution with exponent:

\[
\alpha = 1 + \frac{1}{1 - p}
\]
Preferential attachment: Good news

- Preferential attachment gives power-law in-degrees!
- Intuitively reasonable process
- Can tune model parameter $p$ to get the observed exponent
  - On the web, $P[\text{node has in-degree } d] \sim d^{-2.1}$
  - $2.1 = 1 + 1/(1-p) \Rightarrow p \sim 0.1$
Preferential Attachment: Bad News

- Preferential attachment is not so good at predicting network structure
  - Age-degree correlation
    - Node degree is proportional to its age
    - Solution: Node fitness (virtual degree)
  - Links among high degree nodes:
    - On the web nodes sometimes avoid linking to each other
- Further questions:
  - What is a reasonable model for how people sample network nodes and link to them?
    - Short random walks
Origins of Preferential Attachment

- **Link selection model** -- perhaps the simplest example of a local or random mechanism capable of generating preferential attachment

- **Growth**: At each time step we add a new node to the network

- **Link selection**: We select a link at random and connect the new node to one of the nodes at the two ends of the selected link

- **This simple mechanism generates preferential attachment**
  - Why? Because node is picked with prob. proportional to the number of edges it has
Origins of Preferential Attachment

**Copying model:**

- **(a) Random Connection:** with prob. $p$ the new node links to random $u$
- **(b) Copying:** With prob. $1 - p$ randomly choose an outgoing link of node $u$ and connect the new node to the selected link's target
  - The new node “copies” one of the links of an earlier node
Analysis of the copying model:

- (a) the probability of selecting a node is $1/N$
- (b) is equivalent to selecting a node linked to a randomly selected link. The probability of selecting a degree-$k$ node through the copying process of step (b) is $k/2E$ for undirected networks.
- Again, the likelihood that the new node will connect to a degree-$k$ node follows preferential attachment.

Examples:

- **Social networks**: Copy your friend’s friends.
- **Citation Networks**: Copy references from papers we read.
- **Protein interaction networks**: gene duplication.
Many models lead to Power-Laws

- **Copying mechanism** (directed network)
  - Select a node and an edge of this node
  - Attach to the endpoint of this edge

- **Walking on a network** (directed network)
  - The new node connects to a node, then to every first, second, ... neighbor of this node

- **Attaching to edges**
  - Select an edge and attach to both endpoints of this edge

- **Node duplication**
  - Duplicate a node with all its edges
  - Randomly prune edges of new node
Distances in Preferential Attachment

\[
\bar{h} = \begin{cases} 
\frac{\log n}{\log \log n} & \alpha = 3 \\
\frac{\log \log n}{\log(\alpha - 1)} & 2 < \alpha < 3 \\
\text{const} & \alpha = 2
\end{cases}
\]

Size of the biggest hub is of order \(O(N)\). Most nodes can be connected within two steps, thus the average path length will be independent of the network size \(n\).

The avg. path length increases slower than logarithmically with \(n\). In \(G_{np}\) all nodes have comparable degree, thus most paths will have comparable length. In a scale-free network vast majority of the paths go through the few high degree hubs, reducing the distances between nodes.

Some models produce \(\alpha = 3\). This was first derived by Bollobas et al. for the network diameter in the context of a dynamical model, but it holds for the average path length as well.

The second moment of the distribution is finite, thus in many ways the network behaves as a random network. Hence the average path length follows the result that we derived for the random network model earlier.
Summary: Scale-Free Networks

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

- $\langle k^2 \rangle$ finite
- Average $\langle k \rangle$ finite
- Small world
- The scale-free behavior is relevant
- Behaves like a random network
Summary

- (1) Power-laws in Networks
- (2) Network Robustness
- (3) Preferential Attachment
Evolution of Resilience in Protein Interactomes
Protein Interactomes

- **Protein interactome**: A protein-protein interaction network of a species:
  - Nodes: Species’ proteins
  - Edges: Physical protein-protein interactions (PPI)

- **Tree of life**: Evolutionary history of species:
  - Phylogenetic tree calculated based on similarity of gene sequence information between species:
    - Units: nucleotide substitutions per site
  - **Evolutionary time of a species**: total branch length from the root to the corresponding leaf in the tree
- **Tree of life:** 1,539 bacteria, 111 archaea, and 190 eukarya
- **Protein interactomes:**
  - Located at the *leaves of the tree of life*
  - Separated by millions of years of evolution
  - Ancestral species have gone extinct or evolved into present-day species
    - Older interactomes are lost

[Jure Leskovec, Stanford CS224W: Analysis of Networks]
Protein Failures

- **Protein failure** can occur through:
  - **Removal** of a protein (*e.g.*, nonsense mutation)
  - **Disruption** of a PPI (*e.g.*, environmental factors, such as availability of resources)

- **Resilience** is a critical interactome property:
  - Breakdown of proteins affect the exchange of any biological information between proteins in a cell
  - Protein failures can fragment the interactome and lead to cell death and disease

[Zitnik et al., bioRxiv 454033, 2018]
Does Evolution Mold Interactomes?

Questions for today:
- How do interactomes change through evolution?
- How does natural selection shape the interactomes?
- How do changes in these networks impact species?

Approach:
- Define a network resilience measure
- Use the measure to study resilience of interactomes
- Find a network mechanism of resilience

[Zitnik et al., bioRxiv 454033, 2018]
**Network Resilience: Entropy**

- **Fragmentation of the network** upon node removal:
  - Entropy $H$ on a set of isolated clusters:
    - $p_i = \frac{C_i}{N}$ is the proportion of nodes that belong to cluster $C_i$
      - Probability of seeing a node from $C_i$ when sampling one node from the fragmented network
    - $H$ quantifies **uncertainty in predicting the cluster identity of an individual node** taken at random from the network
  - **Shannon’s diversity index $H_{sh}$**: Normalize w.r.t. network size

**Entropy** $H = -\sum_{C_i} \frac{C_i}{N} \log \frac{C_i}{N}$

$max H = \log N$

Entropy $H = 2.579$

Shannon’s diversity $H_{sh} = H / \max H = 0.584$
Network Resilience: Entropy

$H_{sh} \in [0,1]$:  

0: Connected network  

1: Fragmented network  
  - Each node is its own cluster

**High Shannon’s diversity**  
A large number of small clusters, all of approximately equal size  

Entropy $H = 2.442$  
Shannon’s diversity $H_{sh} = 0.658$

**Low Shannon’s diversity**  
A few large clusters and and only a few small broken-off clusters  

Entropy $H = 1.890$  
Shannon’s diversity $H_{sh} = 0.509$
Resilience: Response of the network to failures across all possible failure rates $f$ (*i.e.*, fraction of nodes removed), $f \in [0, 1]$
Interactomes become more resilient during evolution:
- Natural selection leads to resilient interactomes
- Resilience is evolvable property of life

\[ R^2 = 0.36 \]
\[ p < 9 \times 10^{-10} \]

[Zitnik et al., bioRxiv 454033, 2018]
Next: Identify a network mechanism through which a resilient interactome can arise

Gradual changes of network structure:

1. Structural changes in local network neighborhoods
2. Rewiring of PPIs and network motifs
Decompose each species’ interactome into local protein networks:

- **k-hop subnetwork** centered around each protein (i.e., $N_2(u)$ for node $u$)
  - Local representation of the protein’s direct and nearby interactions
1) Local Network Neighborhoods

**Note 1:** Isolated Clusters metric (IC) is defined as the number of connected components that arise when the central node is removed from the neighborhood, normalized by degree of the central node. Higher value indicates a greater fragmentation of the neighborhood.

**Note 2:** Effective Size metric (ES) captures the bridging potential of the central node, i.e., the “true” size of the node’s neighborhood absent of redundant neighbors.
The number of isolated network clusters and the effective size of the neighborhoods decrease with evolution:

- **Protein neighborhoods become more connected** during evolution
2) Rewiring of PPIs and Motifs

**Approach**

- **Organism 1**
  - Ortholog pair
  - Short evolutionary distance 
    $t < t'$

- **Organism 2**

**Results**

- $IRR(\bullet) = \log \frac{\{\bullet, \bullet, \bullet, \bullet\}}{\{\bullet, \bullet, \bullet, \bullet\}}$
  - $-0.215$ * $p < 10^{-33}$

- $IRR(\circ) = \log \frac{\{\circ, \circ, \circ, \circ\}}{\{\circ, \circ, \circ, \circ\}}$
  - $-0.098$

- $IRR(\triangle) = \log \frac{\{\triangle\}}{\{\triangle\}}$
  - $0.016$

- $IRR(\square) = \log \frac{\{\square\}}{\{\square\}}$
  - $-0.079$

---

Jure Leskovec, Stanford CS224W: Analysis of Networks
What do These Results Mean?

- Do resilient interactomes have any effect on living organisms?
  - Yes!
  - Resilient interactome has an astonishingly beneficial impact on the organism to survive in complex, variable, and competitive environment

Resilient interactome:

- Proteins can interact even in the face of high protein failure rate
- Mutations represent protein failures that are neutral in a given environment, i.e., have no phenotypic effect on the network’s function and are thus invisible to natural selection
- Neutral mutations do not remain neutral indefinitely, and a once neutral mutation may be crucial for survival in a changed environment

Resilient interactome is a reservoir of neutral mutations:

- Important for evolutionary innovation

“Stockpile of gold”

[Zitnik et al., bioRxiv 454033, 2018]