Power Laws & Rich Get Richer

CMSC 498J: Social Media Computing
Department of Computer Science
University of Maryland
Spring 2015

Hadi Amiri
hadi@umd.edu
Lecture Topics

- Popularity as a Network Phenomenon
- Power Laws
- Rich Get Richer model
Popularity

- Popularity can be characterized by **extreme imbalances**!
  - People are known to their immediate social circle!
  - A few people achieve wider visibility!
  - A very, very few achieve global name recognition.

- We need to know:
  - How can we quantify these imbalances?
  - Why do they arise?
Let us focus on the Web in which we can measure popularity accurately!

- Popularity of a page \( \sim \) number of its in-links
  - Easy to count!

Degree Centrality - Cnt.

- A node is central if it has ties to many other nodes
  - Look at the node degree

\[ C(n_i) = \sum_{j=1}^{n} A_{ij} = \sum_{j=1}^{n} A_{i} = \text{degree} \]
• **Question:**
  ▫ As a function of $k$, what fraction of pages on the Web have $k$ in-links?

• **Simple hypothesis:**
  ▫ Normal distribution?
    • the number of pages with $k$ in-links should decrease **exponentially** in $k$, as $k$ grows large
    • $f(k)=1/e^k$
Question:
- As a function of $k$, what fraction of pages on the Web have $k$ in-links?

Research works show something totally different!
- The fraction of Web pages with $k$ in-links is roughly proportional to $1/k^2$.
  - Accords with our intuition about the Web, where there are large number of extremely popular pages.
- $1/k^2$ decreases much more slowly as $k$ increases as compared to the exponential function $1/e^k$. 
• **Question:**
  ▫ As a function of \( k \), what fraction of pages on the Web have \( k \) in-links?

• **Research works show something totally different!**
  ▫ The fraction of Web pages with \( k \)-in-links is roughly proportional to \( 1/k^2 \).
  ▫ Accords with our intuition about the Web, where there are large numbers of extremely popular pages.
  ▫ \( 1/k^2 \) decreases much more slowly as \( k \) increases compared to the exponential function \( 1/e^k \).
Power Law

• A function that decreases as $k$ to some fixed power, e.g. $1/k^c$, is called a power law!
  ▫ It allows to see very large values of $k$ in data!
• Extreme imbalances are likely to arise!

• Power laws seem to dominate in cases when dealing with popularity-like measures.
Power Law- Cnt.

- Histogram of the populations of all US cities with population of 10,000 or more.

Power Law- Cnt.

• **Power law Test**: Given a dataset, test if it exhibits a power law distribution?
  ▫ Compute values wrt popularity measure (e.g. #in-links, #downloads, population of cities, etc.)
  ▫ Test if it's approximately a power law \(1/k^c\) for some \(c\), and if so, estimate the exponent \(c\).
  ▫ Can also bin different values of \(k\) and create a histogram
• What should a power law plot look like?
  ▫ \( f(k) \): the fraction of items that have value \( k \)
  ▫ If power law holds, \( f(k) = a/k^c \) ?
    • for some constant \( c \) and \( a \).
  ▫ \( f(k) = a/k^c = a k^{-c} \)
  ▫ \( \log f(k) = \log a - c \log k \)
    • straight line! “\( \log f(k) \)” as a function of “\( \log k \)”
      • “\( c \)”: slope, and
      • “\( \log a \)”: y-intercept.
    • log-log plot!
Power Law-Cnt.

- If power-law holds, the “log-log” plot should be a straight line.

Figures 1 and 2: In-degree and out-degree distributions subscribe to the power law. The law also holds if only off-site (or "remote-only") edges are considered.
If power-law holds, the “log-log” plot should be a straight line.

Power Laws are everywhere.

Power Law- Cnt.

- But, what is causing Power laws?
Rich Get Richer

**Rich-Get-Richer:** A simple model for the creation of links as a basis for power laws!

1. Pages are created in order and named 1, 2, ..., N.
2. When page $j$ is created, it produces a link to an earlier page $i < j$ according to the following rules:
   a) With probability $p$, page $j$ chooses page $i$ uniformly at random, and creates a link to $i$.
   b) With probability $(1 - p)$, page $j$ chooses page $i$ uniformly at random and creates a link to the page that $i$ points to (copies decision made by $i$).

- Let’s assume that each page creates just 1 link
  - We can extend this model to multiple links as well.
Rich Get Richer- Cnt.

• We observe power law, if we run this model for many pages
  ▫ the fraction of pages with $k$ in-links will be distributed according to a power law $1/k^c$!
  ▫ Value of the exponent $c$ depends on the choice of $p$.
• Correlation between $c$ and $p$?
  ▫ Smaller $p$
    • copying becomes more frequent
      • more likely to see extremely popular pages
        ▫ $c$ gets smaller as well
Rich Get Richer- Cnt.

• Due to copying mechanism: the probability of linking to a page is proportional to the total number of pages that currently link to that page!

• Restarting rule 2 (b):
  ▫ **Rich-Get-Richer rule**: With probability \((1- p)\), page \(j\) chooses page \(i\) with probability **proportional to i’s current number of in-links** and creates a link to \(i\).
  ▫ Also called **preferential attachment**
    • links are formed “preferentially” to pages that already have high popularity.
Rich Get Richer- Cnt.

**Rich-Get-Richer:**

1. Pages are created in order and named 1, 2, ..., N.
2. When page $j$ is created, it produces a link to an earlier page $i < j$ according to the following rules:
   a) With probability $p$, page $j$ chooses page $i$ uniformly at random and creates a link to $i$.
   b) With probability $(1 - p)$, page $j$ chooses page $i$ with probability proportional to $i$’s current number of in-links and creates a link to $i$. 
Rich Get Richer - Cnt.

• Probabilistic model of the rich-get-richer process
  □ $X_j(t)$: number of in-links to node $j$ at a time $t$
• Two points about $X_j(t)$
  1. Initial Condition, value of $X_j(t)$ at time $t=j$
     • $X_j(j) = 0$
     • node $j$ starts with 0 in-link when it’s first created at time $j$!
  2. Expected Change to $X_j(.)$ over time

$$\frac{p}{t} + \frac{(1 - p)X_j(t)}{t}.$$
Rich Get Richer- Cnt.

2. Expected Change to $X_j(.)$ over time
   - Probability that node $j$ gains an in-link in step $t+1$?
     - Happens if the newly created node $t+1$ points to node $j$.
     - Two cases:
       1. With probability $p$, node $t+1$ links to an earlier node chosen uniformly at random:
          - Thus, node $t+1$ links to node $j$ with probability $1/t$.
       2. With probability $1-p$, node $t+1$ links to an earlier node with probability proportional to the node's current number of in-links.
          - At time $t+1$:
            - total number of links in the network?
              - $t$ (one out of each prior node)
            - How many of them point to node $j$?
              - $X_j(t)$ (based on the definition)
            - Thus, node $t+1$ links to node $j$ with probability $X_j(t)/t$.

\[
\frac{p}{t} + \frac{(1-p)X_j(t)}{t}.
\]
Rich Get Richer- Cnt.

• Explain power laws using the Rich-Get-Richer model:
  ▫ Fraction of numbers receiving $k$ calls per day: $1/k^2$
  ▫ Fraction of books bought by $k$ people: $1/k^3$
  ▫ Fraction of papers with $k$ citations: $1/k^3$
  ▫ Fraction of cities with population $k$: $1/k^c$, $c$ constant
    • Cities grow in proportion to their size, simply as a result of people having children!

• Once an item becomes popular, the rich-get-richer dynamics are likely to push it even higher!

• The reasons of why power law is observed is more important than the simple fact that it's there.
Unpredictability of Rich-Get-Richer

• **Initially** rise to popularity is relatively fragile!
  ▫ Random effects early in the process play a role in the future popularity.

• If we replay the history, would the most popular items always be the same?
  ▫ Less likely,
  ▫ But, power-law distribution of popularity would probably be there each of these times!
Unpredictability of Rich-Get-Richer-Cnt.

• Created a music download site
  ▫ 48 obscure songs.
  ▫ Visitors / subjects could listen and download songs
  ▫ “download count" for each song is shown to visitors.
    • the number of times it had been downloaded thus far.
  ▫ Two Experiments:
    1. Visitors upon arrival were being assigned at random to one of eight “parallel” copies of the site.
    2. Visitors upon arrival were being assigned to a copy of the site in which “download counts” info was removed.

Unpredictability of Rich-Get-Richer-Cnt.

- Created a music download site
  - 48 obscure songs.
  - Visitors / subjects could listen and download songs
  - "download count" for each song is shown to visitors.
    - the number of times it had been downloaded thus far.

Figure S1: Schematic of the experimental design.

Unpredictability of Rich-Get-Richer-Cnt.

Experiment 1

- Notes:
  - Each visitor was given information only about the behavior of others in its copy of the site!
  - opportunity to contribute to rich-get-richer dynamics!
  - The parallel copies started out identically
    - same songs, download counts for all songs set to zero.
- Finding: The “market share” of the different songs varied considerably across different parallel sites.

Unpredictability of Rich-Get-Richer-Cnt.

Experiment 2

• Notes:
  ▫ No direct opportunity to contribute to rich-get-richer dynamics!

• Finding: there was significantly less variation in the “market share” of different songs.
  ▫ Feedback produced greater inequality in outcomes!

Unpredictability of Rich-Get-Richer Cnt.

• Popularity of a fake brand!

<table>
<thead>
<tr>
<th>How familiar are you with the following bands?</th>
<th>Don’t know it at all (% of subjects)</th>
<th>Heard of it (% of subjects)</th>
<th>Know it pretty well (% of subjects)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Bands</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guys on Couch</td>
<td>87.9</td>
<td>11.0</td>
<td>1.1</td>
</tr>
<tr>
<td>Grover Dill</td>
<td>88.4</td>
<td>10.5</td>
<td>1.1</td>
</tr>
<tr>
<td>Remnant Soldier</td>
<td>77.2</td>
<td>19.9</td>
<td>2.9</td>
</tr>
<tr>
<td>Fake Band</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peter on Fire</td>
<td>84.5</td>
<td>13.7</td>
<td>1.8</td>
</tr>
</tbody>
</table>

Table S4: Comparing the popularity of the potential bands from our sample to a fake band. Subjects reported being about as familiar with a fake band (Peter on Fire) as three potential bands from our sample. The high recognition rate for Remnant Soldier is likely a question ordering effect — it was asked immediately after the well known band U2.

The Long Tail

- The curve is tailing off slowly downward to the right
The Long Tail - Cnt.

• **Question**: Are most sales being generated by a
  • small set of popular items (**hits**), or
  • large set of less popular items (**niches**)?
• **Answer**: Order products by their sale volume!

Check if this curve is changing shape over time, adding more area under the right at the expense of the left!
• **Zipf's Law**: The frequency of the $k^{\text{th}}$ most common word in English is proportional to $1/k$.

Search and Recommendation

• Search Engine
  ▫ Page popularity based on click data
  ▫ Personalization
    • exposing people to items that may not be popular, but which match user interests as inferred from their history!
Questions?
Reading

• Ch.18 Power Laws and Rich-Get-Richer Phenomena [NCM]