IMA Tutorial (part III):
Generative and probabilistic models of data

May 5, 2003
Probabilistic generative models

- **Observation:** These distributions have the same form:
  1. Fraction of laptops that fail catastrophically during tutorials, by city
  2. Fraction of pairs of shoes that spontaneously de-sole during periods of stress, by city

- **Conclusion:** The distribution arises because the same stochastic process is at work, and this process can be understood beyond the context of each example
Models for Power Laws

- Power laws arise in many different areas of human endeavor, the “hallmark of human activity” (they also occur in nature)
- Can we find the underlying process (processes?) that accounts for this prevalence?

[Mitzenmacher 2002]
An Introduction to the Power Law

- **Definition:** A distribution is said to have a power law if \( \Pr[X \geq x] \rightarrow cx^{-\alpha} \)
- **Normally:** \( 0 < \alpha \leq 2 \) (Var(X) infinite)
- **Sometimes:** \( 0 < \alpha \leq 1 \) (Mean(X) infinite)

![Exponentially-decaying distribution](image1)

![Power law distribution](image2)
Early Observations: Pareto on Income

- [Pareto1897] observed that the random variable $I$ denoting the income of an individual has a power law distribution.
- More strongly, he observed that $\Pr[X>x] = (x/k)^{-\alpha}$
- For density function $f$, note that $\ln f(x) = (-\alpha - 1)\ln(x) + c$ for constant $c$
- Thus, in a plot of log(value) versus log(probability), power laws display a linear tail, while Pareto distributions are linear always.
Early Observations: Yule/Zipf

- [Yule26] observed (and explained) power laws in the context of number of species within a genus.
- [Zipf32] and [Estoup16] studied the relative frequency of words in natural language, beginning a cottage industry that continues to this day.
- A “Yule-Zipf” distribution is typically characterized by rank rather than value:
  - The $i$th most frequent word in English occurs with probability proportional to $1/i$.
- This characterization relies on finite vocabulary.
Early Observations: Lotka on Citations

- [Lotka25] presented the first occurrence of power laws in the context of graph theory, showing a power law for the indegree of the citation graph.
Ranks versus Values

- Commonly encountered phrasings of the power law in the context of word counts:
  1. \( \Pr[\text{word has count } \geq W] \) has some form
  2. Number of words with count \( \geq W \) has some form
  3. The frequency of the word with rank \( r \) has some form

  - The first two forms are clearly identical.
  - What about the third?
Equivalence of rank versus value formulation

- Given: number of words occurring $t$ times $\sim t^{-\alpha}$
- Approach:
  - Consider single most frequent word, with count $T$
  - Characterize word occurring $t$ times in terms of $T$
  - Approximate rank of words occurring $t$ times by counting number of words occurring at each more frequent count.
- Conclusion: Rank-$j$ word occurs $\alpha(cj + d)^{-1/(\alpha-1)}$ times (power law)
- But... high ranks correspond to low values – must keep straight the “head” and the “tail”

[Bookstein90, Adamic99]
Early modeling work

- The characterization of power laws is a limiting statement
- Early modeling work showed approaches that provide the correct form of the tail in the limit
- Later work introduced the rate of convergence of a process to its limiting distribution
A model of Simon

- Following Simon [1955], described in terms of word frequencies
- Consider a book being written. Initially, the book contains a single word, “the.”
- At time $t$, the book contains $t$ words. The process of Simon generates the $t+1^{st}$ word based on the current book.
Constructing a book: snapshot at time $t$

When in the course of human events, it becomes necessary...

Current word frequencies:

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>“the”</td>
<td>1000</td>
</tr>
<tr>
<td>2</td>
<td>“of”</td>
<td>600</td>
</tr>
<tr>
<td>3</td>
<td>“from”</td>
<td>300</td>
</tr>
<tr>
<td>…</td>
<td>“…”</td>
<td>…</td>
</tr>
<tr>
<td>4,791</td>
<td>“necessary”</td>
<td>5</td>
</tr>
<tr>
<td>…</td>
<td>“…”</td>
<td>“…”</td>
</tr>
<tr>
<td>11,325</td>
<td>“neccesary”</td>
<td>1</td>
</tr>
</tbody>
</table>

Let $f(i, t)$ be the number of words of count $i$ at time $t$
The Generative Model

- Assumptions:
  1. Constant probability that a neologism will be introduced at any timestep
  2. Probability of re-using a word of count $i$ is proportional to $if(i,t)$, that is, number of occurrences of count $i$ words.

- Algorithm:
  - With probability $\alpha$, a new word is introduced into the text
  - With remaining probability, a word with count $i$ is introduced with probability proportional to $if(i,t)$
Constructing a book: snapshot at time $t$

Current word frequencies:

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>“the”</td>
<td>1000</td>
</tr>
<tr>
<td>2</td>
<td>“of”</td>
<td>600</td>
</tr>
<tr>
<td>3</td>
<td>“from”</td>
<td>300</td>
</tr>
<tr>
<td>...</td>
<td>“...”</td>
<td>...</td>
</tr>
<tr>
<td>4,791</td>
<td>“necessary”</td>
<td>5</td>
</tr>
<tr>
<td>...</td>
<td>“...”</td>
<td>“...”</td>
</tr>
<tr>
<td>11,325</td>
<td>“neccesary”</td>
<td>1</td>
</tr>
</tbody>
</table>

Pr[“the”] = $(1 - \alpha) \frac{1000}{K}$

Pr[“of”] = $(1 - \alpha) \frac{600}{K}$

Pr[some count-1 word] = $(1 - \alpha) 1 * \frac{f(1,t)}{K}$

$K = \sum if(i,t)$

Let $f(i,t)$ be the number of words of count $i$ at time $t$
What’s going on?

One unique word (which occurs 1 or more times)

Each word in bucket $i$ occurs $i$ times in the current document
What’s going on?

With probability $\alpha$, a new word is introduced into the text.
What’s going on?

With probability $1 - \alpha$, an existing word is reused.

How many times do words in this bucket occur?
What’s going on?

Size of bucket 3 at time $t+1$ depends only on sizes of buckets 2 and 3 at time $t$.

Must show: fraction of balls in 3$^{rd}$ bucket approaches some limiting value.
Models for power laws in the web graph

- Retelling the Simon model: “preferential attachment”
  - Barabasi et al
  - Kumar et al
- Other models for the web graph:
  - [Aiello, Chung, Lu], [Huberman et al]
Why create such a model?

- Evaluate algorithms and heuristics
- Get insight into page creation
- Estimate hard-to-sample parameters
- Help understand web structure
- Cost modeling for query optimization
- To find “surprises” means we must understand what is typical.
## Random graph models

<table>
<thead>
<tr>
<th>Feature</th>
<th>G(n,p)</th>
<th>Web</th>
</tr>
</thead>
<tbody>
<tr>
<td>indeg &gt; 1000</td>
<td>0</td>
<td>100000</td>
</tr>
<tr>
<td>k23's</td>
<td>0</td>
<td>125000</td>
</tr>
<tr>
<td>4-cliques</td>
<td>0</td>
<td>many</td>
</tr>
</tbody>
</table>

Traditional random graphs [Bollobas 85] are not like the web!

Is there a better model?
Desiderata for a graph model

- Succinct description
- Insight into page creation
- No a priori set of "topics", but...
- ... topics should emerge naturally
- Reflect structural phenomena
- Dynamic page arrivals
- Should mirror web's "rich get richer" property, and manifest link correlation.
Page creation on the web

- Some page creators will link to other sites without regard to existing topics, but
- Most page creators will be drawn to pages covering existing topics they care about, and will link to pages within these topics

Model idea: new pages add links by "copying" them from existing pages
Generally, would require…

- Separate processes for:
  - Node creation
  - Node deletion
  - Edge creation
  - Edge deletion
A specific model

- Nodes are created in a sequence of discrete time steps
  - e.g. at each time step, a new node is created with \( d = \Theta(1) \) out-links

- Probabilistic copying
  - links go to random nodes with probability \( \alpha \)
  - copy \( d \) links from a random node with probability \( 1-\alpha \)
Example

With probability $\alpha$, it links to a uniformly-chosen page.
Example

Under copying, your rate of getting new inlinks is proportional to your in-degree. With probability \((1 - 1/g)\), it decides to copy a link.
Degree sequences in this model

\[
\Pr[\text{page has } k \text{ inlinks}] \approx k^{-(2-\alpha)}
\]

\[
(\alpha = 1/11 \text{ matches web})
\]

Heavy-tailed inverse polynomial degree sequences.
Pages like netscape and yahoo exist.
Many cores, cliques, and other dense subgraphs

[Kumar, Raghavan, Rajagopalan, Sivakumar, Tomkins, Upfal 2000]
Model extensions

- Component size distributions.
- More complex copying.
- Tighter lower tail bounds.
- More structure results.
A model of Mandelbrot

- Key idea: Generate frequencies of English words to maximize information transferred per unit cost
- Approach:
  - Say word $i$ occurs with probability $p(i)$
  - Set the transmission cost of word $i$ to be $\log(i)$
  - Average information per word: $-\sum p(i) \log(p(i))$
  - Cost of a word with probability $p(j)$: $\log(j)$
  - Average cost per word: $\sum p(j) \log(j)$
  - Choose probabilities $p(i)$ to maximize information/cost
- Result: $p(j) = c j^{-\beta}$

[Mandelbrot 1953]
Discussion of Mandelbrot’s model

- Trade-offs between communication cost \((\log(p(j)))\) and information.
- Are there other tradeoff-based models that drive similar properties?
### Heuristically Optimized Trade-offs

- **Goal:** construction of trees (note: models to generate trees with power law behavior were first proposed in [Yule26])
- **Idea:** New nodes must trade off connecting to nearby nodes, and connecting to central nodes.
- **Model:**
  - Points arrive uniformly within the unit square
  - New point arrives, and computes two measures for candidate connection points $j$
    - $d(j)$: distance from new node to existing node $j$ (“nearness”)
    - $h(j)$: distance from node $j$ to root of tree (“centrality”)
  - New destination chosen to minimize $\alpha d(j) + h(j)$
- **Result:** for a wide variety of values of $\alpha$, distribution of degrees has a power law

[Fabrikant, Koutsoupias, Papadimitriou 2002]
Monkeys on Typewriters

- Consider a creation model divorced from concerns of information and cost
- Model:
  - Monkey types randomly, hits space bar with probability $q$, character chosen uniformly with remaining probability
- Result:
  - Rank $j$ word occurs with probability $q^j \log(1-q) - 1 = c j^{\beta}$

[Miller 1962]
Other Distributions

- “Power law” means a clean characterization of a particular property on distribution upper tails
- Often used to mean “heavy tailed,” meaning bounded away from an exponentially decaying distribution
- There are other forms of heavy-tailed distributions
- A commonly-occurring example: lognormal distribution
Quick characterization of lognormal distributions

- Let X be a normally-distributed random variable
- Let Y = ln X
- Then Y is lognormal
- Common situations:
  - Multiplicative growth
- Concern: There is a growing sequence of papers dating back several decades questioning whether certain observed values are best described by power law or lognormal (or other) distributions.
One final direction...

- The Central Limit Theorem tells us how sums of independent random variables behave in the limit.
- Example: $\ln X_j = \ln X_0 + \sum \ln F_j$
- $X_j$ well-approximated by a lognormal variable.
- Thus, lognormal variables arise in situations of multiplicative growth.
- Examples in biology, ecology, economics, ...
- Example: [Huberman et al]: growth of web sites.
- Similarly: the product of lognormal variables.
- Each of these generative models is evolutionary.
- What is the role of time?