15-826: Multimedia Databases and Data Mining

Lecture #13: Power laws
Potential causes and explanations
C. Faloutsos

Must-read Material

• Power laws, Pareto distributions and Zipf’s law
Contemporary Physics 46, 323-351 (2005)

Optional Material

Outline

Goal: ‘Find similar / interesting things’
• Intro to DB
• Indexing - similarity search
• Data Mining

Indexing - Detailed outline
• primary key indexing
• secondary key / multi-key indexing
• spatial access methods
  – z-ordering
  – R-trees
  – misc
• fractals
  – intro
  – applications
• text
• fractals
  – intro
  – applications
• text

Indexing - Detailed outline
• fractals
  – intro
  – applications
  • disk accesses for R-trees (range queries)
  • dimensionality reduction
  • selectivity in M-trees
  • dim. curse revisited
  • “fat fractals”
  • quad-tree analysis [Gaede+]
  • nn queries [Belussi+]
This presentation

• Definitions
• Examples and counter-examples
• Generative mechanisms

Definition

• \( p(x) = C x^{-\alpha} \quad (x \geq x_{\text{min}}) \)
• Eg., prob( city pop. between \( x + dx \) )

\[
\log(p(x))
\]

\[
\begin{array}{c}
\log(x_{\text{min}}) \\
\log(x)
\end{array}
\]

For discrete variables

\( p_k = C k^{-\alpha} \quad (k > 0) \)

Or, the Yule distribution:

\[ p_k = C B(k, \alpha) \]
\[ B(k, \alpha) = \frac{\Gamma(k)\Gamma(\alpha)}{\Gamma(k+\alpha)} \approx k^{-\alpha} \]
Estimation for $a$

$$a = 1 + n \left[ \sum_{i=1}^{n} \ln\left( \frac{x_i}{x_{\min}} \right) \right]^{-1}$$

Examples

- Word frequencies
- Citations of scientific papers
- Web hits
- Copies of books sold
- Magnitude of earthquakes
- Diameter of moon craters
- …
[Newman 2005]

Rank-frequency plots
Or Cumulative D.F.

NOT following P.L.

‘abundance' of species

Number of addresses

Cumul. D.F.

Size of forest fires

This presentation

- Definitions
- Examples and counter-examples
- Generative mechanisms
  - Combination of exponentials
  - Inverse
  - Random walk
  - Yule distribution = CRP
  - Percolation
  - Self-organized criticality
  - Other
Combination of exponentials

Let $p(y) = e^{by}$

- e.g., radioactive decay, with half-life $-a$
- (= collection of people, playing russian roulette)

Let $x \sim e^{by}$

- (every time a person survives, we double his capital)

$$p(x) = p(y) \frac{dy}{dx} = 1/b \cdot x^{(-1+a/b)}$$

- I.e., the final capital of each person follows P.L.

Combination of exponentials

- Monkey on a typewriter:
- $m=26$ letters equiprobable;
- space bar has prob. $q_s$
- Freq(x-th most frequent word) = $x^{(-a)}$

see Eq. 47 of [Newman]:

$$a = \frac{2 \ln(m) - \ln (1 - q_s)}{[\ln m - \ln (1 - q_s)]}$$

Inverses of quantities

- $y$ follows $p(y)$ and goes through zero
- $x = 1/y$
- Then $p(x) = \cdots = -p(y)/x^2$
- For $y\sim0$, $x$ has power law tail.
This presentation

- Definitions
- Examples and counter-examples
- Generative mechanisms
  - Combination of exponentials
  - Inverse
  - Random walk
    - Yule distribution = CRP
    - Percolation
    - Self-organized criticality
    - Other

Random walks

Inter-arrival times PDF: \( p(t) \sim t^{-3/2} \)
Random walks

J. G. Oliveira & A.-L. Barabási Human Dynamics: The Correspondence Patterns of Darwin and Einstein.

This presentation

- Definitions
- Examples and counter-examples
- Generative mechanisms
  - Combination of exponentials
  - Inverse
  - Random walk
  - Yule distribution = CRP
  - Percolation
  - Self-organized criticality
  - Other

Yule distribution and CRP

Chinese Restaurant Process (CRP):
Newcomer to a restaurant
- Joins an existing table (preferring large groups)
- Or starts a new table/group of its own, with prob \(1/m\)
a.k.a.: rich get richer; Yule process
Yule distribution and CRP

Then:
\[
\text{Prob(} k \text{ people in a group)} = p_k
\]
\[
= (1 + 1/m) B( k, 2+1/m)
\]
\[
\sim k^{-2+1/m}
\]
(since \( B(a,b) \sim a^{-b} \): power law tail)

Yule distribution and CRP

• Yule process
• Gibrat principle
• Matthew effect
• Cumulative advantage
• Preferential attachement
• ‘rich get richer’

This presentation

• Definitions
• Examples and counter-examples
• Generative mechanisms
  – Combination of exponentials
  – Inverse
  – Random walk
  – Yule distribution = CRP
  – Percolation
  – Self-organized criticality
  – Other
Percolation and forest fires

A burning tree will cause its neighbors to burn next.

Which tree density \( p \) will cause the fire to last longest?
Percolation and forest fires

At $p_c \approx 0.593$:
- No characteristic scale;
- 'patches' of all sizes;
- Korcak-like 'law'.

This presentation

- Definitions
- Examples and counter-examples
- Generative mechanisms
  - Combination of exponentials
  - Inverse
  - Random walk
  - Yule distribution = CRP
  - Percolation
  - Self-organized criticality
  - Other
Self-organized criticality

- Trees appear at random (e.g., seeds, by the wind)
- Fires start at random (e.g., lightning)
- Q1: What is the distribution of size of forest fires?

Self-organized criticality

- A1: Power law-like

Self-organized criticality

- Trees appear at random (e.g., seeds, by the wind)
- Fires start at random (e.g., lightning)
- Q2: what is the average density?
Self-organized criticality

• A2: the critical density \( p_c \approx 0.593 \)

Self-organized criticality

• [Bak]: size of avalanches \( \sim \) power law:
• Drop a grain randomly on a grid
• It causes an avalanche if \( \text{height}(x,y) \) is >1 higher than its four neighbors


This presentation

• Definitions
• Examples and counter-examples
• Generative mechanisms
  – Combination of exponentials
  – Inverse
  – Random walk
  – Yule distribution = CRP
  – Percolation
  – Self-organized criticality
  – Other
Other

- Random multiplication
- Fragmentation
- Lead to lognormals (~ look like power laws)

Others

Random multiplication:
- Start with C dollars; put in bank
- Random interest rate s(t) each year t
- Each year t: C(t) = C(t-1) * (1 + s(t))

- Log(C(t)) = log(C) + log(1 + s(t)) ...
- Gaussian
- Thus C(t) = exp(Gaussian)
- By definition, this is Lognormal
Others

Lognormal:

pdf

0

$

1c

log (pdf)

parabola

log ($)

log (pdf)

parabola

log ($)
Other

- Random multiplication

- Fragmentation

  - lead to lognormals (~ look like power laws)

Other

- Stick of length 1

- Break it at a random point x (0<x<1)

- Break each of the pieces at random

- Resulting distribution: lognormal (why?)

Conclusions

- Power laws and power-law like distributions appear often

- (fractals/self similarity -> power laws)

- Exponentiation/inversion

- Yule process / CRP / rich get richer

- Criticality/percolation/phase transitions

- Fragmentation -> lognormal ~ P.L.