15-826: Multimedia Databases and Data Mining

Lecture #12: Power laws
Potential causes and explanations

C. Faloutsos

Must-read Material


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
  – Why so many power laws?
This presentation

- Definitions
  - Clarification: 3 forms of P.L.
  - Examples and counter-examples
  - Generative mechanisms

Definition

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

\[
\log(p(x)) = \log(x) - \log(x_{\text{min}})
\]

For discrete variables

\[
p_k = C k^{-a} \quad (k > 0)
\]

Or, the Yule distribution:

\[
p_k = C \frac{1}{B(k,a)}
\]

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

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

3 versions of P.L.

PDF = frequency-count plot
Zipf plot = Rank-frequency
NCDF = CCDF

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

Prob( area = x )
area

Prob( area >= x )

Details, and proof sketches:
More power laws: areas – Korcak’s law

Scandinavian lakes area vs complementary cumulative count (log-log axes)

3 versions of P.L.

PDF

Prob( area = x )

NCDF = CCDF

Prob( area >= x )

NCDF = CCDF

Prob( area >= x )
3 versions of P.L.

PDF

NCDF = CCDF

\[
\text{Prob}( \text{area } = x ) = -a - 1
\]

\[
\text{Prob}( \text{area } \geq x ) = -a
\]

Zipf plot = Rank-frequency

\[
\text{NCDF} = \text{CCDF}
\]
3 versions of P.L.

PDF

Zipf plot = NCDF = CCDF
Rank-frequency

Prob( area = x )
area

Prob( area >= x )

Prob( area = x )

area

Prob( area >= x )

Prob( area = x )

area

Prob( area >= x )
### 3 versions of P.L.

PDF

$\text{Zipf plot} = \frac{1}{x^{1/a}}$

NCDF = CCDF

Prob( area = x )

\[ \text{Prob (area} \geq x) \]

- $a - 1$

\[ x \]

\[ \text{frequency} \]

\[ \text{area} \]

\[ \text{rank} \]

- $1/a$

\[ x \]

\[ \text{Prob (area} \geq x) \]

- $a$

\[ x \]

\[ \text{Rank-frequency} \]

Zipf plot = Rank-frequency

NCDF = CCDF

Prob( area = x )

\[ \text{Prob (area} = x) \]

- $a - 1$

\[ x \]

\[ \text{frequency} \]

\[ \text{count} \]

- $1/a$

\[ x \]

\[ \text{frequency} \]

\[ \text{Prob (area} \geq x) \]

- $a$

\[ x \]

\[ \text{Rank-frequency} \]

Zipf plot = Rank-frequency

NCDF = CCDF

Prob( area = x )

\[ \text{Prob (area} = x) \]

- $a - 1$

\[ x \]

\[ \text{frequency} \]

\[ \text{count} \]

- $1/a$

\[ x \]

\[ \text{frequency} \]

\[ \text{Prob (area} \geq x) \]

- $a$

\[ x \]

\[ \text{Rank-frequency} \]

Zipf plot = Rank-frequency

NCDF = CCDF

Prob( area = x )

\[ \text{Prob (area} = x) \]

- $a - 1$

\[ x \]

\[ \text{frequency} \]

\[ \text{count} \]

- $1/a$

\[ x \]

\[ \text{frequency} \]
Sanity check:

- Zipf showed that if
  - Slope of rank-frequency is \(-1\)
  - Then slope of freq-count is \(-2\)

- Check it!
3 versions of P.L.

PDF = frequency-count plot
Zipf plot = Rank-frequency
NCDF = CCDF

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

\[
\text{Prob( area } = x ) \quad \text{area} \quad \text{Prob( area } \geq x )
\]

This presentation

- Definitions
- Clarification: 3 forms of P.L.
- Examples and counter-examples
- Generative mechanisms

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

Let \( p(y) = e^{ay} \)
- eg., 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)
  \( p(x) = p(y) \cdot \frac{dy}{dx} = \frac{1}{b} x^{(-1-a/b)} \)
- ie, the final capital of each person follows P.L.

Combination of exponentials

- Monkey on a typewriter:
  - \( m = 26 \) letters equiprobable;
  - space bar has prob. \( q_s \)
  \textbf{THEN:} \( \text{Freq}(x\text{-th most frequent word}) = x^{(-a)} \)
  see Eq. 47 of [Newman]:
  \[ a = \frac{2 \ln(m) - \ln (1-q_s)}{\ln m - \ln (1-q_s)} \]

This presentation

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

- \( y \) follows \( p(y) \) and goes through zero
- \( x = 1/y \)
- Then \( p(x) = \ldots = -p(y)/x^2 \)
- For \( y \to 0 \), \( x \) has power law tail.

This presentation

- Definitions
- Clarification
- 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: \( p(t) \sim \ldots \)
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. 

Figure 1: 

This presentation

- Definitions - clarification
- Examples and counter-examples
- Generative mechanisms
  - Combination of exponentials
  - Inverse
  - Random walk
  - Yale 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

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

Yule distribution and CRP

• Yule process
• Gibrat principle
• Matthew effect
• Cumulative advantage
• Preferential attachment
• ‘rich get richer’
This presentation

- Definitions - clarification
- 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 $p_c \sim 0.593$:
- No characteristic scale;
- 'patches' of all sizes;
- Korcak-like 'law'.
This presentation

• Definitions - clarification
• 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

CCDF

Area of cluster s
Self-organized criticality

• Trees appear at random (eg., seeds, by the wind)
• Fires start at random (eg., 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 - clarification
- 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(..) + log(..) ... -> Gaussian
Others

Random multiplication:

- \( \log(C(t)) = \log(C) + \log(...) + \log(...) \ldots \rightarrow \text{Gaussian} \)

- Thus \( C(t) = \exp(\text{Gaussian}) \)

- By definition, this is Lognormal

Lognormal:

\[
\log(p) = e^h
\]

\( h = \text{body height} \)

Lognormal:

\[
\log(p) = \text{parabola}
\]

\( \log(S) \)
Others

Lognormal:

\[ \log(\text{pdf}) \]

parabola

$1c$ \hspace{1cm} \log($) 

Other

- Random multiplication
- Fragmentation
- \( \rightarrow \) 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.

References

- Zipf, Power-laws, and Pareto - a ranking tutorial, Lada A. Adamic
- Human Behavior and Principle of Least Effort, G.K. Zipf, Addison Wesley (1949)