



**15-826: Multimedia Databases
and Data Mining**

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



Must-read Material

- Mark E.J. Newman: *Power laws, Pareto distributions and Zipf's law*, Contemporary Physics 46, 323-351 (2005), or <http://arxiv.org/abs/cond-mat/0412004v3>

15-826 Copyright: C. Faloutsos (2012) 2



Optional Material

- (optional, but very useful: Manfred Schroeder *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise* W.H. Freeman and Company, 1991) – ch. 15.

15-826 Copyright: C. Faloutsos (2012) 3



CMU SCS

Outline

Goal: 'Find **similar / interesting** things'

- Intro to DB
- ➔ • Indexing - similarity search
- Data Mining

15-826 Copyright: C. Faloutsos (2012) 4



CMU SCS

Indexing - Detailed outline

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

15-826 Copyright: C. Faloutsos (2012) 5



CMU SCS

Indexing - Detailed outline

- fractals
 - intro
 - applications
 - disk accesses for R-trees (range queries)
 - ...
 - dim. curse revisited
 - ...
- ➔ – Why so many power laws?

15-826 Copyright: C. Faloutsos (2012) 6

CMU SCS

This presentation

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

15-826 Copyright: C. Faloutsos (2012) 7

CMU SCS

Definition

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

$\log(p(x))$

$\log(x_{\min}) \quad \log(x)$

15-826 Copyright: C. Faloutsos (2012) 8

CMU SCS

For discrete variables

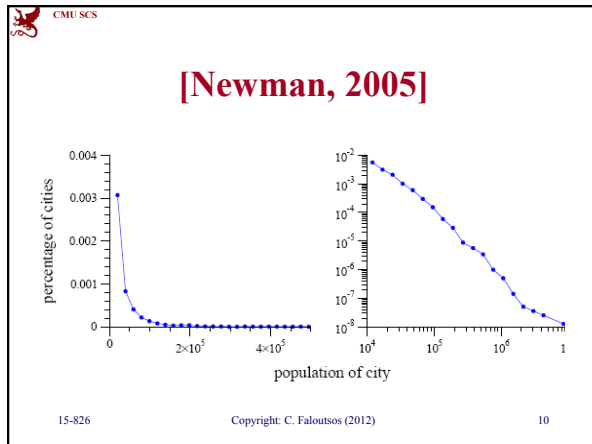
$$p_k = C k^{-a} \quad (k > 0)$$

Or, the Yule distribution:

$$p_k = C B(k, a)$$

$$B(k, a) = \Gamma(k)\Gamma(a)/\Gamma(k + a) \approx k^{-a}$$

15-826 Copyright: C. Faloutsos (2012) 9



CMU SCS

Estimation for a

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

15-826 Copyright: C. Faloutsos (2012) 11

- CMU SCS
- ### This presentation
- Definitions
 - ➔ • Clarification: 3 forms of P.L.
 - Examples and counter-examples
 - Generative mechanisms
- 15-826 Copyright: C. Faloutsos (2012) 12

CMU SCS

Jumping to the conclusion:

15-826 Copyright: C. Faloutsos (2011) 13

CMU SCS

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)

$-a-1$

x

area

$-1/a$

rank

Prob(area >= x)

$-a$

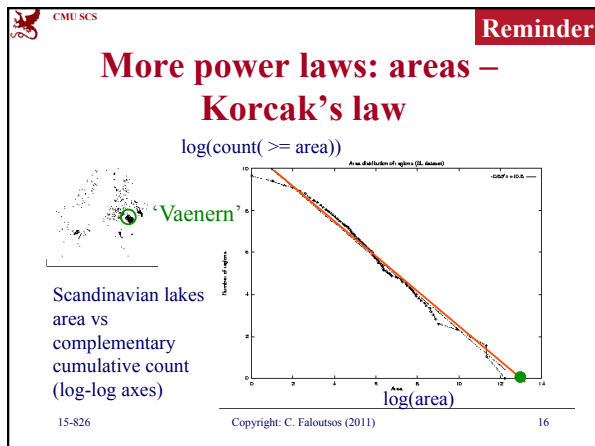
x

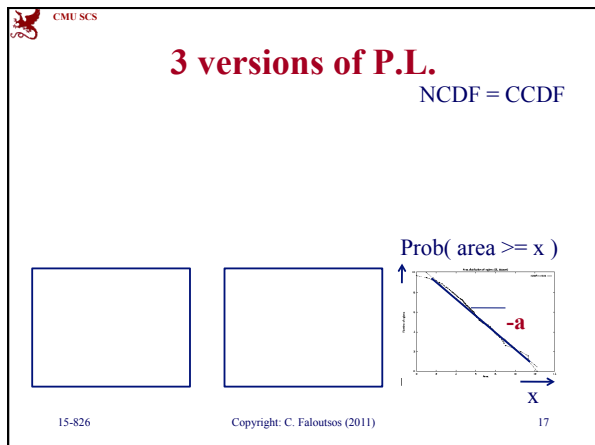
15-826 Copyright: C. Faloutsos (2011) 14

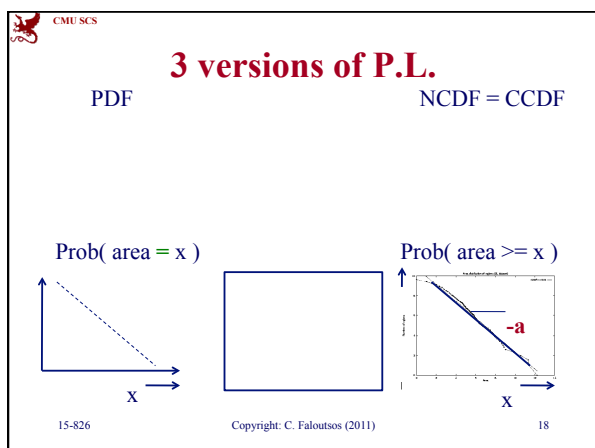
CMU SCS

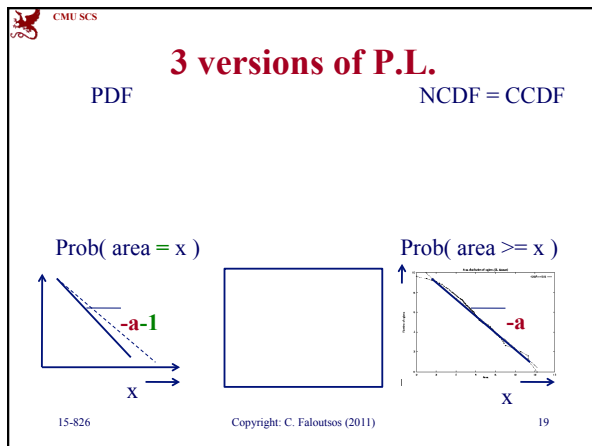
Details, and proof sketches:

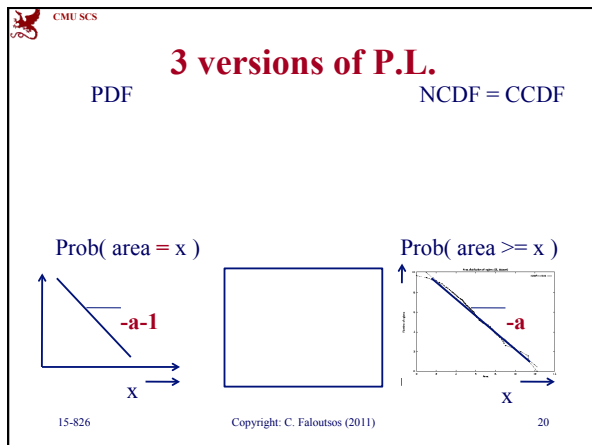
15-826 Copyright: C. Faloutsos (2011) 15

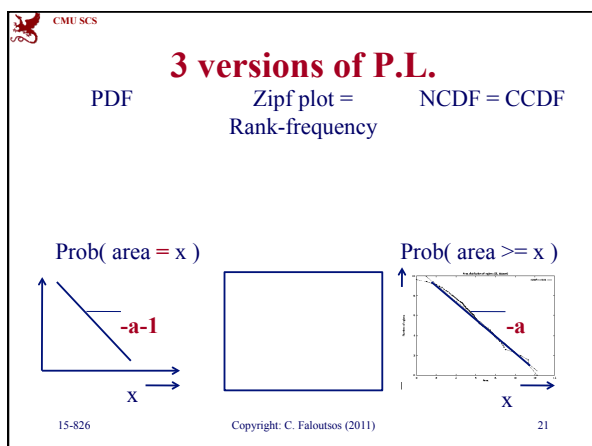


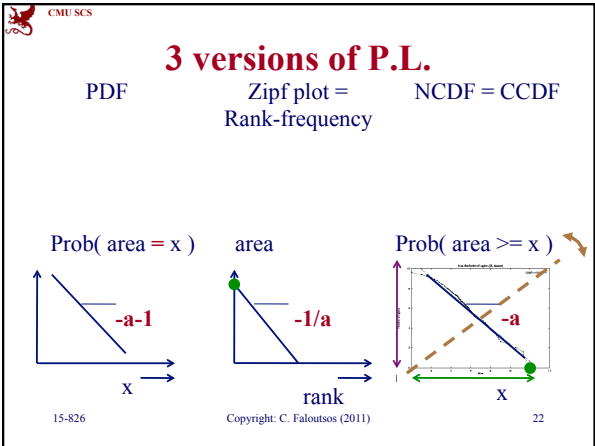


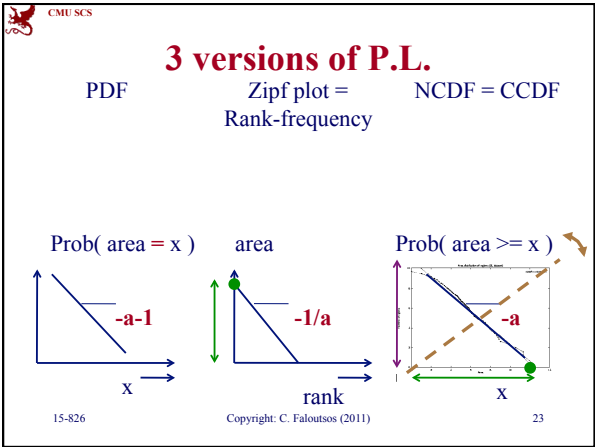


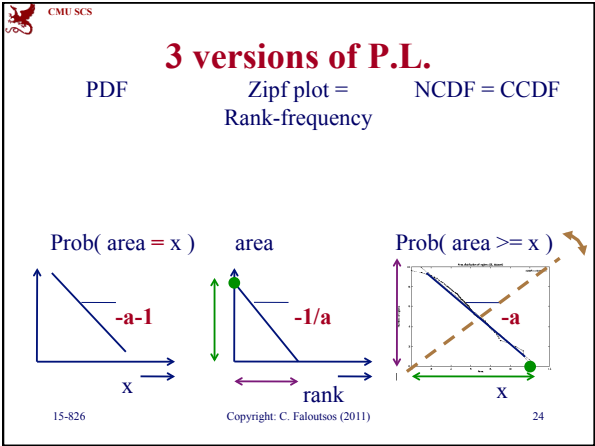


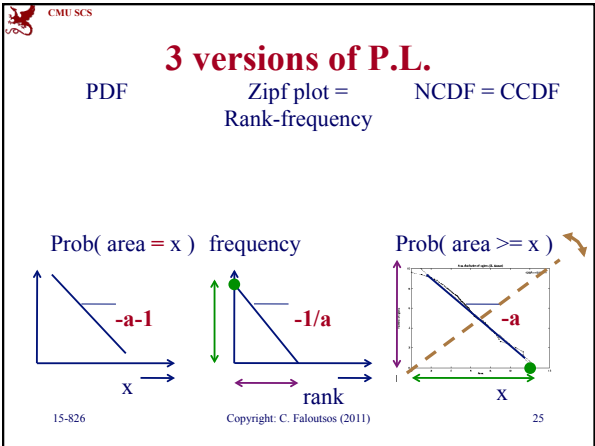


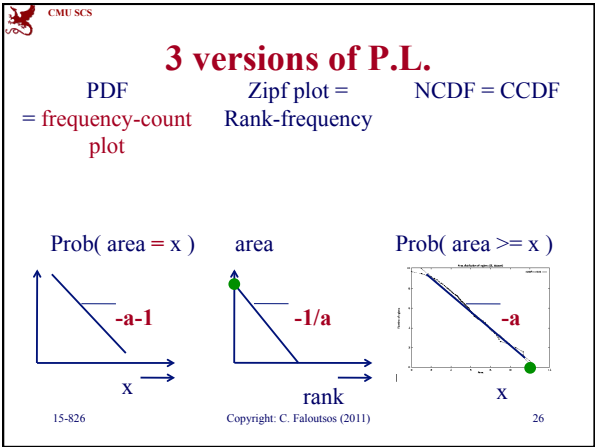


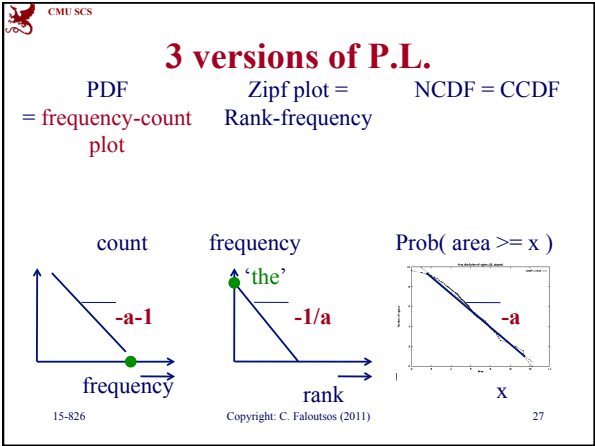












CMU SCS

Sanity check:

- Zipf showed that if
 - Slope of rank-frequency is -1
 - Then slope of freq-count is -2
- Check it!

15-826 Copyright: C. Faloutsos (2011) 28

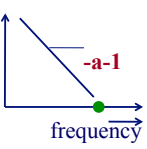
CMU SCS

3 versions of P.L.

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

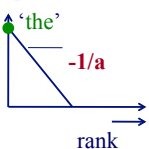
slope = -2 $\iff$ slope = -1

count



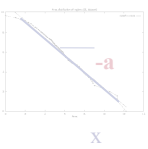
frequency

frequency



rank

Prob(area $\geq x$)



x

15-826 Copyright: C. Faloutsos (2011) 29

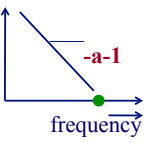
CMU SCS

3 versions of P.L.

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

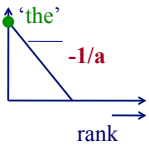
✓ slope = -2 $\iff$ slope = -1

count



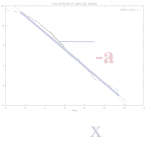
frequency

frequency



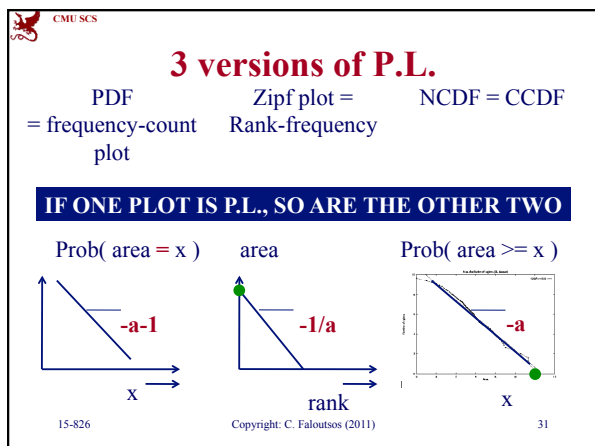
rank

Prob(area $\geq x$)



x

15-826 Copyright: C. Faloutsos (2011) 30



CMU SCS

This presentation

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

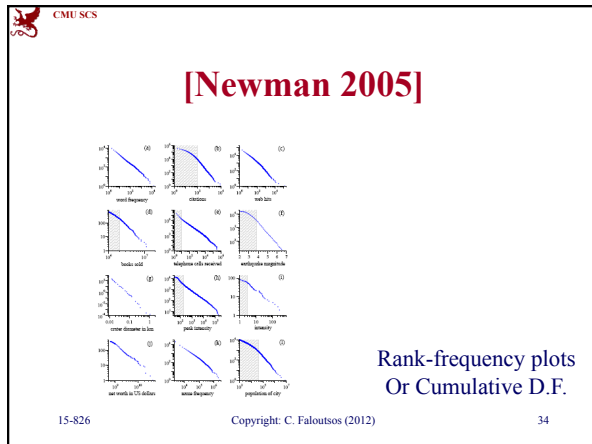
15-826 Copyright: C. Faloutsos (2012) 32

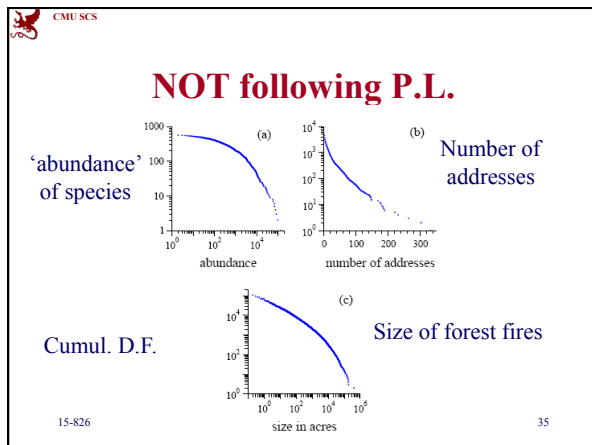
CMU SCS

Examples

- Word frequencies
- Citations of scientific papers
- Web hits
- Copies of books sold
- Magnitude of earthquakes
- Diameter of moon craters
- ...

15-826 Copyright: C. Faloutsos (2012) 33





CMU SCS

This presentation

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

15-826 Copyright: C. Faloutsos (2012) 36



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)

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

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

15-826 Copyright: C. Faloutsos (2012) 37



Combination of exponentials

- Monkey on a typewriter:
- $m=26$ letters equiprobable;
- space bar has prob. q_s

THEN: Freq(x -th most frequent word) $= x^{(-a)}$
 see Eq. 47 of [Newman]:

$$a = [2 \ln(m) - \ln(1 - q_s)] / [\ln m - \ln(1 - q_s)]$$

15-826 Copyright: C. Faloutsos (2012) 38



This presentation

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

15-826 Copyright: C. Faloutsos (2012) 39

CMU SCS

Inverses of quantities

- y follows $p(y)$ and goes through zero $y \rightarrow \text{speed}$
- $x = 1/y$ $x \rightarrow \text{travel time}$
- Then $p(x) = \dots = -p(y) / x^2$
- For $y \sim 0$, x has power law tail.

15-826 Copyright: C. Faloutsos (2012) 40

CMU SCS

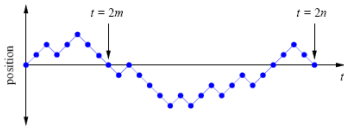
This presentation

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

15-826 Copyright: C. Faloutsos (2012) 41

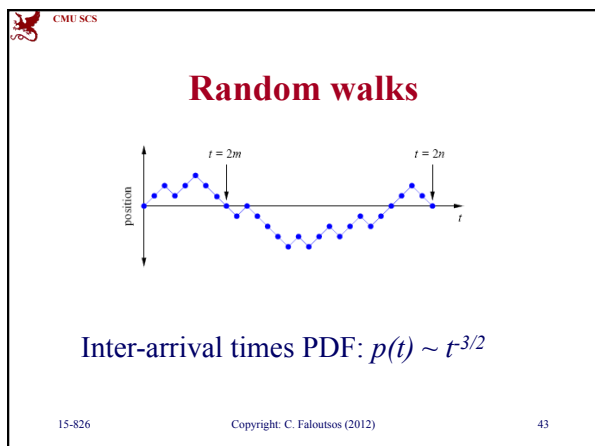
CMU SCS

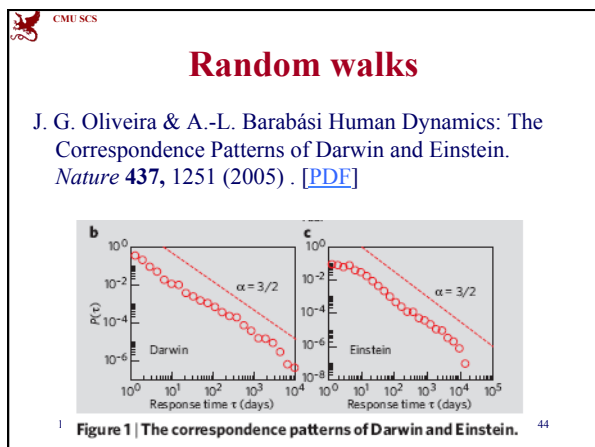
Random walks



Inter-arrival times PDF: $p(t) \sim ??$

15-826 Copyright: C. Faloutsos (2012) 42





CMU SCS

This presentation

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

15-826 Copyright: C. Faloutsos (2012) 45



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

15-826 Copyright: C. Faloutsos (2012) 46



Yule distribution and CRP

Then:
Prob(k 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)

15-826 Copyright: C. Faloutsos (2012) 47



Yule distribution and CRP

- Yule process
- Gibrat principle
- Matthew effect
- Cumulative advantage
- Preferential attachment
- ‘rich get richer’

15-826 Copyright: C. Faloutsos (2012) 48

CMU SCS

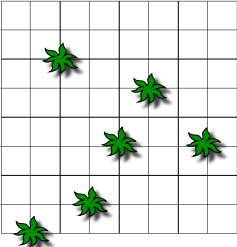
This presentation

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

15-826 Copyright: C. Faloutsos (2012) 49

CMU SCS

Percolation and forest fires



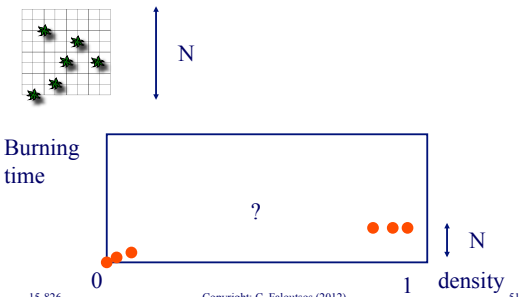
A burning tree will cause its neighbors to burn next.

Which tree density p will cause the fire to last longest?

15-826 Copyright: C. Faloutsos (2012) 50

CMU SCS

Percolation and forest fires

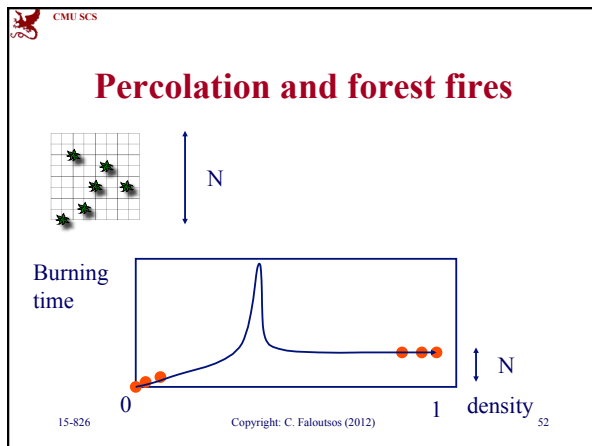


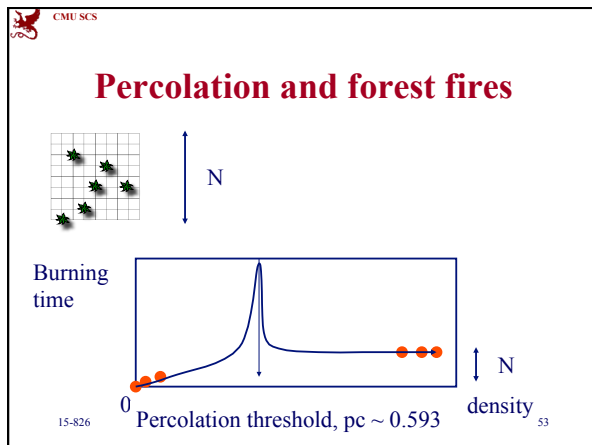
Burning time

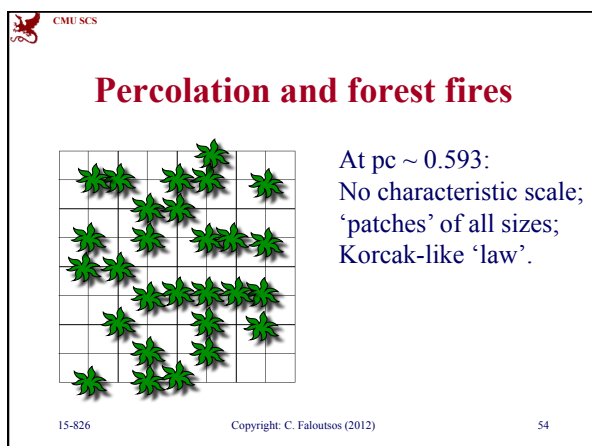
?

0 1 density

15-826 Copyright: C. Faloutsos (2012) 51







CMU SCS

This presentation

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

15-826

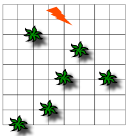
Copyright: C. Faloutsos (2012)

55

CMU SCS

Self-organized criticality

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



15-826

Copyright: C. Faloutsos (2012)

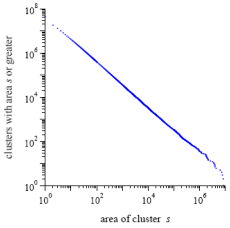
56

CMU SCS

Self-organized criticality

- A1: Power law-like

CCDF



15-826

Copyright: C. Faloutsos

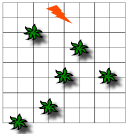
Area of cluster s

57

CMU SCS

Self-organized criticality

- Trees appear at random (eg., seeds, by the wind)
- Fires start at random (eg., lightning)
- Q2: what is the average density?



15-826 Copyright: C. Faloutsos (2012) 58

CMU SCS

Self-organized criticality

- A2: the critical density $p_c \sim 0.593$

15-826 Copyright: C. Faloutsos (2012) 59

CMU SCS

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

[Per Bak: *How Nature works*, 1996]

15-826 Copyright: C. Faloutsos (2012) 60



CMU SCS

This presentation

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

15-826 Copyright: C. Faloutsos (2012) 61



CMU SCS

Other

- Random multiplication
- Fragmentation

-> lead to lognormals (~ look like power laws)

15-826 Copyright: C. Faloutsos (2012) 62



CMU SCS

Others

Random multiplication:

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

• $\log(C(t)) = \log(C) + \log(\dots) + \log(\dots) \dots \rightarrow$
Gaussian

15-826 Copyright: C. Faloutsos (2012) 63



CMU SCS

Others

Random multiplication:

- $\text{Log}(C(t)) = \text{log}(C) + \text{log}(\dots) + \text{log}(\dots) \dots \rightarrow \text{Gaussian}$
- Thus $C(t) = \exp(\text{Gaussian})$
- By definition, this is Lognormal

15-826

Copyright: C. Faloutsos (2012)

64

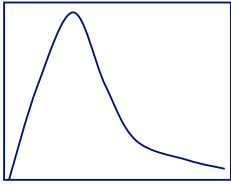


CMU SCS

Others

Lognormal:

pdf



pdf



$h = \text{body height}$

15-826

Copyright: C. Faloutsos (2012)

65

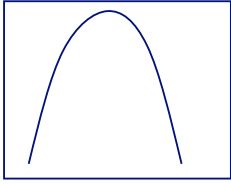


CMU SCS

Others

Lognormal:

log(pdf)



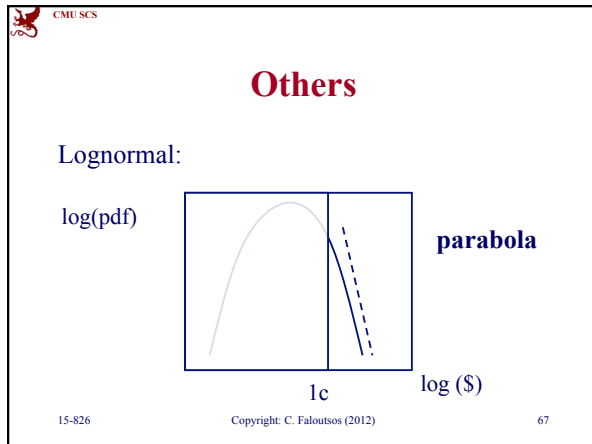
parabola

15-826

Copyright: C. Faloutsos (2012)

66

22



CMU SCS

Other

- Random multiplication
- ➔ Fragmentation
- > lead to lognormals (~ look like power laws)

15-826 Copyright: C. Faloutsos (2012) 68

CMU SCS

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?)

15-826 Copyright: C. Faloutsos (2012) 69

CMU SCS

Fragmentation -> lognormal

15-826 Copyright: C. Faloutsos (2012) 70

CMU SCS

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.

15-826 Copyright: C. Faloutsos (2012) 71

CMU SCS

References

- *Zipf, Power-laws, and Pareto - a ranking tutorial*, Lada A. Adamic
www.hpl.hp.com/research/idl/papers/ranking/ranking.html
- L.A. Adamic and B.A. Huberman, *'Zipf's law and the Internet'*, *Glottometrics* 3, 2002, 143-150
- *Human Behavior and Principle of Least Effort*, G.K. Zipf, Addison Wesley (1949)

15-826 Copyright: C. Faloutsos (2011) 72
