

CarnegieMellon

15-826: Multimedia Databases and Data Mining

Lecture #11: Power laws
Potential causes and explanations

C. Faloutsos

CarnegieMellon

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>

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.

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


Indexing - Detailed outline

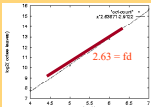
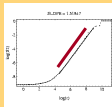
- fractals
 - intro
 - applications
 - disk accesses for R-trees (range queries)
 - ...
 - dim. curse revisited
 - ...
-  – Why so many power laws?

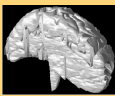

CarnegieMellon

Problem



- Why so many power-laws? 









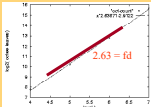
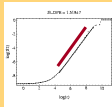
15-826 Copyright (c) 2019 C. Faloutsos 7

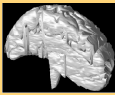

CarnegieMellon

Conclusion



- Why so many power-laws? 
- Many reasons:
 - Self similarity
 - rich-get-richer
 - etc

15-826 Copyright (c) 2019 C. Faloutsos 8

CarnegieMellon

This presentation

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

15-826

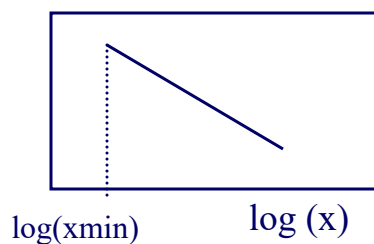
Copyright (c) 2019 C. Faloutsos

9

CarnegieMellon

Definition

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

 $\log(p(x))$ 

15-826

Copyright (c) 2019 C. Faloutsos

10

CarnegieMellon

For discrete variables

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

Or, the Yule distribution:

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

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

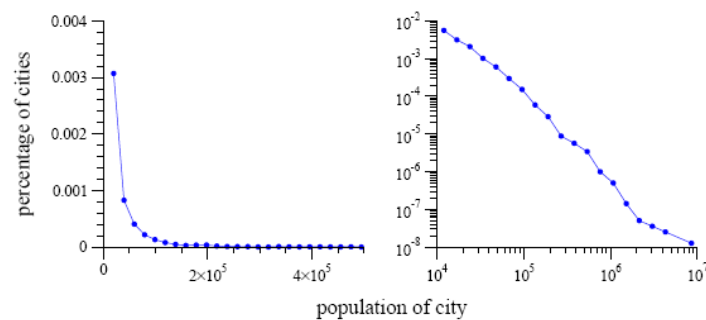
15-826

Copyright (c) 2019 C. Faloutsos

11

CarnegieMellon

[Newman, 2005]

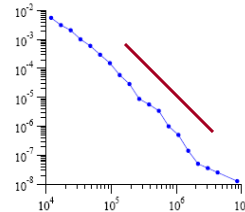


15-826

Copyright (c) 2019 C. Faloutsos

12

Estimation for a



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

This presentation

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

CarnegieMellon

Jumping to the conclusion:

15-826
Copyright (c) 2019 C. Faloutsos
15

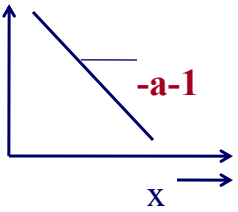
CarnegieMellon

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)

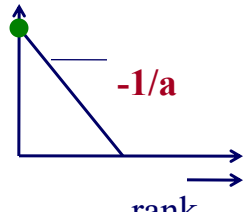


$-a-1$

x

15-826

area

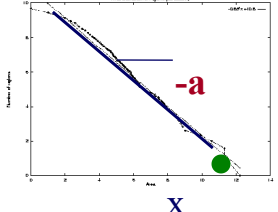


$-1/a$

rank

Copyright (c) 2019 C. Faloutsos

Prob(area \geq x)



$-a$

x

16

CarnegieMellon


Details, and proof sketches:

15-826 Copyright (c) 2019 C. Faloutsos 17

CarnegieMellon Reminder

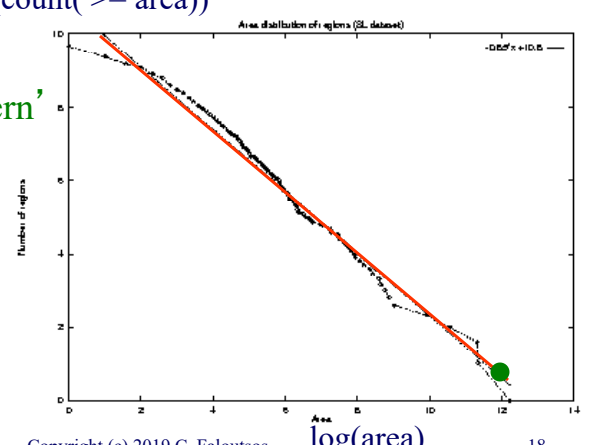
More power laws: areas – Korcak's law

$\log(\text{count}(\geq \text{area}))$



'Vaenem'

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




15-826 Copyright (c) 2019 C. Faloutsos 18


CarnegieMellon

3 versions of P.L.

NCDF = CCDF

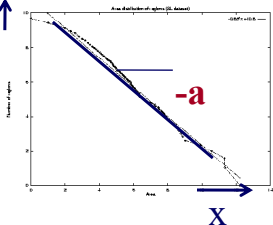


15-826



Copyright (c) 2019 C. Faloutsos

Prob(area $\geq x$)



19

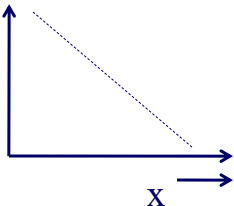
CarnegieMellon

3 versions of P.L.


PDF

NCDF = CCDF

Prob(area = x)

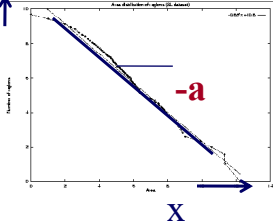


15-826

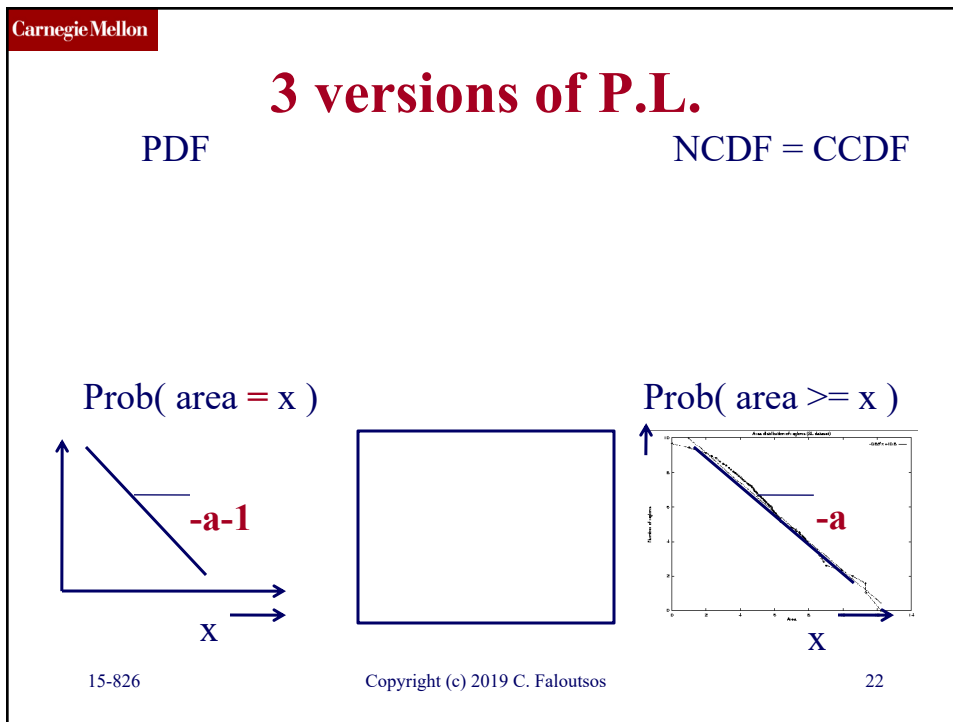
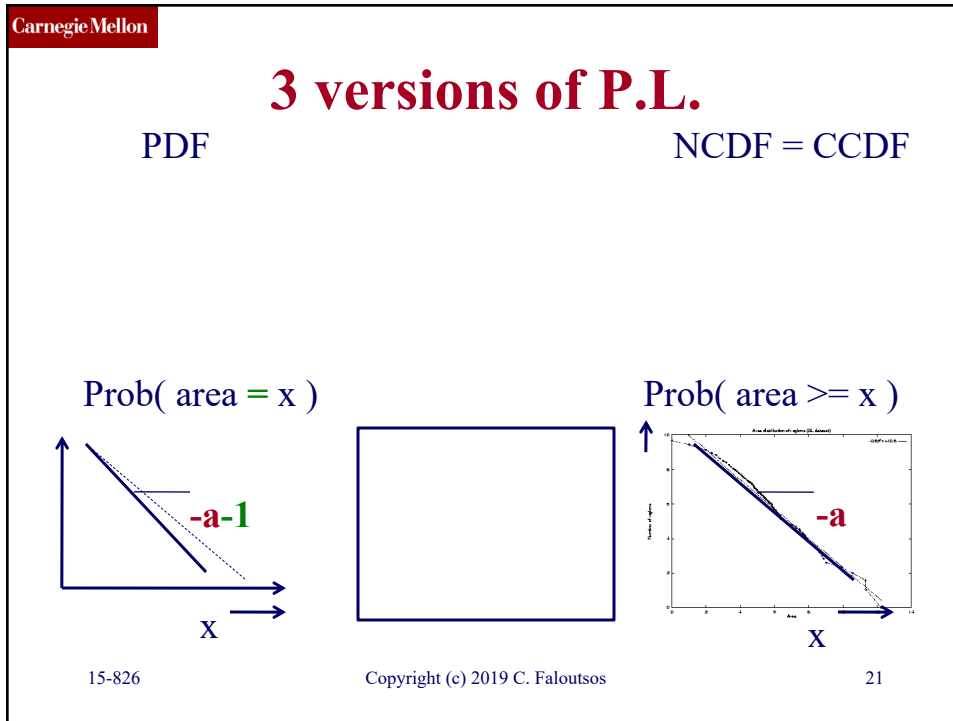


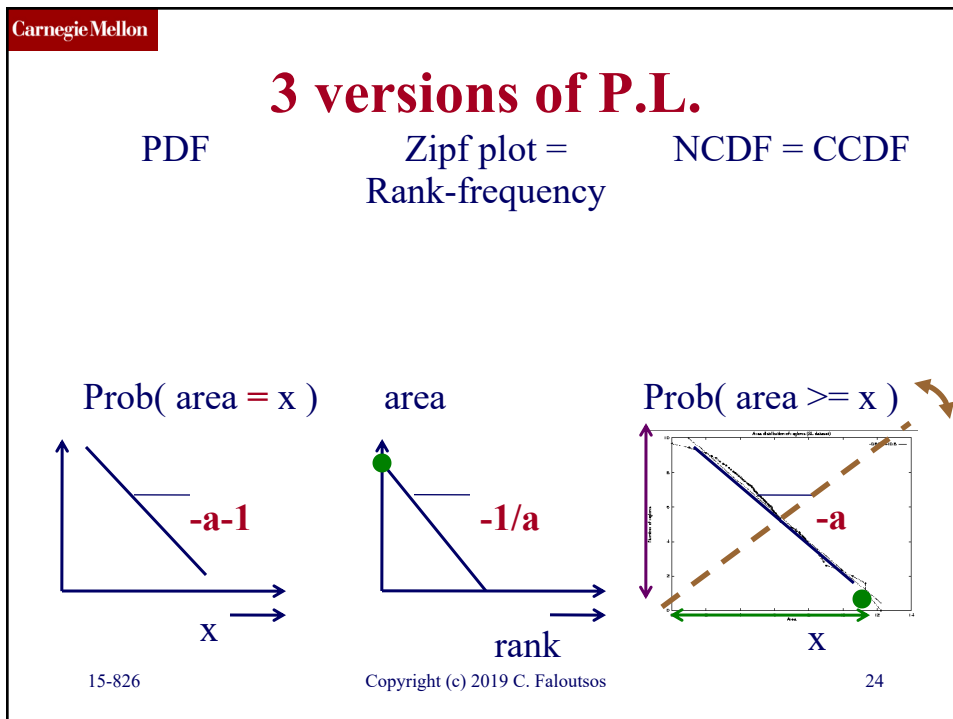
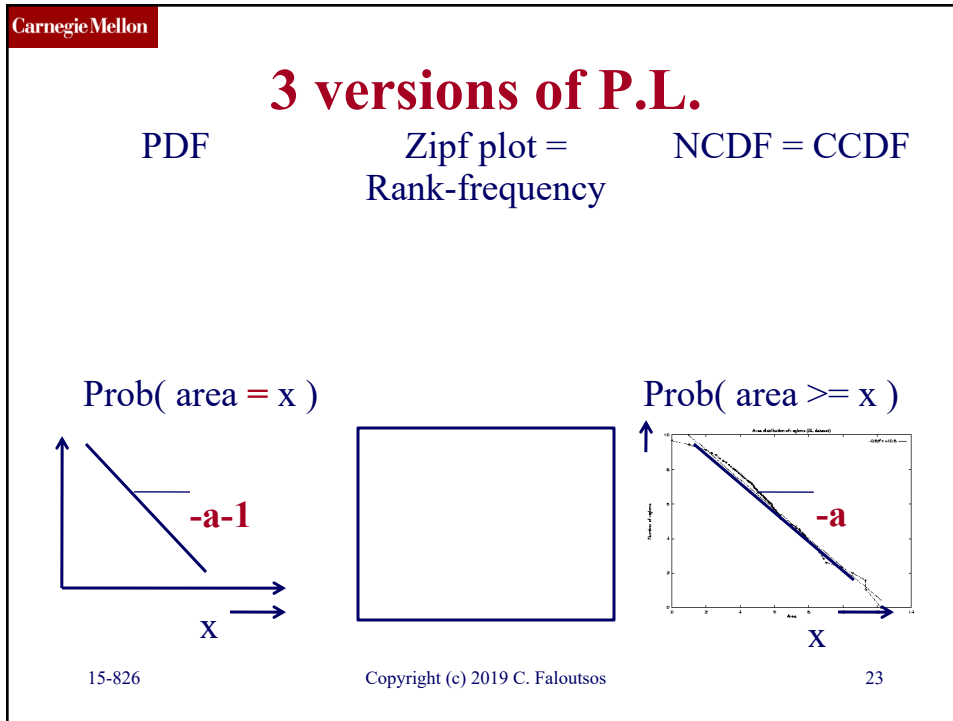
Copyright (c) 2019 C. Faloutsos

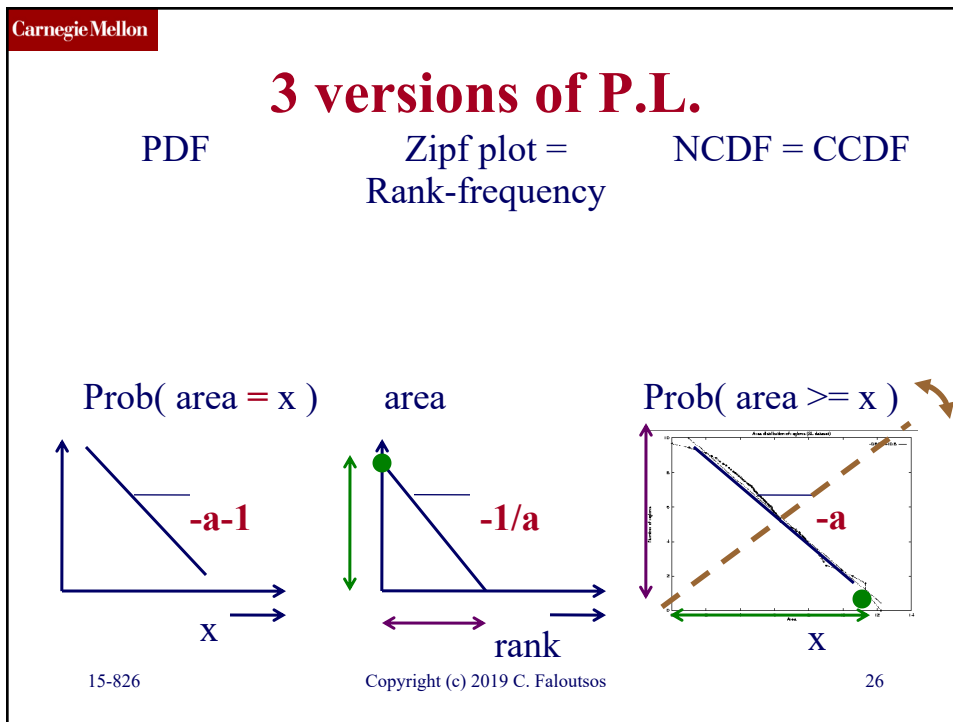
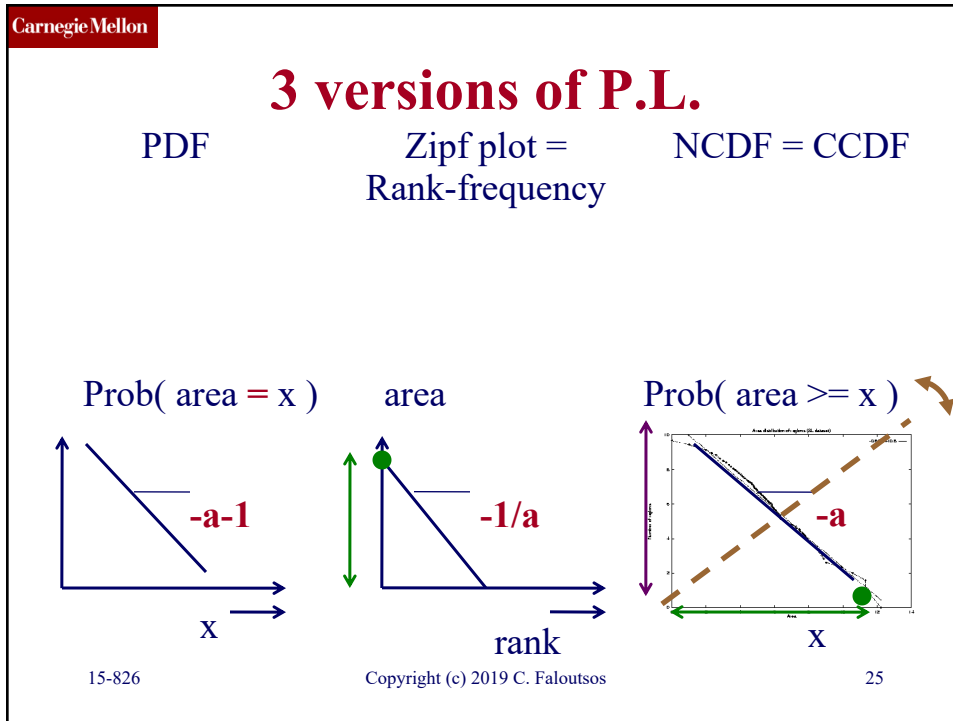
Prob(area $\geq x$)

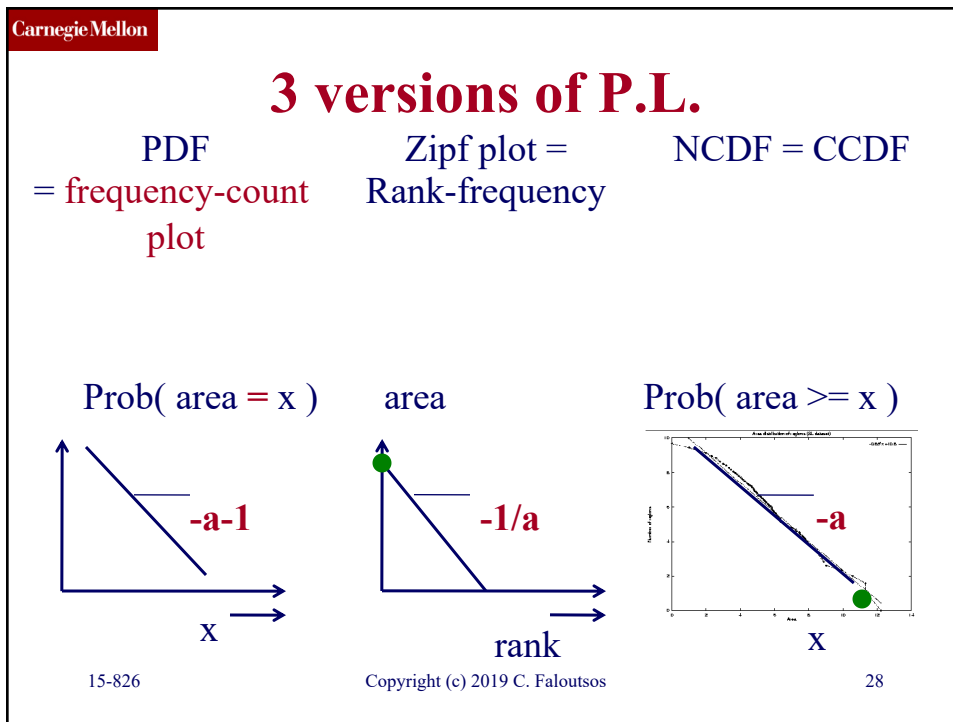
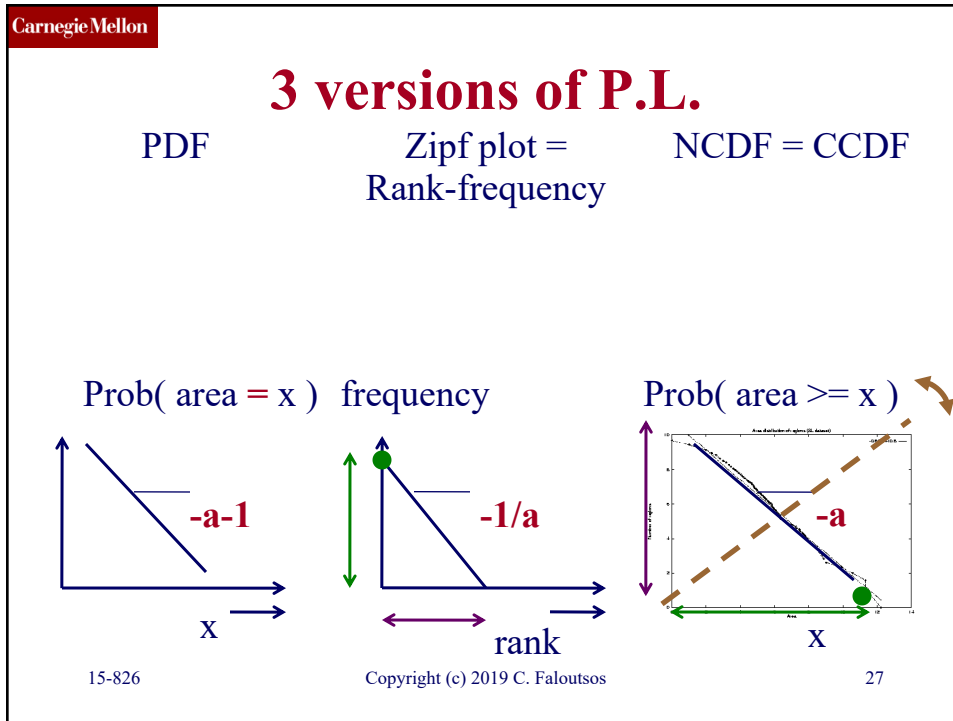


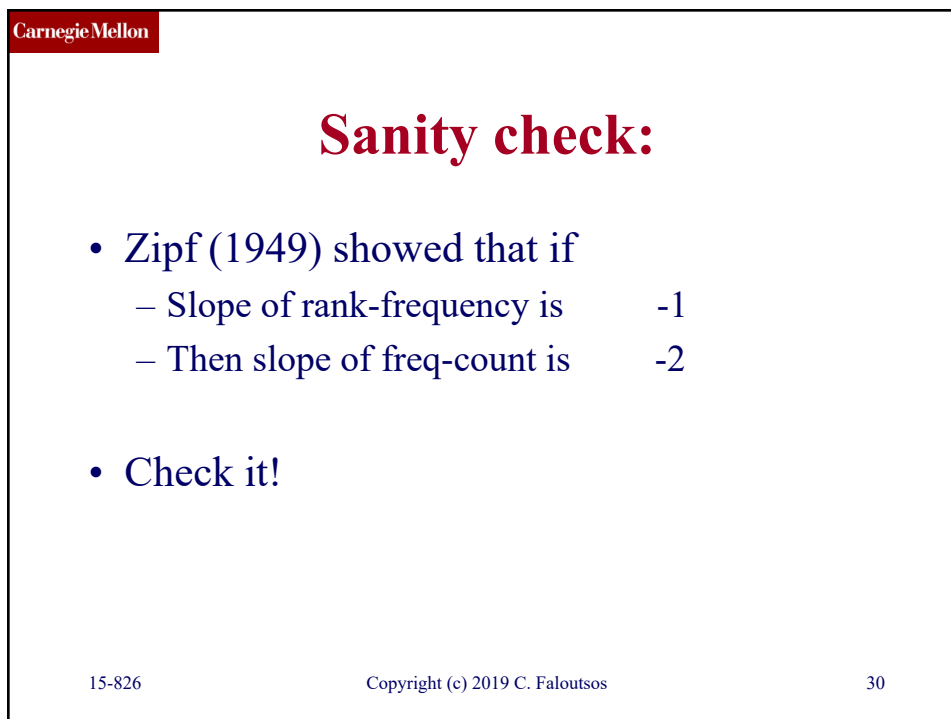
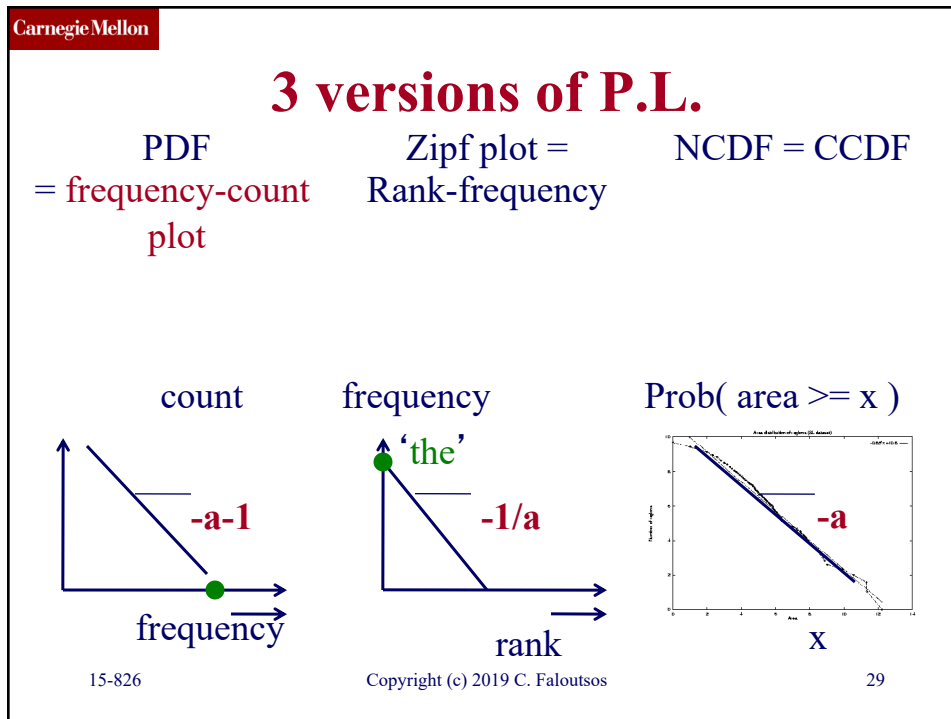
20

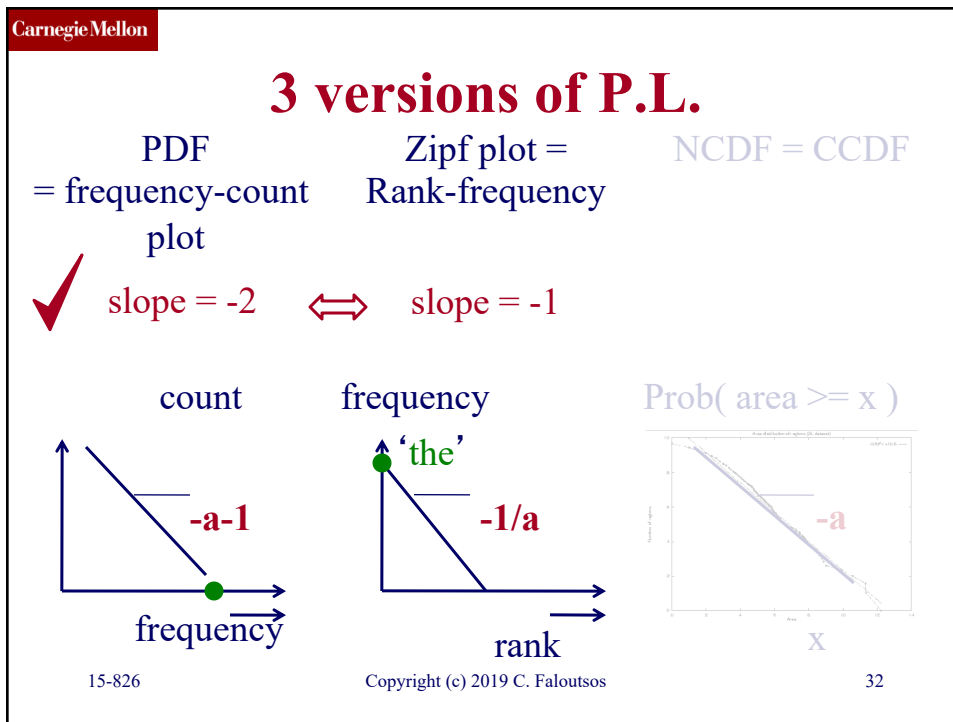
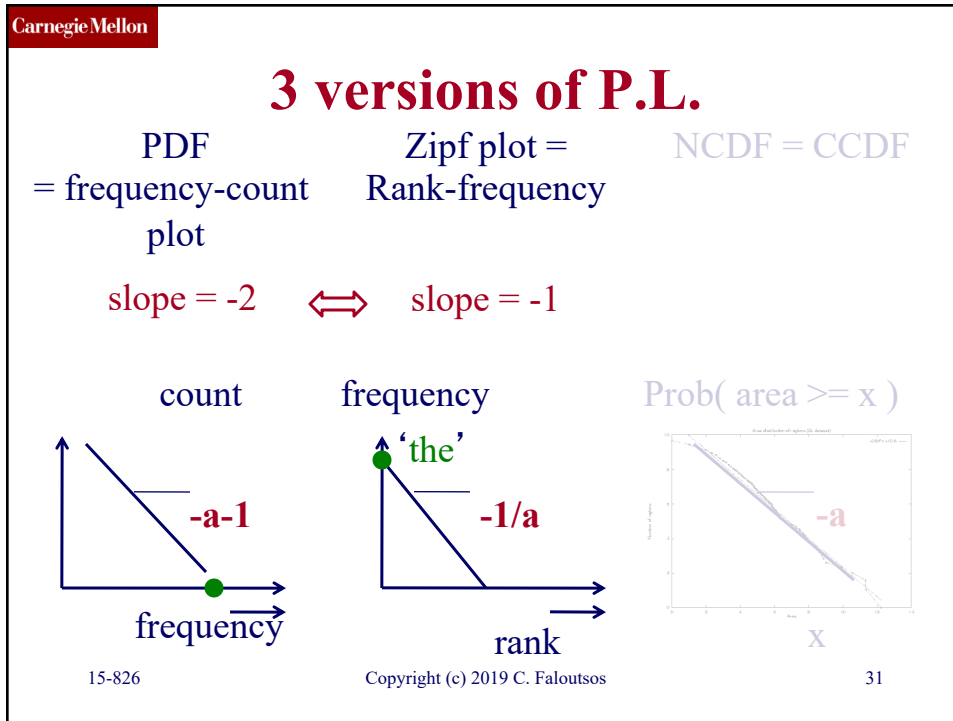












CarnegieMellon

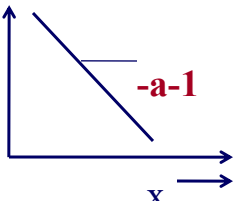
3 versions of P.L.

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

plot

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

Prob(area = x)

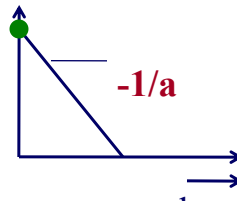


$-a-1$

x

15-826

area

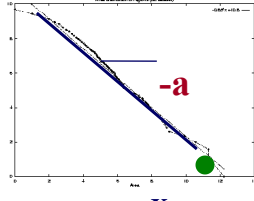


$-1/a$

rank

Copyright (c) 2019 C. Faloutsos

Prob(area \geq x)



$-a$

x

33

CarnegieMellon

This presentation

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

15-826 Copyright (c) 2019 C. Faloutsos 34

CarnegieMellon

Examples

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

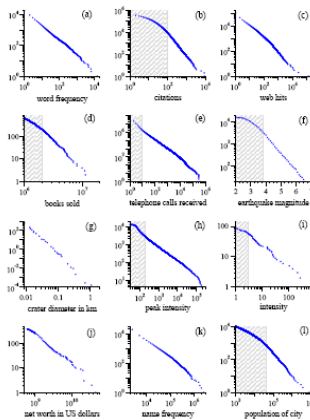
15-826

Copyright (c) 2019 C. Faloutsos

35

CarnegieMellon

[Newman 2005]

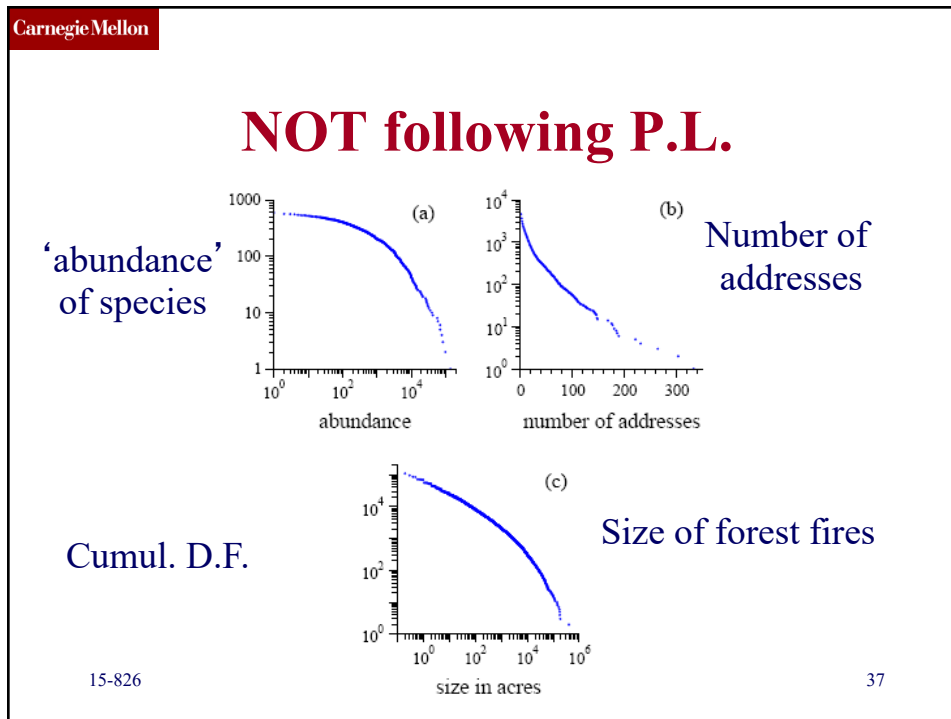


Rank-frequency plots
Or (complementary)
Cumulative D.F.

15-826

Copyright (c) 2019 C. Faloutsos

36



CarnegieMellon

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) 2019 C. Faloutsos 38

CarnegieMellon

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) 2019 C. Faloutsos

39

CarnegieMellon

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) 2019 C. Faloutsos

40

CarnegieMellon

Combination of exponentials

- Most freq 'words' ?



15-826

Copyright (c) 2019 C. Faloutsos

41

CarnegieMellon

Combination of exponentials

- Most freq 'words' ?
- a, b, \dots, z
- $aa, ab, \dots, az, ba, \dots, bz, \dots, zz$
- ...



15-826

Copyright (c) 2019 C. Faloutsos

42

CarnegieMellon

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) 2019 C. Faloutsos 43



CarnegieMellon

Inverses of quantities

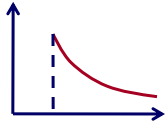
- y follows $p(y)$ and goes through zero
- $x = 1/y$
- Then $p(x) = \dots = -p(y) / x^2$
- For $y \sim 0$, x has power law tail.

y : ██████████

0mph.....1mph

count



Travel time

$y \rightarrow$ speed
 $x \rightarrow$ travel time

15-826 Copyright (c) 2019 C. Faloutsos 44

This presentation

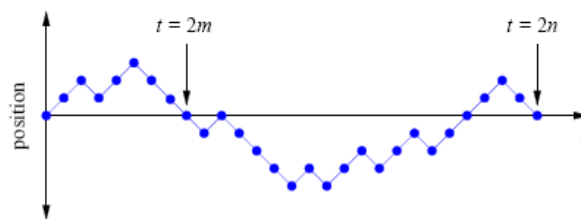
- Definitions
- Clarification
- Examples and counter-examples
- Generative mechanisms
 - Combination of exponentials
 - Inverse
 - ➔ – Random walk
 - Yule distribution = CRP
 - Percolation
 - Self-organized criticality
 - Other

15-826

Copyright (c) 2019 C. Faloutsos

45

Random walks



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

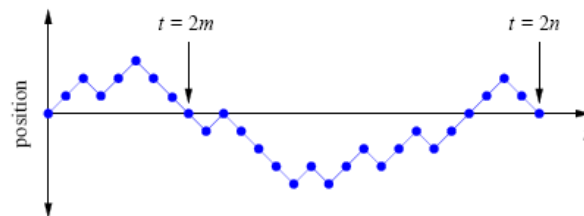
15-826

Copyright (c) 2019 C. Faloutsos

46

CarnegieMellon

Random walks



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

William Feller: *An introduction to probability theory and its applications*, Vol. 1, Wiley 1971
p. 78 Eq (3.7) and Stirling's approx (p. 75, Eq(2.4))

CarnegieMellon

Random walks

J. G. Oliveira & A.-L. Barabási Human Dynamics: The Correspondence Patterns of Darwin and Einstein.
Nature **437**, 1251 (2005) . [[PDF](#)]

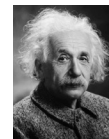
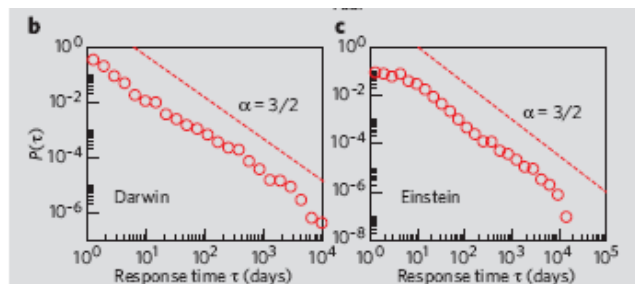


Figure 1 | The correspondence patterns of Darwin and Einstein.

15-826

Copyright (c) 2019 C. Faloutsos

48

This presentation

- Definitions - clarification
- 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

CarnegieMellon

Yule distribution and CRP

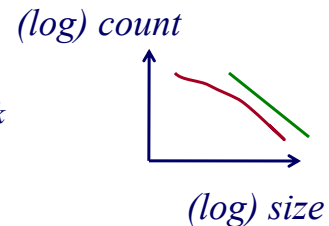
Then:

$$\text{Prob}(k \text{ people in a group}) = p_k$$

$$= (1 + 1/m) B(k, 2+1/m)$$

$$\sim k^{-(2+1/m)}$$

(since $B(a,b) \sim a^{-b}$: power law tail)



15-826

Copyright (c) 2019 C. Faloutsos

51

CarnegieMellon

Yule distribution and CRP

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

15-826

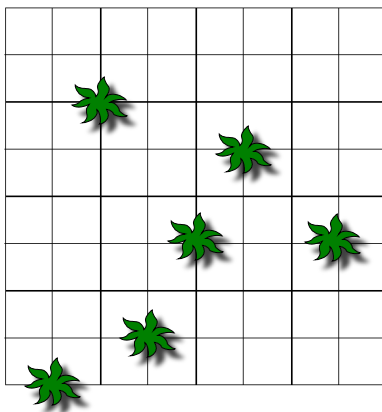
Copyright (c) 2019 C. Faloutsos

52

This presentation

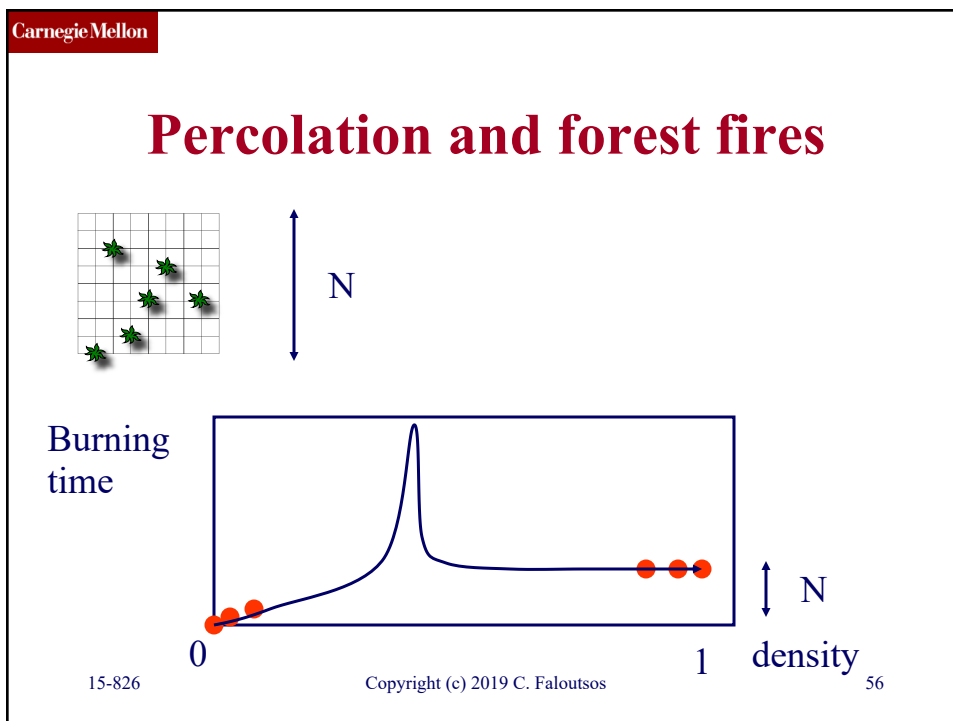
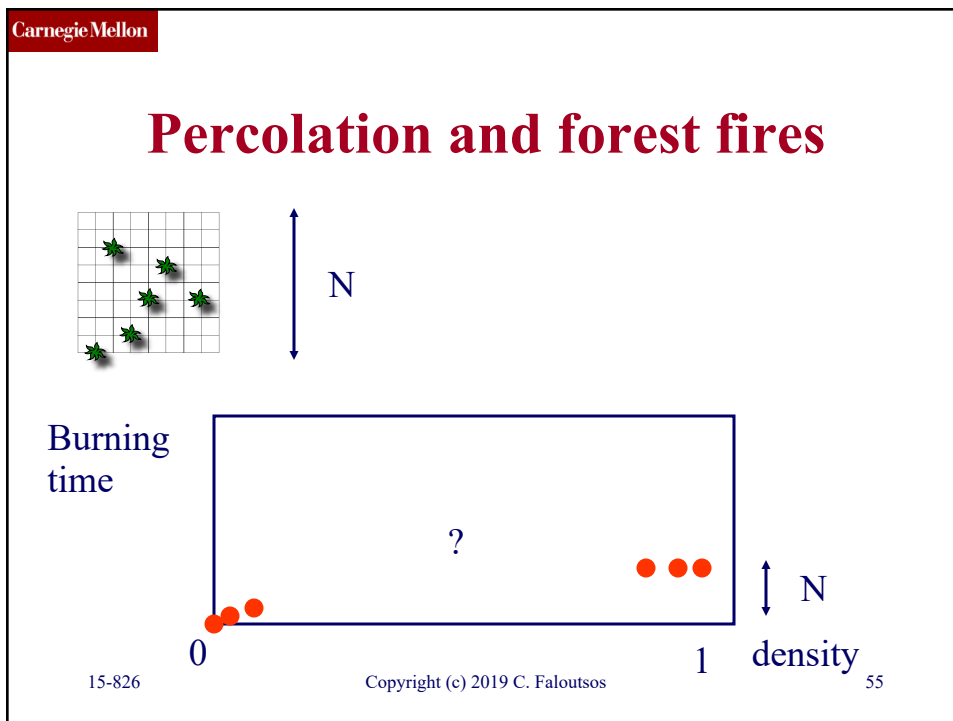
- Definitions - clarification
- Examples and counter-examples
- Generative mechanisms
 - Combination of exponentials
 - Inverse
 - Random walk
 - Yule distribution = CRP
 - ➔ – Percolation
 - Self-organized criticality
 - Other

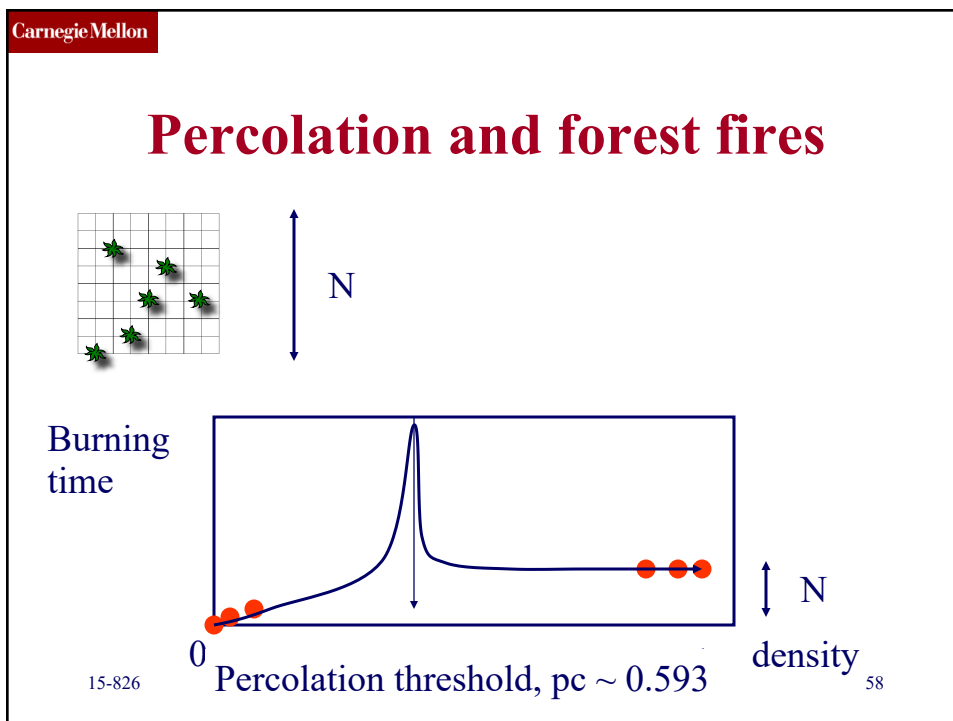
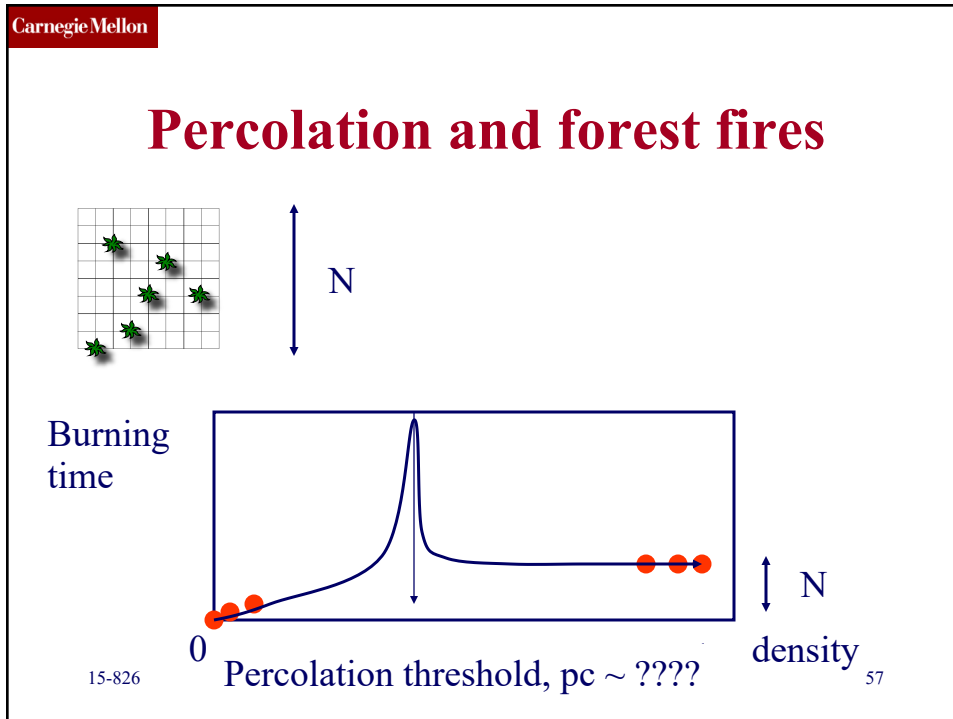
Percolation and forest fires



A burning tree will cause its neighbors to burn next.

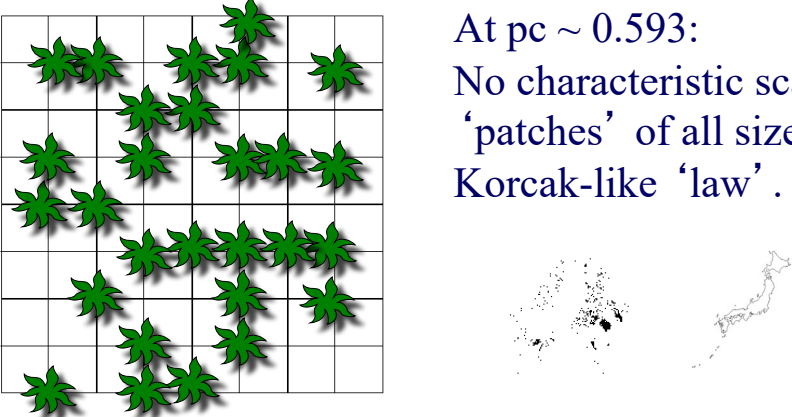
Which tree density p will cause the fire to last longest?





CarnegieMellon

Percolation and forest fires



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

15-826 Copyright (c) 2019 C. Faloutsos 59

CarnegieMellon

This presentation

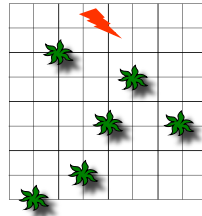
- Definitions - clarification
- Examples and counter-examples
- Generative mechanisms
 - Combination of exponentials
 - Inverse
 - Random walk
 - Yule distribution = CRP
 - Percolation
 - ➔ – Self-organized criticality
 - Other

15-826 Copyright (c) 2019 C. Faloutsos 60

CarnegieMellon

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) 2019 C. Faloutsos

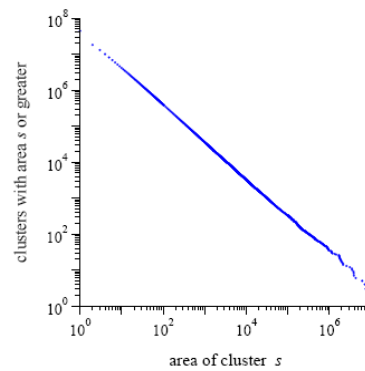
61

CarnegieMellon

Self-organized criticality

- A1: Power law-like

CCDF



15-826

Copyright (c) 2019 C. F

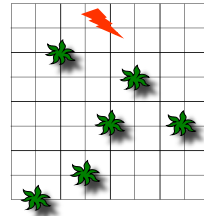
Area of cluster s

62

CarnegieMellon

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) 2019 C. Faloutsos

63

CarnegieMellon

Self-organized criticality

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

15-826

Copyright (c) 2019 C. Faloutsos

64

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]

This presentation

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

Other - lognormal

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

Other - lognormal

Random multiplication:

- Start with C dollars; put in bank
- Random interest rate $s(t)$ each year t
- Each year t : $C(t) = C(t-1) * (1 + s(t))$
- $\text{Log}(C(t)) = \log(C) + \log(..) + \log(..) \dots \rightarrow$
Gaussian

CarnegieMellon

Other - lognormal

Random multiplication:

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

15-826

Copyright (c) 2019 C. Faloutsos

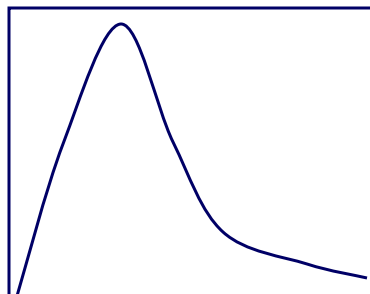
69

CarnegieMellon

Other - lognormal

Lognormal:

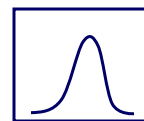
pdf



0

 $\$ = e^h$

pdf

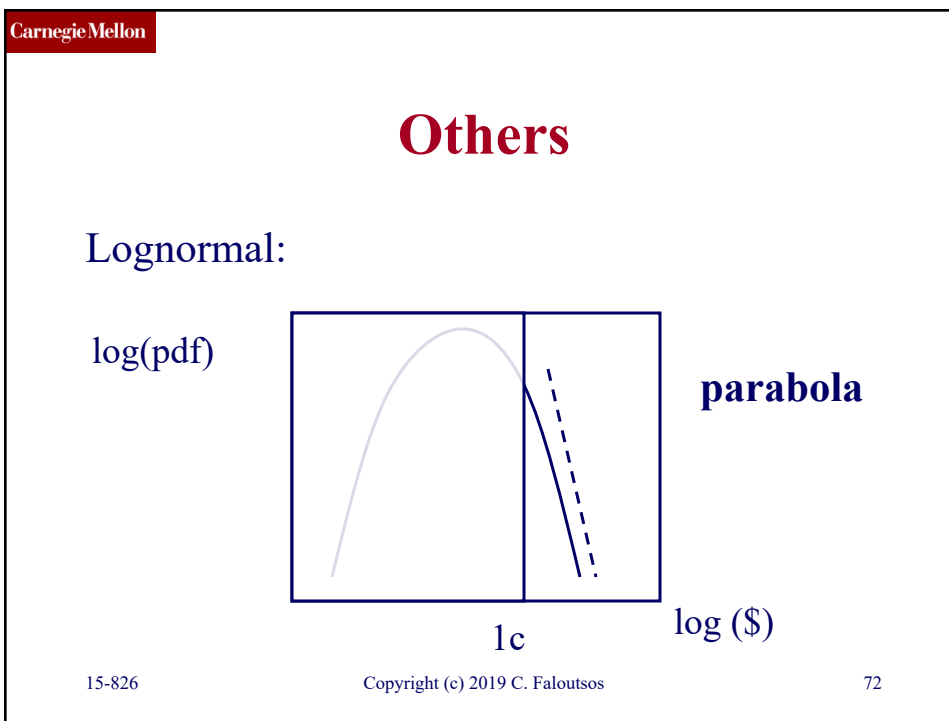
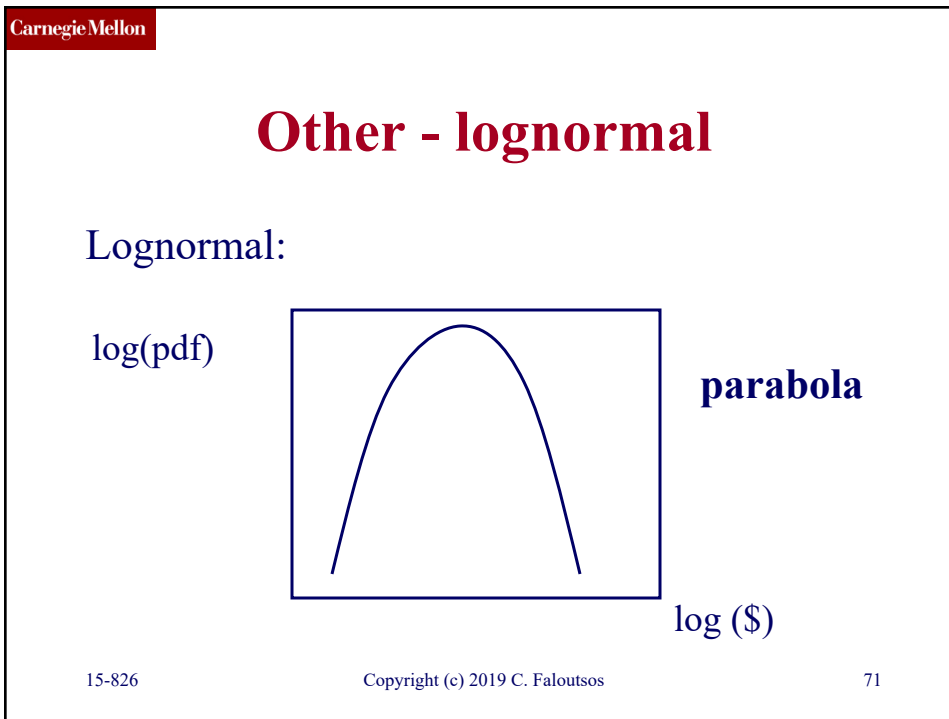


$h = \text{body}$
 height

15-826

Copyright (c) 2019 C. Faloutsos

70



Other - lognormal

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

Other - lognormal

- 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?)

CarnegieMellon

Fragmentation -> lognormal

15-826 Copyright (c) 2019 C. Faloutsos 75

CarnegieMellon

This presentation

- Definitions - clarification
- Examples and counter-examples
- Generative mechanisms
 - Combination of exponentials
 - Inverse
 - Random walk
 - Yule distribution = CRP
 - Percolation
 - Self-organized criticality
 - Other – lognormal
 - ➔ – Other – log-logistic (repeated, from lecture on graph mining)

15-826 Copyright (c) 2019 C. Faloutsos 76

CarnegieMellon

‘TLaC: Lazy Contractor’

- The longer a task (phonecall) has taken,
- The even longer it will take

Odds ratio=

Casualties(<x):
Survivors(>=x)

== power law

15-826 (c) C. Faloutsos, 2019 77

CarnegieMellon

Log-logistic distribution

- $CDF(t)/(1 - CDF(t)) == OR(t)$
- For log-logistic: $\log[OR(t)] = \beta + \rho * \log(t)$

Odds ratio=

Casualties(<x):
Survivors(>=x)

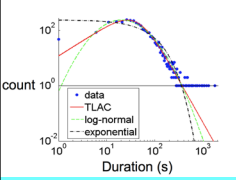
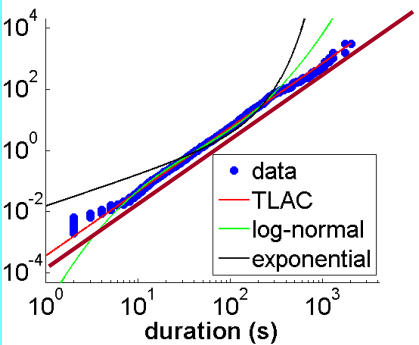
== power law

15-826 (c) C. Faloutsos, 2019 78

CarnegieMellon

Log-logistic distribution

- $CDF(t)/(1 - CDF(t)) == OR(t)$
- For log-logistic: $\log[OR(t)] = \beta + \rho * \log(t)$

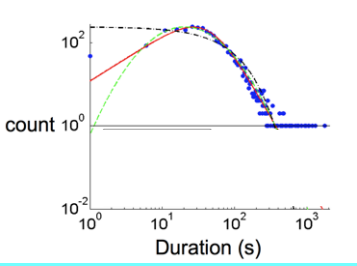
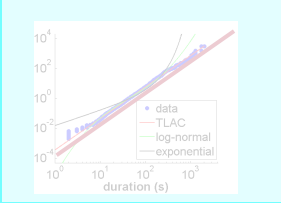



15-826
(c) C. Faloutsos, 2019
79

CarnegieMellon

Log-logistic distribution

- $CDF(t)/(1 - CDF(t)) == OR(t)$
- For log-logistic: $\log[OR(t)] = \beta + \rho * \log(t)$

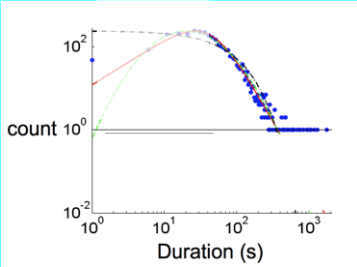
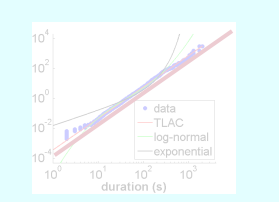
- PDF looks like hyperbola

15-826
(c) C. Faloutsos, 2019
80

CarnegieMellon

Log-logistic distribution

- $\text{CDF}(t)/(1 - \text{CDF}(t)) == \text{OR}(t)$
- For log-logistic: $\log[\text{OR}(t)] = \beta + \rho * \log(t)$

- PDF looks like hyperbola;
- and, if clipped, like power-law

15-826 (c) C. Faloutsos, 2019 81

CarnegieMellon

**Attention:
Phase 1**

Log-logistic distribution

Nice 1 page description: section II of

Pravallika Devineni, Danai Koutra, Michalis Faloutsos, and Christos Faloutsos.

[*If walls could talk: Patterns and anomalies in Facebook wallposts.*](#)

ASONAM 2015, pp 367-374.

15-826 (c) C. Faloutsos, 2019 82

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 \sim P.L.

15-826

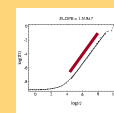
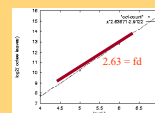
Copyright (c) 2019 C. Faloutsos

83

Conclusions



- Why so many power-laws?
- Many reasons:
 - Self similarity
 - rich-get-richer
 - etc



15-826

Copyright (c) 2019 C. Faloutsos

84

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)