Carnegie Mellon

15-826: Multimedia (Databases) and Data Mining

Lecture #11: Power laws

Potential causes and explanations

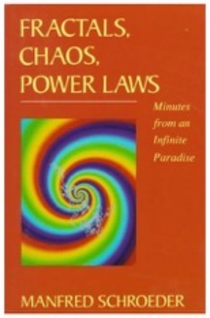
C. Faloutsos

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

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

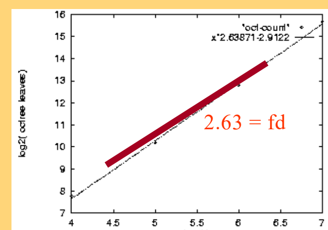
- primary key indexing
- secondary key / multi-key indexing
- spatial access methods
  - z-ordering
  - R-trees
  - misc
- fractals
  - intro
  - applications
- text

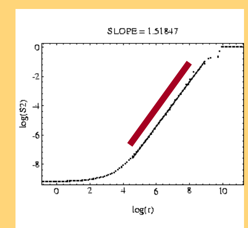
Indexing - Detailed outline

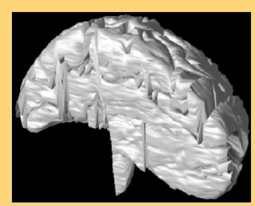
- fractals
  - intro
  - applications
    - disk accesses for R-trees (range queries)
    - …
    - dim. curse revisited
    - …
  - Why so many power laws?

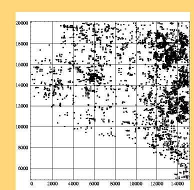
Problem

- Why so many power-laws?



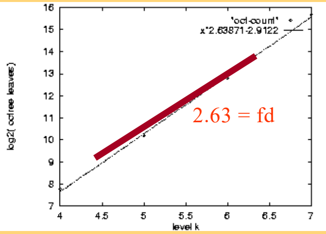


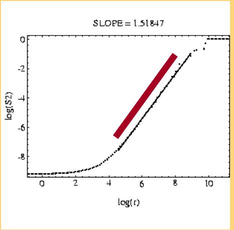


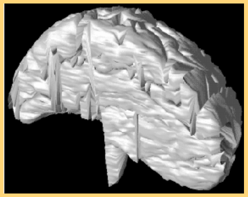


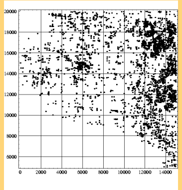
Conclusion

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



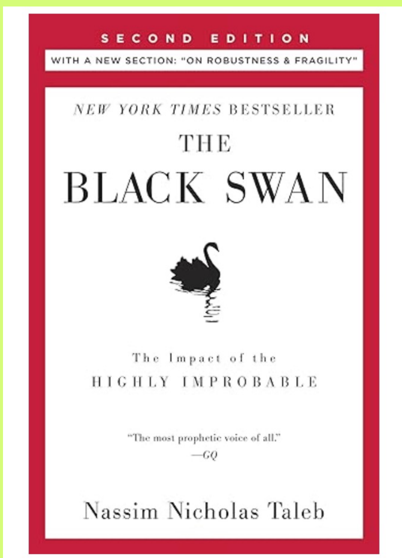






Reminder

Why ‘black swan’?



The black swan, by Nassim Nicholas Taleb, 2010

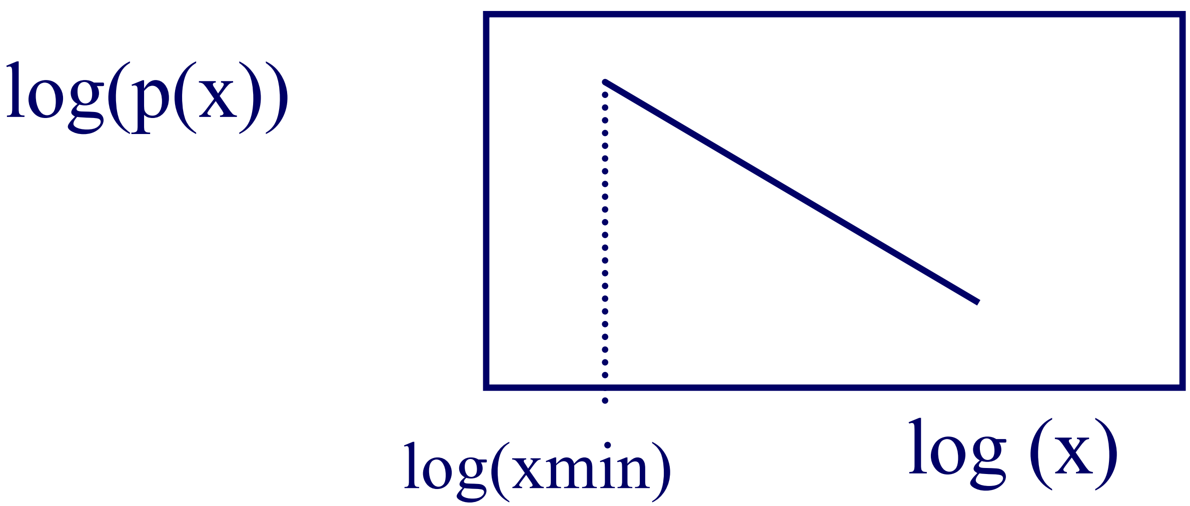
(power laws in multiple settings; leading to investment strategy (!))

This presentation

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

Definition

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



For discrete variables

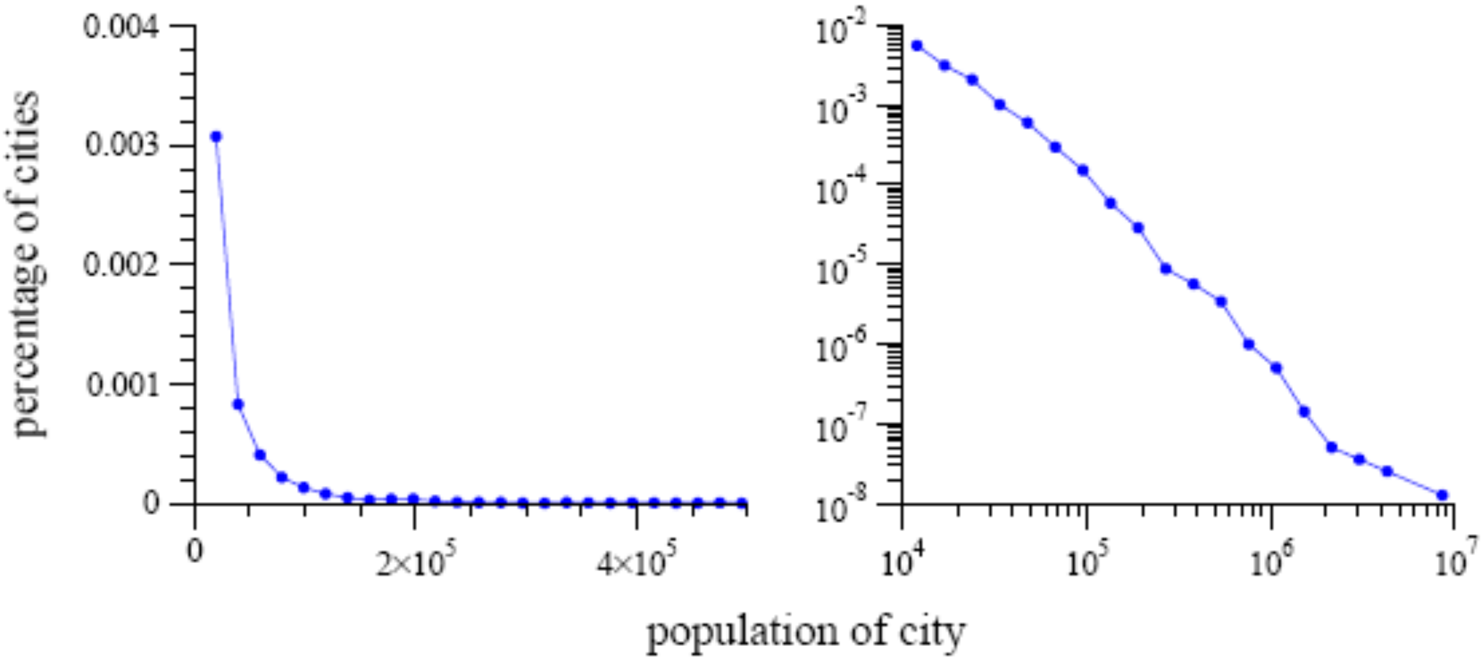
$$p_k = C k^{-a} \qquad (k > 0)$$

Or, the Yule distribution:

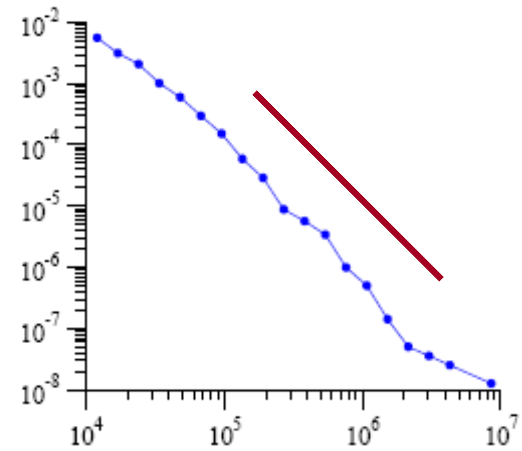
$$p_k = C\, B(k,a)$$

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

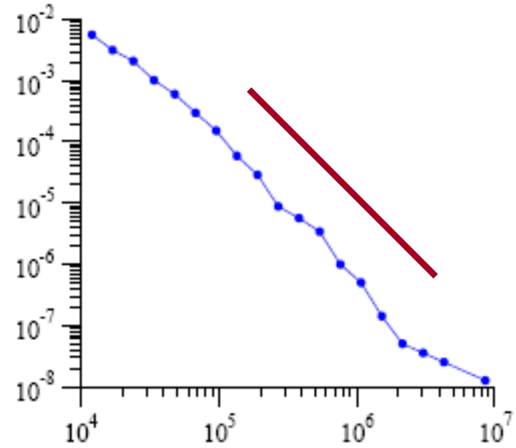
[Newman, 2005]


0.004
0.003
0.002
0.001
0
percentage of cities
0
2×10^5
4×10^5
10^-2
10^-3
10^-4
10^-5
10^-6
10^-7
10^-8
10^4
10^5
10^6
10^7
population of city

Estimation for *a*



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

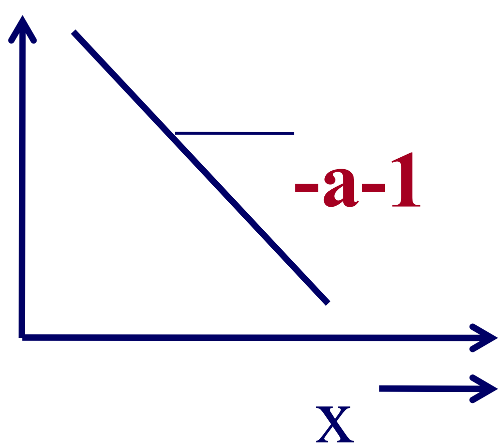
Jumping to the conclusion:

3 versions of P.L.

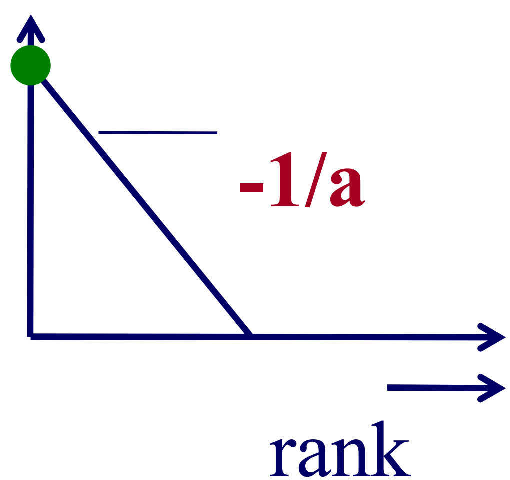
PDF = frequency-count plot	Zipf plot = Rank-frequency	NCDF = CCDF

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

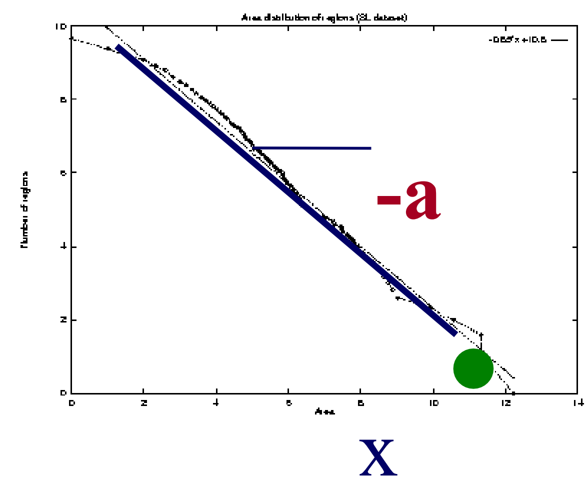
Prob(area = x)



area



Prob(area >= x)



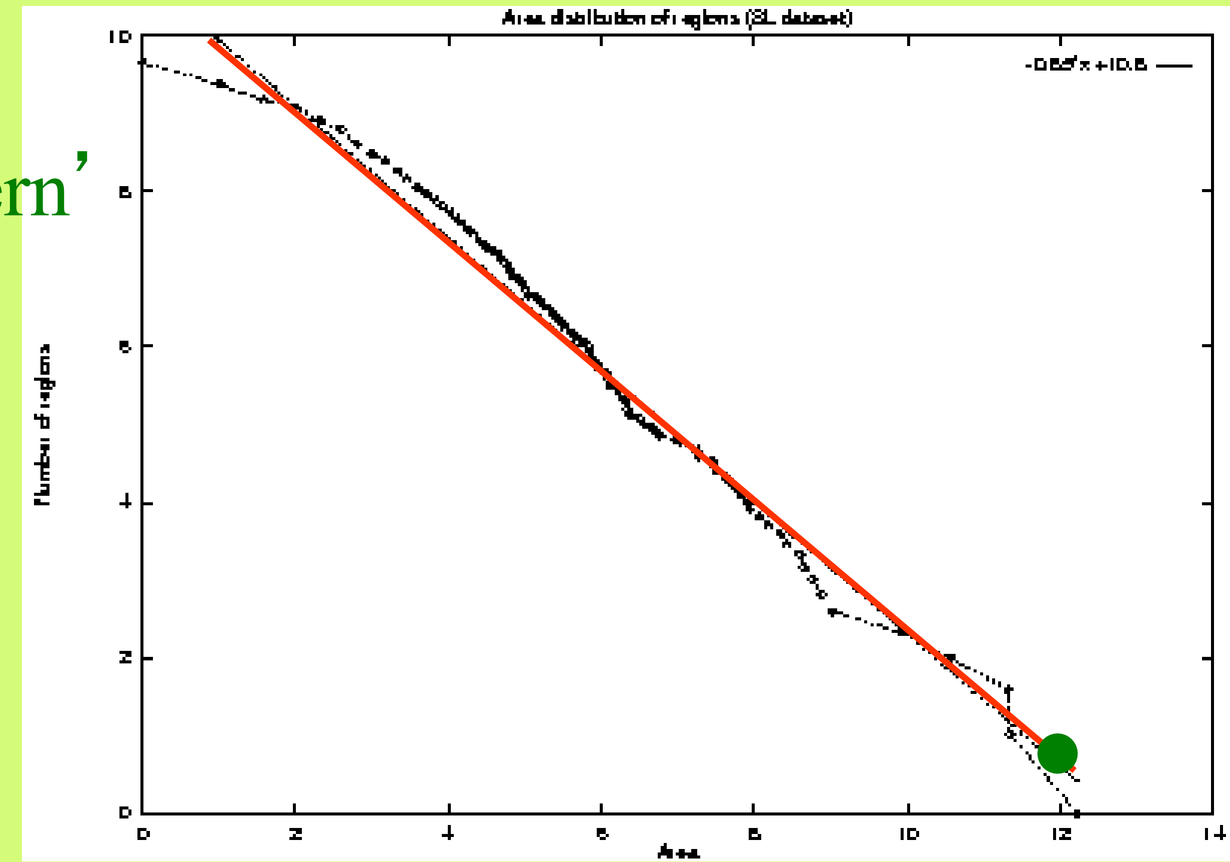
Details, and proof sketches:

Reminder

More power laws: areas – Korcak' s law

log(count(>= area))

'Vaenern'



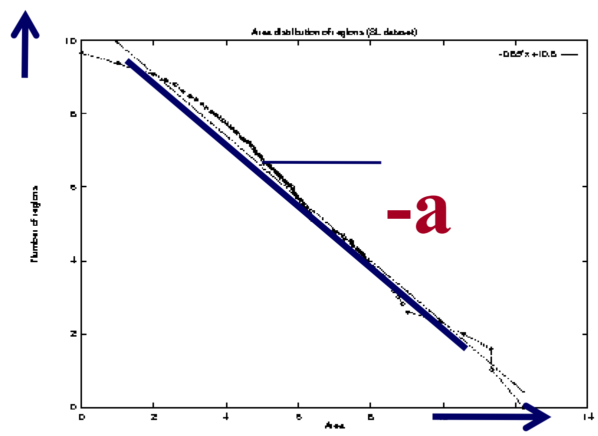
log(area)

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

3 versions of P.L.

NCDF = CCDF

Prob(area >= x)



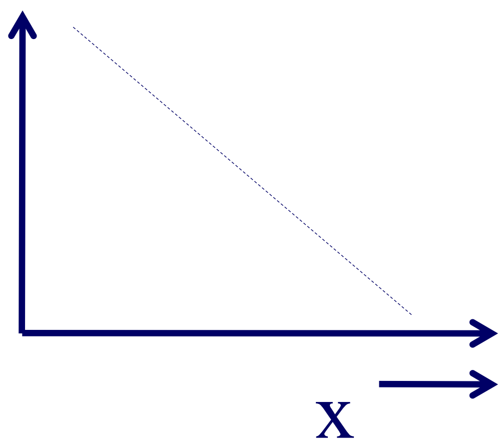
x

3 versions of P.L.

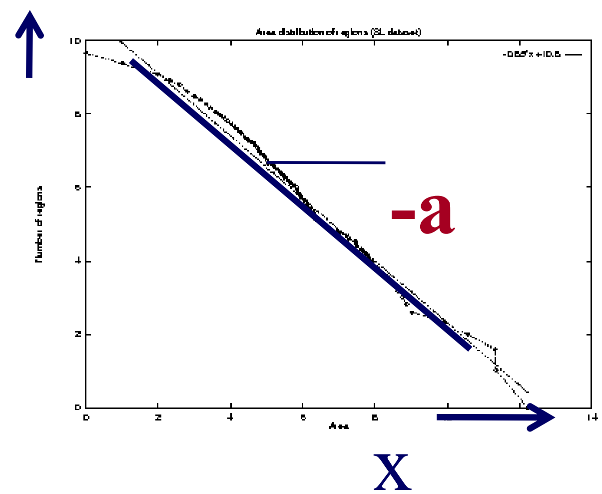
PDF

NCDF = CCDF

Prob(area = x)



Prob(area >= x)

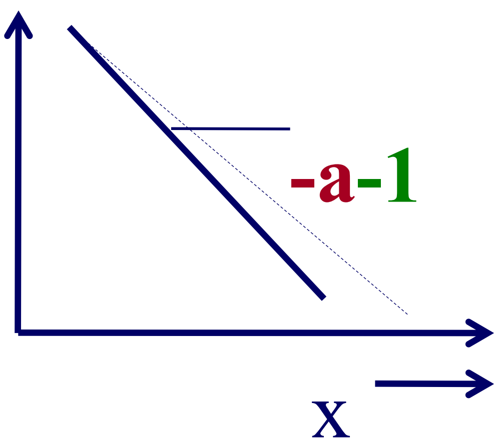


3 versions of P.L.

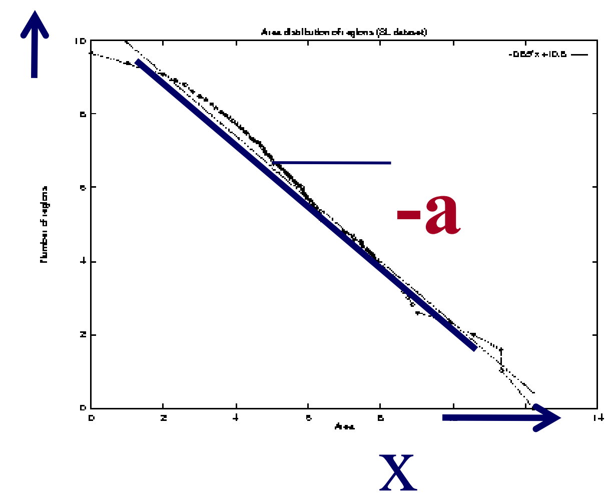
PDF

NCDF = CCDF

Prob(area = x)



Prob(area >= x)

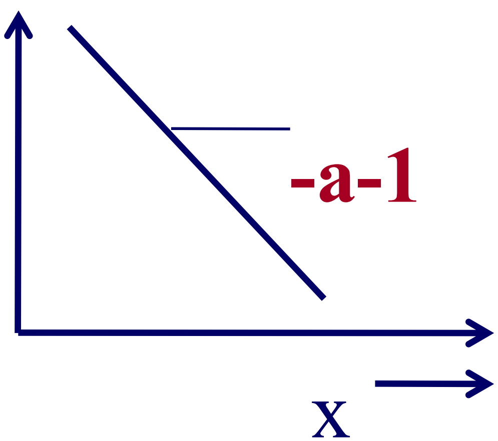


3 versions of P.L.

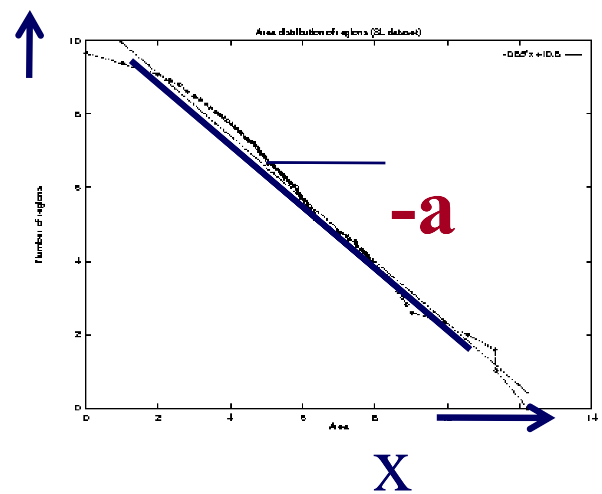
PDF

NCDF = CCDF

Prob(area = x)



Prob(area >= x)



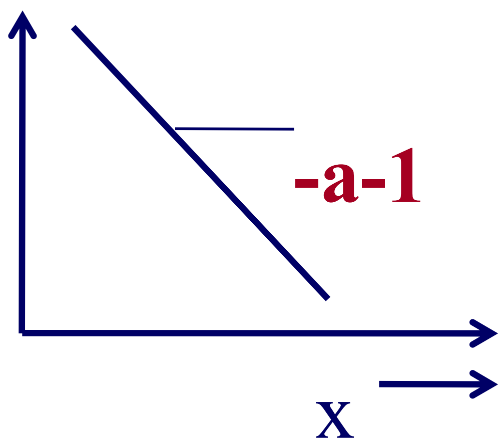
3 versions of P.L.

PDF

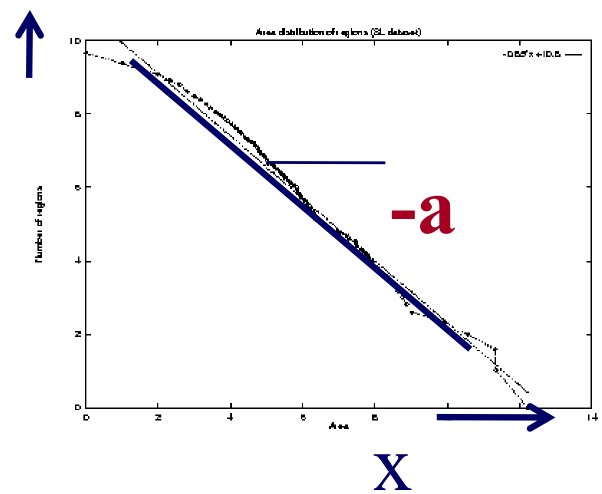
Zipf plot = Rank-frequency

NCDF = CCDF

Prob(area = x)



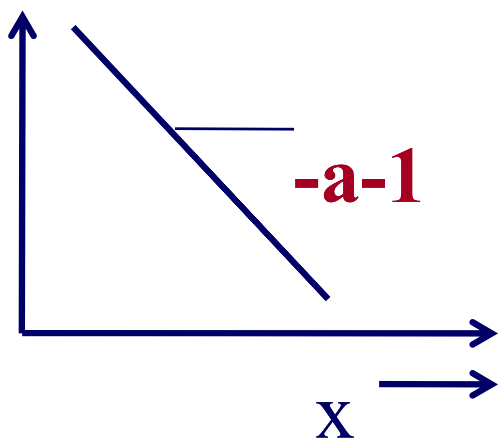
Prob(area >= x)



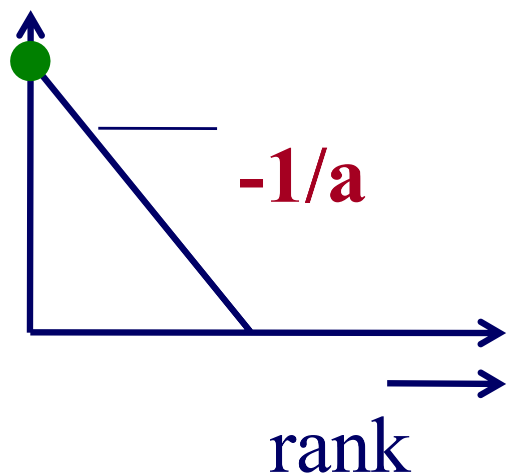
3 versions of P.L.

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

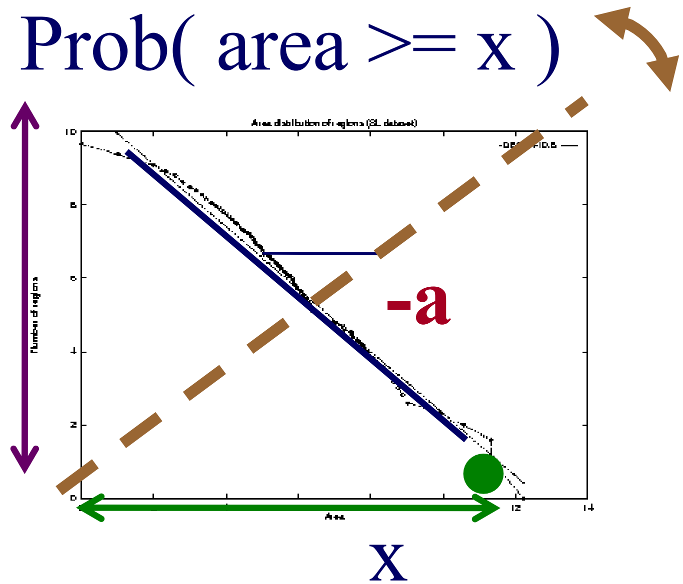
Prob(area = x)



area



Prob(area >= x)

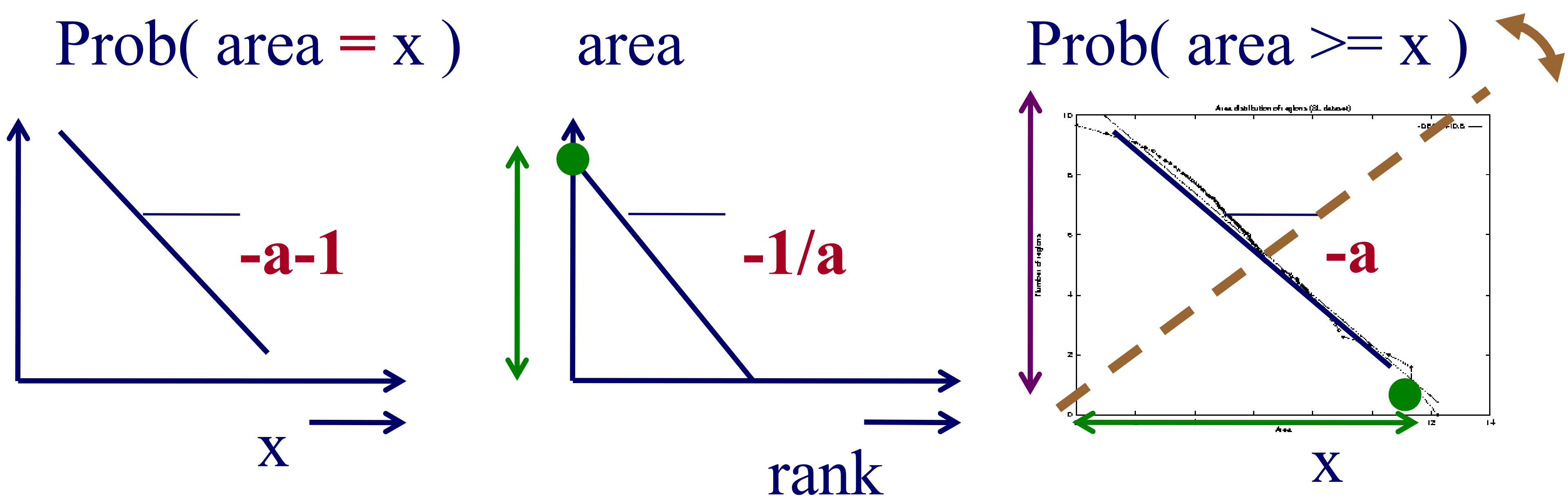


3 versions of P.L.

PDF

Zipf plot =
Rank-frequency

NCDF = CCDF



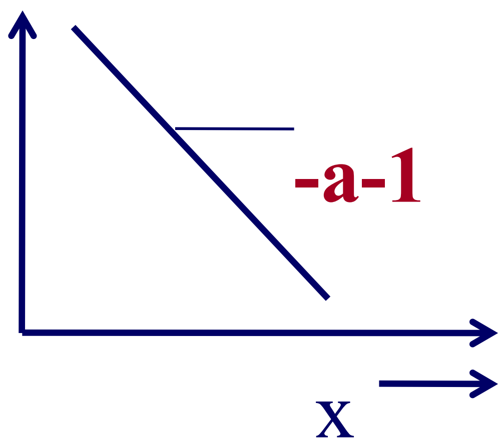
3 versions of P.L.

PDF

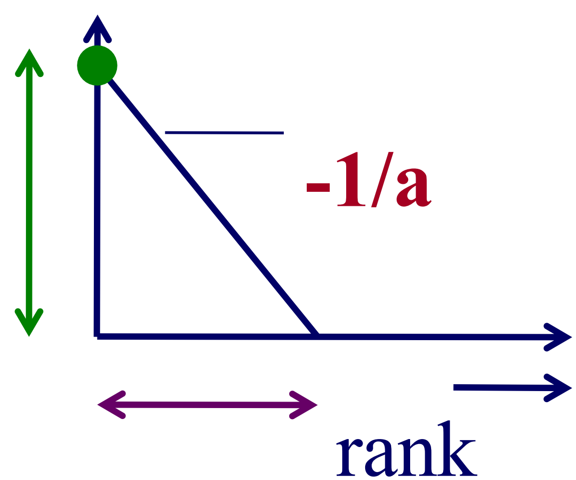
Zipf plot =
Rank-frequency

NCDF = CCDF

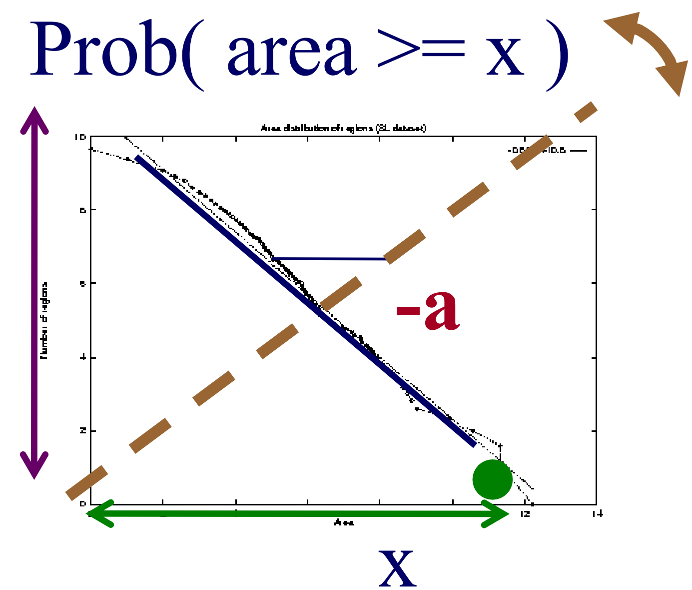
Prob(area = x)



area



Prob(area >= x)



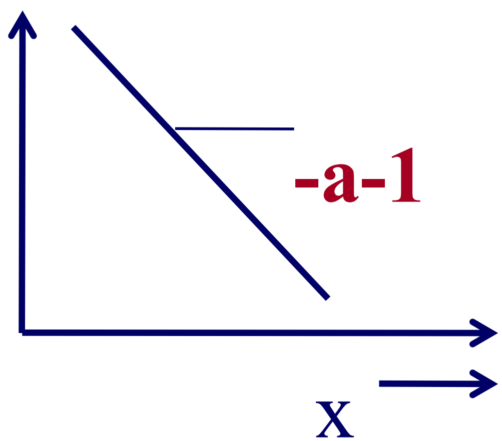
3 versions of P.L.

PDF

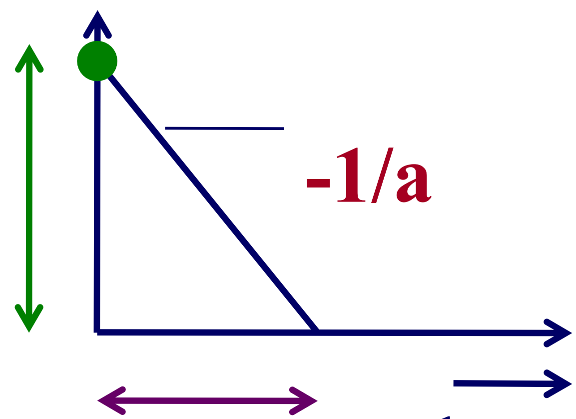
Zipf plot = Rank-frequency

NCDF = CCDF

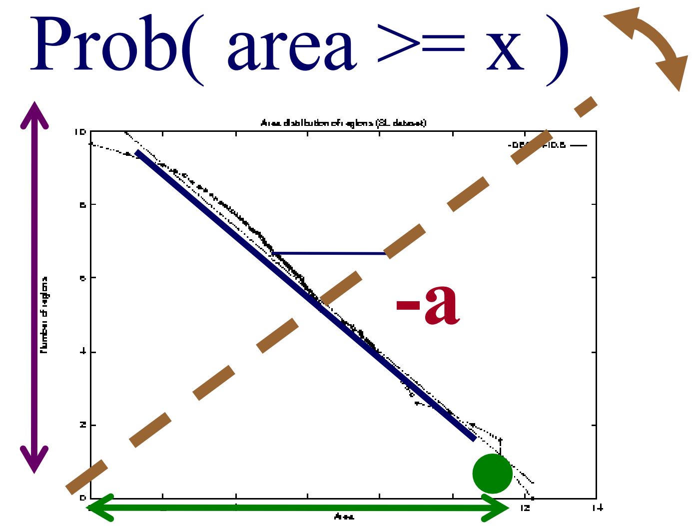
Prob(area = x)



frequency



Prob(area >= x)



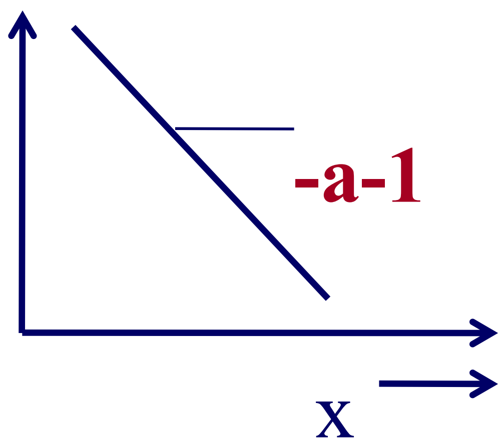
3 versions of P.L.

PDF = frequency-count plot

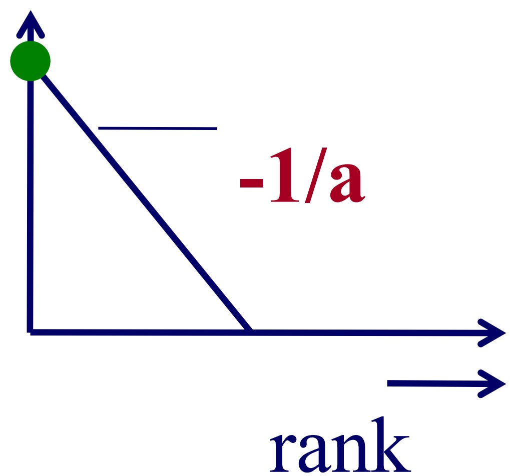
Zipf plot = Rank-frequency

NCDF = CCDF

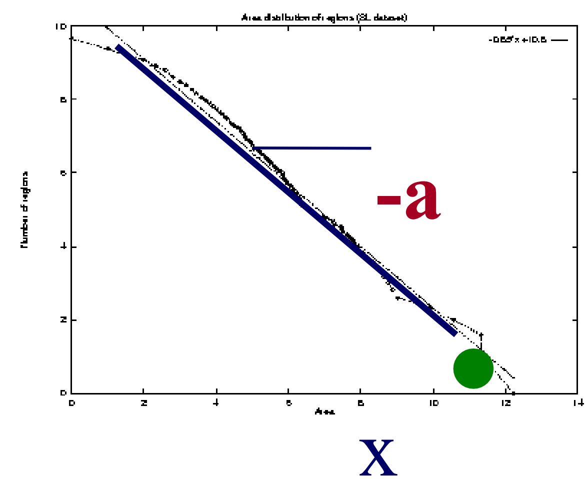
Prob(area = x)



area



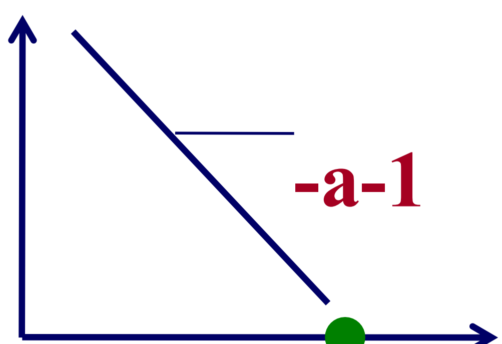
Prob(area >= x)



3 versions of P.L.

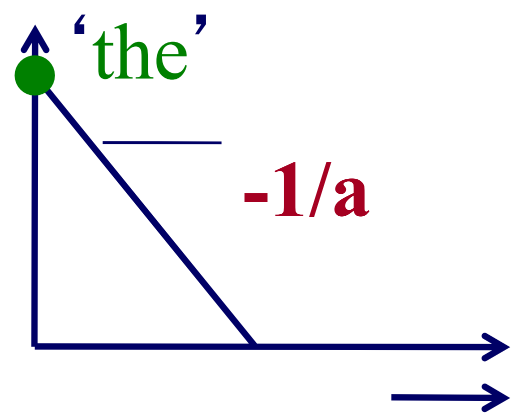
PDF = frequency-count plot	Zipf plot = Rank-frequency	NCDF = CCDF

count



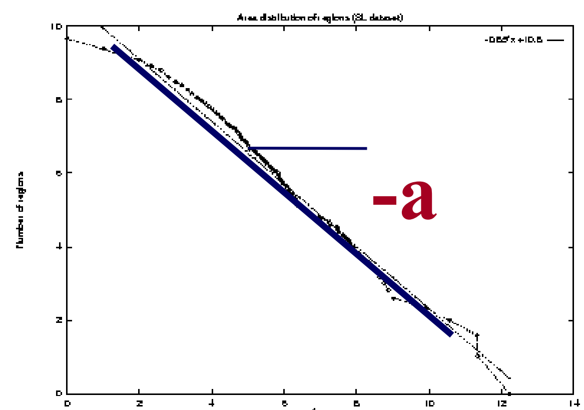
frequency

frequency



rank

Prob(area >= x)



x

Sanity check:

- Zipf (1949) showed that if
  - Slope of rank-frequency is -1
  - Then slope of freq-count is -2

- Check it!

3 versions of P.L.

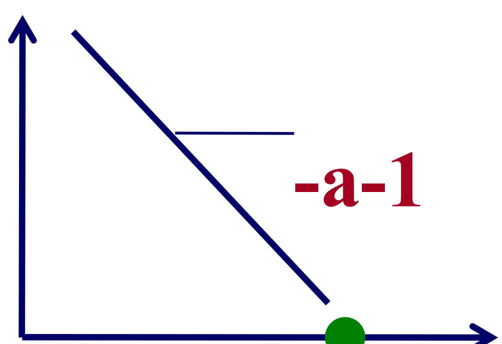
PDF
= frequency-count
plot

Zipf plot =
Rank-frequency

NCDF = CCDF

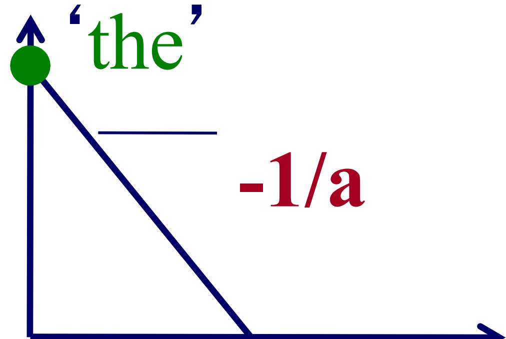
slope = -2 ⟺ slope = -1

count



frequency

frequency



rank

Prob(area >= x)

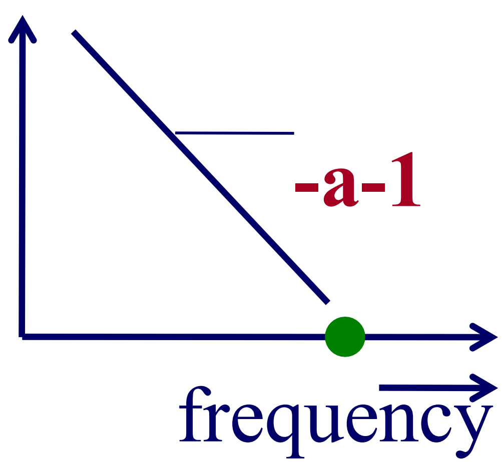
x

3 versions of P.L.

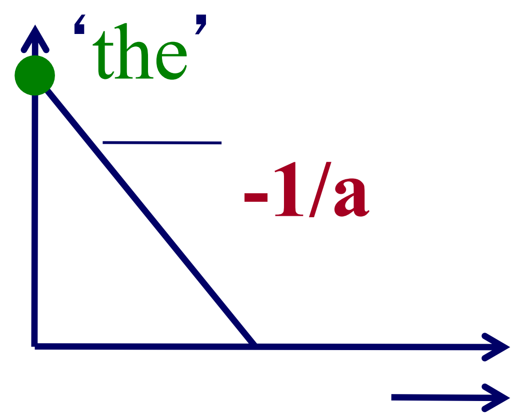
PDF = frequency-count plot | Zipf plot = Rank-frequency | NCDF = CCDF

✓ slope = -2 ⟺ slope = -1

count



frequency



Prob(area >= x)

3 versions of P.L.

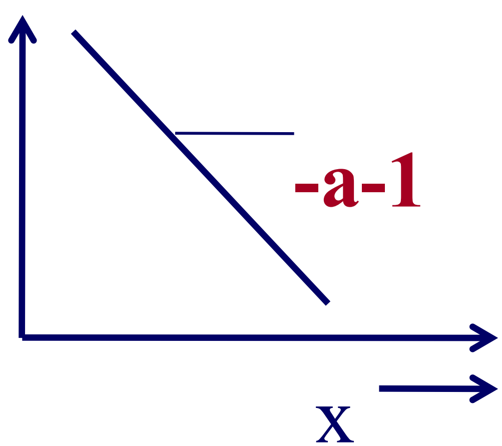
PDF = frequency-count plot

Zipf plot = Rank-frequency

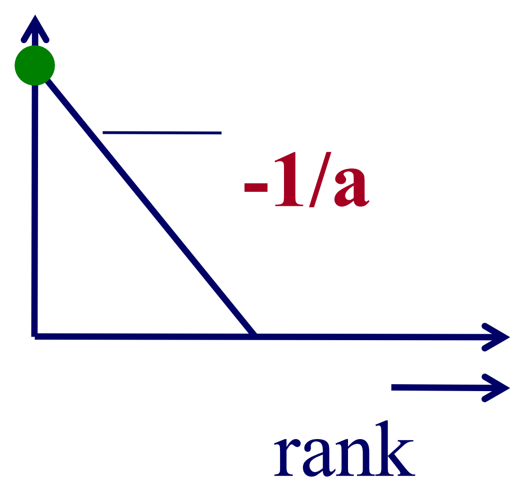
NCDF = CCDF

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

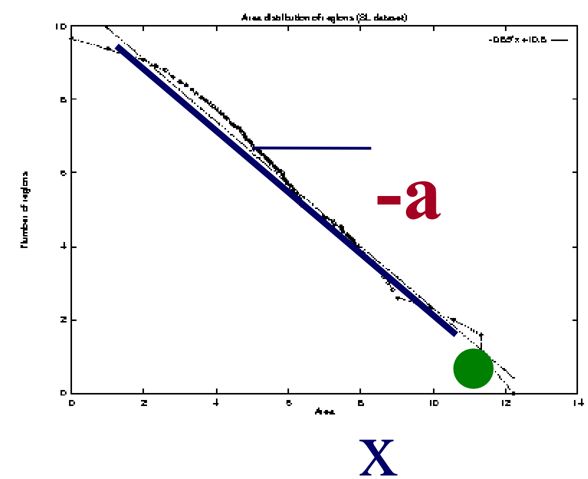
Prob(area = x)



area



Prob(area >= x)



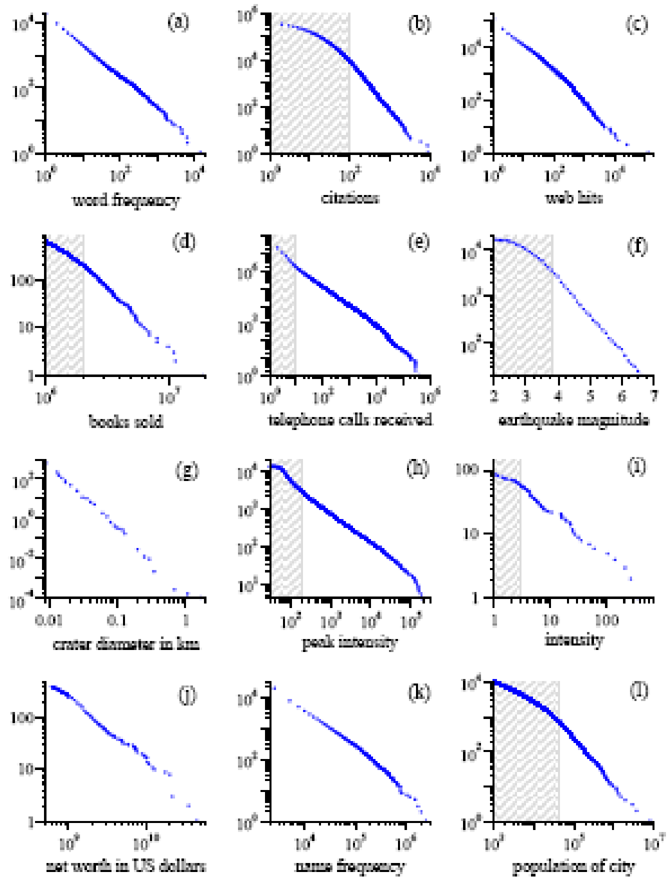
This presentation

- 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]

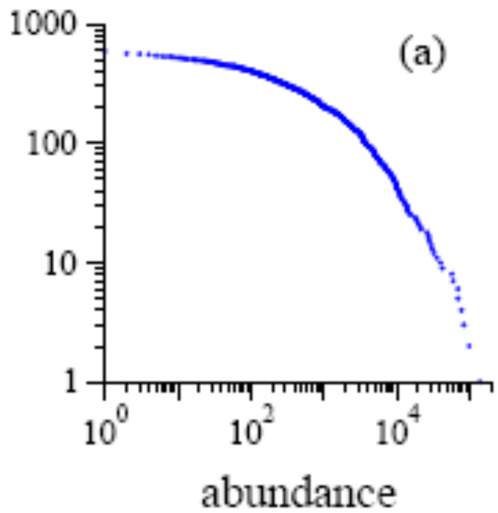


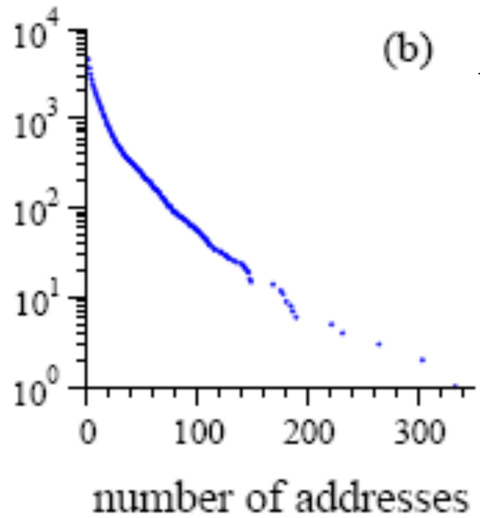
word freq; web hits;
books sold;
earthquake magnitude;
crater diameter;
…

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

NOT following P.L.

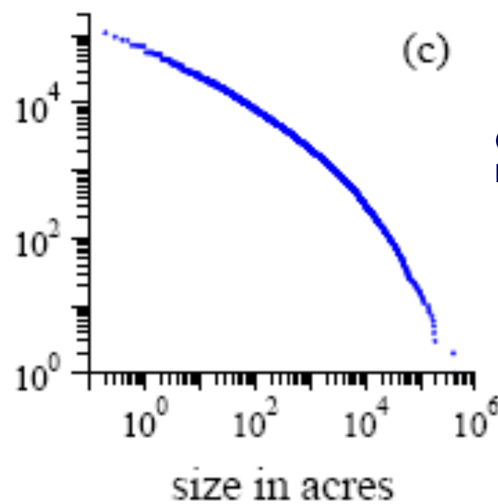
‘abundance’ of species





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 $p(y) = e^{ay}$ [Prob(survive y time-ticks)]

- eg., radioactive decay, with half-life –a
- (= collection of people, playing russian roulette)

Let $x \sim e^{by}$ (capital multiplies, every time tick)

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

Final capital distribution:

2x

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

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

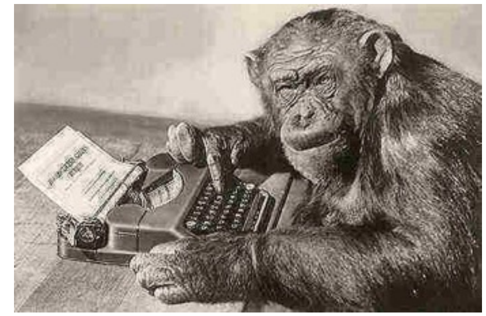
Combination of exponentials

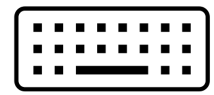
- Q: What simple mechanism could generate Zipf's law?

B. Mandelbrot

- A: Monkey on a typewriter:

Combination of exponentials





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

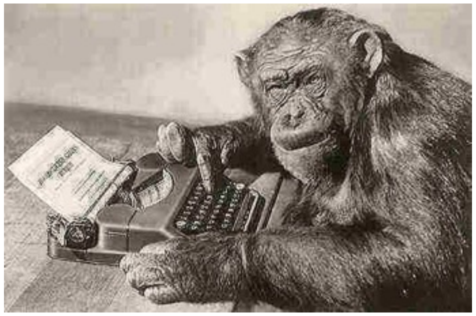
THEN: Freq(x-th most frequent word) = $x^{(-a)}$

see Eq. 47 of [Newman]:

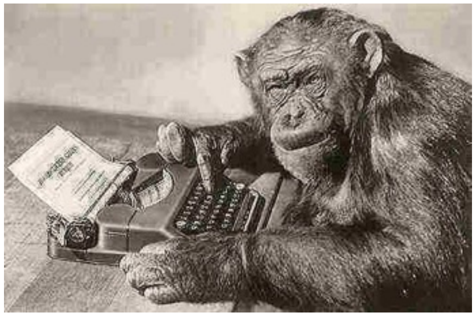
$$a = [2 \ln(m) - \ln (1 - q_s)] / [\ln m - \ln (1 - q_s)]$$

Combination of exponentials

- Most freq 'words'?



Combination of exponentials



- Most freq ‘words’ ?
- *a, b , ... z*
- *aa, ab, ... az, ba, ... bz, ... zz*
- ...

This presentation

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

Carnegie Mellon

Rare case

Inverses of quantities

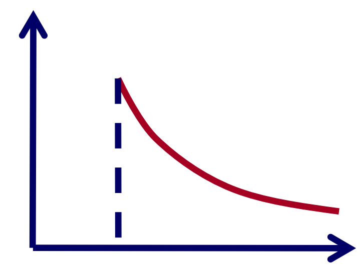
- y follows $p(y)$ and goes through zero
- $x = 1/y$
- Then $p(x) = \ldots = - p(y) / x^2$
- For $y \sim 0$, x has power law tail.

y-> speed
x-> travel time

y:

0mph……..1mph


count

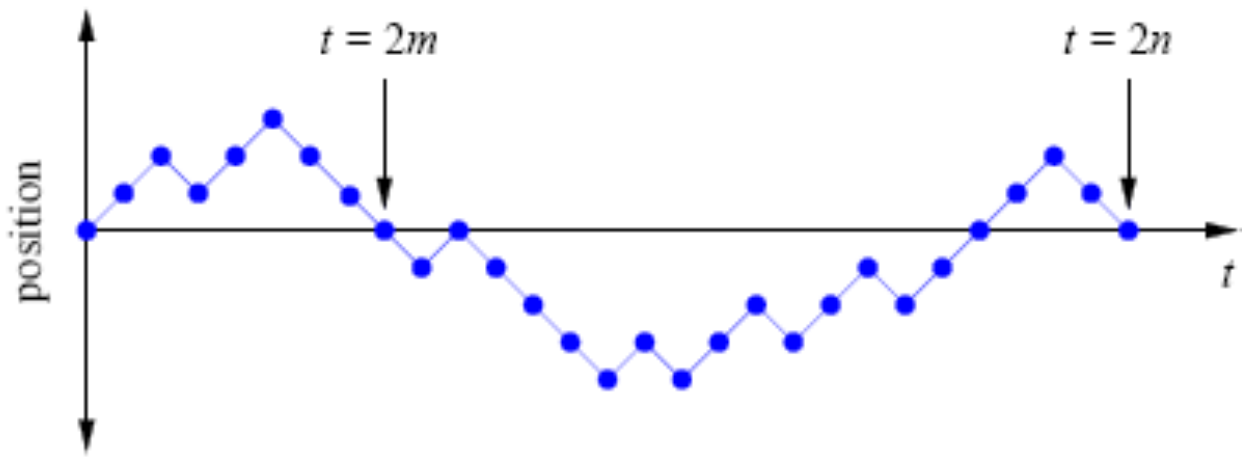


Travel time

This presentation

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

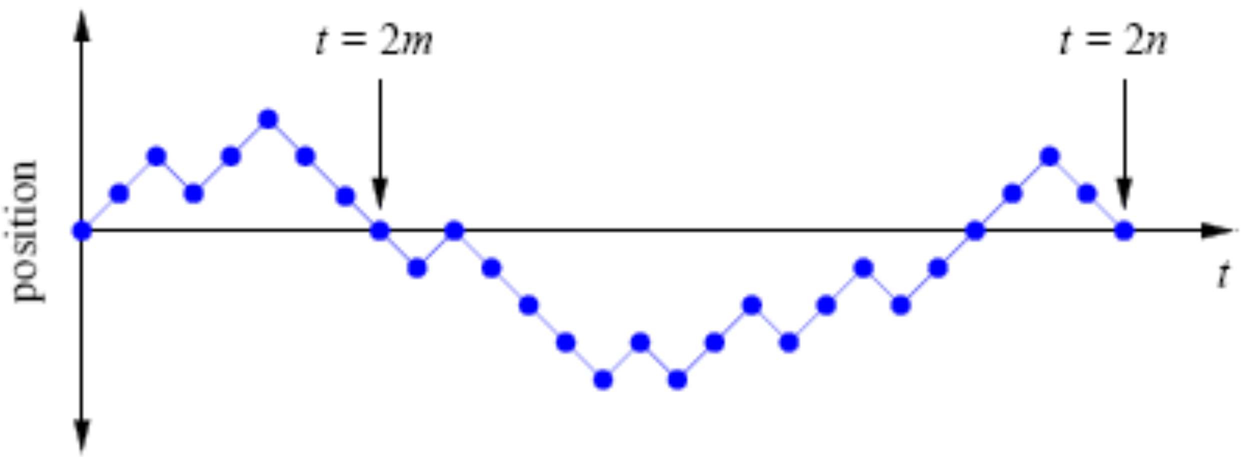
Random walks



Inter-arrival times PDF: *p(t)* ~ ??

Carnegie Mellon

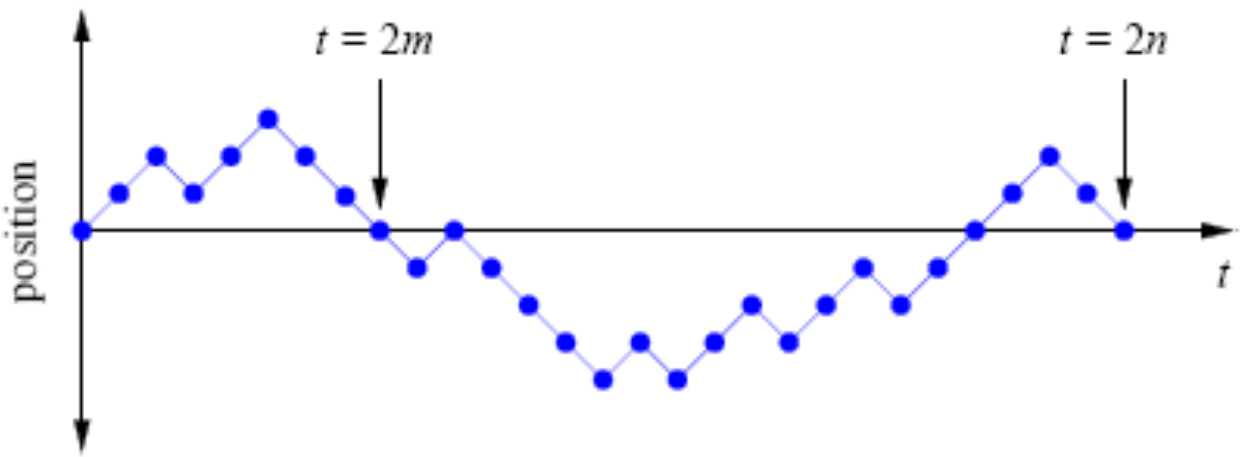
Random walks



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

a=??

Random walks



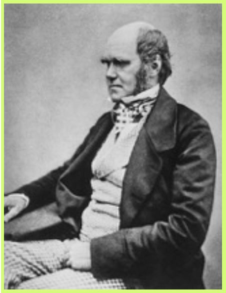
Inter-arrival times PDF: $p(t) \sim t^{-3/2}$

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))

Reminder

Random walks

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



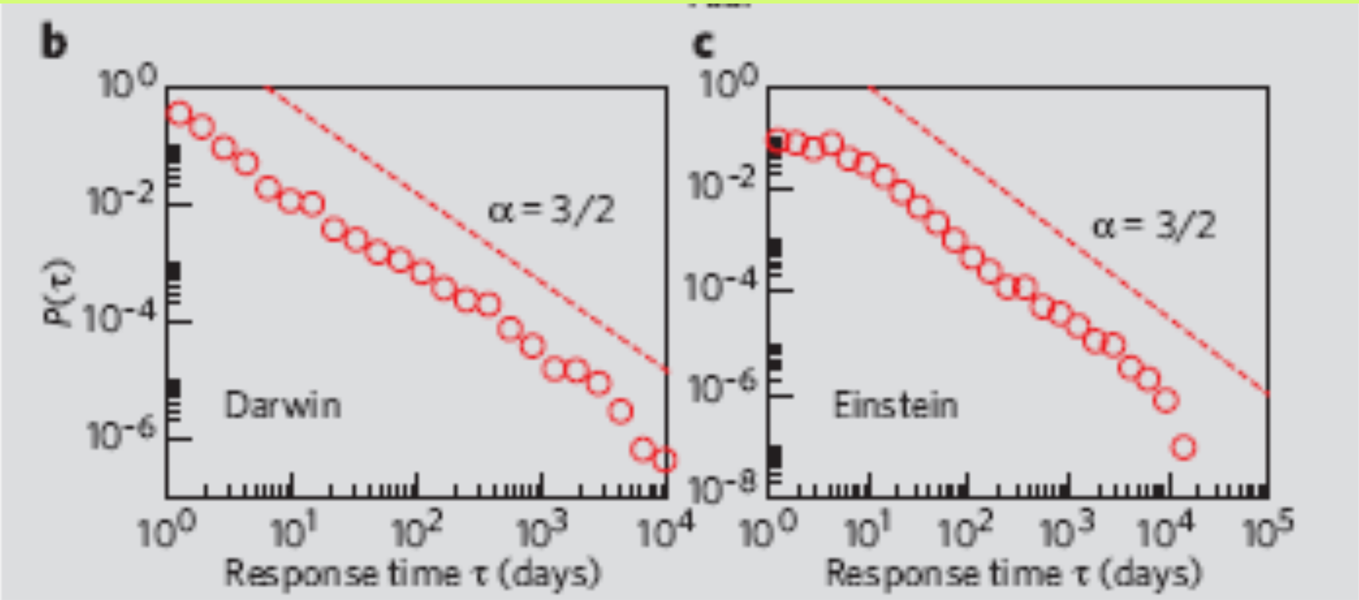
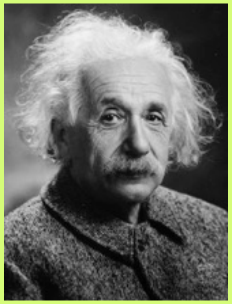


Figure 1 | The correspondence patterns of Darwin and Einstein.

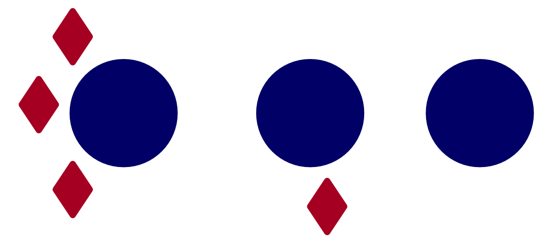


This presentation

- 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

Yule distribution and CRP

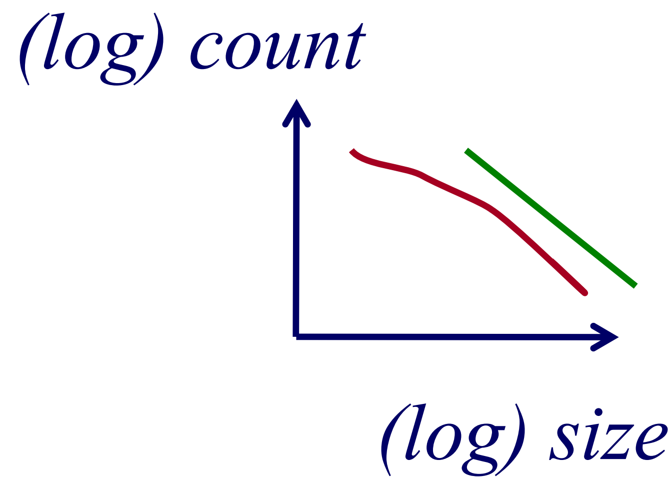
Then:

Prob(k people in a group) = p_k

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

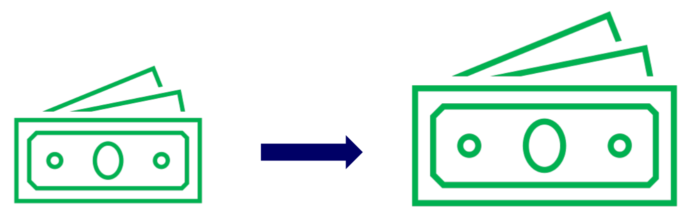
$\sim k^{-(2+1/m)}$

(since B(a,b) ~ a ** (-b) : power law tail)



Yule distribution and CRP

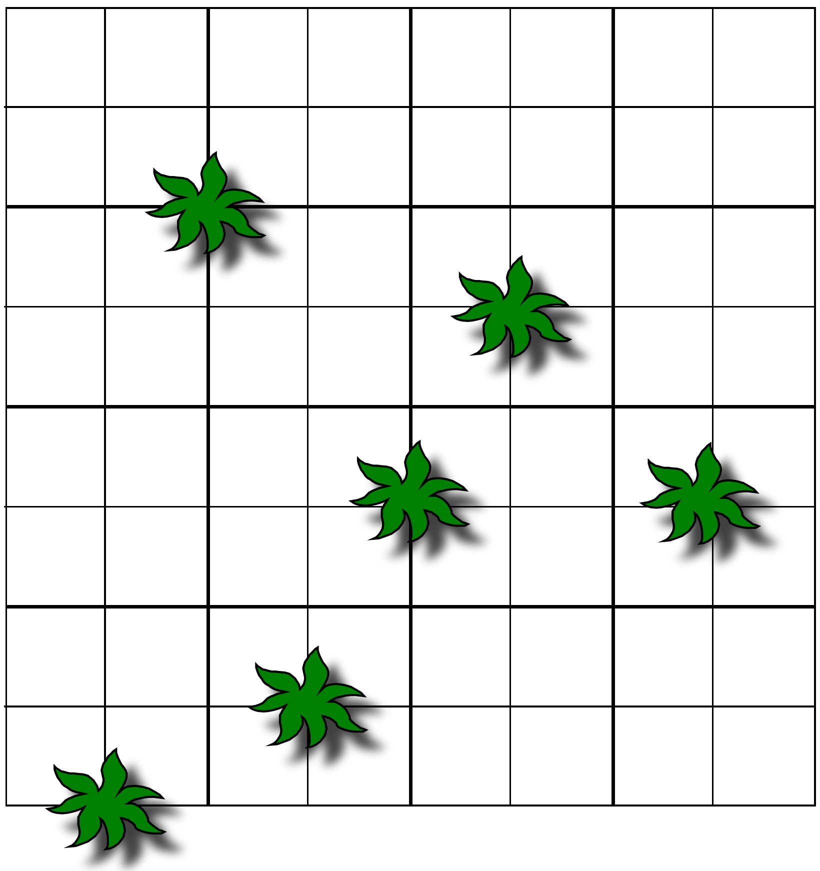
- Yule process
- Gibrat principle
- Matthew effect
- Cumulative advantage
- Preferential attachement
- ‘rich get richer’



This presentation

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

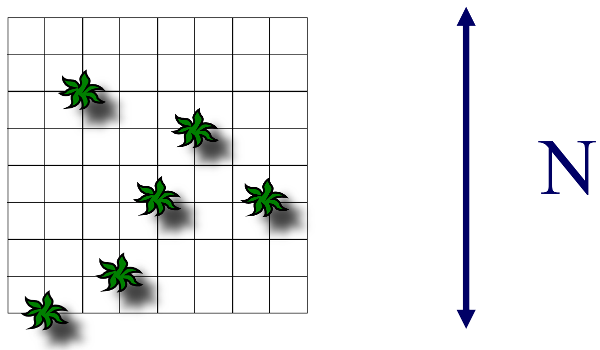
Percolation and forest fires

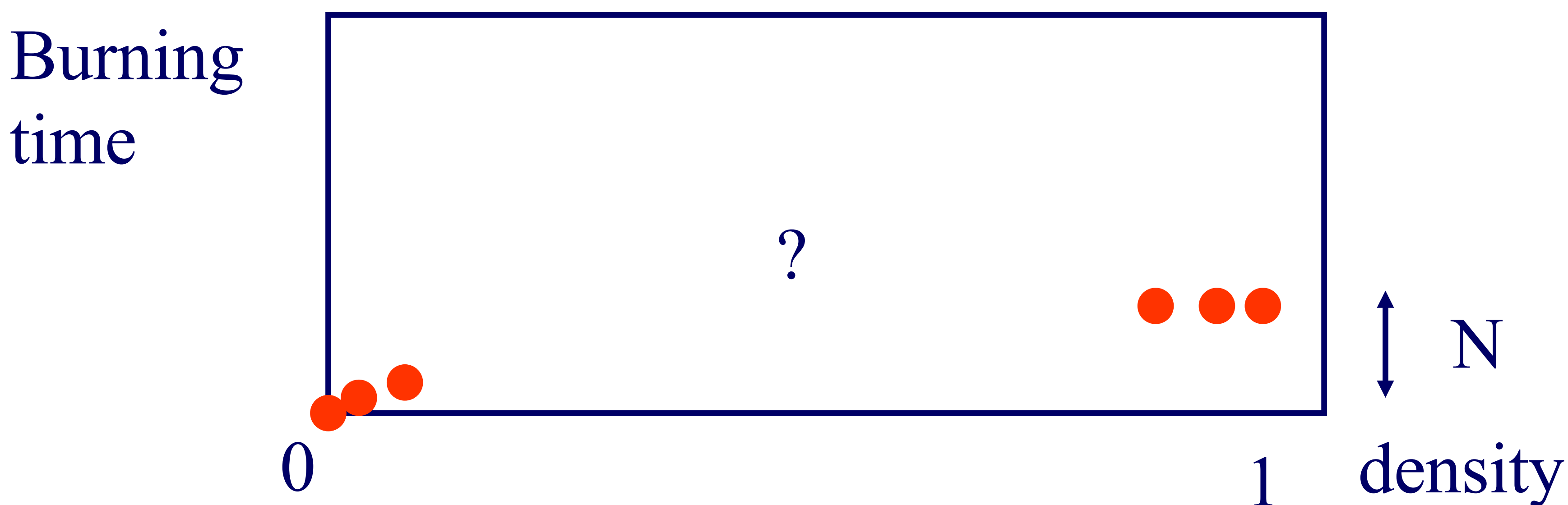


A burning tree will cause its neighbors to burn next.

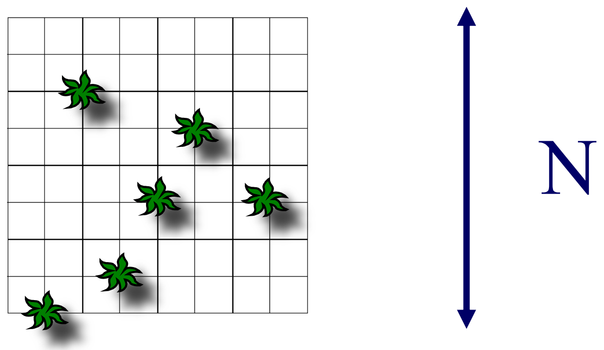
Which tree density *p* will cause the fire to last longest?

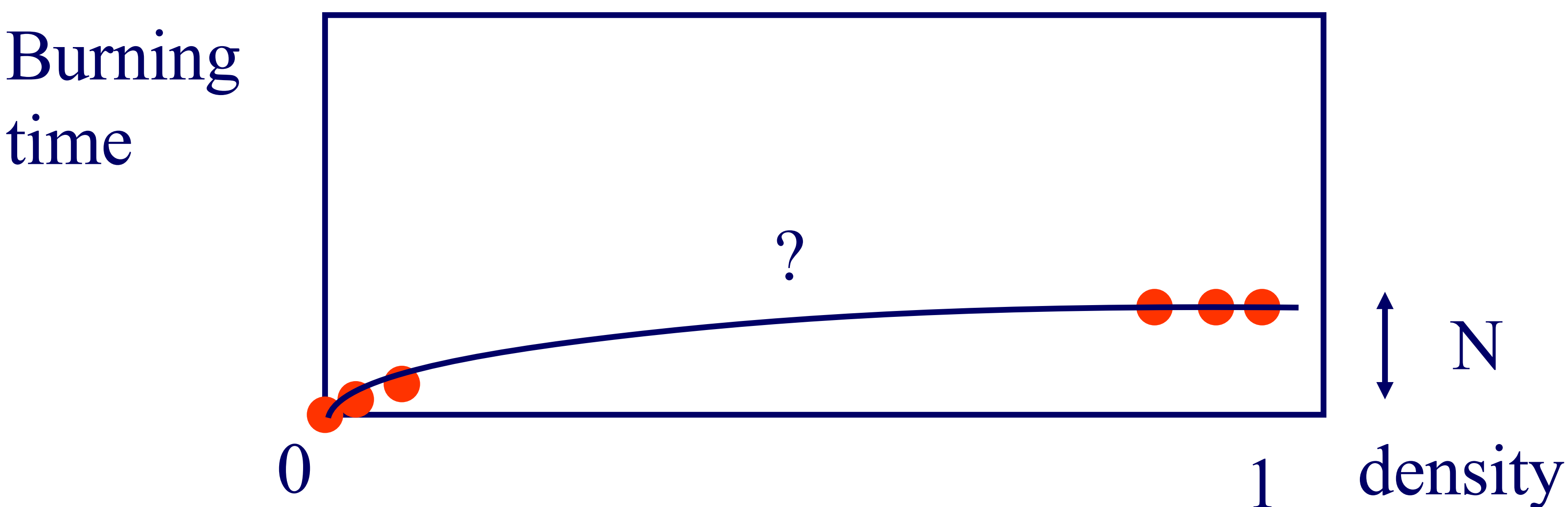
Percolation and forest fires



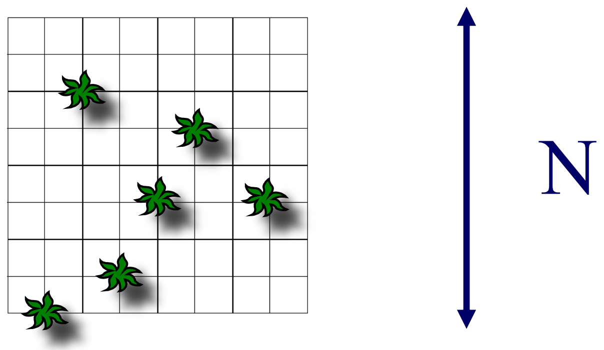


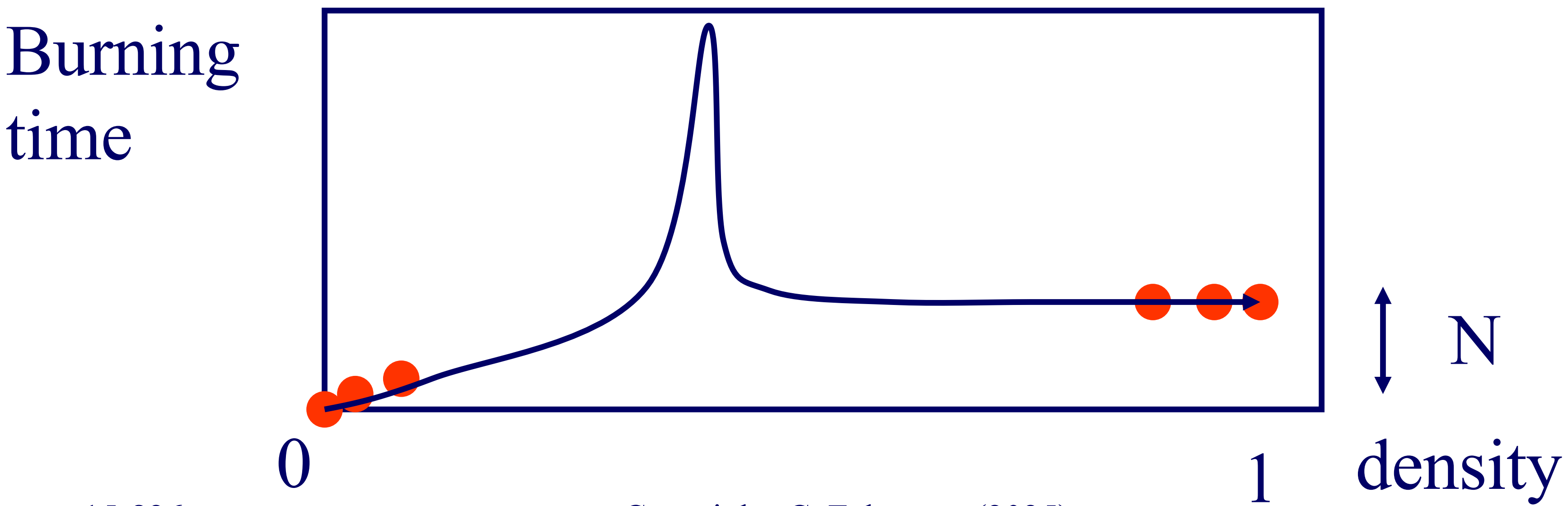
Percolation and forest fires



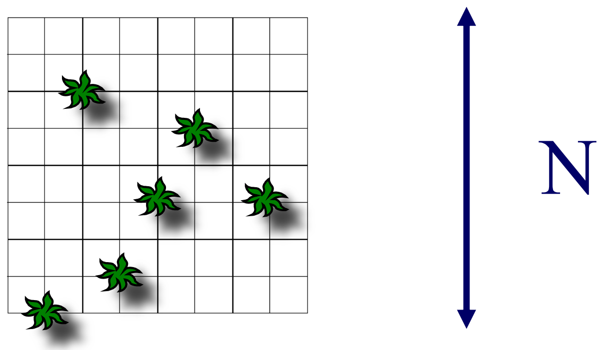


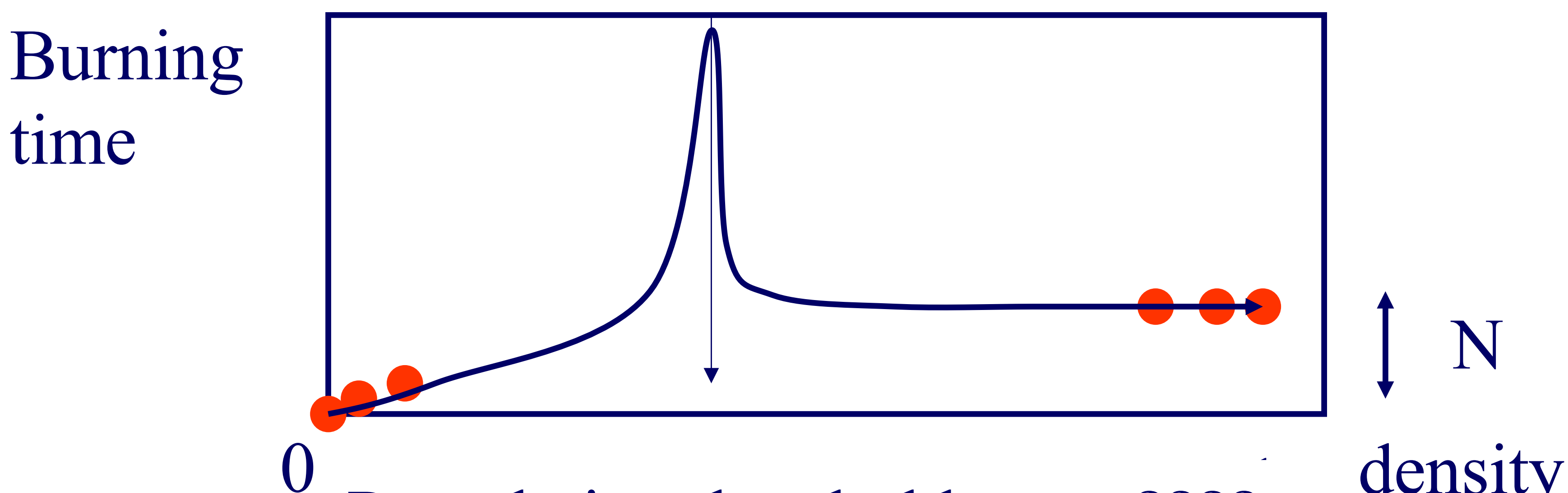
Percolation and forest fires





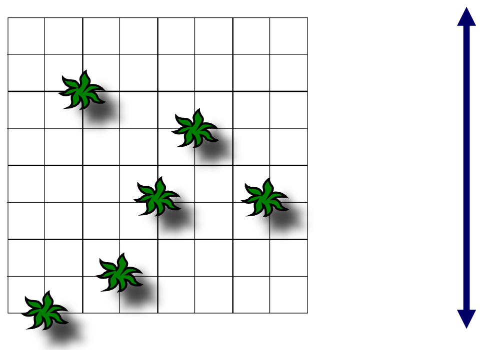
Percolation and forest fires

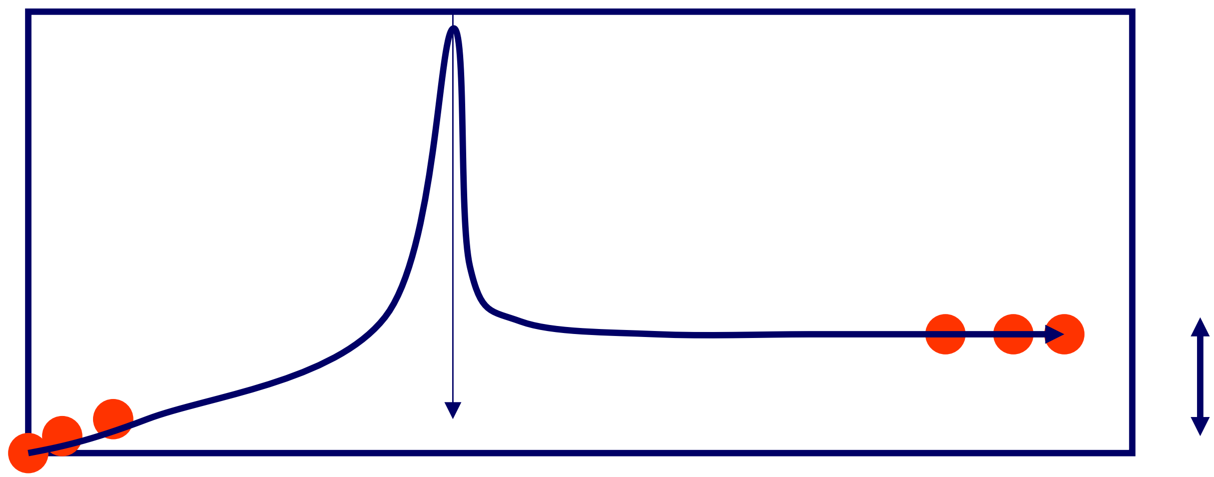




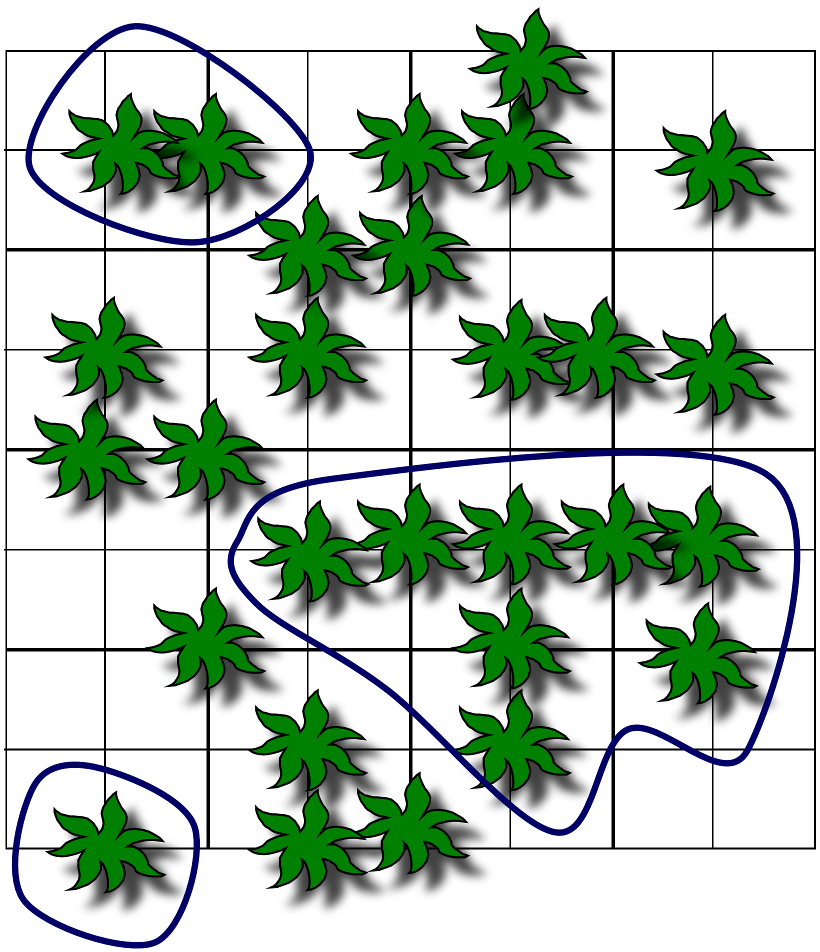
Percolation threshold, pc ~ ????

Percolation and forest fires





Percolation and forest fires



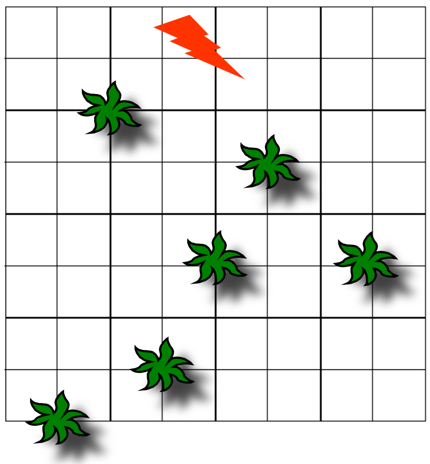
At pc ~ 0.593:
No characteristic scale;
'patches' of all sizes;
Korcak-like 'law'.

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

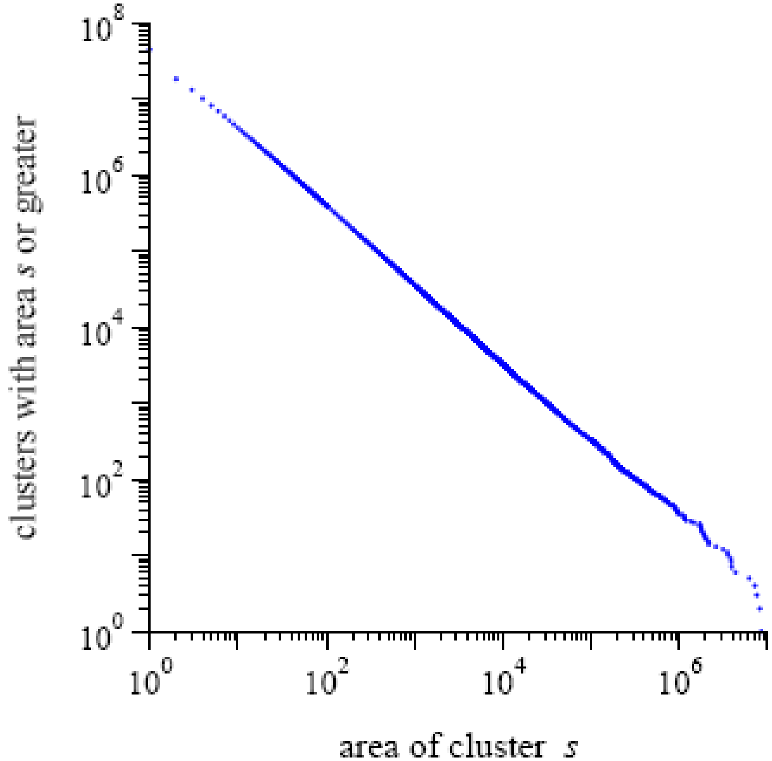
- Trees appear at random (eg., seeds, by the wind)
- Fires start at random (eg., lightning)
- Q1: What is the distribution of size of forest fires?



Self-organized criticality

- A1: Power law-like

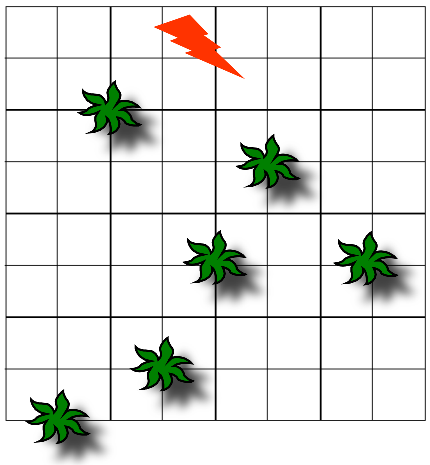
CCDF



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?



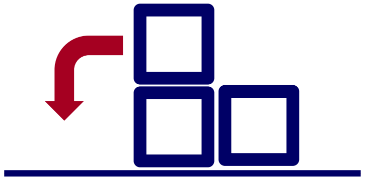
Self-organized criticality

- A2: the critical density pc ~ 0.593

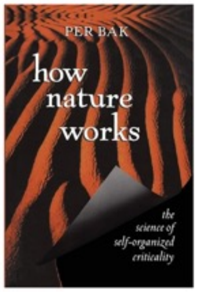
Carnegie Mellon

Self-organized criticality

- [Bak]: size of avalanches ~ power law:
- Drop a grain randomly on a grid
- It causes an avalanche if 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))

- Log(C(t)) = log(C) + log(..) + log(..) … -> Gaussian

Other - lognormal

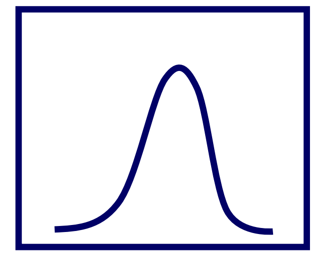
Random multiplication:

- Log(C(t)) = log(C) + log(..) + log(..) … -> Gaussian

- Thus C(t) = exp(Gaussian)
- By definition, this is Lognormal

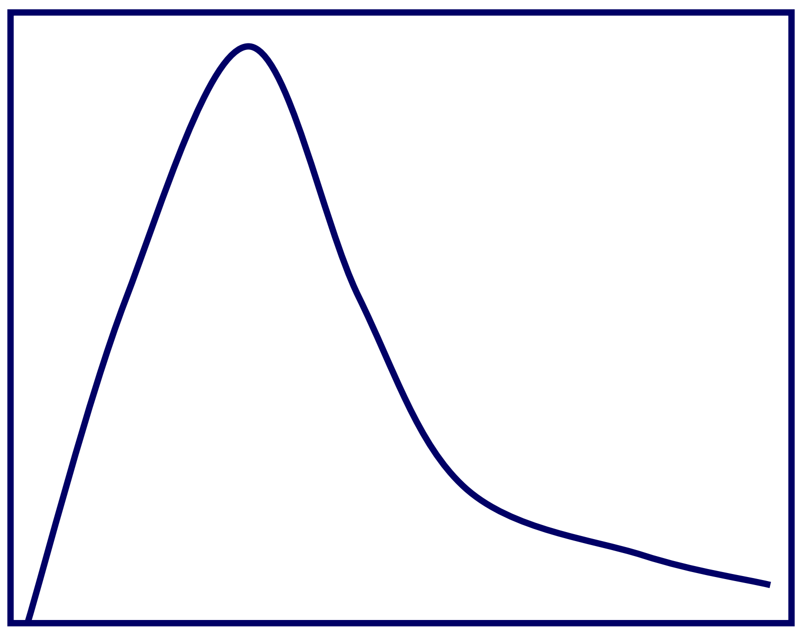
Other - lognormal

pdf



Lognormal:

pdf



h = body height

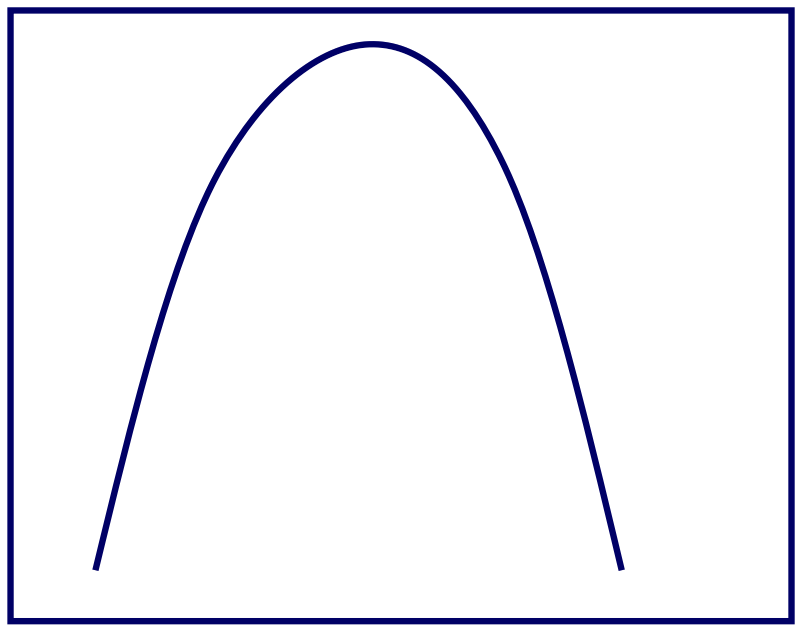
$\$ = e^h$

0

Other - lognormal

Lognormal:

log(pdf)

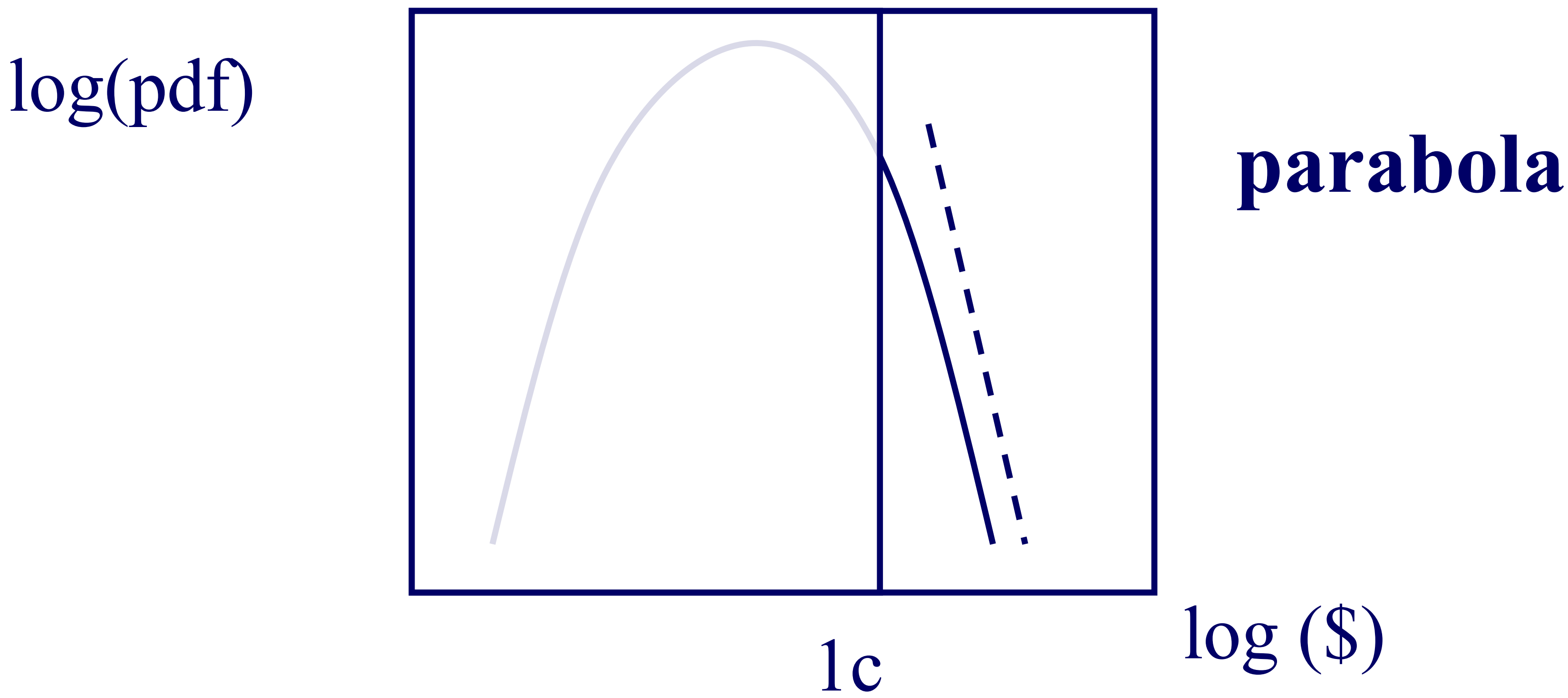


parabola

log ($)

Others

Lognormal:



Other - lognormal

- Random multiplication
- Fragmentation

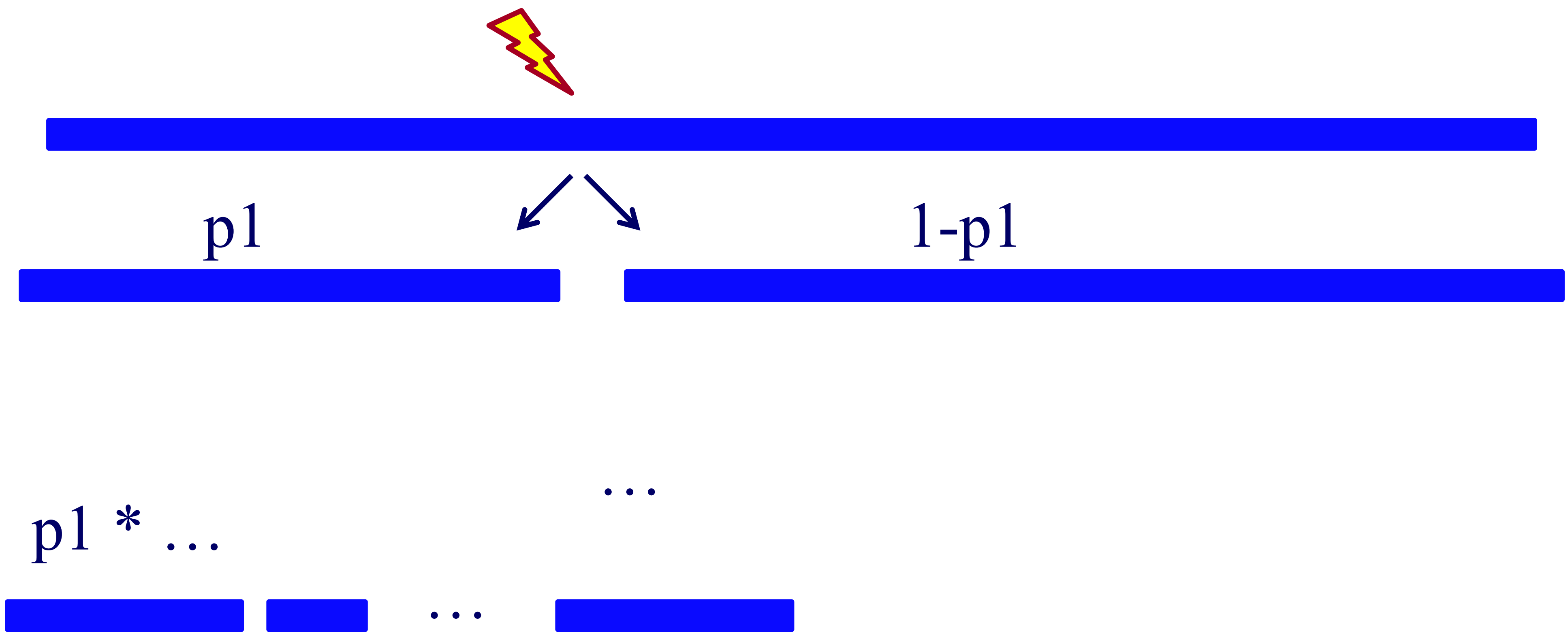
-> lead to lognormals (~ look like power laws)

Other - lognormal

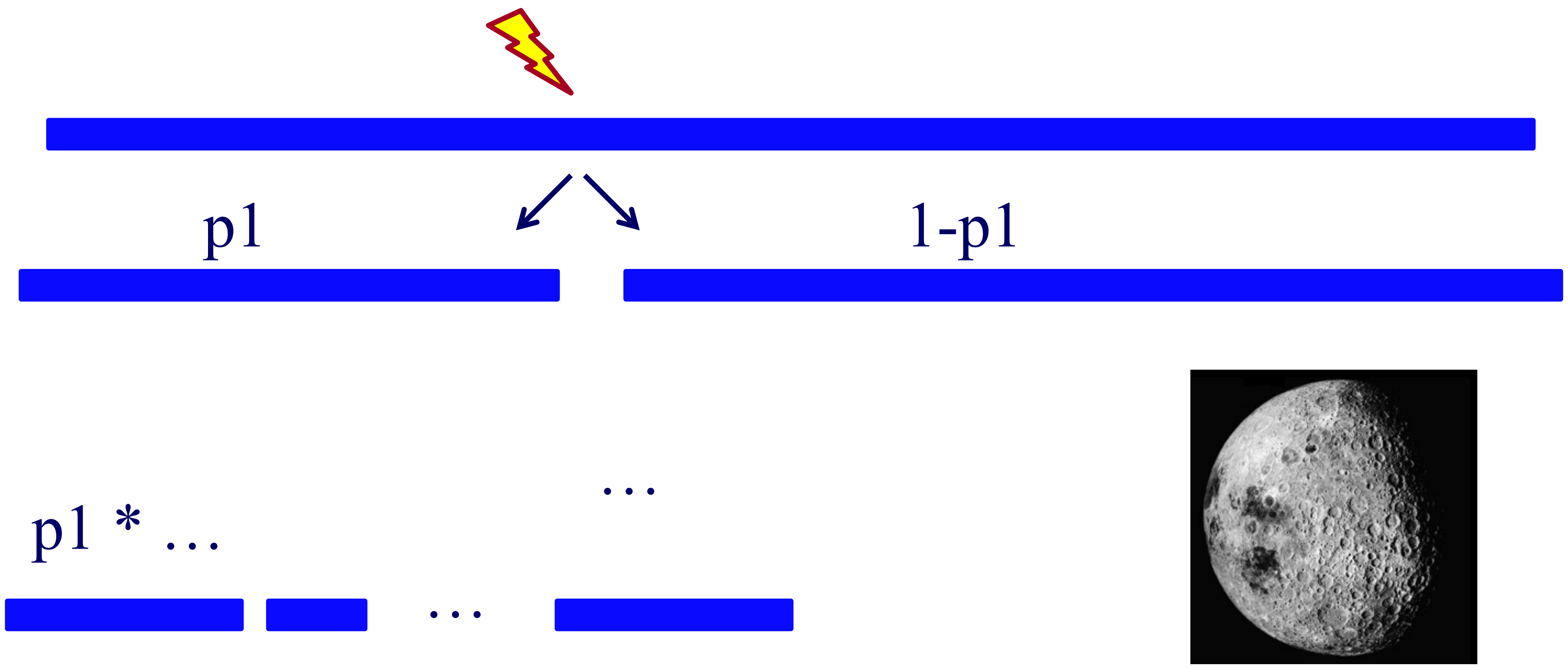
- Stick of length 1
- Break it at a random point x (0<x<1)
- Break each of the pieces at random

- Resulting distribution: lognormal (why?)

Fragmentation -> lognormal



Fragmentation -> lognormal



This presentation

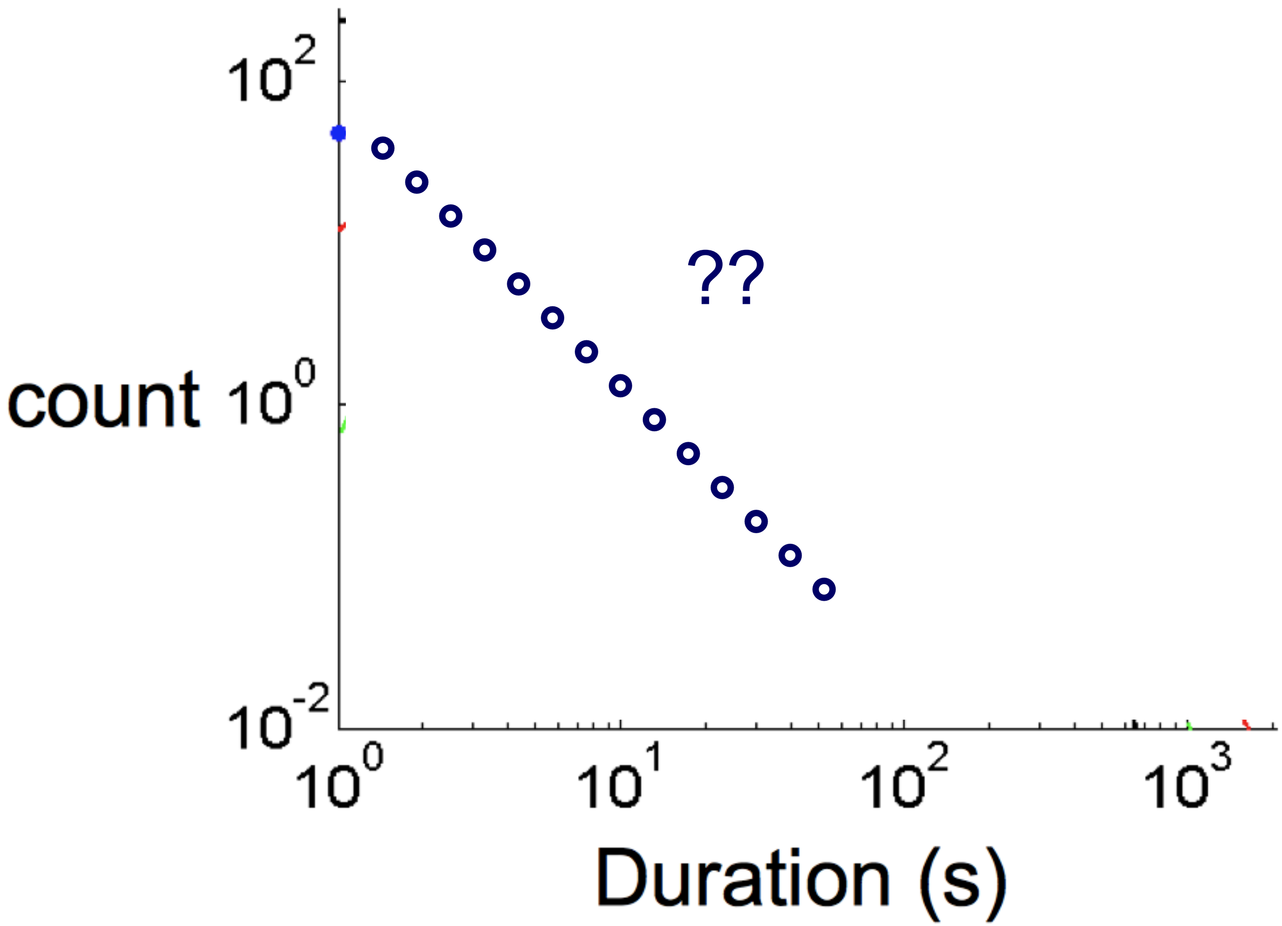
- Definitions - clarification
- Examples and counter-examples
- Generative mechanisms
  - Combination of exponentials
  - Inverse
  - Random walk
  - Yule distribution = CRP
  - Percolation
  - Self-organized criticality
  - Other – lognormal
  - Other – log-logistic (NOT in [Newman 2005])

Duration of phonecalls

Pedro O. S. Vaz de Melo, Leman Akoglu, Christos Faloutsos, Antonio Alfredo Ferreira Loureiro: *Surprising Patterns for the Call Duration Distribution of Mobile Phone Users*. ECML/PKDD 2010

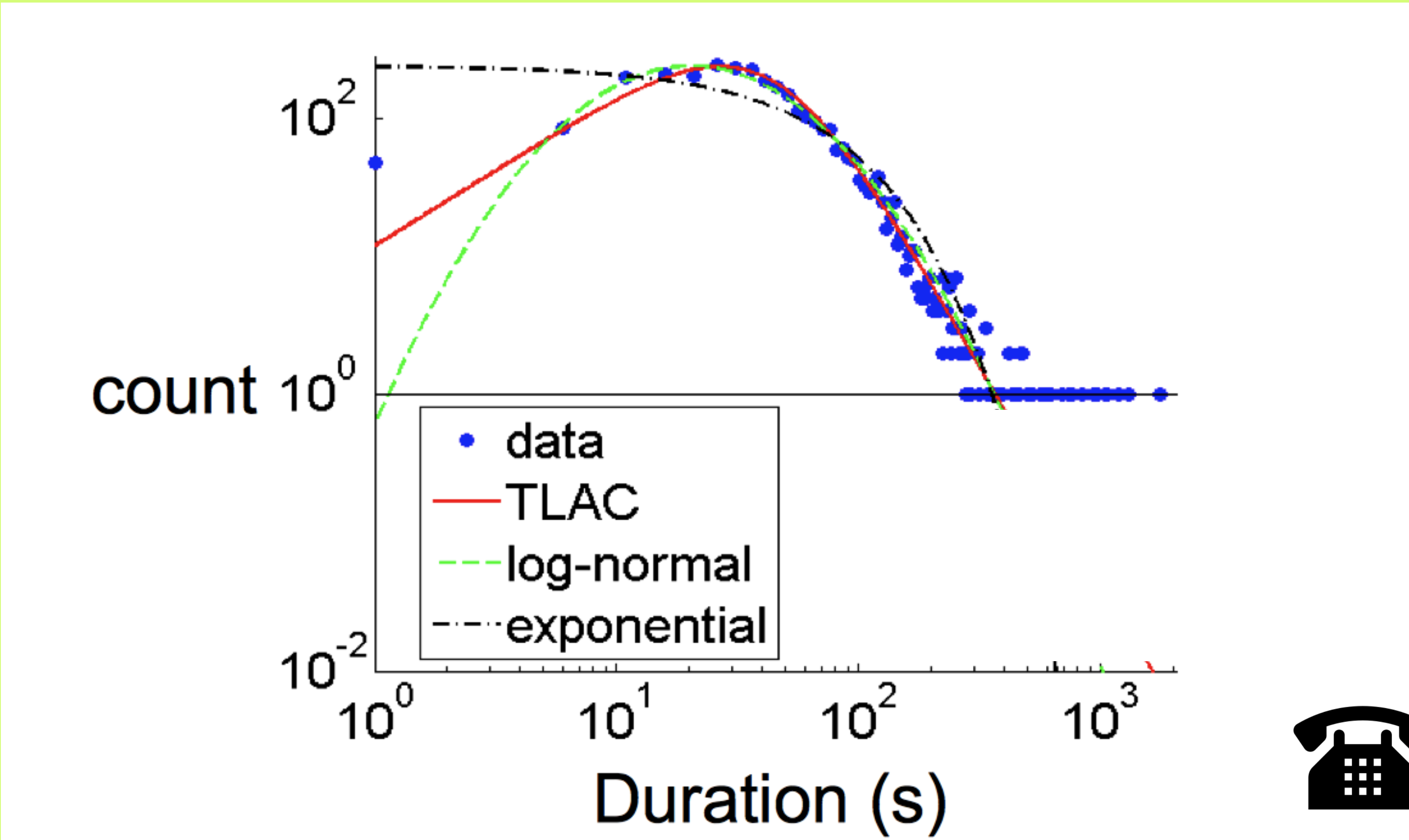
Reminder

Probably, power law (?)



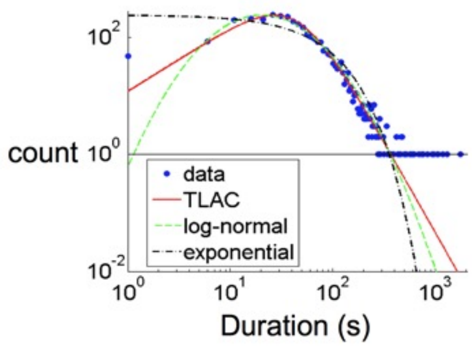
Reminder

No Power Law!



'TLaC: Lazy Contractor'

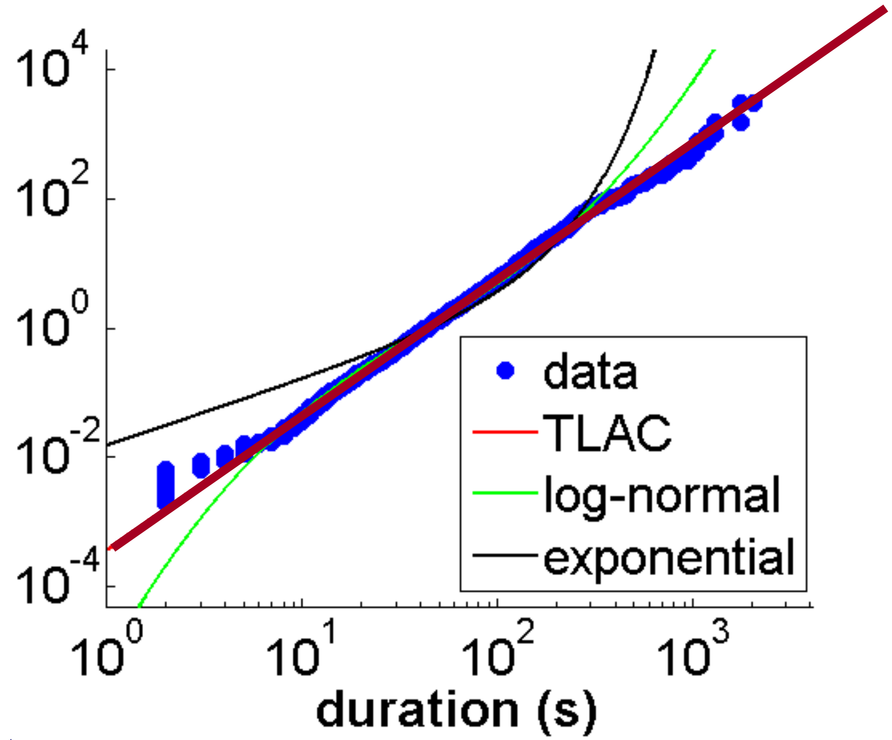
- The longer a task (phonecall) has taken,
- The even longer it will take



Odds ratio=

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

== power law



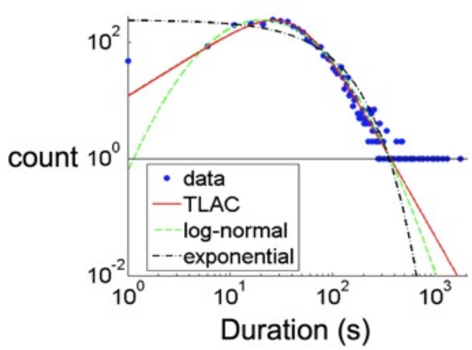
Log-logistic distribution

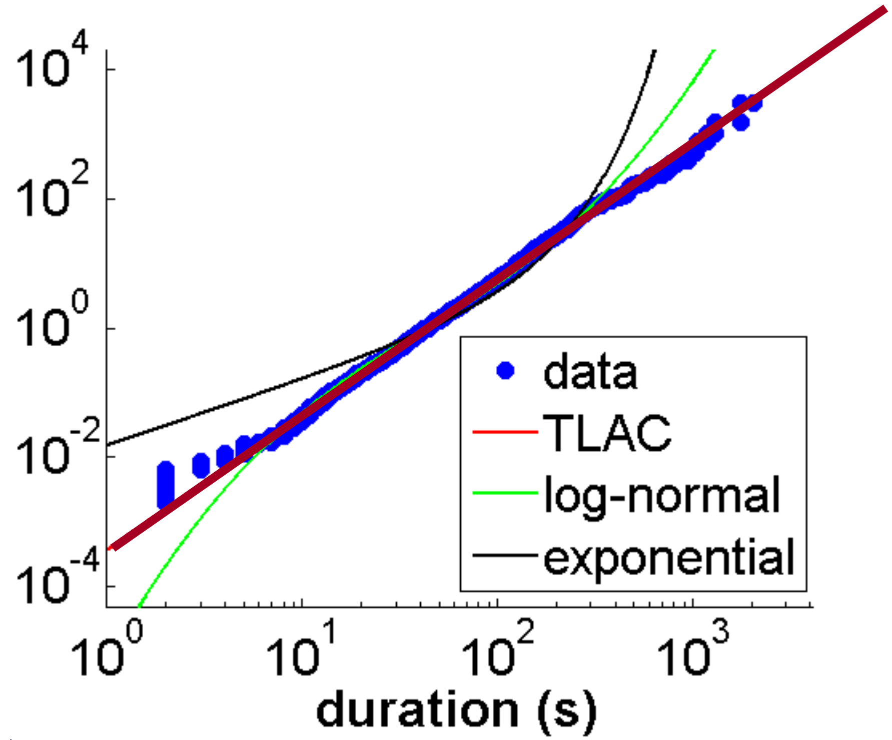
- CDF(t)/(1- CDF(t)) == OR(t)
- For log-logistic: log[OR(t)] = β + ρ*log(t)

Odds ratio=

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

== power law

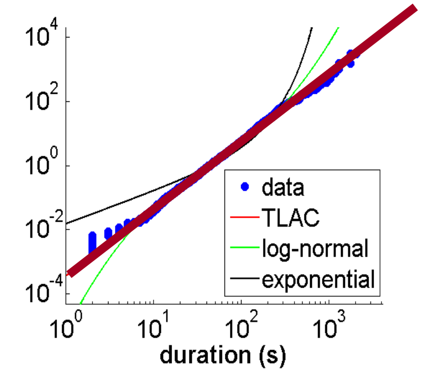


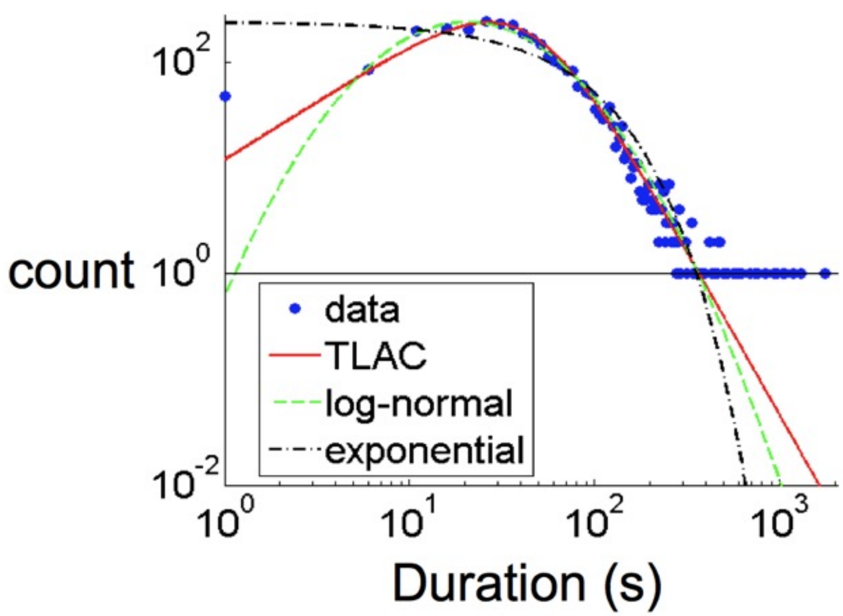


Log-logistic distribution

- CDF(t)/(1- CDF(t)) == OR(t)
- For log-logistic: log[OR(t)] = β + ρ*log(t)

OR(t)

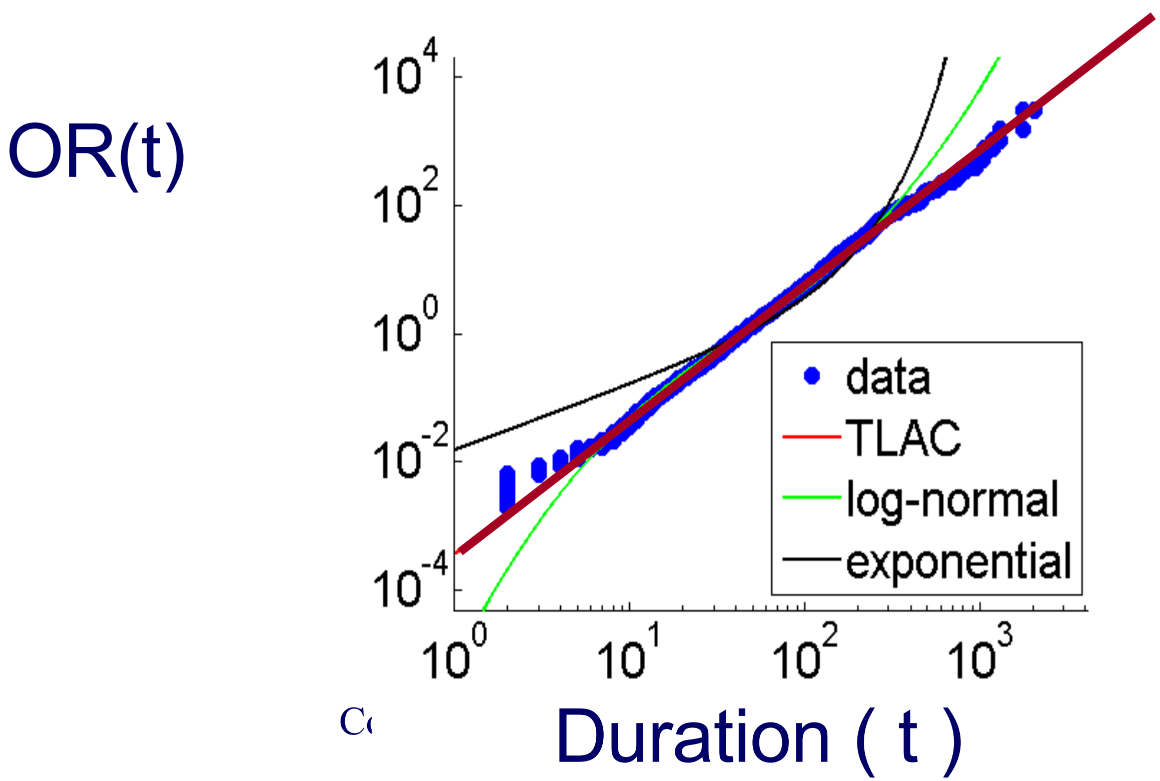




- PDF looks like hyperbola;
- and, if clipped, like power-law

Log-logistic distribution

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



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.

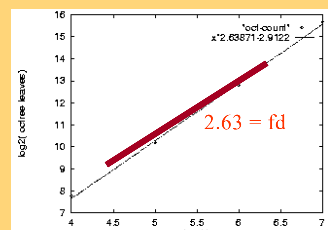
ASONAM 2015, pp 367-374.

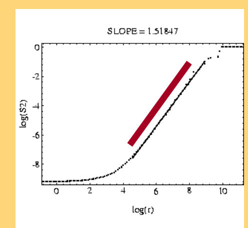
Conclusions

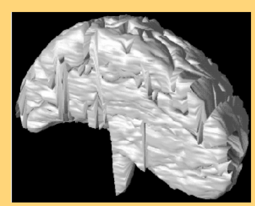
- Power laws and power-law like distributions appear often
- (fractals/self similarity -> power laws)
- Exponentiation/inversion
- Yule process / CRP / rich get richer
- Criticality/percolation/phase transitions
- Fragmentation -> lognormal ~ P.L.

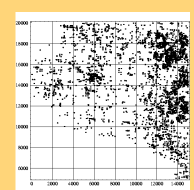
Conclusions - 1

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





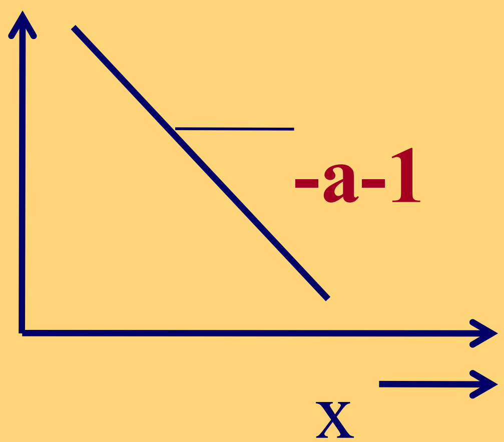


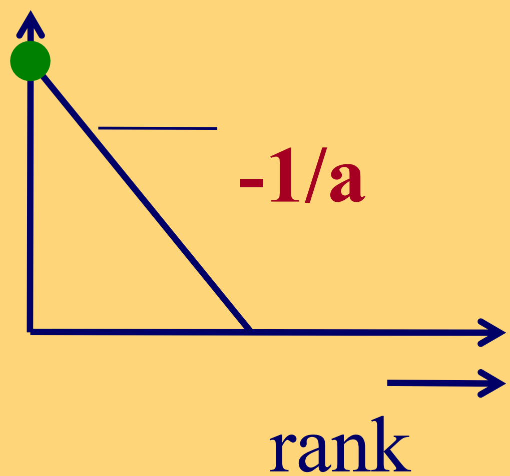


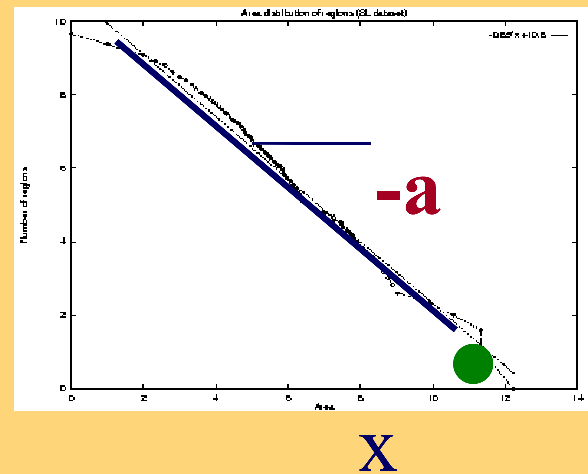
Conclusions 2: 3 versions of P.L.

PDF = frequency-count plot	Zipf plot = Rank-frequency	NCDF = CCDF
Prob(area = x)	area	Prob(area >= x)

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

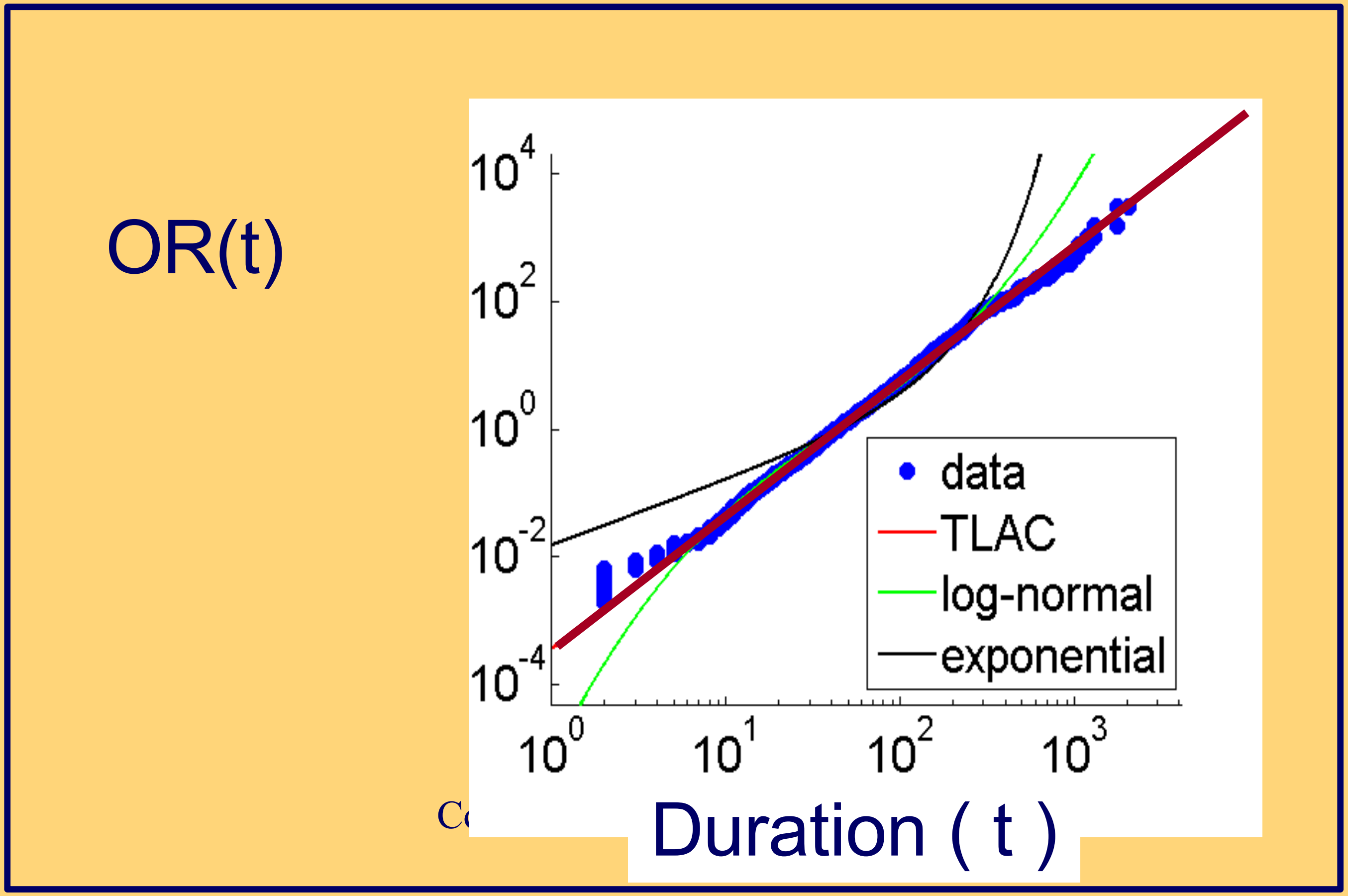






Conclusions-3: Odds ratio

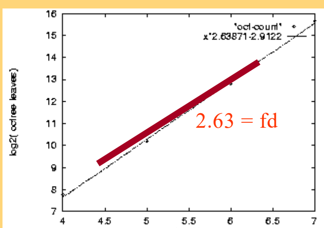
- CDF(t)/(1- CDF(t)) == OR(t)
- For log-logistic: $\boxed{\log[OR(t)] = \beta + \rho * \log(t)}$

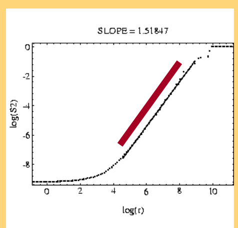


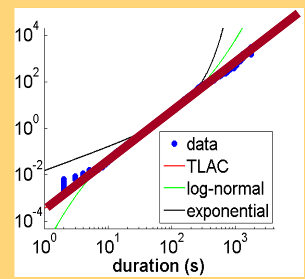
Conclusions 1-3:

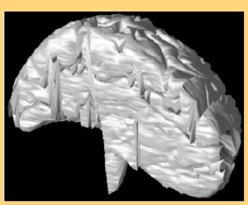
Take logarithms

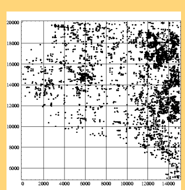
of PDF, or CCDF or Odds-ratio

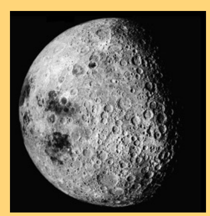












References

- *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)