CMU SCS

# 15-826: Multimedia Databases and Data Mining

Lecture #13: Power laws

Potential causes and explanations

*C. Faloutsos*

CMU SCS

## 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



CMU SCS

## 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.



CMU SCS

## Outline

Goal: 'Find similar / interesting things'

- Intro to DB
- ➡ Indexing - similarity search
- Data Mining

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
- ...



CMU SCS

# Indexing - Detailed outline

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



CMU SCS

# This presentation

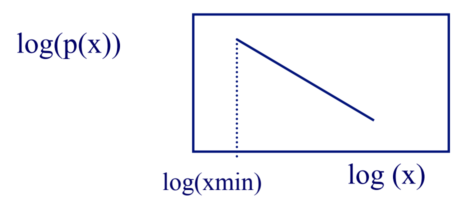
- Definitions
- Clarification: 3 forms of P.L.
- Examples and counter-examples
- Generative mechanisms



CMU SCS

# Definition

- p(x) = C x ^ (-a) (x >= xmin)
- Eg., prob( city pop. between x + dx)



CMU SCS

## For discrete variables

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

Or, the Yule distribution:

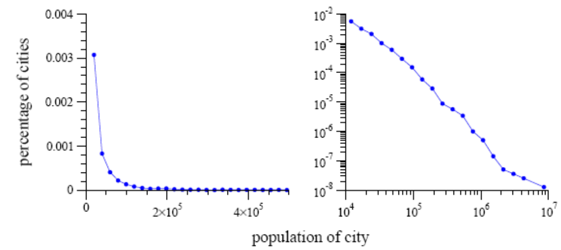
$$p_k = C\, B(k,a)$$

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

15-826 Copyright: C. Faloutsos (2017) 9

CMU SCS

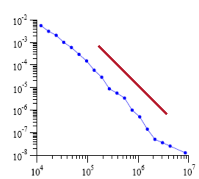
## [Newman, 2005]




15-826 Copyright: C. Faloutsos (2017) 10

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 (2017) 11

CMU SCS

## This presentation

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

15-826 Copyright: C. Faloutsos (2017) 12

CMU SCS

## Jumping to the conclusion:



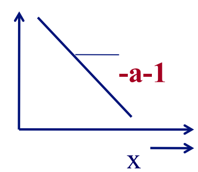
CMU SCS

## 3 versions of P.L.

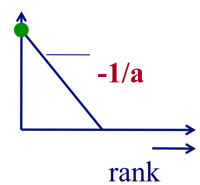
PDF = frequency-count plot | Zipf plot = Rank-frequency | NCDF = CCDF

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

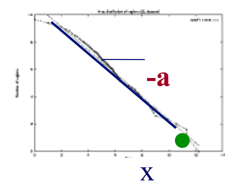
Prob( area = x )




area




Prob( area >= x )






CMU SCS

## Details, and proof sketches:

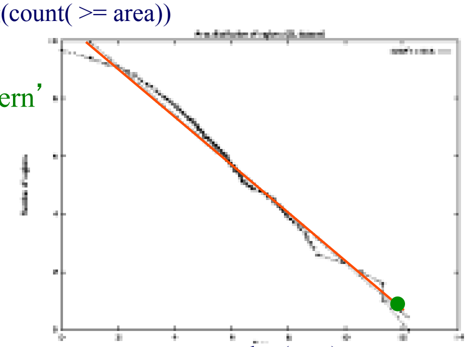


CMU SCS

**Reminder**

## More power laws: areas – Korcak's law


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

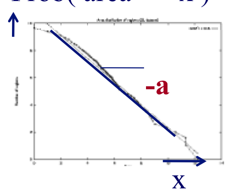


CMU SCS

# 3 versions of P.L.

NCDF = CCDF

Prob( area >= x )





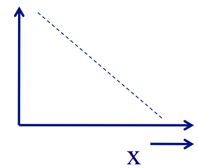

CMU SCS

# 3 versions of P.L.

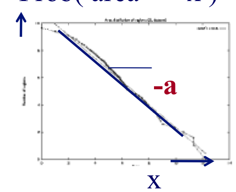
PDF

NCDF = CCDF

Prob( area = x )




Prob( area >= x )





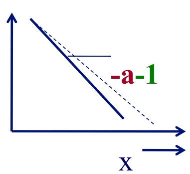

CMU SCS

# 3 versions of P.L.

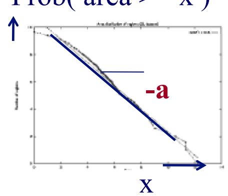
PDF

NCDF = CCDF

Prob( area = x )




Prob( area >= x )





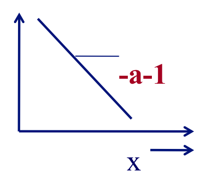

CMU SCS

# 3 versions of P.L.

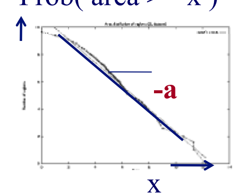
PDF

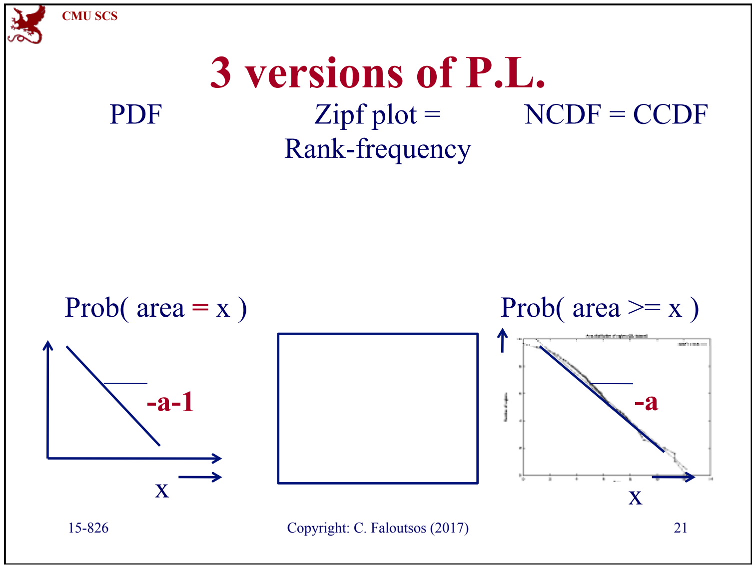
NCDF = CCDF

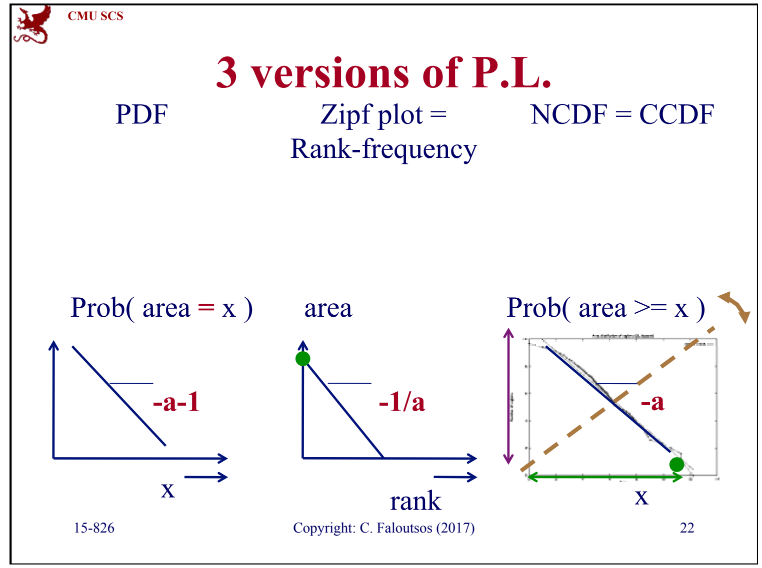
Prob( area = x )

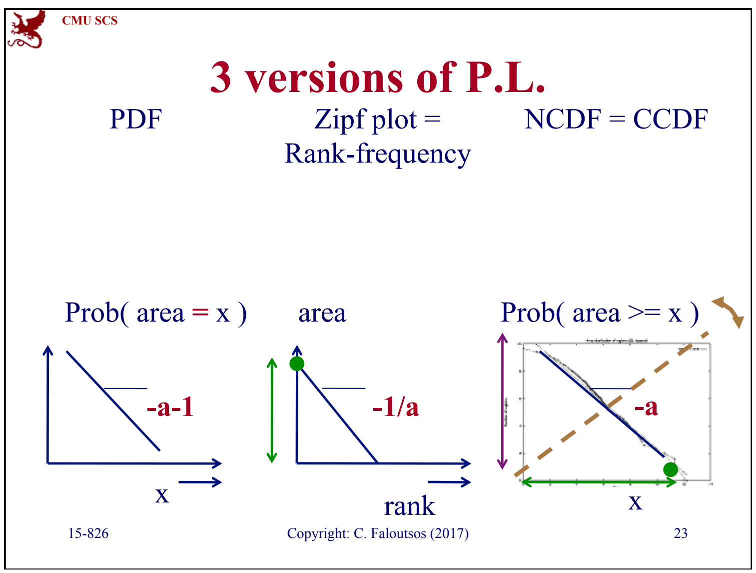


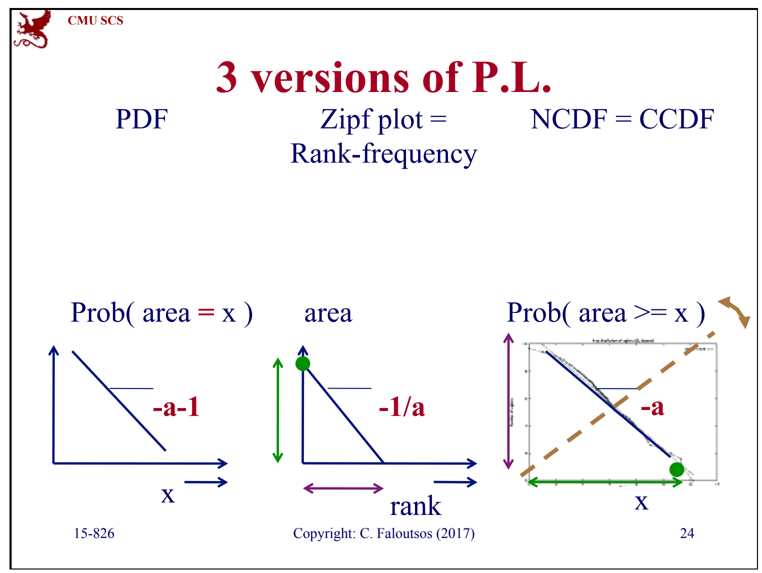

Prob( area >= x )




CMU SCS
3 versions of P.L.
PDF
Zipf plot =
Rank-frequency
NCDF = CCDF
Prob( area = x )
Prob( area >= x )
-a-1
-a
x
x
15-826
Copyright: C. Faloutsos (2017)
21


CMU SCS
3 versions of P.L.
PDF
Zipf plot =
Rank-frequency
NCDF = CCDF
Prob( area = x )
area
Prob( area >= x )
-a-1
-1/a
-a
x
rank
x
15-826
Copyright: C. Faloutsos (2017)
22

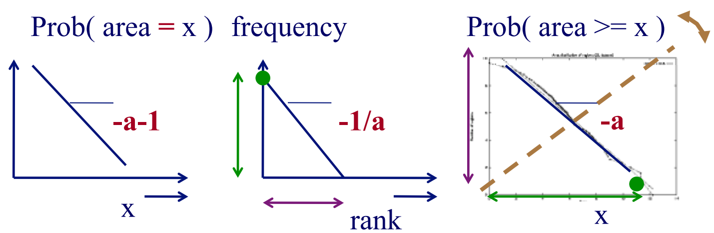

CMU SCS
3 versions of P.L.
PDF
Zipf plot =
Rank-frequency
NCDF = CCDF
Prob( area = x )
area
Prob( area >= x )
-a-1
-1/a
-a
x
rank
x
15-826
Copyright: C. Faloutsos (2017)
23


CMU SCS
3 versions of P.L.
PDF
Zipf plot =
Rank-frequency
NCDF = CCDF
Prob( area = x )
area
Prob( area >= x )
-a-1
-1/a
-a
x
rank
x
15-826
Copyright: C. Faloutsos (2017)
24

CMU SCS

# 3 versions of P.L.

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

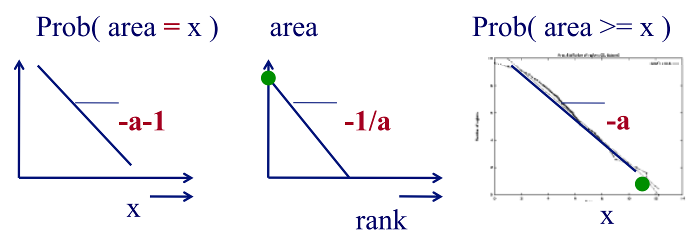



15-826 Copyright: C. Faloutsos (2017) 25

CMU SCS

# 3 versions of P.L.

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

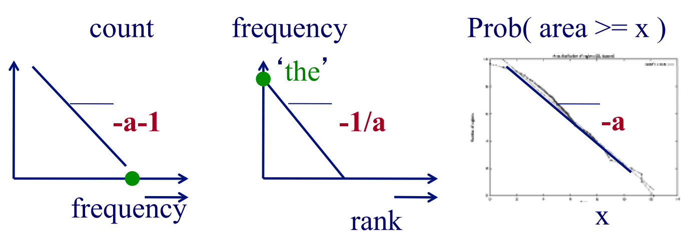



15-826 Copyright: C. Faloutsos (2017) 26

CMU SCS

# 3 versions of P.L.

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




15-826 Copyright: C. Faloutsos (2017) 27

CMU SCS

# Sanity check:

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

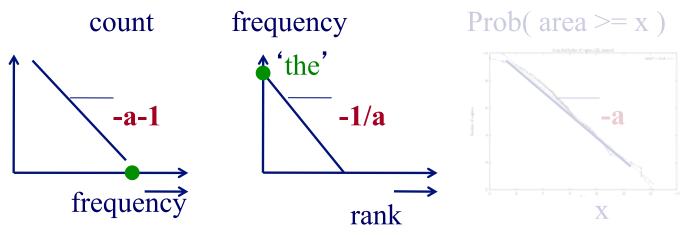
15-826 Copyright: C. Faloutsos (2017) 28

CMU SCS

## 3 versions of P.L.

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

slope = -2 ⟺ slope = -1



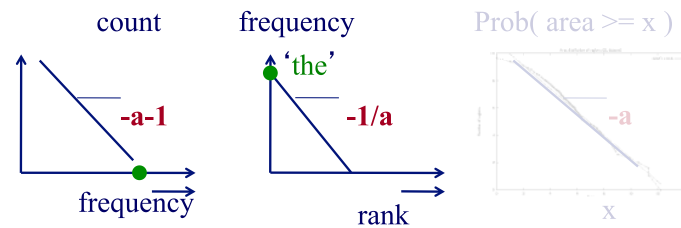



CMU SCS

## 3 versions of P.L.

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

✓ slope = -2 ⟺ slope = -1



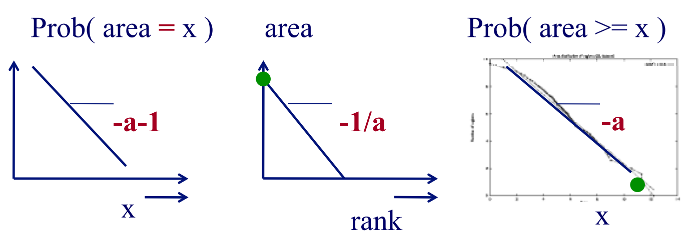



CMU SCS

## 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**






CMU SCS

## This presentation

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

CMU SCS

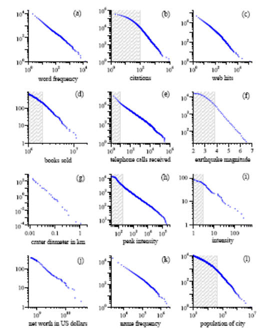
# Examples

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

15-826 Copyright: C. Faloutsos (2017) 33

CMU SCS

# [Newman 2005]



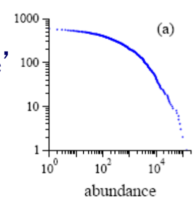
Rank-frequency plots
Or Cumulative D.F.

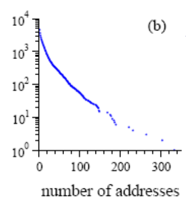
15-826 Copyright: C. Faloutsos (2017) 34

CMU SCS

# NOT following P.L.

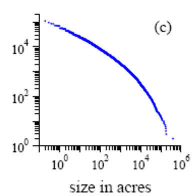
'abundance' of species




Number of addresses

Cumul. D.F.




Size of forest fires

15-826 35

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 (2017) 36

CMU SCS

## 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)

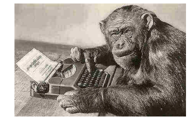
$p(x) = p(y)*dy/dx = 1/b\ x^{(-1+a/b)}$

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

15-826 Copyright: C. Faloutsos (2017) 37

CMU SCS

## 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]:

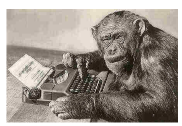
$a = [2 \ln(m) - \ln (1 - q_s)] / [\ln m - \ln (1 - q_s)]$

15-826 Copyright: C. Faloutsos (2017) 38

CMU SCS

## Combination of exponentials

- Most freq 'words'?

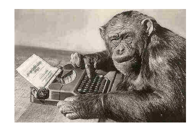


15-826 Copyright: C. Faloutsos (2017) 39

CMU SCS

## Combination of exponentials

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



15-826 Copyright: C. Faloutsos (2017) 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  41

CMU SCS

## 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.

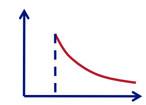
*y-> speed*

*x-> travel time*

*y*:

0mph…..…1mph

count



Travel time

 42

CMU SCS

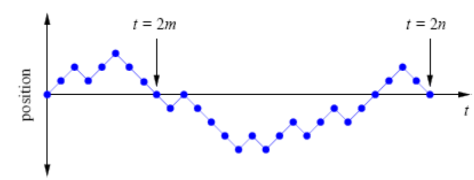
## This presentation

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

15-826  43

CMU SCS

## Random walks

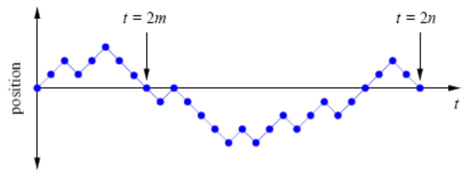



Inter-arrival times PDF: $p(t) \sim$ ??

15-826  44

CMU SCS

# 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))

CMU SCS

# Random walks

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

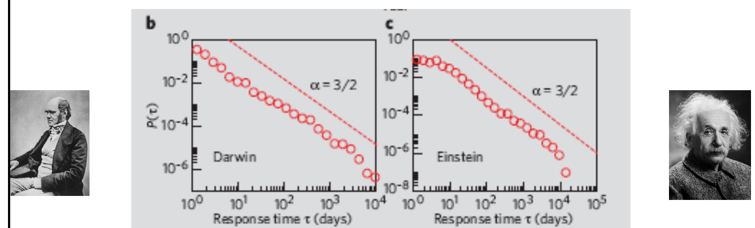



**Figure 1 | The correspondence patterns of Darwin and Einstein.**



CMU SCS

# This presentation

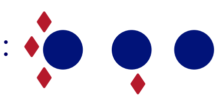
- Definitions - clarification
- Examples and counter-examples
- Generative mechanisms
  - Combination of exponentials
  - Inverse
  - Random walk
  - ➡ Yule distribution = CRP
  - Percolation
  - Self-organized criticality
  - Other



CMU SCS

# 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

CMU SCS

## Yule distribution and CRP

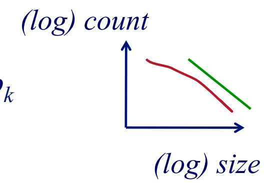
Then:

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)




15-826 Copyright: C. Faloutsos (2017) 49

CMU SCS

## Yule distribution and CRP

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

15-826 Copyright: C. Faloutsos (2017) 50

CMU SCS

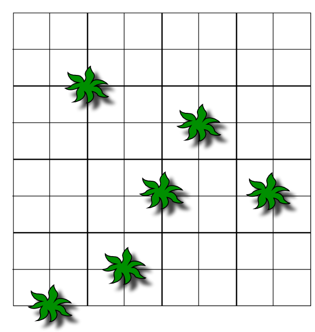
## This presentation

- 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 (2017) 51

CMU SCS

## Percolation and forest fires



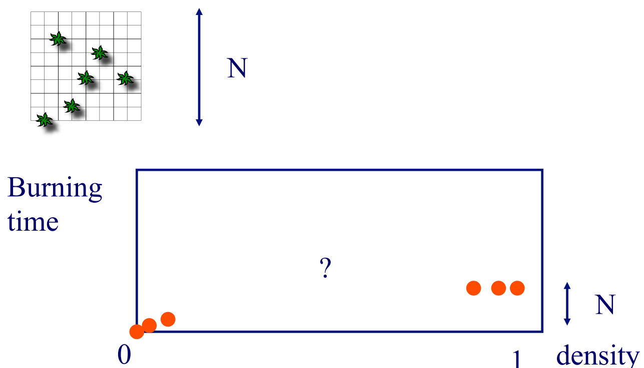
A burning tree will cause its neighbors to burn next.

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

15-826 Copyright: C. Faloutsos (2017) 52

CMU SCS

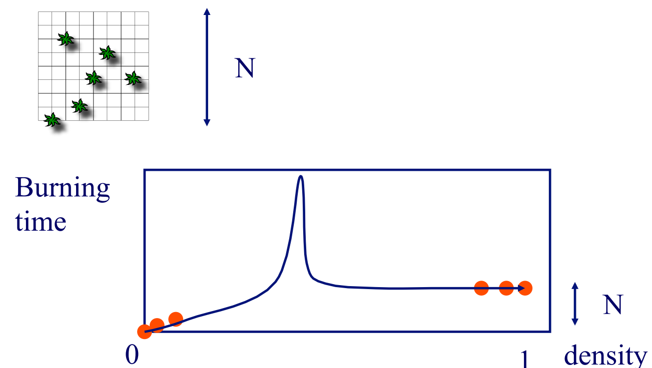
## Percolation and forest fires



CMU SCS

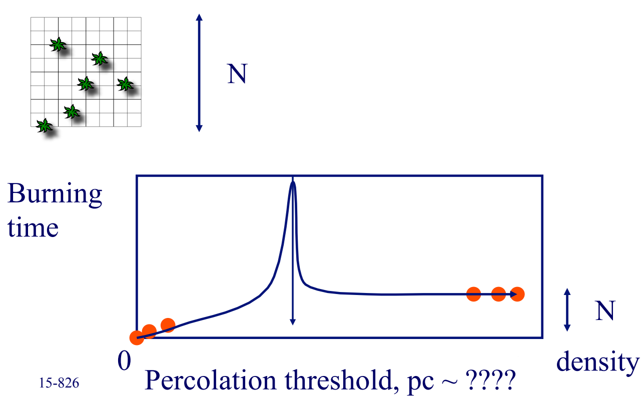
## Percolation and forest fires



CMU SCS

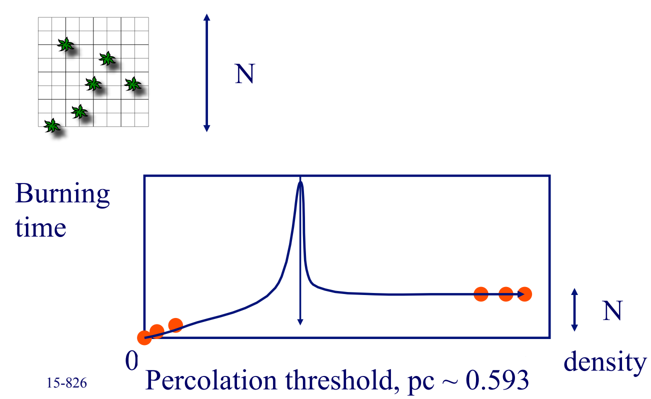
## Percolation and forest fires



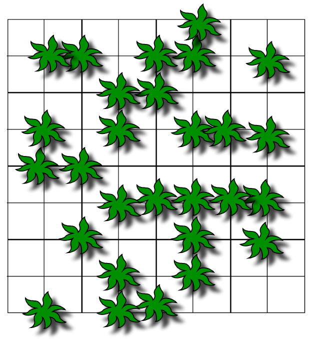



CMU SCS

## Percolation and forest fires



## Percolation and forest fires



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





## This presentation

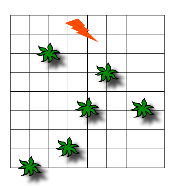
- Definitions - clarification
- Examples and counter-examples
- Generative mechanisms
  - Combination of exponentials
  - Inverse
  - Random walk
  - Yule distribution = CRP
  - Percolation
  - ➡ Self-organized criticality
  - Other





## 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?



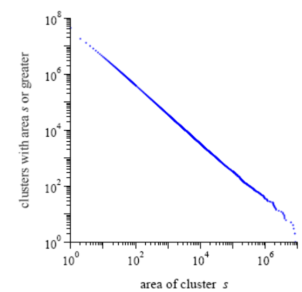




## Self-organized criticality

- A1: Power law-like

CCDF

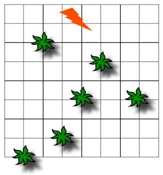



Area of cluster s

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?





CMU SCS

## Self-organized criticality

- A2: the critical density pc ~ 0.593



CMU SCS

## 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]



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

CMU SCS

## Other

- Random multiplication
- Fragmentation

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



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(..) + log(..) … -> Gaussian



CMU SCS

## Others

Random multiplication:

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

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

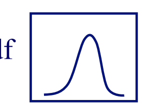


CMU SCS

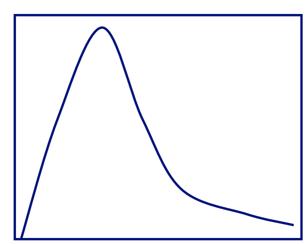
## Others

Lognormal:

pdf



pdf



0

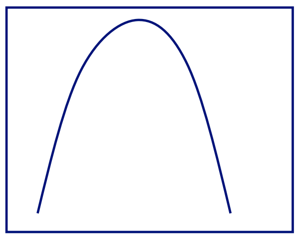
$\$ = e^h$

$h$ = body height

CMU SCS

# Others

Lognormal:

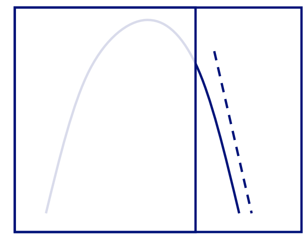





CMU SCS

# Others

Lognormal:






CMU SCS

# Other

- Random multiplication
- ➡ Fragmentation

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



CMU SCS

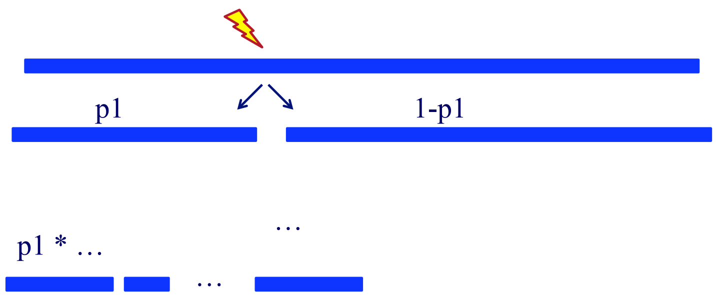
# Other

- 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?)

CMU SCS

## Fragmentation -> lognormal






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.



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)