

CMU SCS

# 15-826: Multimedia Databases and Data Mining

*Fractals - introduction*  
C. Faloutsos

---

---


---

---

---

---

---



CMU SCS

## Outline

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

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

15-826 Copyright: C. Faloutsos (2007) 2

---

---


---

---

---

---

---



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 (2007) 3

---

---

---

---

---

---

---

CMU SCS

## Intro to fractals - outline

➔

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems
- More examples and tools
- Discussion - putting fractals to work!
- Conclusions – practitioner's guide
- Appendix: gory details - boxcounting plots

15-826 Copyright: C. Faloutsos (2007) 4

---

---

---

---

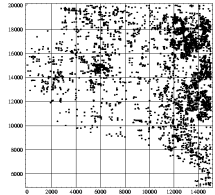
---

---

---

CMU SCS

## Problem #1: GIS - points



Road end-points of Montgomery county:

- Q1: how many d.a. for an R-tree?
- Q2 : distribution?
  - not uniform
  - not Gaussian
  - no rules??

15-826 Copyright: C. Faloutsos (2007) 5

---

---

---

---

---

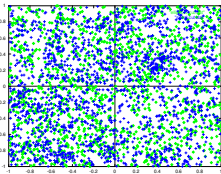
---

---

CMU SCS

## Problem #2 - spatial d.m.

Galaxies (Sloan Digital Sky Survey w/ B. Nichol)



- 'spiral' and 'elliptical' galaxies  
(stores and households ...)
- patterns?
- attraction/repulsion?
- how many 'spi' within  $r$  from an 'ell'?

15-826 Copyright: C. Faloutsos (2007) 6

---

---


---

---

---

---

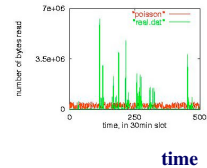
---



CMU SCS

### Problem #3: traffic

# bytes



time

- disk trace (from HP - J. Wilkes); Web traffic - fit a model
- how many explosions to expect?
- queue length distr.?

15-826

Copyright: C. Faloutsos (2007)

7

---

---


---

---

---

---

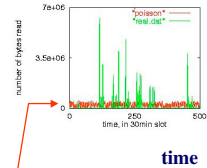
---



CMU SCS

### Problem #3: traffic

# bytes



time

Poisson  
indep.,  
ident. distr

15-826

Copyright: C. Faloutsos (2007)

8

---

---


---

---

---

---

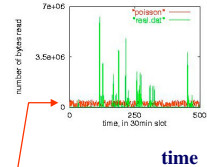
---



CMU SCS

### Problem #3: traffic

# bytes



time

~~Poisson~~  
indep.,  
ident. distr

15-826

Copyright: C. Faloutsos (2007)

9

---

---

---

---

---

---

---

CMU SCS

## Problem #3: traffic

# bytes

time

~~Poisson  
indep.  
ident. distr~~

Q: Then, how to generate such bursty traffic?

15-826 Copyright: C. Faloutsos (2007) 10

---

---

---

---

---

---

---

---

CMU SCS

## Common answer:

- Fractals / self-similarities / power laws
- Seminal works from Hilbert, Minkowski, Cantor, Mandelbrot, (Hausdorff, Lyapunov, Ken Wilson, ...)

15-826 Copyright: C. Faloutsos (2007) 11

---

---

---

---

---

---

---

---

CMU SCS

## Road map

- Motivation – 3 problems / case studies
- ➔ • Definition of fractals and power laws
- Solutions to posed problems
- More examples and tools
- Discussion - putting fractals to work!
- Conclusions – practitioner's guide
- Appendix: gory details - boxcounting plots

15-826 Copyright: C. Faloutsos (2007) 12

---

---

---

---

---

---

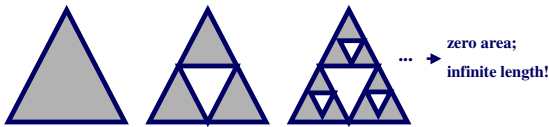
---

---

CMU SCS

## What is a fractal?

= self-similar point set, e.g., Sierpinski triangle:



... → zero area;  
infinite length!

15-826 Copyright: C. Faloutsos (2007) 13

---

---

---

---

---

---

---

---

CMU SCS

## Definitions (cont'd)

- Paradox: Infinite perimeter ; Zero area!
- 'dimensionality': between 1 and 2
- actually:  $\log(3)/\log(2) = 1.58\dots$

15-826 Copyright: C. Faloutsos (2007) 14

---

---

---

---

---

---

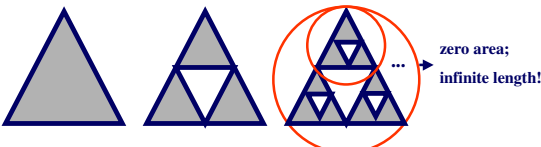
---

---

CMU SCS

## Dfn of fd:

ONLY for a perfectly self-similar point set:



... → zero area;  
infinite length!

$=\log(n)/\log(f) = \log(3)/\log(2) = 1.58$

15-826 Copyright: C. Faloutsos (2007) 15

---

---

---


---

---

---

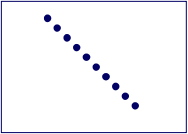
---

---

 CMU SCS

### Intrinsic (‘fractal’) dimension

- Q: fractal dimension of a line?
- A: 1 ( $= \log(2)/\log(2)!$ )



15-826      Copyright: C. Faloutsos (2007)      16

---

---


---

---

---

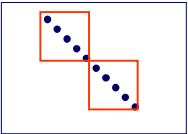
---

---

 CMU SCS

### Intrinsic (‘fractal’) dimension

- Q: fractal dimension of a line?
- A: 1 ( $= \log(2)/\log(2)!$ )



15-826      Copyright: C. Faloutsos (2007)      17

---

---


---

---

---

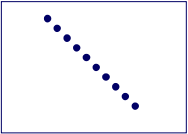
---

---

 CMU SCS

### Intrinsic (‘fractal’) dimension

- Q: dfn for a given set of points?



x	y
5	1
4	2
3	3
2	4

15-826      Copyright: C. Faloutsos (2007)      18

---

---


---

---

---

---

---

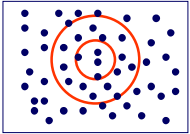
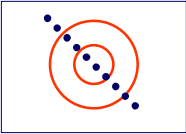


CMU SCS

### Intrinsic ('fractal') dimension

- Q: fractal dimension of a line?
- A:  $nn( \leq r ) \sim r^1$   
( 'power law':  $y=x^a$  )

- Q: fd of a plane?
- A:  $nn( \leq r ) \sim r^2$   
 $fd == \text{slope of } (\log(nn) \text{ vs } \log(r))$



15-826

Copyright: C. Faloutsos (2007)

19

---

---


---

---

---

---

---



CMU SCS

### Intrinsic ('fractal') dimension

- Algorithm, to estimate it?

Notice

- $avg\ nn(\leq r)$  is exactly  $\frac{tot\#pairs(\leq r)}{N}$

including 'mirror' pairs

15-826

Copyright: C. Faloutsos (2007)

20

---

---


---

---

---

---

---

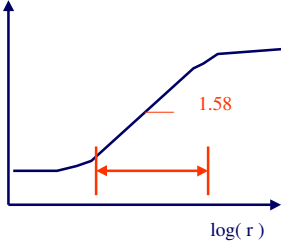



CMU SCS

### Sierpinsky triangle

== 'correlation integral'

$\log(\#pairs \text{ within } \leq r)$



15-826

Copyright: C. Faloutsos (2007)

21

---

---

---

---

---

---

---

7

CMU SCS

## Observations:

- Euclidean objects have **integer** fractal dimensions
  - point: 0
  - lines and smooth curves: 1
  - smooth surfaces: 2
- fractal dimension  $\rightarrow$  roughness of the periphery

15-826 Copyright: C. Faloutsos (2007) 22

---

---

---

---

---

---


---

---

CMU SCS

## Important properties

- fd = embedding dimension  $\rightarrow$  uniform pointset
- a point set may have several fd, depending on scale



15-826 Copyright: C. Faloutsos (2007) 23

---

---

---

---

---

---


---

---

CMU SCS

## Important properties

- fd = embedding dimension  $\rightarrow$  uniform pointset
- a point set may have several fd, depending on scale



2-d

15-826 Copyright: C. Faloutsos (2007) 24

---

---

---

---

---

---

---

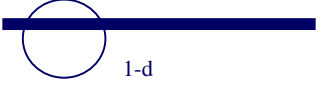
---



CMU SCS

## Important properties

- $fd$  = embedding dimension  $\rightarrow$  uniform pointset
- a point set may have several  $fd$ , depending on scale



15-826 Copyright: C. Faloutsos (2007) 25

---

---

---

---

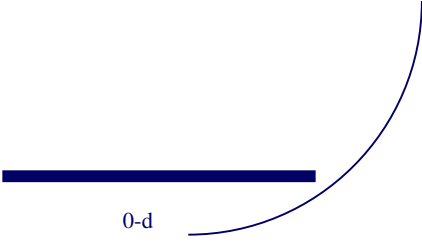
---

---

---

CMU SCS

## Important properties



15-826 Copyright: C. Faloutsos (2007) 26

---

---

---

---

---

---

---

CMU SCS

## Road map

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- ➔ • Solutions to posed problems
- More examples and tools
- Discussion - putting fractals to work!
- Conclusions – practitioner's guide
- Appendix: gory details - boxcounting plots

15-826 Copyright: C. Faloutsos (2007) 27

---

---

---

---

---

---

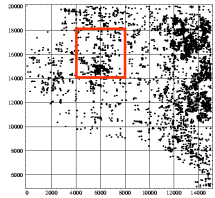
---

CMU SCS

## Problem #1: GIS points

Cross-roads of Montgomery county:

- any rules?



15-826 Copyright: C. Faloutsos (2007) 28

---

---

---

---

---

---

---

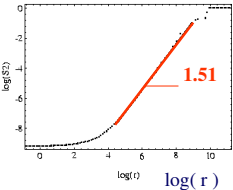
---

CMU SCS

## Solution #1

$\log(\text{\#pairs}(\text{within} \leq r))$

SLOPE = 1.51847



1.51

$\log(r)$

A: self-similarity ->

- $\Leftrightarrow$  fractals
- $\Leftrightarrow$  scale-free
- $\Leftrightarrow$  power-laws  
( $y=x^a$ ,  $F=C*r^{(-2)}$ )
- avg#neighbors( $\leq r$ )  
 $= r^D$

15-826 Copyright: C. Faloutsos (2007) 29

---

---

---

---

---

---

---

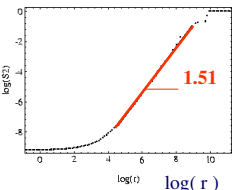
---

CMU SCS

## Solution #1

$\log(\text{\#pairs}(\text{within} \leq r))$

SLOPE = 1.51847



1.51

$\log(r)$

A: self-similarity

- avg#neighbors( $\leq r$ )  
 $\sim r^{(1.51)}$

15-826 Copyright: C. Faloutsos (2007) 30

---

---

---


---

---

---

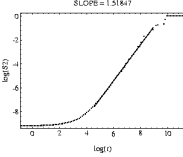
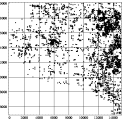
---

---

CMU SCS

### Examples:MG county

- Montgomery County of MD (road end-points)



15-826

Copyright: C. Faloutsos (2007)

31

---

---


---

---

---

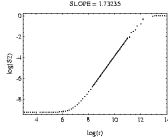
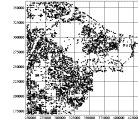
---

---

CMU SCS

### Examples:LB county

- Long Beach county of CA (road end-points)



15-826

Copyright: C. Faloutsos (2007)

32

---

---


---

---

---

---

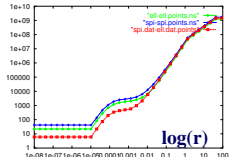
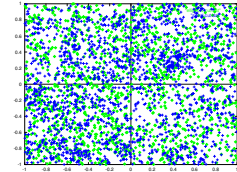
---

CMU SCS

### Solution#2: spatial d.m.

Galaxies ( ‘BOPS’ plot - [sigmod2000])

**log(#pairs)**



15-826

Copyright: C. Faloutsos (2007)

33

---

---

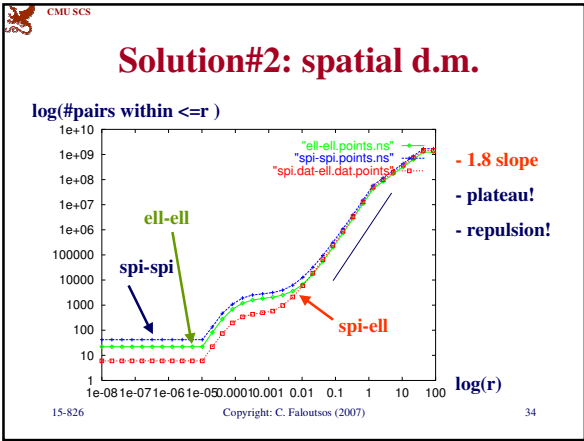
---

---

---

---

---



---

---

---

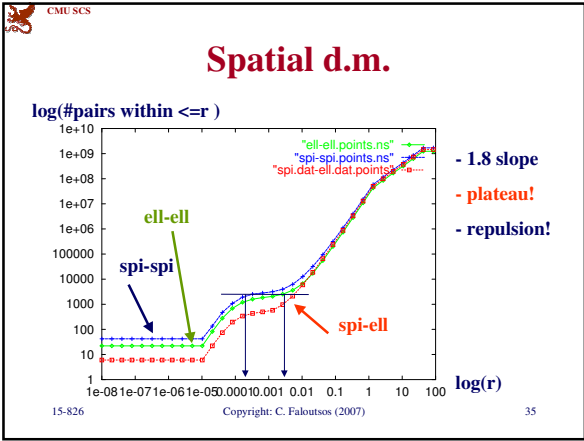
---

---

---

---

---



---

---

---

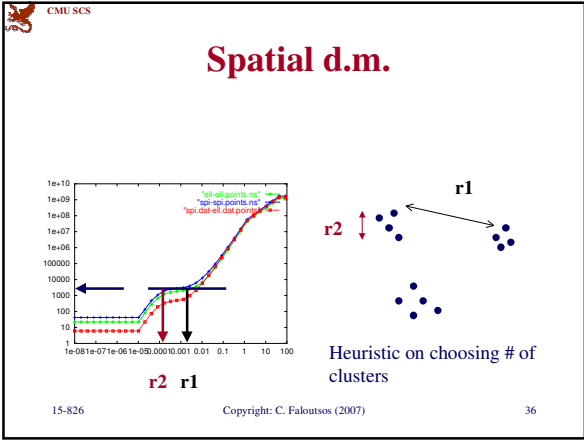
---

---

---

---

---



---

---

---

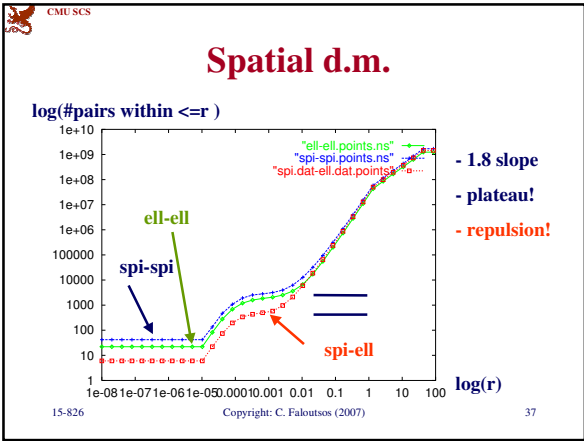
---

---

---

---

---



---

---

---

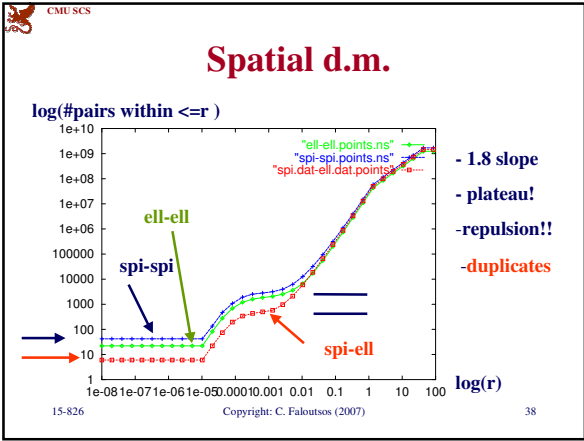
---

---

---

---

---



---

---

---

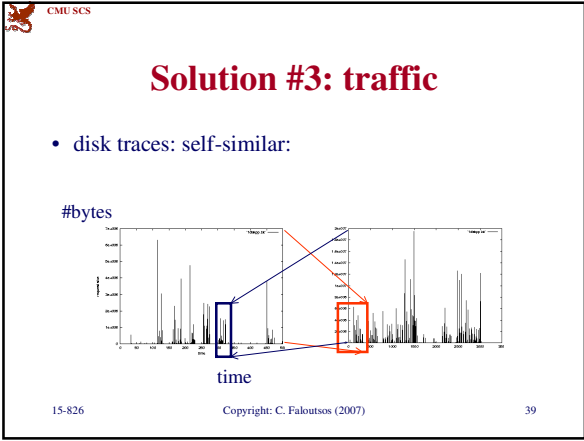
---

---

---

---

---



---

---

---


---

---

---

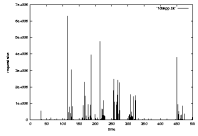
---

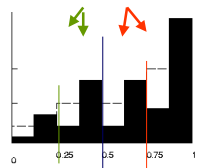
---

CMU SCS

### Solution #3: traffic

- disk traces (80-20 'law' = 'multifractal')

#bytes  
  
time

  
20% 80%

15-826Copyright: C. Faloutsos (2007)40

---

---

---


---

---

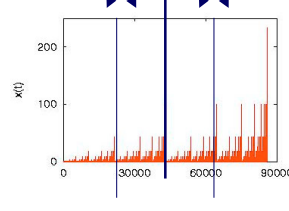
---

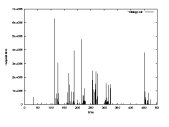
---

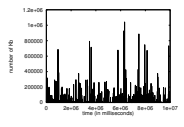
---

CMU SCS

### 80-20 / multifractals

20 80  
  
x(t)

  
number of bytes

  
time (in milliseconds)

15-826Copyright: C. Faloutsos (2007)41

---

---

---


---

---

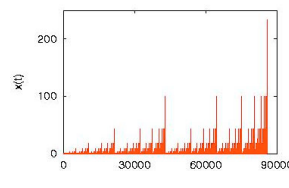
---

---

---

CMU SCS

### 80-20 / multifractals

20 80  
  
x(t)

- $p : (1-p)$  in general
- yes, there are dependencies

15-826Copyright: C. Faloutsos (2007)42

---

---

---

---

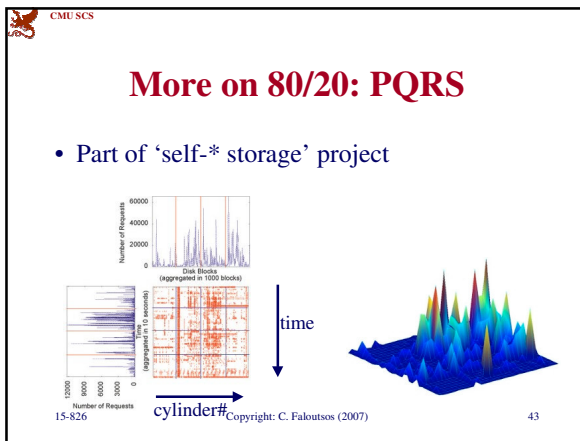
---

---

---

---

14




---

---

---

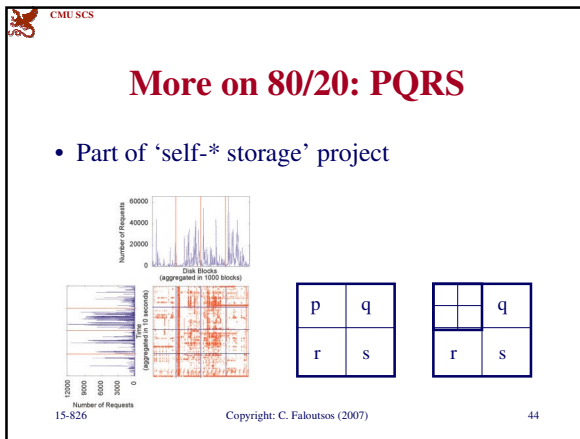
---

---

---

---

---




---

---

---

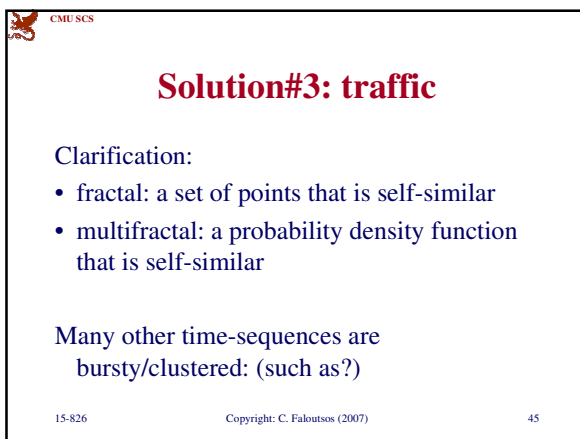
---

---

---

---

---




---

---

---


---

---

---

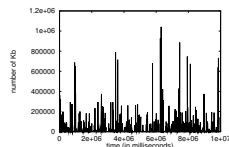
---

---

CMU SCS

### Example:

- network traffic



<http://repository.cs.vt.edu/lbl-conn-7.tar.Z>

15-826Copyright: C. Faloutsos (2007)46

---

---

---


---

---

---

---

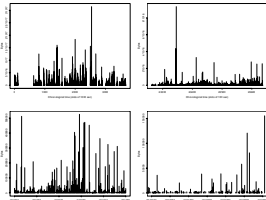
---

CMU SCS

### Web traffic

- [Crovella Bestavros, SIGMETRICS'96]

1000 sec; 100sec  
10sec; 1sec



15-826Copyright: C. Faloutsos (2007)47

---

---

---


---

---

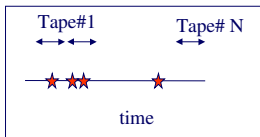
---

---

---

CMU SCS

### Tape accesses



# tapes needed, to retrieve  $n$  records?  
(# days down, due to failures / hurricanes / communication noise...)

15-826Copyright: C. Faloutsos (2007)48

---

---

---

---

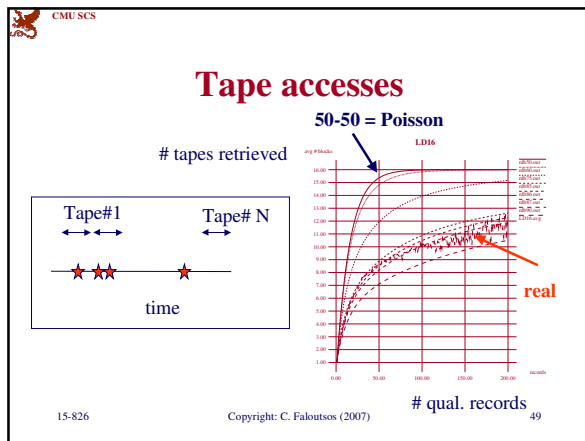
---

---

---

---






---

---

---

---

---

---

---

---

CMU SCS

## Road map

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems
- ➔ • More **tools** and examples
- Discussion - putting fractals to work!
- Conclusions – practitioner's guide
- Appendix: gory details - boxcounting plots

15-826 Copyright: C. Faloutsos (2007) 50

---

---

---

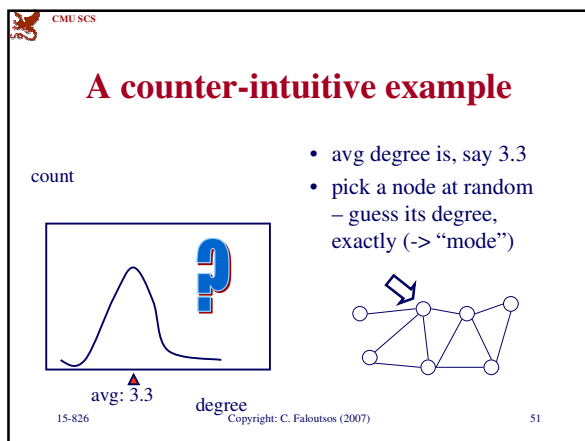
---

---

---

---

---




---

---

---

---

---

---

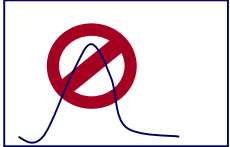
---

---

CMU SCS

## A counter-intuitive example

count



avg: 3.3 degree

15-826 Copyright: C. Faloutsos (2007) 52

- avg degree is, say 3.3
- pick a node at random
  - guess its degree, exactly (-> “mode”)
- A: 1!!

---

---

---

---

---

---

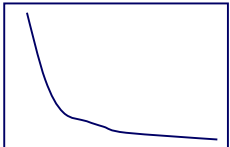
---

---

CMU SCS

## A counter-intuitive example

count



avg: 3.3 degree

15-826 Copyright: C. Faloutsos (2007) 53

- avg degree is, say 3.3
- pick a node at random
  - what is the degree you expect it to have?
- A: 1!!
- A': very skewed distr.
- Corollary: **the mean is meaningless!**
- (and std -> infinity (!))

---

---

---

---

---

---

---

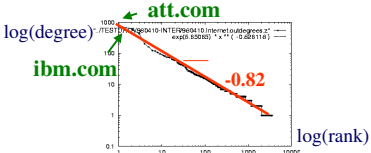
---

CMU SCS

## Rank exponent $R$

- Power law in the degree distribution [SIGCOMM99]

internet domains



log(rank)

log(degree)

att.com

ibm.com

-0.82

15-826 Copyright: C. Faloutsos (2007) 54

---

---

---

---

---

---

---

---



CMU SCS

## More tools

- Zipf’s law
- Korcak’s law / “fat fractals”

15-826

Copyright: C. Faloutsos (2007)

55

---

---


---

---

---

---

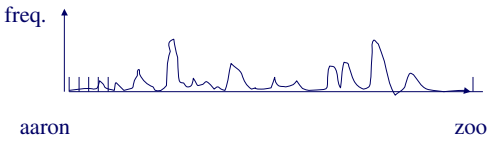
---



CMU SCS

## A famous power law: Zipf’s law

- Q: vocabulary word frequency in a document - any pattern?



15-826

Copyright: C. Faloutsos (2007)

56

---

---


---

---

---

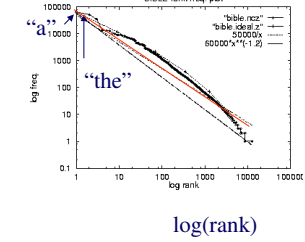
---

---



CMU SCS

## A famous power law: Zipf’s law



- Bible - rank vs frequency (log-log)

15-826

Copyright: C. Faloutsos (2007)

57

---

---


---

---

---

---

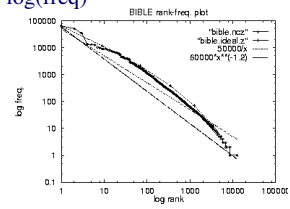
---



CMU SCS

### A famous power law: Zipf's law

log(freq)



log(rank)

- Bible - rank vs frequency (log-log)
- similarly, in **many other** languages; for customers and sales volume; city populations etc etc

15-826

Copyright: C. Faloutsos (2007)

58

---

---

---


---

---

---

---

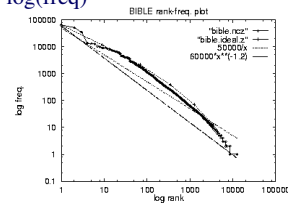
---



CMU SCS

### A famous power law: Zipf's law

log(freq)



log(rank)

- Zipf distr:  
 $\text{freq} = 1 / \text{rank}$
- generalized Zipf:  
 $\text{freq} = 1 / (\text{rank})^a$

15-826

Copyright: C. Faloutsos (2007)

59

---

---

---


---

---

---

---

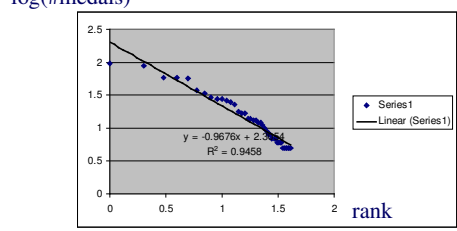
---



CMU SCS

### Olympic medals (Sidney):

log(#medals)



rank

15-826

Copyright: C. Faloutsos (2007)

60

---

---

---

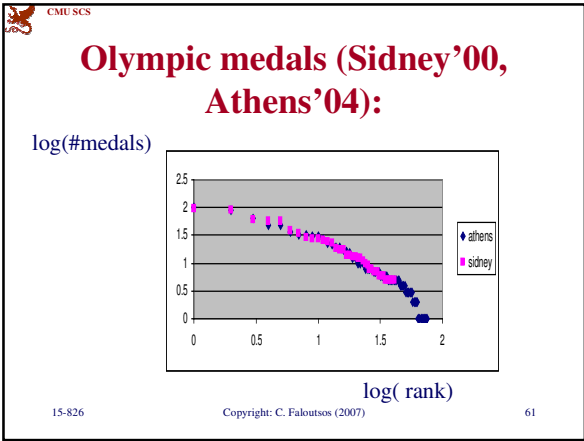
---

---

---

---

---



---

---

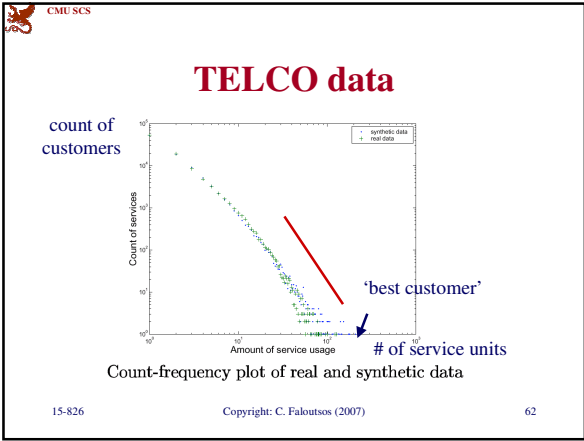
---

---

---

---

---



---

---

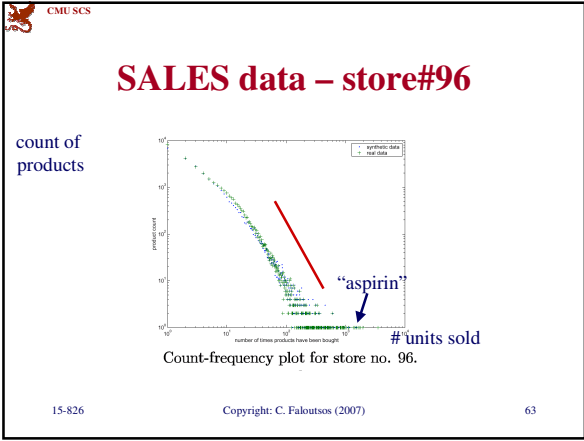
---

---

---

---

---



---

---


---

---

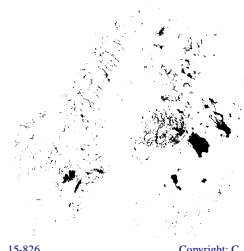
---

---

---

CMU SCS

# More power laws: areas – Korcak’s law



Scandinavian lakes  
Any pattern?

15-826

Copyright: C. Faloutsos (2007)

64

---

---


---

---


---

---

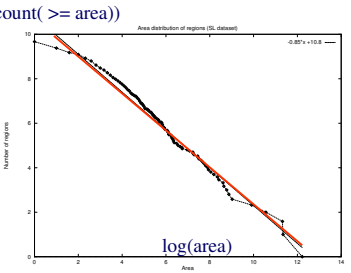
---

CMU SCS

# More power laws: areas – Korcak’s law



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



log(count(  $\geq$  area))

Area distribution of regions (SL dataset)

$-0.987 \pm 0.018$

log(area)

15-826

Copyright: C. Faloutsos (2007)

65

---

---


---

---


---

---

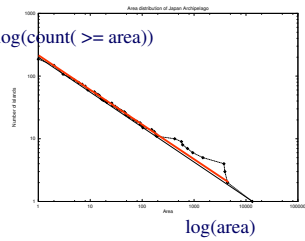
---

CMU SCS

# More power laws: Korcak



Japan islands;  
area vs cumulative  
count (log-log axes)



log(count(  $\geq$  area))

Area distribution of Japan Archipelago

log(area)

15-826

Copyright: C. Faloutsos (2007)

66

---

---


---

---


---

---

---

CMU SCS

### (Korcak’s law: Aegean islands)



15-826

Copyright: C. Faloutsos (2007)

67

---

---


---

---


---

---

---

CMU SCS

### Korcak’s law & “fat fractals”



How to generate such regions?

15-826

Copyright: C. Faloutsos (2007)

68

---

---


---

---

---

---

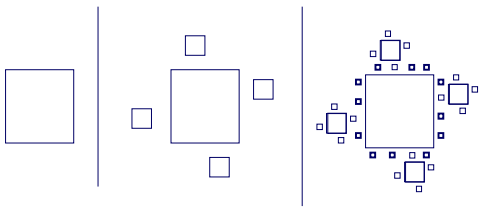
---

CMU SCS

### Korcak’s law & “fat fractals”

Q: How to generate such regions?

A: recursively, from a single region



15-826

Copyright: C. Faloutsos (2007)

69

---

---

---

---

---

---

---

CMU SCS

### so far we've seen:

- concepts:
  - fractals, multifractals and fat fractals
- tools:
  - correlation integral (= pair-count plot)
  - rank/frequency plot (Zipf's law)
  - CCDF (Korcak's law)

15-826 Copyright: C. Faloutsos (2007) 70

---

---

---

---

---

---

---

---

CMU SCS

### so far we've seen:

- concepts:
  - fractals, multifractals and fat fractals
- tools:
  - correlation integral (= pair-count plot)
  - rank/frequency plot (Zipf's law)
  - CCDF (Korcak's law)

same info

15-826 Copyright: C. Faloutsos (2007) 71

---

---

---

---

---

---

---

---

CMU SCS

### Road map

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems
- ➔ • More tools and **examples**
- Discussion - putting fractals to work!
- Conclusions – practitioner's guide
- Appendix: gory details - boxcounting plots

15-826 Copyright: C. Faloutsos (2007) 72

---

---

---

---

---

---

---


---



CMU SCS

## Other applications: Internet

- How does the internet look like?



15-826 Copyright: C. Faloutsos (2007) 73

---

---

---

---

---


---

---

CMU SCS

## Other applications: Internet

- How does the internet look like?
- Internet routers: how many neighbors within  $h$  hops?



15-826 Copyright: C. Faloutsos (2007) 74

---

---

---

---

---

---

---

CMU SCS

## (reminder: our tool-box:)

- concepts:
  - fractals, multifractals and fat fractals
- tools:
  - correlation integral (= pair-count plot)
  - rank/frequency plot (Zipf's law)
  - CCDF (Korczak's law)

15-826 Copyright: C. Faloutsos (2007) 75

---

---

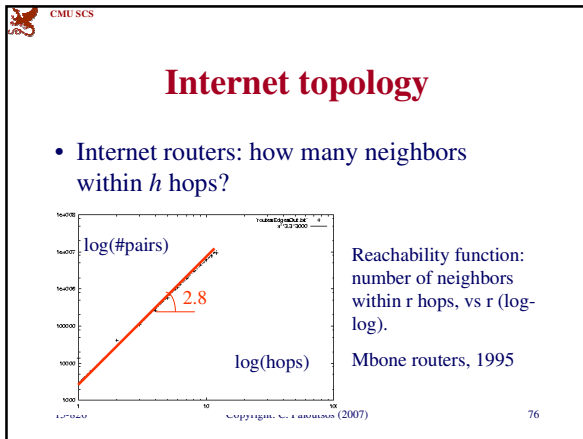
---

---

---

---

---




---

---

---

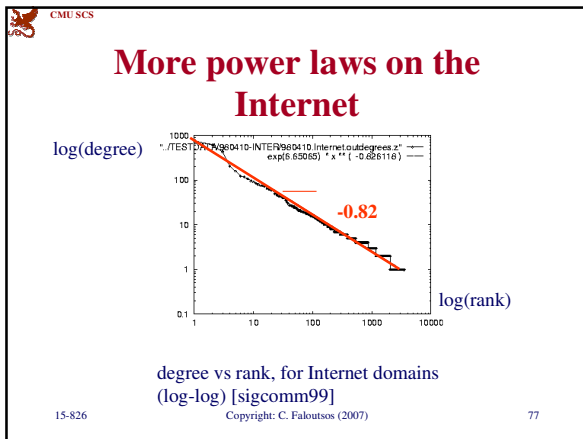
---

---

---

---

---




---

---

---

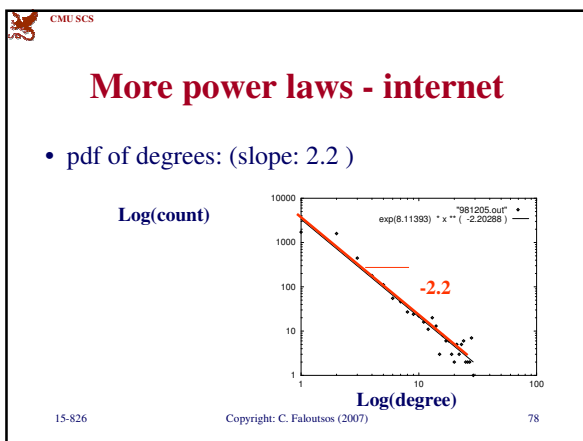
---

---

---

---

---




---

---

---


---

---

---

---

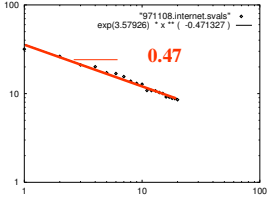
---



CMU SCS

## Even more power laws on the Internet

log( i-th eigenvalue)



0.47

log(i)

Scree plot for Internet domains (log-log) [sigcomm99]

Copyright: C. Faloutsos (2007)

15-826

79

---

---


---

---

---

---

---



CMU SCS

## Fractals & power laws:

appear in numerous settings:

- **medical**
- geographical / geological
- social
- computer-system related

15-826

Copyright: C. Faloutsos (2007)

80

---

---


---

---

---

---

---

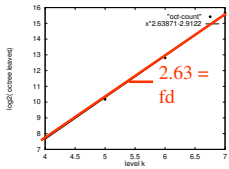


CMU SCS

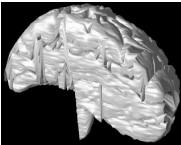
## More apps: Brain scans

- Oct-trees; brain-scans

Log(#octants)



2.63 = fd



octree levels

15-826

Copyright: C. Faloutsos (2007)

81

---

---


---

---

---

---

---



CMU SCS

## More apps: Medical images

[Burdett et al, SPIE '93]:

- benign tumors:  $fd \sim 2.37$
- malignant:  $fd \sim 2.56$

15-826 Copyright: C. Faloutsos (2007) 82

---

---


---

---

---

---


---



CMU SCS

## More fractals:

- cardiovascular system: 3 (!)
- lungs: 2.9



15-826 Copyright: C. Faloutsos (2007) 83

---

---


---

---

---

---

---



CMU SCS

## Fractals & power laws:

appear in numerous settings:

- medical
- **geographical / geological**
- social
- computer-system related

15-826 Copyright: C. Faloutsos (2007) 84

---

---


---

---

---

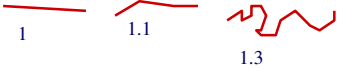
---

---

CMU SCS

### More fractals:

- Coastlines: 1.2-1.58



15-826

Copyright: C. Faloutsos (2007)

85

---

---


---


---

---

---

---

CMU SCS



15-826

Copyright: C. Faloutsos (2007)

86

---

---


---

---

---

---

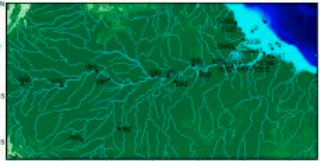
---

CMU SCS

### More fractals:

- the fractal dimension for the Amazon river is 1.85 (Nile: 1.4)

[[ems.gphys.unc.edu/nonlinear/fractals/examples.html](http://ems.gphys.unc.edu/nonlinear/fractals/examples.html)]



15-826

Copyright: C. Faloutsos (2007)

87

---

---

---

---

---

---

---

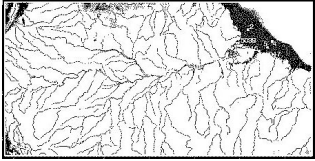
29

CMU SCS

## More fractals:

- the fractal dimension for the Amazon river is 1.85 (Nile: 1.4)

[ems.gphys.unc.edu/nonlinear/fractals/examples.html]



15-826 Copyright: C. Faloutsos (2007) 88

---

---

---

---

---

---

---

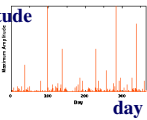
---

CMU SCS

## More power laws

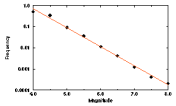
- Energy of earthquakes (Gutenberg-Richter law) [simscience.org]

amplitude



day

log(freq)



magnitude

15-826 Copyright: C. Faloutsos (2007) 89

---

---

---

---

---

---

---

---

CMU SCS

## Fractals & power laws:

appear in numerous settings:

- medical
- geographical / geological
- social
- computer-system related

15-826 Copyright: C. Faloutsos (2007) 90

---

---

---


---

---

---

---

---




CMU SCS


More fractals:

stock prices (LYCOS) - random walks: 1.5

1 year



2 years



15-826

Copyright: C. Faloutsos (2007)

91

---

---


---

---

---

---

---



CMU SCS

Even more power laws:

• Income distribution (Pareto’s law)

• size of firms

• publication counts (Lotka’s law)

15-826

Copyright: C. Faloutsos (2007)

92

---

---


---

---

---

---

---

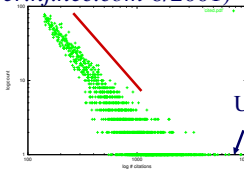


CMU SCS

Even more power laws:

library science (Lotka’s law of publication count); and citation counts:  
([citeseer.nj.nec.com](http://citeseer.nj.nec.com) 6/2001)

log(count)



Ullman

log(#citations)

15-826

Copyright: C. Faloutsos (2007)

93

---

---

---

---

---

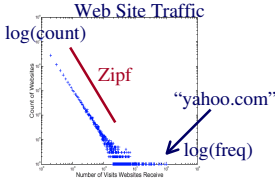
---

---

CMU SCS

## Even more power laws:

- web hit counts [w/ A. Montgomery]



15-826 Copyright: C. Faloutsos (2007) 94

---

---

---

---

---

---

---

---

CMU SCS

## Fractals & power laws:

appear in numerous settings:

- medical
- geographical / geological
- social
- computer-system related**

15-826 Copyright: C. Faloutsos (2007) 95

---

---

---

---

---

---

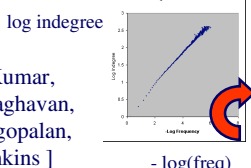
---

---

CMU SCS

## Power laws, cont'd

- In- and out-degree distribution of web sites [Barabasi], [IBM-CLEVER]



from [Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins]

15-826 Copyright: C. Faloutsos (2007) 96

---

---

---

---

---

---

---

---



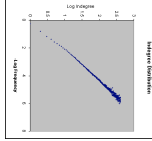
CMU SCS

## Power laws, cont'd

- In- and out-degree distribution of web sites [Barabasi], [IBM-CLEVER]

log(freq)

from [Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins]



log indegree

15-826 Copyright: C. Faloutsos (2007) 97

---

---

---

---

---

---

---

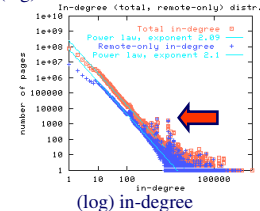
---

CMU SCS

## "Foiled by power law"

- [Broder+, WWW'00]

(log) count



In-degree (total, remote-only) distr.

Power law, exponent 2.09

Power law, exponent 2.1

Power law, exponent 2.1

“The anomalous bump at 120 on the x-axis is due a large clique formed by a single spammer”

15-826 Copyright: C. Faloutsos (2007) 98

---

---

---

---

---

---

---

---

CMU SCS

## Power laws, cont'd

- In- and out-degree distribution of web sites [Barabasi], [IBM-CLEVER]
- length of file transfers [Crovella+Bestavros '96]
- duration of UNIX jobs [Harchol-Balter]

15-826 Copyright: C. Faloutsos (2007) 99

---

---

---


---

---

---

---

---



CMU SCS

## Even more power laws:

- Distribution of UNIX file sizes
- web hit counts [Huberman]

15-826 Copyright: C. Faloutsos (2007) 100

---

---


---

---

---

---

---



CMU SCS

## Road map

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems
- More examples and tools
- ➔ • Discussion - putting fractals to work!
- Conclusions – practitioner's guide
- Appendix: gory details - boxcounting plots

15-826 Copyright: C. Faloutsos (2007) 101

---

---


---

---

---

---

---



CMU SCS

## What else can they solve?

- ✓ • separability [KDD'02]
  - forecasting [CIKM'02]
  - dimensionality reduction [SBBD'00]
  - non-linear axis scaling [KDD'02]
- ✓ • disk trace modeling [PEVA'02]
  - selectivity of spatial/multimedia queries [PODS'94, VLDB'95, ICDE'00]
  - ...

15-826 Copyright: C. Faloutsos (2007) 102

---

---


---

---

---

---

---

CMU SCS

### Settings for fractals:

Points; areas (-> fat fractals), eg:

15-826

Copyright: C. Faloutsos (2007)

103

---

---


---

---

---

---

---

CMU SCS

### Settings for fractals:

Points; areas, eg:

- cities/stores/hospitals, over earth’s surface
- time-stamps of events (customer arrivals, packet losses, criminal actions) over time
- regions (sales areas, islands, patches of habitats) over space

15-826

Copyright: C. Faloutsos (2007)

104

---

---


---

---

---

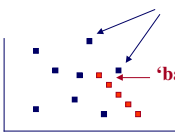
---

---

CMU SCS

### Settings for fractals:

- customer feature vectors (age, income, frequency of visits, amount of sales per visit)



‘good’ customers

‘bad’ customers

15-826

Copyright: C. Faloutsos (2007)

105

---

---

---

---

---

---

---



## Some uses of fractals:

- Detect non-existence of rules (if points are uniform)
- Detect non-homogeneous regions (eg., legal login time-stamps may have different fd than intruders')
- Estimate number of neighbors / customers / competitors within a radius

15-826

Copyright: C. Faloutsos (2007)

106

---

---

---

---

---

---

---



## Multi-Fractals

Setting: points or objects, w/ some value, eg:

- cities w/ populations
- positions on earth and amount of gold/water/oil underneath
- product ids and sales per product
- people and their salaries
- months and count of accidents

15-826

Copyright: C. Faloutsos (2007)

107

---

---

---

---

---

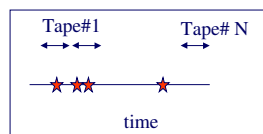
---

---



## Use of multifractals:

- Estimate tape/disk accesses
  - *how many of the 100 tapes contain my 50 phonecall records?*
  - *how many days without an accident?*



15-826

Copyright: C. Faloutsos (2007)

108

---

---


---

---

---

---

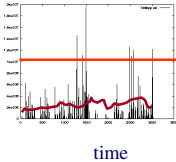
---

CMU SCS

### Use of multifractals

- how often do we exceed the threshold?

#bytes



time

15-826 Copyright: C. Faloutsos (2007) 109

---

---

---


---

---

---

---

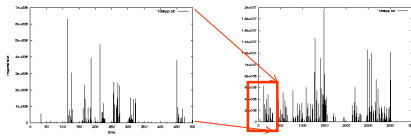
---

CMU SCS

### Use of multifractals cont'd

- Extrapolations for/from samples

#bytes



time

15-826 Copyright: C. Faloutsos (2007) 110

---

---

---


---

---

---

---

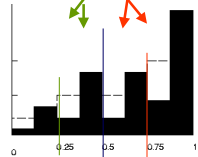
---

CMU SCS

### Use of multifractals cont'd

- How many distinct products account for 90% of the sales?

20% ↗ 80%



0 0.25 0.5 0.75 1

15-826 Copyright: C. Faloutsos (2007) 111

---

---

---

---

---

---

---

---

CMU SCS

## Road map

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems
- More examples and tools
- Discussion - putting fractals to work!
- ➔ • Conclusions – practitioner's guide
- Appendix: gory details - boxcounting plots

15-826 Copyright: C. Faloutsos (2007) 112

---

---

---

---

---

---

---

---

CMU SCS

## Conclusions

- Real data often **disobey** textbook assumptions (Gaussian, Poisson, uniformity, independence)

15-826 Copyright: C. Faloutsos (2007) 113

---

---

---

---

---

---

---

---

CMU SCS

## Conclusions - cont'd

Self-similarity & power laws: appear in **many** cases

Bad news:

lead to skewed distributions  
(no Gaussian, Poisson,  
uniformity, independence,  
mean, variance)

Good news:

- 'correlation integral' for separability
- rank/frequency plots
- 80-20 (multifractals)
- (Hurst exponent, strange attractors, renormalization theory, ++)

15-826 Copyright: C. Faloutsos (2007) 114

---

---

---

---

---

---

---

---

CMU SCS

## Conclusions

- **tool#1: (for points) ‘correlation integral’:**  
(#pairs within  $\leq r$ ) vs (distance  $r$ )
- **tool#2: (for categorical values) rank-frequency plot (a’la Zipf)**
- **tool#3: (for numerical values) CCDF:**  
Complementary cumulative distr. function  
(#of elements with value  $\geq a$ )

15-826 Copyright: C. Faloutsos (2007) 115

---

---

---

---

---

---

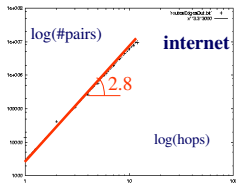
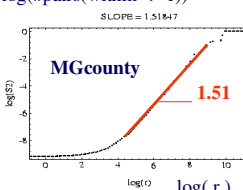
---

---

CMU SCS

## Practitioner’s guide:

- **tool#1: #pairs vs distance, for a set of objects,**  
with a distance function (slope = intrinsic dimensionality)

15-826 Copyright: C. Faloutsos (2007) 116

---

---

---

---

---

---

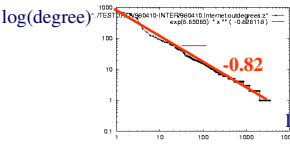
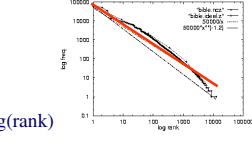
---

---

CMU SCS

## Practitioner’s guide:

- **tool#2: rank-frequency plot (for categorical attributes)**

15-826 Copyright: C. Faloutsos (2007) 117

---

---

---

---

---

---

---

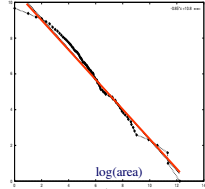
---

CMU SCS

## Practitioner's guide:

- **tool#3: CCDF**, for (skewed) **numerical attributes**, eg. areas of islands/lakes, UNIX jobs...)

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



scandinavian lakes

15-826 Copyright: C. Faloutsos (2007) 118

---

---

---

---

---

---

---

---

CMU SCS

## Resources:

- Software for fractal dimension
  - <http://www.cs.cmu.edu/~christos>
  - [christos@cs.cmu.edu](mailto:christos@cs.cmu.edu)

15-826 Copyright: C. Faloutsos (2007) 119

---

---

---

---

---

---

---

---

CMU SCS

## Books

- Strongly recommended intro book:
  - Manfred Schroeder *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise* W.H. Freeman and Company, 1991
- Classic book on fractals:
  - B. Mandelbrot *Fractal Geometry of Nature*, W.H. Freeman, 1977

15-826 Copyright: C. Faloutsos (2007) 120

---

---

---

---


---

---

---

---





CMU SCS

References

- [vldb95] Alberto Belussi and Christos Faloutsos, *Estimating the Selectivity of Spatial Queries Using the 'Correlation' Fractal Dimension* Proc. of VLDB, p. 299-310, 1995
- [Broder+'00] Andrei Broder, Ravi Kumar , Farzin Maghoul1, Prabhakar Raghavan , Sridhar Rajagopalan , Raymie Stata, Andrew Tomkins , Janet Wiener, *Graph structure in the web* , WWW'00
- M. Crovella and A. Bestavros, *Self similarity in World wide web traffic: Evidence and possible causes* , SIGMETRICS '96.

15-826

Copyright: C. Faloutsos (2007)

121

---

---


---

---

---

---

---



CMU SCS

References

- [ieeeTN94] W. E. Leland, M.S. Taqqu, W. Willinger, D.V. Wilson, *On the Self-Similar Nature of Ethernet Traffic*, IEEE Transactions on Networking, 2, 1, pp 1-15, Feb. 1994.
- [pods94] Christos Faloutsos and Ibrahim Kamel, *Beyond Uniformity and Independence: Analysis of R-trees Using the Concept of Fractal Dimension*, PODS, Minneapolis, MN, May 24-26, 1994, pp. 4-13

15-826

Copyright: C. Faloutsos (2007)

122

---

---


---

---

---

---

---



CMU SCS

References

- [vldb96] Christos Faloutsos, Yossi Matias and Avi Silberschatz, *Modeling Skewed Distributions Using Multifractals and the '80-20 Law'* Conf. on Very Large Data Bases (VLDB), Bombay, India, Sept. 1996.

15-826

Copyright: C. Faloutsos (2007)

123

---

---


---

---

---

---

---



CMU SCS

## References

- [vldb96] Christos Faloutsos and Volker Gaede *Analysis of the Z-Ordering Method Using the Hausdorff Fractal Dimension* VLD, Bombay, India, Sept. 1996
- [sigcomm99] Michalis Faloutsos, Petros Faloutsos and Christos Faloutsos, *What does the Internet look like? Empirical Laws of the Internet Topology*, SIGCOMM 1999

15-826 Copyright: C. Faloutsos (2007) 124

---

---


---

---

---

---

---



CMU SCS

## References

- [icde99] Guido Proietti and Christos Faloutsos, *I/O complexity for range queries on region data stored using an R-tree* International Conference on Data Engineering (ICDE), Sydney, Australia, March 23-26, 1999
- [sigmod2000] Christos Faloutsos, Bernhard Seeger, Agma J. M. Traina and Caetano Traina Jr., *Spatial Join Selectivity Using Power Laws*, SIGMOD 2000

15-826 Copyright: C. Faloutsos (2007) 125

---

---


---

---

---

---

---



CMU SCS

## Appendix - Gory details

- Bad news: There are more than one fractal dimensions
  - Minkowski fd; Hausdorff fd; Correlation fd; Information fd
- Great news:
  - they can all be computed fast!
  - they usually have nearby values

15-826 Copyright: C. Faloutsos (2007) 126

---

---


---

---

---

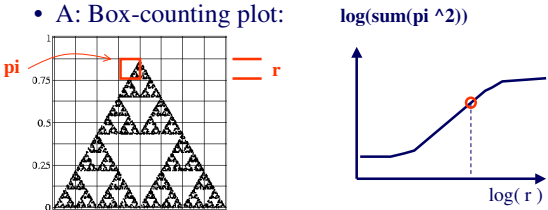
---

---

CMU SCS

### Fast estimation of $fd(s)$ :

- How, for the (correlation) fractal dimension?
- A: Box-counting plot:



15-826 Copyright: C. Faloutsos (2007) 127

---

---


---

---

---

---

---

CMU SCS

### Definitions

- $pi$  : the percentage (or count) of points in the  $i$ -th cell
- $r$ : the side of the grid

15-826 Copyright: C. Faloutsos (2007) 128

---

---


---

---

---

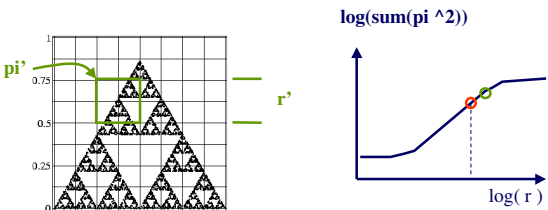
---

---

CMU SCS

### Fast estimation of $fd(s)$ :

- compute  $\text{sum}(pi^2)$  for another grid side,  $r'$



15-826 Copyright: C. Faloutsos (2007) 129

---

---

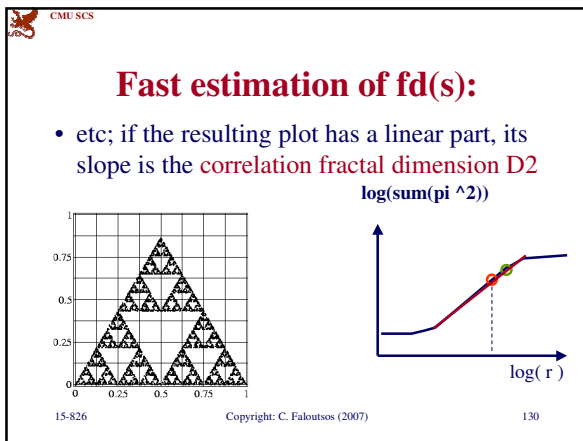
---

---

---

---

---




---

---

---

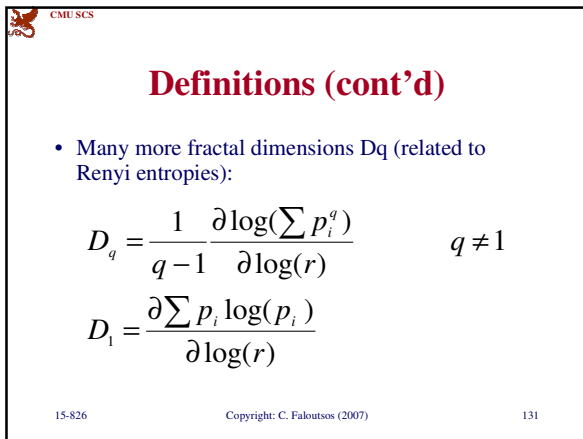
---

---

---

---

---




---

---

---

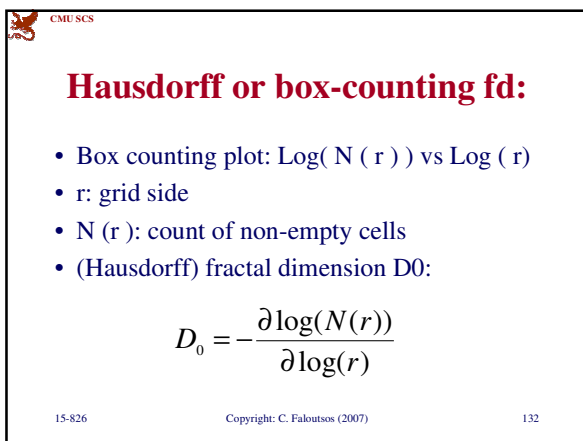
---

---

---

---

---




---

---

---

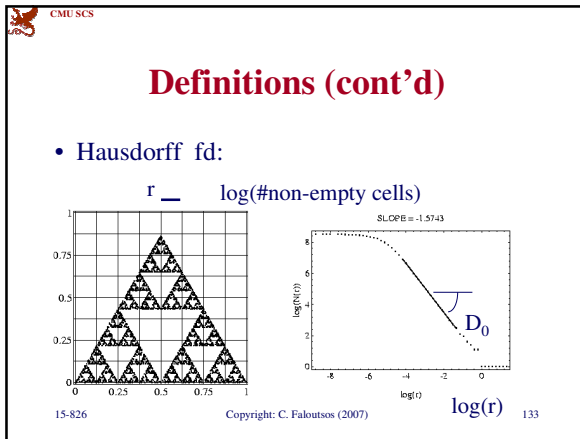
---

---

---

---

---




---

---

---

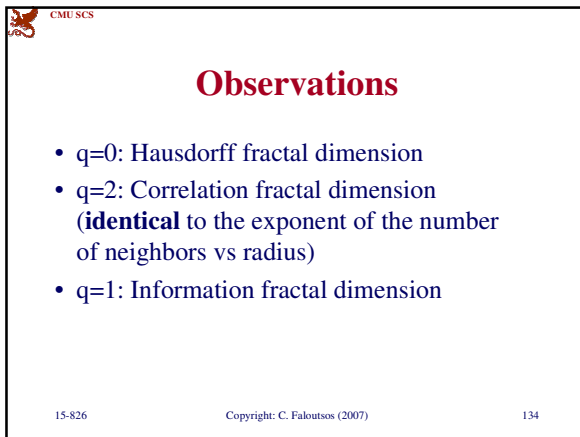
---

---

---

---

---




---

---

---

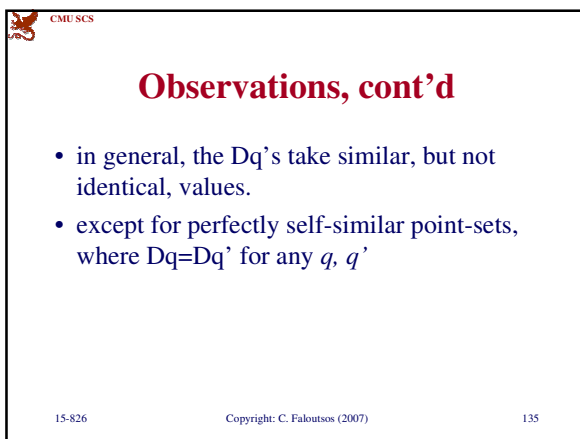
---

---

---

---

---




---

---

---


---

---

---

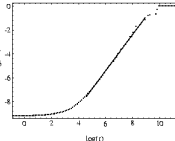
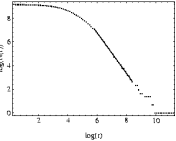
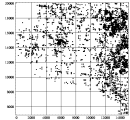
---

---

CMU SCS

### Examples:MG county

- Montgomery County of MD (road end-points)



15-826

Copyright: C. Faloutsos (2007)

136

---

---

---


---

---

---

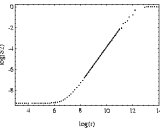
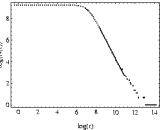
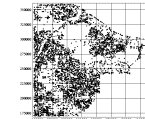
---

---

CMU SCS

### Examples:LB county

- Long Beach county of CA (road end-points)



15-826

Copyright: C. Faloutsos (2007)

137

---

---

---


---

---

---

---

---

CMU SCS

### Conclusions

- many fractal dimensions, with nearby values
- can be computed quickly ( $O(N)$  or  $O(N \log(N))$ )
- (code: on the web)

15-826

Copyright: C. Faloutsos (2007)

138

---

---

---

---

---

---

---

---