

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 (2006) 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 (2006) 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 (2006) 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 (2006) 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 (2006) 6

**CMU SCS**

### Problem #3: traffic

# bytes

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

15-826 Copyright: C. Faloutsos (2006) 7

**CMU SCS**

### Problem #3: traffic

# bytes

Poisson  
indep.,  
ident. distr

15-826 Copyright: C. Faloutsos (2006) 8

**CMU SCS**

### Problem #3: traffic

# bytes

Poisson  
indep.,  
ident. distr

15-826 Copyright: C. Faloutsos (2006) 9

**CMU SCS**

### Problem #3: traffic

# bytes

Poisson  
indep.,  
ident. distr

Q: Then, how to generate such bursty traffic?

15-826 Copyright: C. Faloutsos (2006) 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 (2006) 11

**CMU SCS**

### Road map

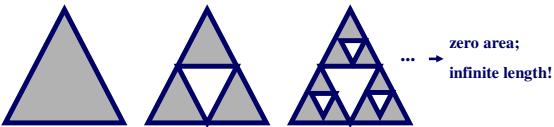
- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems
- More examples and tools
- Discussion - putting fractals to work!
- Conclusions – practitioner's guide
- Appendix: gory details - boxcounting plots

15-826 Copyright: C. Faloutsos (2006) 12

 CMU SCS

## What is a fractal?

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



... → zero area;  
infinite length!

15-826      Copyright: C. Faloutsos (2006)      13

 CMU SCS

## Definitions (cont'd)

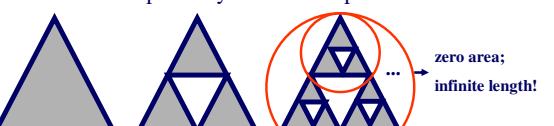
- Paradox: Infinite perimeter ; Zero area!
- ‘dimensionality’: between 1 and 2
- actually:  $\text{Log}(3)/\text{Log}(2) = 1.58\dots$

15-826      Copyright: C. Faloutsos (2006)      14

 CMU SCS

## Dfn of fd:

ONLY for a perfectly self-similar point set:



... → zero area;  
infinite length!

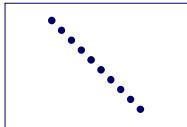
$= \log(n)/\log(f) = \log(3)/\log(2) = 1.58$

15-826      Copyright: C. Faloutsos (2006)      15

 CMU SCS

## Intrinsic ('fractal') dimension

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

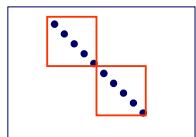


15-826      Copyright: C. Faloutsos (2006)      16

 CMU SCS

## Intrinsic ('fractal') dimension

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

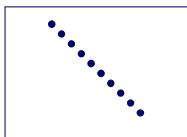


15-826      Copyright: C. Faloutsos (2006)      17

 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 (2006)      18

**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 == slope of  $(\log(nn) \text{ vs } \log(r))$

15-826 Copyright: C. Faloutsos (2006) 19

**Intrinsic ('fractal') dimension**

- Algorithm, to estimate it?
- Notice
- $\text{avg } nn(\leq r)$  is exactly  $\text{tot\#pairs}(\leq r) / N$

including 'mirror' pairs

15-826 Copyright: C. Faloutsos (2006) 20

**Sierpinsky triangle**

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

$\text{== 'correlation integral'}$

15-826 Copyright: C. Faloutsos (2006) 21

**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 (2006) 22

**Important properties**

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

15-826 Copyright: C. Faloutsos (2006) 23

**Important properties**

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

15-826 Copyright: C. Faloutsos (2006) 24

**CMU SCS**

## Important properties

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



15-826 Copyright: C. Faloutsos (2006) 25

**CMU SCS**

## Important properties



15-826 Copyright: C. Faloutsos (2006) 26

**CMU SCS**

## Road map

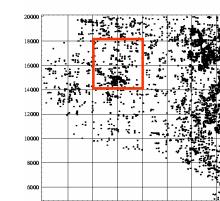
- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems 
- More examples and tools
- Discussion - putting fractals to work!
- Conclusions – practitioner's guide
- Appendix: gory details - boxcounting plots

15-826 Copyright: C. Faloutsos (2006) 27

**CMU SCS**

## Problem #1: GIS points

Cross-roads of Montgomery county:  
• any rules?

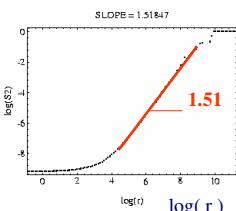


15-826 Copyright: C. Faloutsos (2006) 28

**CMU SCS**

## Solution #1

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



A: self-similarity ->

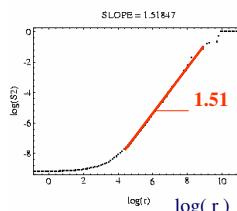
- $\Leftrightarrow$  fractals
- $\Leftrightarrow$  scale-free
- $\Leftrightarrow$  power-laws ( $y=x^{\alpha}$ ,  $F=C \cdot r^{(-2)}$ )
- avg#neighbors( $\leq r$ ) =  $r^D$

15-826 Copyright: C. Faloutsos (2006) 29

**CMU SCS**

## Solution #1

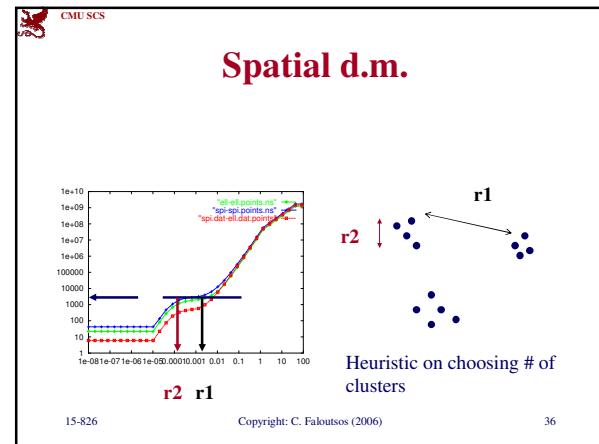
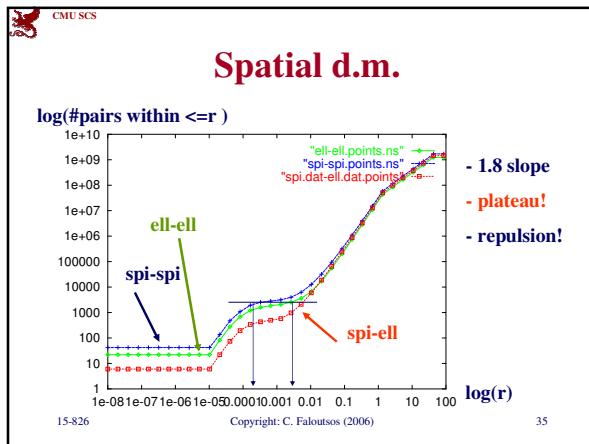
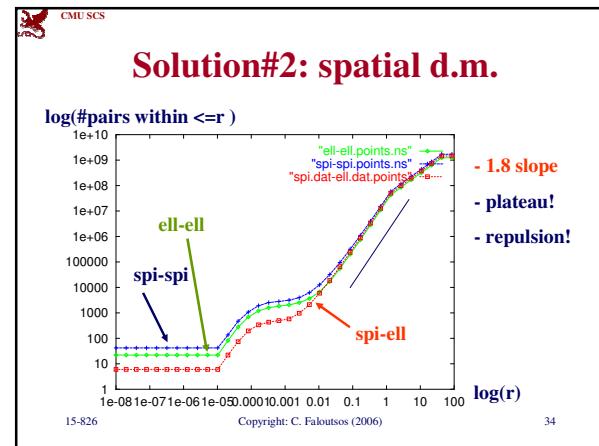
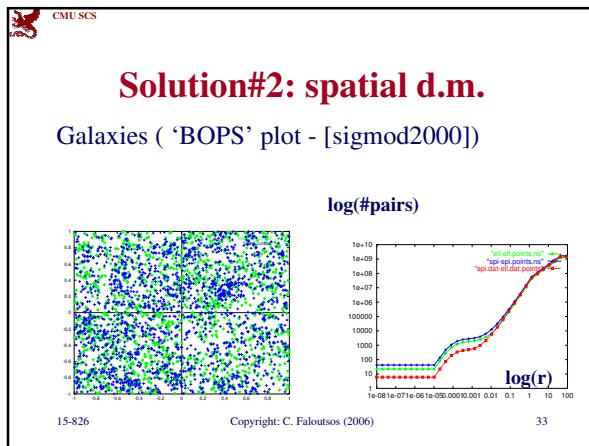
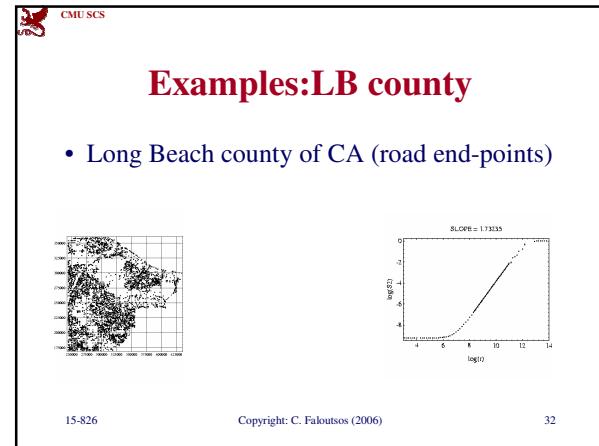
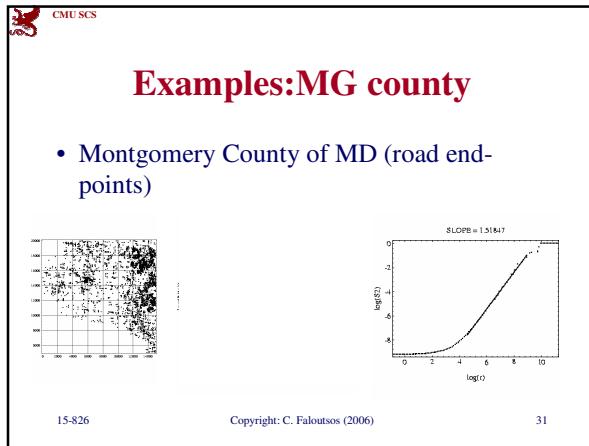
$\log(\#\text{pairs}(\text{within } \leq r))$

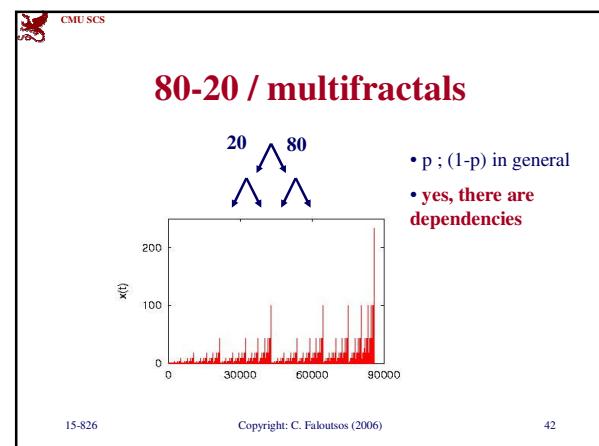
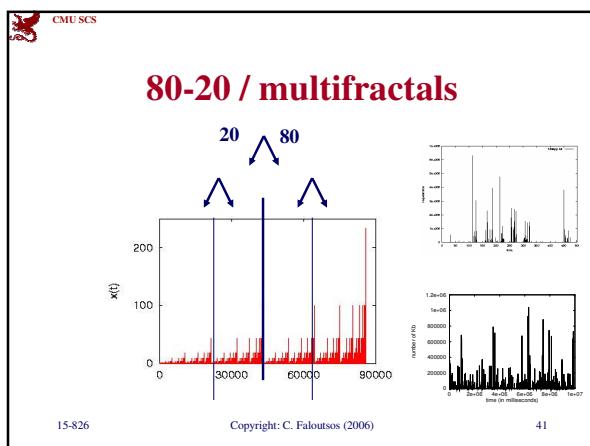
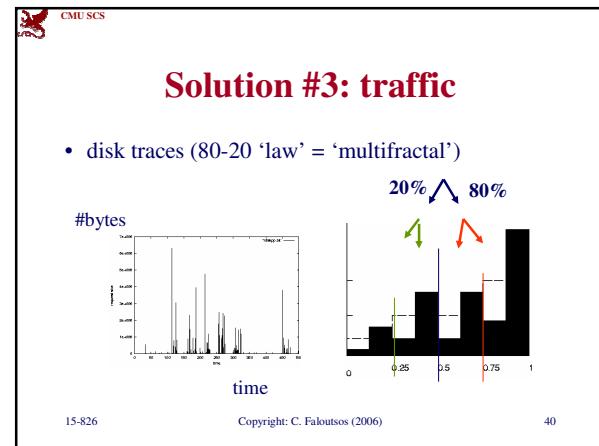
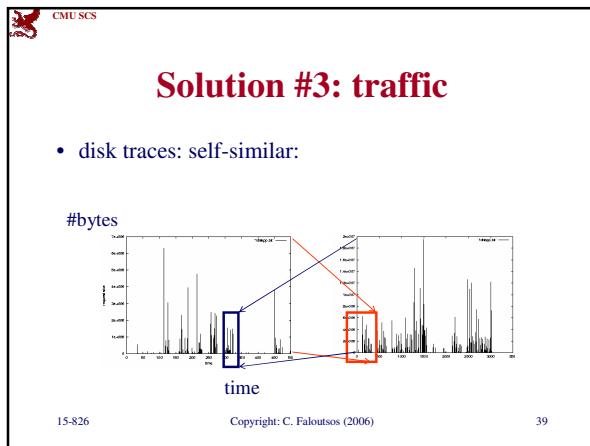
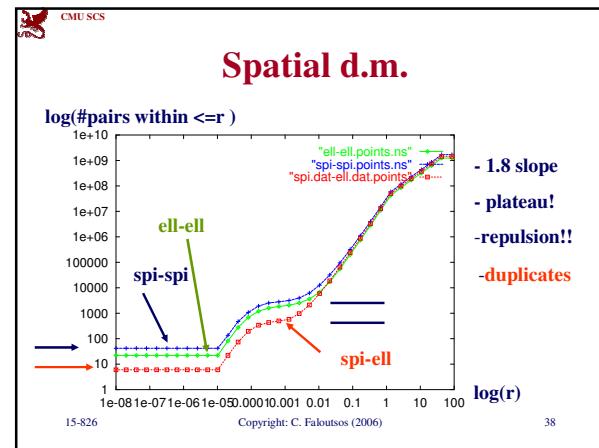
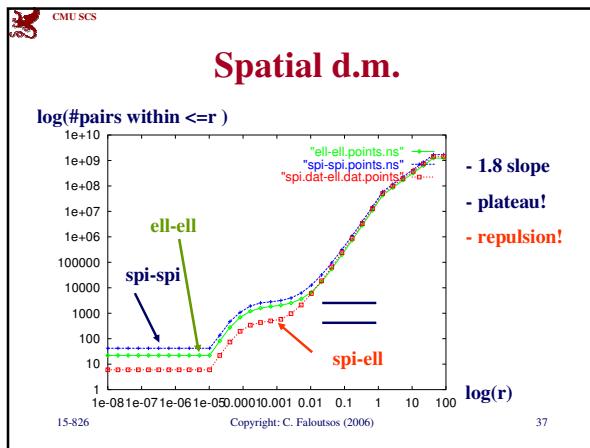


A: self-similarity

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

15-826 Copyright: C. Faloutsos (2006) 30





**CMU SCS**

## More on 80/20: PQRS

- Part of ‘self-\* storage’ project

Copyright: C. Faloutsos (2006) 43

**CMU SCS**

## More on 80/20: PQRS

- Part of ‘self-\* storage’ project

Copyright: C. Faloutsos (2006) 44

**CMU SCS**

## Solution#3: traffic

Clarification:

- fractal: a set of points that is self-similar
- multiplicative: a probability density function that is self-similar

Many other time-sequences are bursty/clustered: (such as?)

15-826 Copyright: C. Faloutsos (2006) 45

**CMU SCS**

## Example:

- network traffic

<http://repository.cs.vt.edu/lbl-conn-7.tar.Z> 15-826 Copyright: C. Faloutsos (2006) 46

**CMU SCS**

## Web traffic

- [Crovella Bestavros, SIGMETRICS’96]

1000 sec; 100sec  
10sec; 1sec

15-826 Copyright: C. Faloutsos (2006) 47

**CMU SCS**

## Tape accesses

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

15-826 Copyright: C. Faloutsos (2006) 48

**Tape accesses**

**50-50 = Poisson**

The diagram shows a horizontal timeline with two tapes, Tape #1 and Tape #N, represented by double-headed arrows. Below the timeline, three red stars indicate retrieval points. To the right is a plot titled "50-50 = Poisson" showing the number of blocks (y-axis, 0.00 to 16.00) versus time in seconds (x-axis, 0.00 to 200.00). The plot includes several curves: a solid blue curve labeled "LD16", a dashed green curve, a dotted red curve, and a dashed blue curve. A red arrow points to the solid blue curve, which is labeled "real".

# tapes retrieved  
Tape#1      Tape# N  
time  
15-826      Copyright: C. Faloutsos (2006)

# qual. records 49

**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 (2006)      50

**A counter-intuitive example**

count

A graph with "count" on the y-axis and "degree" on the x-axis. A bell-shaped curve peaks at an average of 3.3. A red 'P' is shown above the peak, and a red 'no' symbol is overlaid on the graph.

avg: 3.3      degree  
15-826      Copyright: C. Faloutsos (2006)      51

**A counter-intuitive example**

count

A graph with "count" on the y-axis and "degree" on the x-axis. A bell-shaped curve peaks at an average of 3.3. A red 'no' symbol is overlaid on the graph.

avg: 3.3      degree  
15-826      Copyright: C. Faloutsos (2006)      52

**A counter-intuitive example**

count

A graph with "count" on the y-axis and "degree" on the x-axis, showing a long tail characteristic of a power-law distribution.

avg: 3.3      degree  
15-826      Copyright: C. Faloutsos (2006)      53

**Rank exponent  $R$**

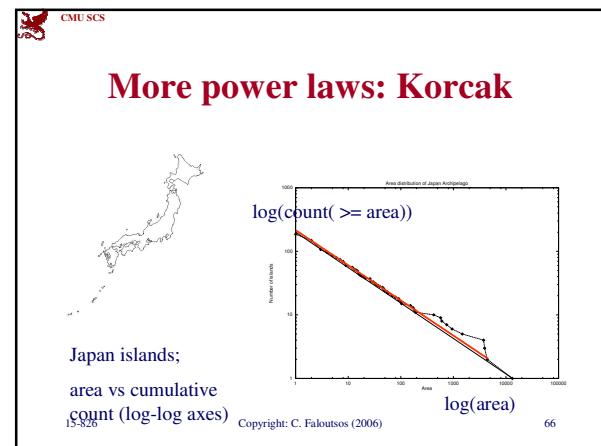
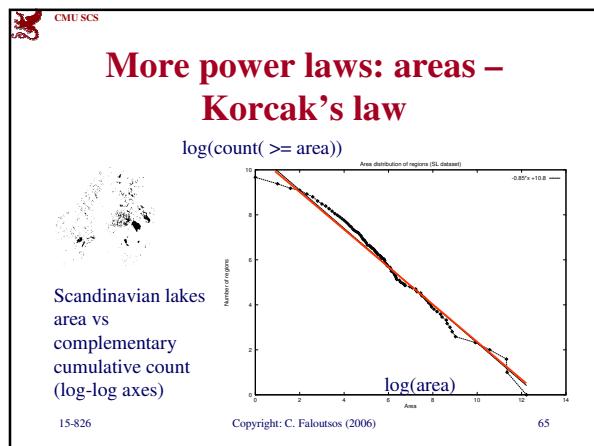
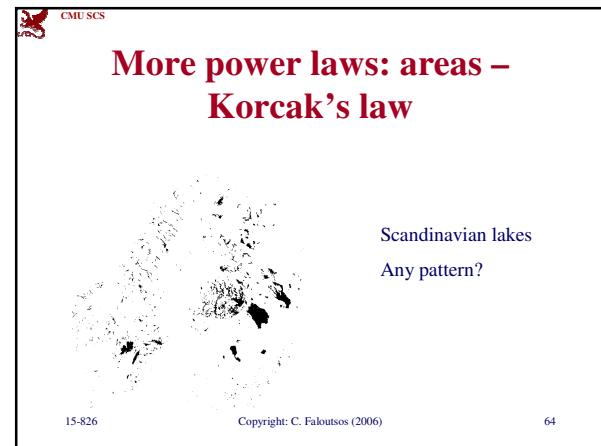
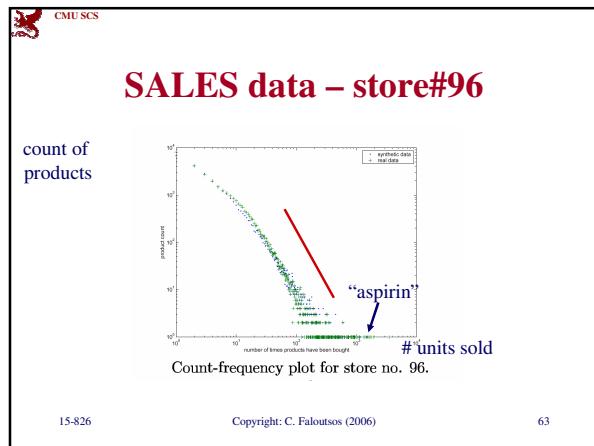
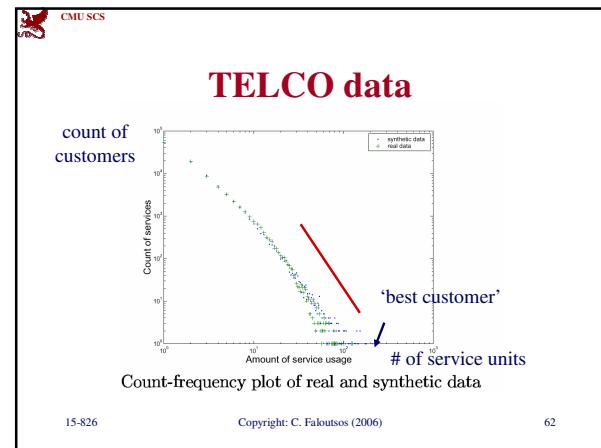
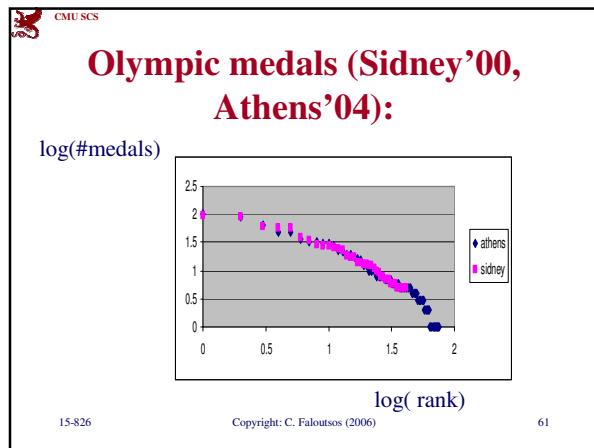
- Power law in the degree distribution [SIGCOMM99]

internet domains

A log-log plot showing the relationship between log(degree) on the y-axis and log(rank) on the x-axis for internet domains. The plot shows a linear decay with a slope of -0.82. Data points for "att.com" and "ibm.com" are labeled.

log(degree)      att.com  
log(rank)      ibm.com  
-0.82  
15-826      Copyright: C. Faloutsos (2006)      54





**(Korcak's law: Aegean islands)**

15-826      Copyright: C. Faloutsos (2006)      67

**Korcak's law & “fat fractals”**

15-826      Copyright: C. Faloutsos (2006)      68

**Korcak's law & “fat fractals”**

Q: How to generate such regions?  
A: recursively, from a single region

15-826      Copyright: C. Faloutsos (2006)      69

**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 (2006)      70

**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 (2006)      71

**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 (2006)      72

CMU SCS

## Other applications: Internet

- How does the internet look like?

15-826 Copyright: C. Faloutsos (2006) 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 (2006) 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 (2006) 75

CMU SCS

## Internet topology

- Internet routers: how many neighbors within  $h$  hops?

Reachability function:  
number of neighbors  
within r hops, vs r (log-log).  
Mbone routers, 1995

Copyright: C. Faloutsos (2006) 76

CMU SCS

## More power laws on the Internet

log(degree)     $\text{exp}(8.11393) * x^{-0.82}$   
log(rank)

degree vs rank, for Internet domains (log-log) [sigcomm99]

15-826 Copyright: C. Faloutsos (2006) 77

CMU SCS

## More power laws - internet

- pdf of degrees: (slope: 2.2 )

Log(count)     $\text{exp}(8.11393) * x^{-2.2}$   
Log(degree)

15-826 Copyright: C. Faloutsos (2006) 78

**Even more power laws on the Internet**

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

log(i-th eigenvalue)

log(i)

0.47

971198.internet.svals

exp(3.57926) \* x<sup>-0.471327</sup>

15-826 Copyright: C. Faloutsos (2006) 79

**Fractals & power laws:**

appear in numerous settings:

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

15-826 Copyright: C. Faloutsos (2006) 80

**More apps: Brain scans**

- Oct-trees; brain-scans

Log(#octants)

octree levels

2.63 = fd

15-826 Copyright: C. Faloutsos (2006) 81

**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 (2006) 82

**More fractals:**

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

15-826 Copyright: C. Faloutsos (2006) 83

**Fractals & power laws:**

appear in numerous settings:

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

15-826 Copyright: C. Faloutsos (2006) 84

**CMU SCS**

## More fractals:

- Coastlines: 1.2-1.58

15-826      Copyright: C. Faloutsos (2006)      85



**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 (2006)      87

**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 (2006)      88

**CMU SCS**

## More power laws

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

15-826      Copyright: C. Faloutsos (2006)      89

**CMU SCS**

## Fractals & power laws:

appear in numerous settings:

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

15-826      Copyright: C. Faloutsos (2006)      90

**More fractals:**

stock prices (LYCOS) - random walks: 1.5

1 year      2 years

15-826      Copyright: C. Faloutsos (2006)      91

**Even more power laws:**

- Income distribution (Pareto's law)
- size of firms
- publication counts (Lotka's law)

15-826      Copyright: C. Faloutsos (2006)      92

**Even more power laws:**

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

log(count)      log(#citations)

Ullman

15-826      Copyright: C. Faloutsos (2006)      93

**Even more power laws:**

- web hit counts [w/ A. Montgomery]

Web Site Traffic

log(count)      log(freq)

“yahoo.com”

15-826      Copyright: C. Faloutsos (2006)      94

**Fractals & power laws:**

appear in numerous settings:

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

15-826      Copyright: C. Faloutsos (2006)      95

**Power laws, cont'd**

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

Indegree Distribution

log indegree      - log(freq)

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

15-826      Copyright: C. Faloutsos (2006)      96

**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 (2006)      97

**CMU SCS**

## “Foiled by power law”

- [Broder+, WWW'00]

(log) count

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

Number of pages

in-degree

Copyright: C. Faloutsos (2006)

15-826      Copyright: C. Faloutsos (2006)      98

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

**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 (2006)      99

**CMU SCS**

## Even more power laws:

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

15-826      Copyright: C. Faloutsos (2006)      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 (2006)      101

**CMU SCS**

## What else can they solve?

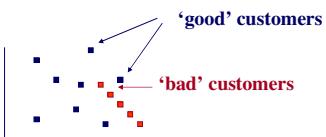
- ✓ separability [KDD'02]
- forecasting [CIKM'02]
- dimensionality reduction [SBB'D'00]
- non-linear axis scaling [KDD'02]
- ✓ disk trace modeling [PEVA'02]
- selectivity of spatial/multimedia queries [PODS'94, VLDB'95, ICDE'00]
- ...

15-826      Copyright: C. Faloutsos (2006)      102

 CMU SCS

## Settings for fractals:

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



15-826 Copyright: C. Faloutsos (2006) 103

 CMU SCS

## Settings for fractals:

Points; areas, eg:

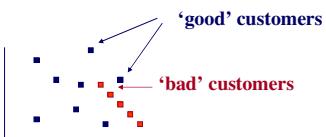
- cities/stores/hospitals, over earth's surface
- time-stamps of events (customer arrivals, packet losses, criminal actions) over time
- regions (sales areas, islands, patches of habitats) over space

15-826 Copyright: C. Faloutsos (2006) 104

 CMU SCS

## Settings for fractals:

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



15-826 Copyright: C. Faloutsos (2006) 105

 CMU SCS

## 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 (2006) 106

 CMU SCS

## Multi-Fractals

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

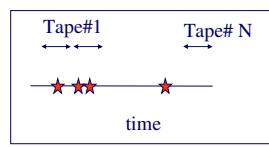
- cities w/ populations
- positions on earth and amount of gold/water/oil underneath
- product ids and sales per product
- people and their salaries
- months and count of accidents

15-826 Copyright: C. Faloutsos (2006) 107

 CMU SCS

## 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 (2006) 108

**CMU SCS**

## Use of multifractals

- how often do we exceed the threshold?

15-826      Copyright: C. Faloutsos (2006)      109

**CMU SCS**

## Use of multifractals cont'd

- Extrapolations for/from samples

15-826      Copyright: C. Faloutsos (2006)      110

**CMU SCS**

## Use of multifractals cont'd

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

15-826      Copyright: C. Faloutsos (2006)      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 (2006)      112

**CMU SCS**

## Conclusions

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

15-826      Copyright: C. Faloutsos (2006)      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, 114 • ++)
---	---

15-826      Copyright: C. Faloutsos (2006)      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 (2006)      115

**CMU SCS**

## Practitioner’s guide:

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

log(#pairs(within  $\leq r$ ))

internet

log(#pairs)      log(hops)

SLOPB = 2.8

MGcounty

log(r)      log( r )

SLOPB = 1.51

15-826      Copyright: C. Faloutsos (2006)      116

**CMU SCS**

## Practitioner’s guide:

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

internet domains

Bible

log(freq)      log(rank)

log(degree)      log(rank)

-0.82

15-826      Copyright: C. Faloutsos (2006)      117

**CMU SCS**

## Practitioner’s guide:

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

log(count( $\geq a$ ))

scandinavian lakes

log(area)

15-826      Copyright: C. Faloutsos (2006)      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 (2006)      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 (2006)      120

 CMU SCS

## References

- [vl95] 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 Maghoul, 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 (2006)      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 (2006)      122

 CMU SCS

## References

- [vl96] 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 (2006)      123

 CMU SCS

## References

- [vl96] Christos Faloutsos and Volker Gaede *Analysis of the Z-Ordering Method Using the Hausdorff Fractal Dimension* VLDB, 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 (2006)      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 (2006)      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 (2006)      126

**CMU SCS**

### Fast estimation of fd(s):

- How, for the (correlation) fractal dimension?
- A: Box-counting plot:  $\log(\text{sum}(\pi^2))$

15-826 Copyright: C. Faloutsos (2006) 127

**CMU SCS**

### Definitions

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

15-826 Copyright: C. Faloutsos (2006) 128

**CMU SCS**

### Fast estimation of fd(s):

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

15-826 Copyright: C. Faloutsos (2006) 129

**CMU SCS**

### Fast estimation of fd(s):

- etc; if the resulting plot has a linear part, its slope is the correlation fractal dimension  $D_2$

15-826 Copyright: C. Faloutsos (2006) 130

**CMU SCS**

### Definitions (cont'd)

- Many more fractal dimensions  $D_q$  (related to Renyi entropies):

$$D_q = \frac{1}{q-1} \frac{\partial \log(\sum p_i^q)}{\partial \log(r)} \quad q \neq 1$$

$$D_1 = \frac{\partial \sum p_i \log(p_i)}{\partial \log(r)}$$

15-826 Copyright: C. Faloutsos (2006) 131

**CMU SCS**

### Hausdorff or box-counting fd:

- Box counting plot:  $\log(N(r))$  vs  $\log(r)$
- $r$ : grid side
- $N(r)$ : count of non-empty cells
- (Hausdorff) fractal dimension  $D_0$ :

$$D_0 = -\frac{\partial \log(N(r))}{\partial \log(r)}$$

15-826 Copyright: C. Faloutsos (2006) 132

**CMU SCS**

## Definitions (cont'd)

- Hausdorff fd:

$r \equiv \log(\# \text{non-empty cells})$

15-826 Copyright: C. Faloutsos (2006) 133

**CMU SCS**

## Observations

- $q=0$ : Hausdorff fractal dimension
- $q=2$ : Correlation fractal dimension (**identical** to the exponent of the number of neighbors vs radius)
- $q=1$ : Information fractal dimension

15-826 Copyright: C. Faloutsos (2006) 134

**CMU SCS**

## Observations, cont'd

- in general, the  $D_q$ 's take similar, but not identical, values.
- except for perfectly self-similar point-sets, where  $D_q = D_{q'}$  for any  $q, q'$

15-826 Copyright: C. Faloutsos (2006) 135

**CMU SCS**

## Examples: MG county

- Montgomery County of MD (road end-points)

15-826 Copyright: C. Faloutsos (2006) 136

**CMU SCS**

## Examples: LB county

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

15-826 Copyright: C. Faloutsos (2006) 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 (2006) 138