


CMU SCS

15-826: Multimedia Databases and Data Mining

Lecture #9: Fractals – examples & algo's
C. Faloutsos




CMU SCS

Must-read Material

- Christos Faloutsos and Ibrahim Kamel, [*Beyond Uniformity and Independence: Analysis of R-trees Using the Concept of Fractal Dimension*](#), Proc. ACM SIGACT-SIGMOD-SIGART PODS, May 1994, pp. 4-13, Minneapolis, MN.

15-826 Copyright: C. Faloutsos (2016) 2



CMU SCS

Recommended Material

optional, but **very** useful:

- Manfred Schroeder *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise* W.H. Freeman and Company, 1991
 - Chapter 10: boxcounting method
 - Chapter 1: Sierpinski triangle

15-826 Copyright: C. Faloutsos (2016) 3



CMU SCS

Outline

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

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

15-826 Copyright: C. Faloutsos (2016) 4

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 (2016) 5

CMU SCS

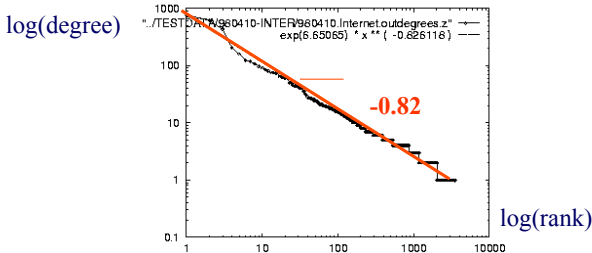
Road map

- 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 (2016) 6

CMU SCS

More power laws on the Internet



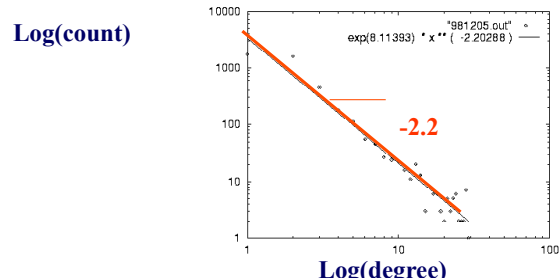
degree vs rank, for Internet domains (log-log) [sigcomm99]

15-826 Copyright: C. Faloutsos (2016) 7

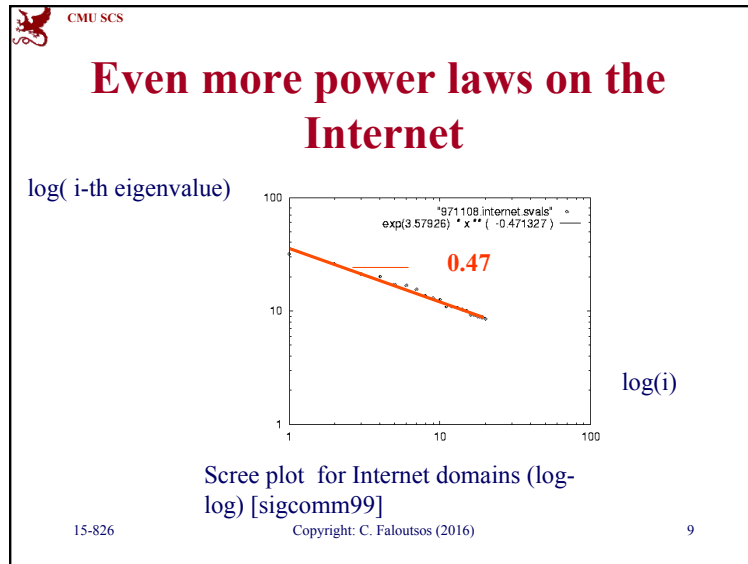
CMU SCS

More power laws - internet

- pdf of degrees: (slope: 2.2)



15-826 Copyright: C. Faloutsos (2016) 8

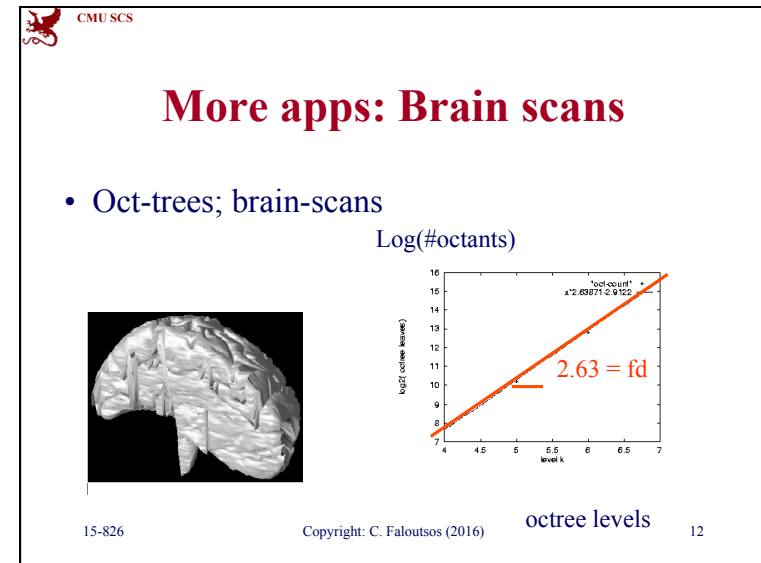
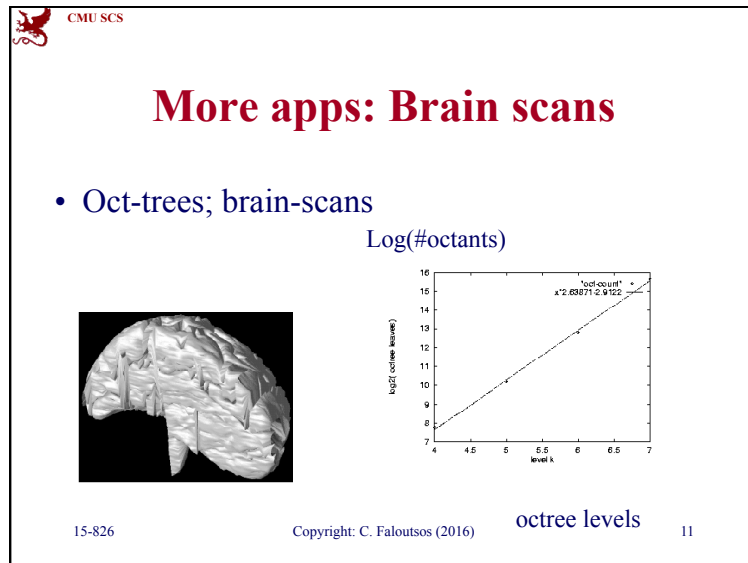


Fractals & power laws:

appear in numerous settings:

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

15-826 Copyright: C. Faloutsos (2016) 10



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 (2016) 13

CMU SCS

More fractals:

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



15-826 Copyright: C. Faloutsos (2016) 14

CMU SCS

Fractals & power laws:

appear in numerous settings:


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

15-826 Copyright: C. Faloutsos (2016) 15

CMU SCS

More fractals:

- Coastlines: 1.2-1.58



15-826 Copyright: C. Faloutsos (2016) 16



15-826

Copyright: C. Faloutsos (2016)

More fractals:

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

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

A topographic map of the Amazon basin, showing the extensive and highly branched river network. The map uses a color gradient from green (low elevation) to blue (sea level) to show the terrain. The river network is clearly visible as a dense web of lines.

15-826

Copyright: C. Faloutsos (2016)

18

More fractals:

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

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

A black and white fractal image of a river network, showing the complex, branching structure of the river system. The image is a high-contrast, black-and-white representation of the river network, highlighting its fractal nature.

15-826

Copyright: C. Faloutsos (2016)

19

More power laws

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

amplitude

Maximum Amplitude

day

log(freq)

Frequency

magnitude

15-826

Copyright: C. Faloutsos (2016)

20

CMU SCS

Fractals & power laws:

appear in numerous settings:

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


15-826 Copyright: C. Faloutsos (2016) 21

CMU SCS


More fractals:

stock prices (LYCOS) - random walks: 1.5

1 year



2 years



15-826 Copyright: C. Faloutsos (2016) 22

CMU SCS

Even more power laws:

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

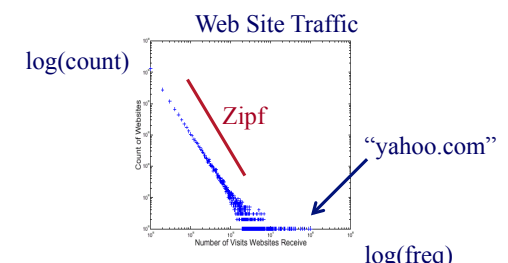
15-826 Copyright: C. Faloutsos (2016) 23

CMU SCS

Even more power laws:

- web hit counts [w/ A. Montgomery]

Web Site Traffic



15-826 Copyright: C. Faloutsos (2016) 24

CMU SCS

Fractals & power laws:

appear in numerous settings:

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

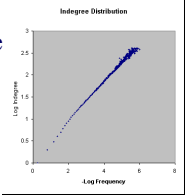
15-826 Copyright: C. Faloutsos (2016) 25

CMU SCS

Power laws, cont'd

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

log indegree



- log(freq)

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

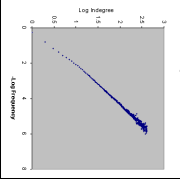
15-826 Copyright: C. Faloutsos (2016) 26

CMU SCS

Power laws, cont'd

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

log(freq)



log indegree

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

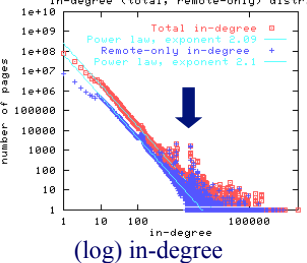
15-826 Copyright: C. Faloutsos (2016) 27

CMU SCS

“Foiled by power law”

- [Broder+, WWW'00]

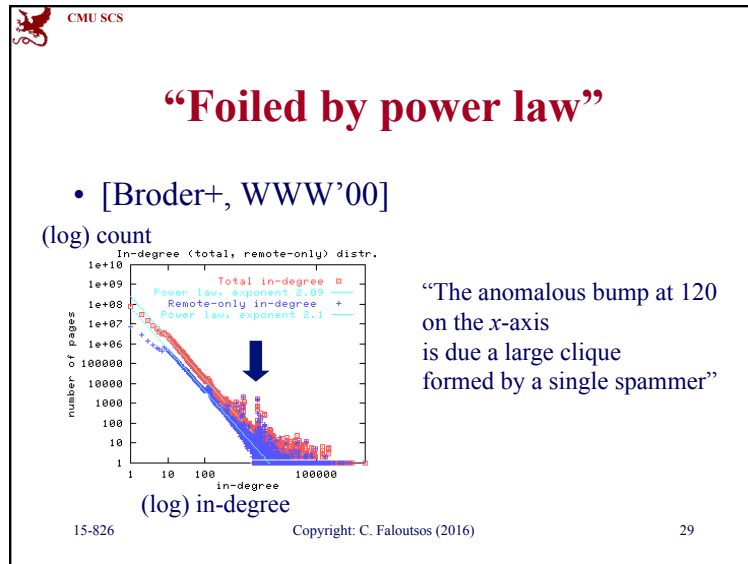
(log) count



number of pages

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

15-826 Copyright: C. Faloutsos (2016) 28



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 (2016) 30

CMU SCS

Even more power laws:

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

15-826 Copyright: C. Faloutsos (2016) 31

CMU SCS

Road map

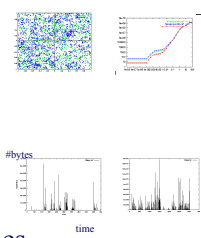
- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems
- More examples and tools
- ➔ Discussion - putting fractals to work!
- Conclusions – practitioner’s guide
- Appendix: gory details - boxcounting plots

15-826 Copyright: C. Faloutsos (2016) 32

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 [Wang+'02]
- selectivity of spatial/multimedia queries [PODS'94, VLDB'95, ICDE'00]
- ...



15-826 Copyright: C. Faloutsos (2016) 33

CMU SCS

Settings for fractals:

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

15-826 Copyright: C. Faloutsos (2016) 34

CMU SCS

Settings for fractals:

Points; areas, eg:

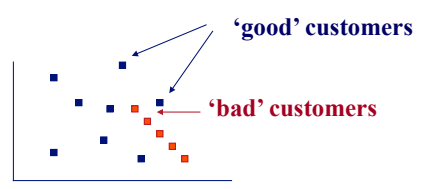
- cities/stores/hospitals, over earth's surface
- time-stamps of events (customer arrivals, packet losses, criminal actions) over time
- regions (sales areas, islands, patches of habitats) over space

15-826 Copyright: C. Faloutsos (2016) 35

CMU SCS

Settings for fractals:

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



15-826 Copyright: C. Faloutsos (2016) 36

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 (2016) 37

CMU SCS

Multi-Fractals

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

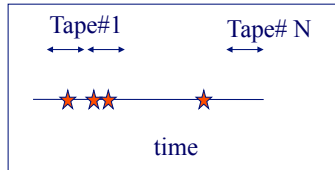
- cities w/ populations
- positions on earth and amount of gold/water/oil underneath
- product ids and sales per product
- people and their salaries
- months and count of accidents

15-826 Copyright: C. Faloutsos (2016) 38

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?*



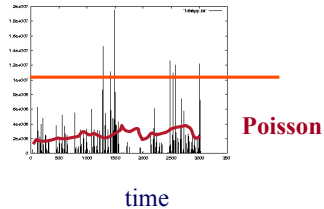
The diagram shows a horizontal line representing time. Above the line, two double-headed arrows labeled 'Tape#1' and 'Tape# N' indicate access intervals. Below the line, three stars represent events occurring at different points in time.

15-826 Copyright: C. Faloutsos (2016) 39

CMU SCS

Use of multifractals

- how often do we exceed the threshold?



The graph plots '#bytes' on the y-axis against 'time' on the x-axis. A horizontal orange line represents a threshold. The data points are shown as a series of vertical spikes. A red curve labeled 'Poisson' is overlaid on the data, representing a model fit.

15-826 Copyright: C. Faloutsos (2016) 40

CMU SCS

Use of multifractals cont'd

- Extrapolations for/from samples

#bytes

time

15-826 Copyright: C. Faloutsos (2016) 41

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 (2016) 42

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 (2016) 43

CMU SCS

Conclusions

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

15-826 Copyright: C. Faloutsos (2016) 44

CMU SCS

Conclusions

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

15-826 Copyright: C. Faloutsos (2016) 45

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)


15-826 Copyright: C. Faloutsos (2016) 46

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 (2016) 47

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 (2016) 48

CMU SCS

Practitioner's guide:

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

log(#pairs) internet

log(hops)

2.8

log(#pairs(within <= r))

SLOPE = 1.51847

MGcounty

log(r)

1.51

15-826 Copyright: C. Faloutsos (2016) 49

CMU SCS

Practitioner's guide:

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

internet domains

log(degree)

log(rank)

-0.82

Bible

log(freq)

log(rank)

15-826 Copyright: C. Faloutsos (2016) 50

CMU SCS

Practitioner's guide:

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

log(count(>= area))

scandinavian lakes

log(area)


15-826 Copyright: C. Faloutsos (2016) 51

CMU SCS

Resources:

- Software for fractal dimension
 - www.cs.cmu.edu/~christos/software.html
 - And specifically 'fdnq_h':
 - www.cs.cmu.edu/~christos/SRC/fdnq_h.zip
- Also, in 'R': 'fdim' package

15-826 Copyright: C. Faloutsos (2016) 52




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 (2016) 53




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 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 (2016) 54




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 (2016) 55




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 (2016) 56




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 (2016) 57




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 (2016) 58




CMU SCS

References

- [Wang+02] Mengzhi Wang, Anastassia Ailamaki and Christos Faloutsos, [Capturing the spatio-temporal behavior of real traffic data](#) Performance 2002 (IFIP Int. Symp. on Computer Performance Modeling, Measurement and Evaluation), Rome, Italy, Sept. 2002

15-826 Copyright: C. Faloutsos (2016) 59



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 (2016) 60

CMU SCS

Fast estimation of fd(s):

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

15-826 Copyright: C. Faloutsos (2016) 61

CMU SCS

Definitions

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

15-826 Copyright: C. Faloutsos (2016) 62

CMU SCS

Fast estimation of fd(s):

- compute $\sum(\pi_i^2)$ for another grid side, r'

15-826 Copyright: C. Faloutsos (2016) 63

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 (2016) 64

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 (2016) 65

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 (2016) 66

CMU SCS

Definitions (cont'd)

- Hausdorff fd:

r — $\log(\# \text{non-empty cells})$

SLOPE = -1.5743

D_0

15-826 Copyright: C. Faloutsos (2016) 67

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 (2016) 68

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 (2016) 69

CMU SCS

Examples: MG county

- Montgomery County of MD (road end-points)

q=0
SLOPE = -1.71945

q=2
SLOPE = 1.51847

15-826 Copyright: C. Faloutsos (2016) 70

CMU SCS

Examples: LB county

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

q=0
SLOPE = -1.72775

q=2
SLOPE = 1.73235

15-826 Copyright: C. Faloutsos (2016) 71

CMU SCS

Conclusions

- many fractal dimensions, with nearby values
- can be computed quickly ($O(N)$ or $O(N \log(N))$)
- (code: on the web:
 - www.cs.cmu.edu/~christos/SRC/fdnq_h.zip
 - Or 'R' ('fdim' package)

15-826 Copyright: C. Faloutsos (2016) 72