CMU SCS

15-826: Multimedia Databases and Data Mining

Lecture #8: Fractals - introduction

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.

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 (on reserve in the library)
  - Chapter 10: boxcounting method
  - Chapter 1: Sierpinski triangle

Outline

Goal: 'Find similar / interesting things'

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

Indexing - Detailed outline

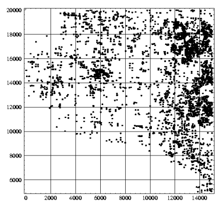
- primary key indexing
- secondary key / multi-key indexing
- spatial access methods
  - z-ordering
  - R-trees
  - misc
- fractals
  - ➡ intro
  - applications
- text

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

CMU SCS

Problem #1: GIS - points



Road end-points of Montgomery county:

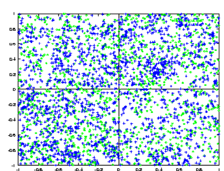
- •Q1: how many d.a. for an R-tree?
- •Q2 : distribution?
  - •not uniform
  - •not Gaussian
  - •no rules??

15-826 Copyright: C. Faloutsos (2013) 7

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?

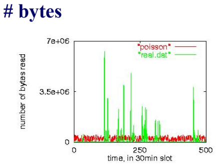
- how many 'spi' within r from an 'ell'?

15-826 Copyright: C. Faloutsos (2013) 8

CMU SCS

Problem #3: traffic

 # bytes



time

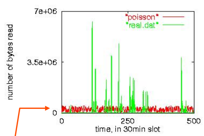
- disk trace (from HP - J. Wilkes); Web traffic - fit a model
- how many explosions to expect?
- queue length distr.?

15-826 Copyright: C. Faloutsos (2013) 9

CMU SCS

Problem #3: traffic

bytes



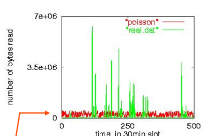
time

Poisson
indep.,
ident. distr

CMU SCS

Problem #3: traffic

bytes



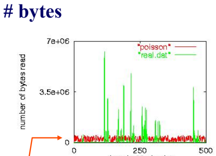
time

~~Poisson~~
indep.,
ident. distr

CMU SCS

Problem #3: traffic

bytes



time

~~Poisson~~
~~indep.,~~
ident. distr

Q: Then, how to generate such bursty traffic?

CMU SCS

Common answer:

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

CMU SCS

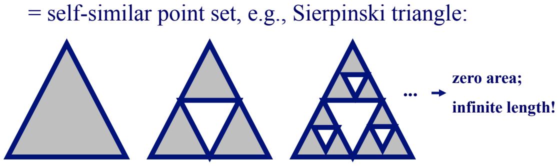
Road map

- 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

CMU SCS

What is a fractal?

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



Dimensionality??

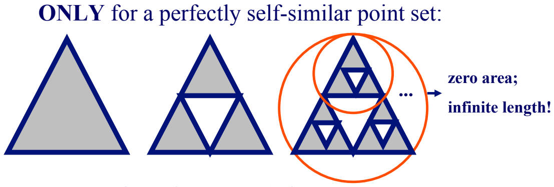
Definitions (cont'd)

- Paradox: Infinite perimeter ; Zero area!
- 'dimensionality': between 1 and 2
- actually: Log(3)/Log(2) = 1.58...

CMU SCS

Dfn of fd:

ONLY for a perfectly self-similar point set:

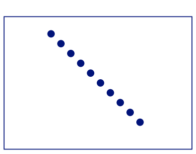


=log(n)/log(f) = log(3)/log(2) = 1.58

CMU SCS

Intrinsic ('fractal') dimension

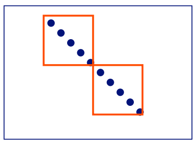
- Q: fractal dimension of a line?
- A: 1 (= log(2)/log(2)!)



CMU SCS

Intrinsic ('fractal') dimension

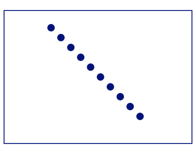
- Q: fractal dimension of a line?
- A: 1 (= log(2)/log(2)!)



CMU SCS

Intrinsic ('fractal') dimension

- Q: dfn for a given set of points?



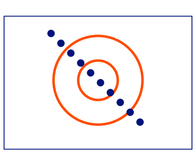
x	y
5	1
4	2
3	3
2	4

CMU SCS

Intrinsic ('fractal') dimension

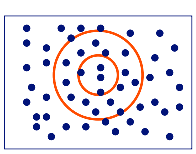
- Q: fractal dimension of a line?
- A: nn (<= r) ~ r^1

('power law': y=x^a)



- Q: fd of a plane?
- A: nn (<= r) ~ r^2

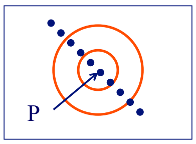
fd== slope of (log(nn) vs log (r))



EXPLANATIONS

Intrinsic ('fractal') dimension

- **Local** fractal dimension of point 'P'?
- A: nn_P (<= r) ~ r^1

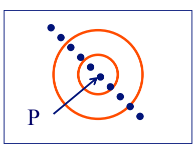


- If this equation holds for several values of r,
- Then, the **local fractal dimension** of point P:
- Local fd = exp = 1

EXPLANATIONS

Intrinsic ('fractal') dimension

- **Local** fractal dimension of point 'A'?
- A: nn_P (<= r) ~ r^1

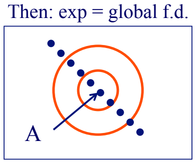


- If this is true for all points of the cloud
- Then the exponent is the **global** f.d.
- Or simply the f.d.

EXPLANATIONS

Intrinsic ('fractal') dimension

- **Global** fractal dimension?
- A: if
- sum_{all_P} [nn_P (<= r)] ~ r^1
- Then: exp = global f.d.



- If this is true for all points of the cloud
- Then the exponent is the **global** f.d.
- Or simply the f.d.

CMU SCS

EXPLANATIONS

Intrinsic ('fractal') dimension

- Algorithm, to estimate it?

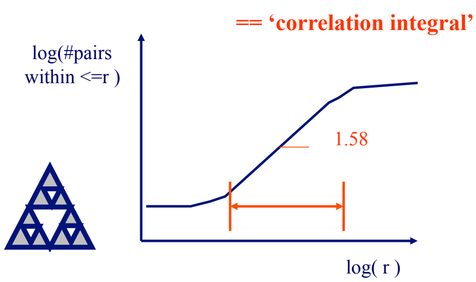
Notice

- Sum_{all_P} *[* nn_P *(<=r)]* is exactly *tot#pairs(<=r)*

including 'mirror' pairs

CMU SCS

Sierpinsky triangle



CMU SCS

Observations:

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

CMU SCS

Important properties

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

CMU SCS

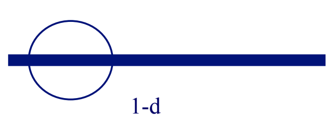
Important properties

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

CMU SCS

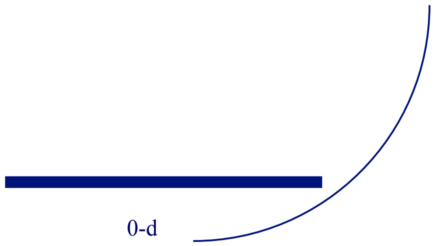
Important properties

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



CMU SCS

Important properties



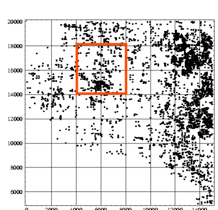
CMU SCS

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

CMU SCS

Problem #1: GIS points

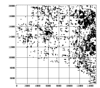


Cross-roads of Montgomery county:

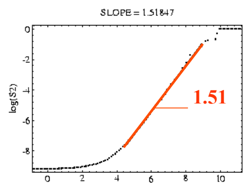
•any rules?

CMU SCS

Solution #1



log(#pairs(within <= r))



log(r)

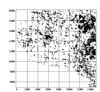
A: self-similarity ->

- <=> fractals
- <=> scale-free
- <=> power-laws (y=x^a, F=C*r^(-2))
- avg#neighbors(<= r) = r^D

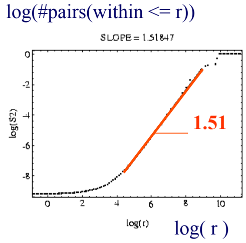
15-826 Copyright: C. Faloutsos (2013) 34

CMU SCS

Solution #1



log(#pairs(within <= r))



log(r)

A: self-similarity

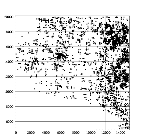
- avg#neighbors(<= r) ~ r^(1.51)

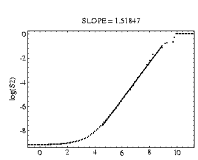
15-826 Copyright: C. Faloutsos (2013) 35

CMU SCS

Examples:MG county

- Montgomery County of MD (road end-points)



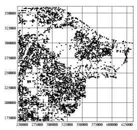


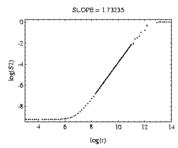
15-826 Copyright: C. Faloutsos (2013) 36

CMU SCS

Examples:LB county

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

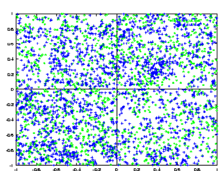


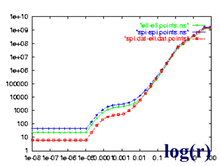


CMU SCS

Solution#2: spatial d.m.

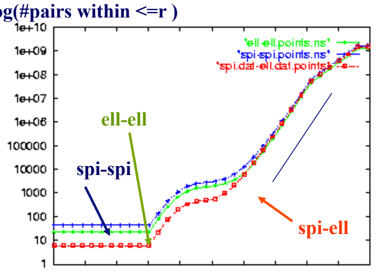
Galaxies (‘BOPS’ plot - [sigmod2000])





CMU SCS

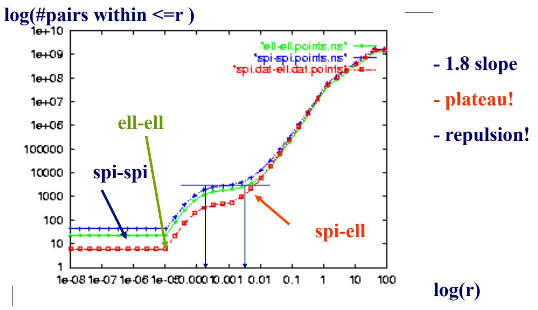
Solution#2: spatial d.m.



- 1.8 slope
- plateau!
- repulsion!

CMU SCS

Spatial d.m.

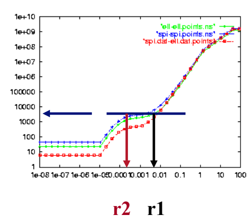


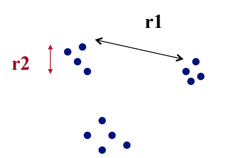
- 1.8 slope
- plateau!
- repulsion!

15-826 Copyright: C. Faloutsos (2013) 40

CMU SCS

Spatial d.m.



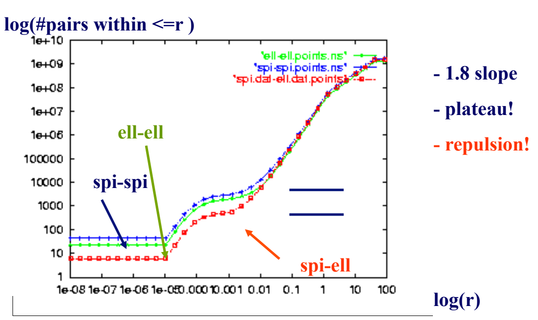


Heuristic on choosing # of clusters

15-826 Copyright: C. Faloutsos (2013) 41

CMU SCS

Spatial d.m.

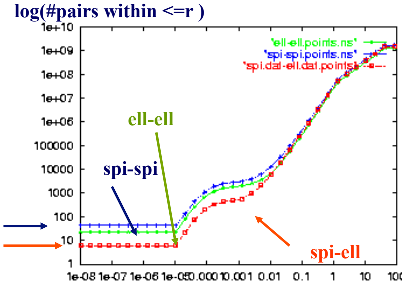


- 1.8 slope
- plateau!
- repulsion!

15-826 Copyright: C. Faloutsos (2013) 42

CMU SCS

Spatial d.m.



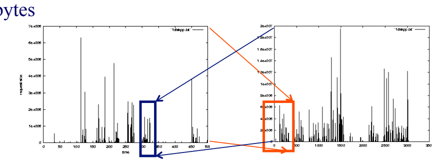
15-826 Copyright: C. Faloutsos (2013) 43

CMU SCS

Solution #3: traffic

- disk traces: self-similar:

#bytes



time

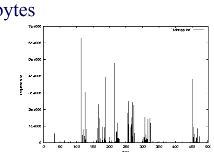
15-826 Copyright: C. Faloutsos (2013) 44

CMU SCS

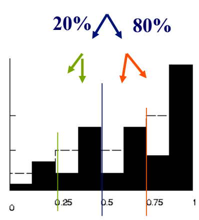
Solution #3: traffic

- disk traces (80-20 ‘law’ = ‘multifractal’)

#bytes



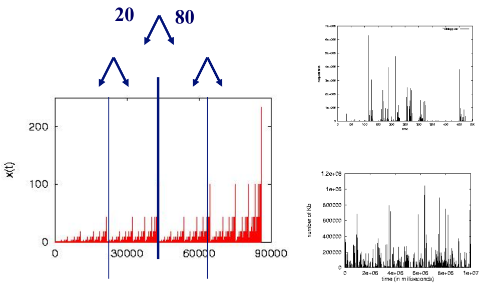
time



15-826 Copyright: C. Faloutsos (2013) 45

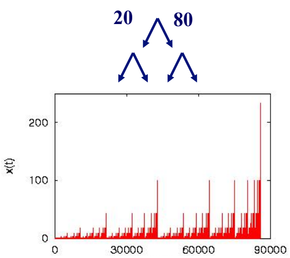
CMU SCS

80-20 / multifractals



CMU SCS

80-20 / multifractals

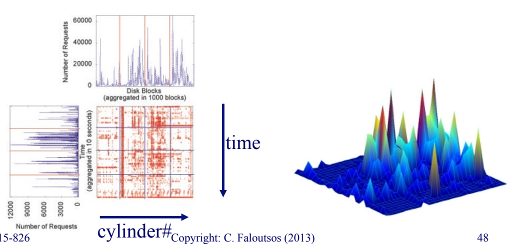


- p ; (1-p) in general
- **yes, there are dependencies**

CMU SCS

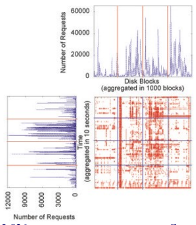
More on 80/20: PQRS

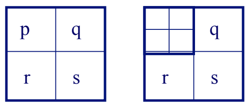
- Part of 'self-* storage' project [Wang+'02]



More on 80/20: PQRS

- Part of 'self-* storage' project [Wang+'02]





Solution#3: traffic

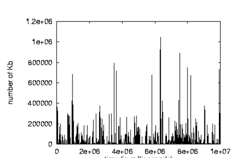
Clarification:

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

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

Example:

- network traffic



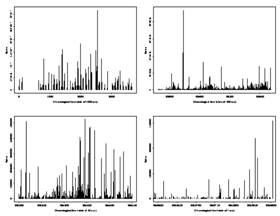
http://repository.cs.vt.edu/lbl-conn-7.tar.Z

CMU SCS

Web traffic

- [Crovella Bestavros, SIGMETRICS'96]

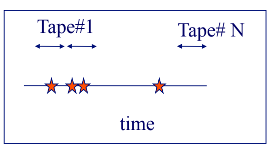
1000 sec; 100sec
10sec; 1sec



15-826 Copyright: C. Faloutsos (2013) 52

CMU SCS

Tape accesses



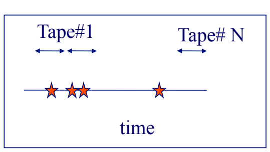
tapes needed, to retrieve *n* records?

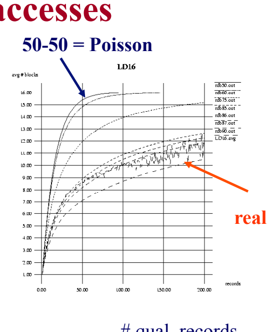
(# days down, due to failures / hurricanes / communication noise...)

15-826 Copyright: C. Faloutsos (2013) 53

CMU SCS

Tape accesses





15-826 Copyright: C. Faloutsos (2013) 54

CMU SCS

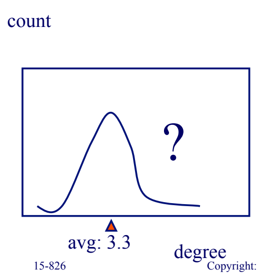
Road map

- 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

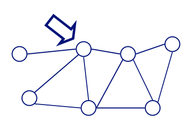
15-826 Copyright: C. Faloutsos (2013) 55

CMU SCS

A counter-intuitive example



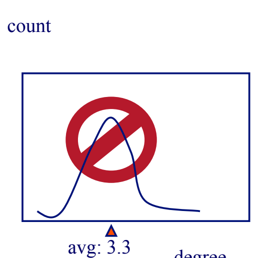
- avg degree is, say 3.3
- pick a node at random – guess its degree, exactly (-> "mode")



15-826 Copyright: C. Faloutsos (2013) 56

CMU SCS

A counter-intuitive example

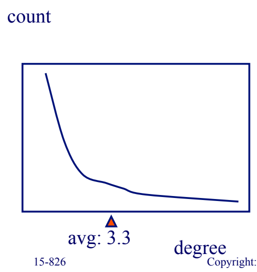


- avg degree is, say 3.3
- pick a node at random – guess its degree, exactly (-> "mode")
- A: 1!!

15-826 Copyright: C. Faloutsos (2013) 57

CMU SCS

A counter-intuitive example



- 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 (!))

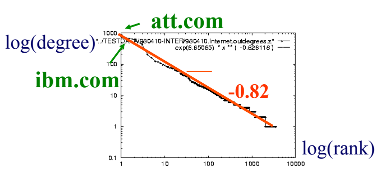
15-826 Copyright: C. Faloutsos (2013) 58

CMU SCS

Rank exponent *R*

- Power law in the degree distribution [SIGCOMM99]

internet domains



15-826 Copyright: C. Faloutsos (2013) 59

CMU SCS

More tools

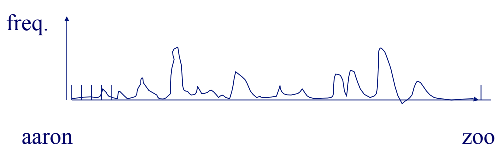
- Zipf's law
- Korcak's law / "fat fractals"

15-826 Copyright: C. Faloutsos (2013) 60

CMU SCS

A famous power law: Zipf's law

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

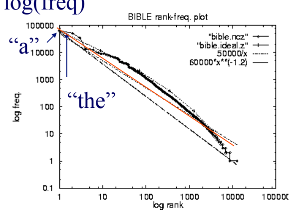


15-826 Copyright: C. Faloutsos (2013) 61

CMU SCS

A famous power law: Zipf's law

log(freq)



log(rank)

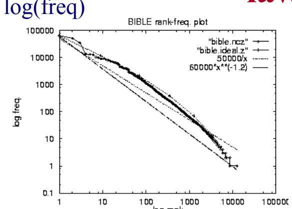
- Bible - **rank** vs **frequency** (log-log)

15-826 Copyright: C. Faloutsos (2013) 62

CMU SCS

A famous power law: Zipf's law

log(freq)



log(rank)

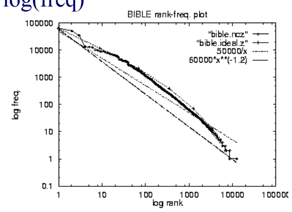
- Bible - rank vs frequency (log-log)
- similarly, in **many other** languages; for customers and sales volume; city populations etc etc

15-826 Copyright: C. Faloutsos (2013) 63

CMU SCS

A famous power law: Zipf's law

log(freq)



•Zipf distr:

freq = 1/ rank

•generalized Zipf:

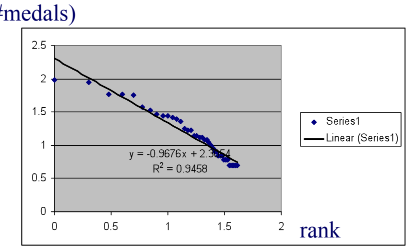
freq = 1 / (rank)^a

log(rank)

CMU SCS

Olympic medals (Sydney):

log(#medals)

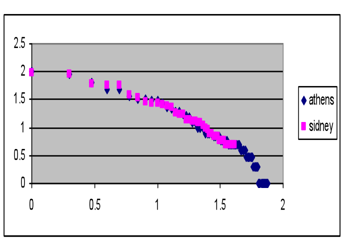


rank

CMU SCS

Olympic medals (Sydney'00, Athens'04):

log(#medals)

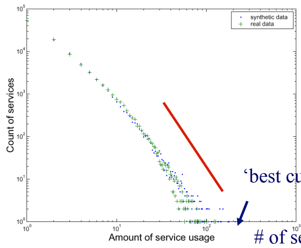


log(rank)

CMU SCS

TELCO data

count of customers



'best customer'

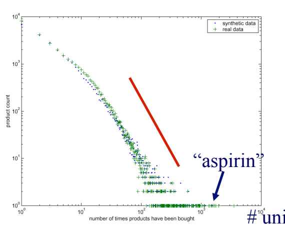
# of service units

Count-frequency plot of real and synthetic data

CMU SCS

SALES data – store#96

count of products



"aspirin"

units sold

Count-frequency plot for store no. 96.

CMU SCS

More power laws: areas – Korcak's law

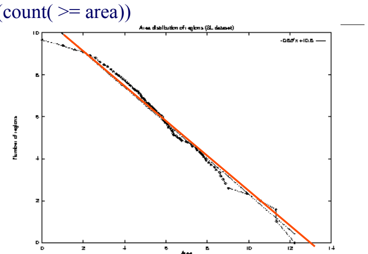
Scandinavian lakes

Any pattern?

CMU SCS

More power laws: areas – Korcak's law

log(count(>= area))



log(area)

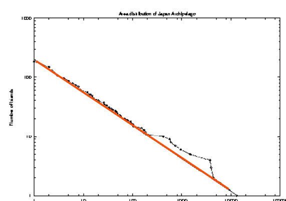
Scandinavian lakes area vs complementary cumulative count (log-log axes)

15-826 Copyright: C. Faloutsos (2013) 70

CMU SCS

More power laws: Korcak

log(count(>= area))



log(area)

Japan islands;

area vs cumulative count (log-log axes)

15-826 Copyright: C. Faloutsos (2013) 71

CMU SCS

(Korcak's law: Aegean islands)

15-826 Copyright: C. Faloutsos (2013) 72

CMU SCS

Korcak's law & "fat fractals"

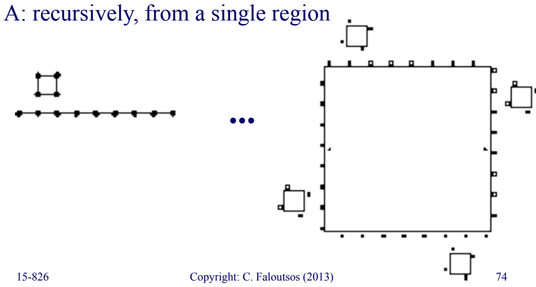
How to generate such regions?

15-826 Copyright: C. Faloutsos (2013) 73

CMU SCS

Korcak's law & "fat fractals"

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



15-826 Copyright: C. Faloutsos (2013) 74

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 (2013) 75

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

CMU SCS

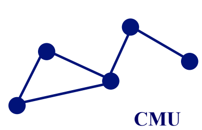
Road map

- 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

CMU SCS

Other applications: Internet

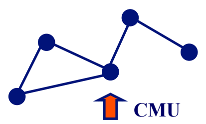
- How does the internet look like?



CMU SCS

Other applications: Internet

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



CMU SCS

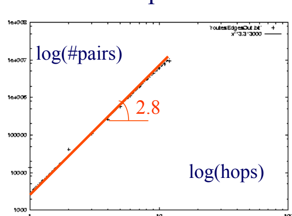
(reminder: our tool-box:)

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

CMU SCS

Internet topology

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

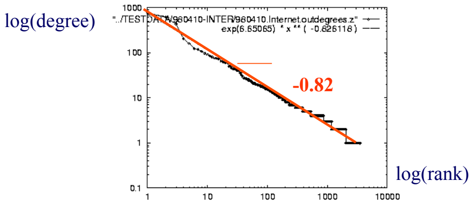


Reachability function: number of neighbors within r hops, vs r (log-log).

Mbone routers, 1995

CMU SCS

More power laws on the Internet

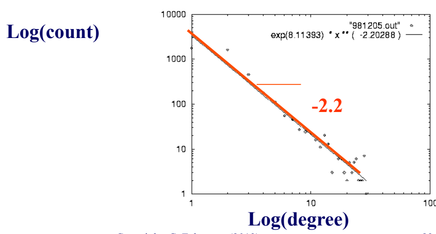


degree vs rank, for Internet domains (log-log) [sigcomm99]

CMU SCS

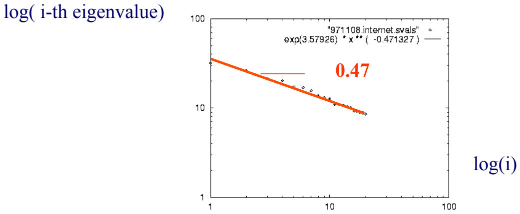
More power laws - internet

- pdf of degrees: (slope: 2.2)



CMU SCS

Even more power laws on the Internet



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

CMU SCS

Fractals & power laws:

appear in numerous settings:

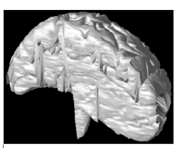
- **medical**
- geographical / geological
- social
- computer-system related

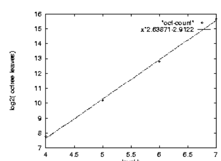
CMU SCS

More apps: Brain scans

- Oct-trees; brain-scans

Log(#octants)





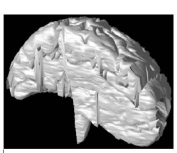
octree levels

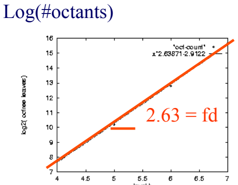
CMU SCS

More apps: Brain scans

- Oct-trees; brain-scans

Log(#octants)





octree levels

More apps: Medical images

[Burdett et al, SPIE ‘93]:

- benign tumors: fd ~ 2.37
- malignant: fd ~ 2.56

More fractals:

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

Fractals & power laws:

appear in numerous settings:

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

CMU SCS

More fractals:

- Coastlines: 1.2-1.58

CMU SCS

CMU SCS

More fractals:

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

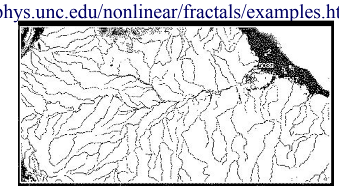
[ems.gphys.unc.edu/nonlinear/fractals/examples.html]

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]

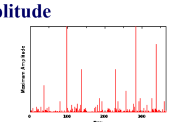


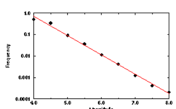
15-826 Copyright: C. Faloutsos (2013) 94

CMU SCS

More power laws

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





15-826 Copyright: C. Faloutsos (2013) 95

CMU SCS

Fractals & power laws:

appear in numerous settings:

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

15-826 Copyright: C. Faloutsos (2013) 96

CMU SCS

More fractals:

stock prices (LYCOS) - random walks: 1.5

1 year

2 years

CMU SCS

Even more power laws:

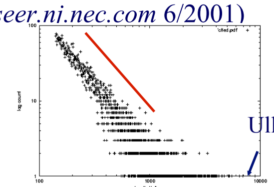
- Income distribution (Pareto's law)
- size of firms
- publication counts (Lotka's law)

CMU SCS

Even more power laws:

library science (Lotka's law of publication count); and citation counts: (*citeseer.ni.nec.com* 6/2001)

log(count)



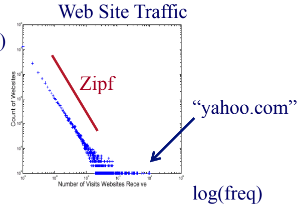
Ullman

log(#citations)

CMU SCS

Even more power laws:

- web hit counts [w/ A. Montgomery]



15-826 Copyright: C. Faloutsos (2013) 100

CMU SCS

Fractals & power laws:

appear in numerous settings:

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

15-826 Copyright: C. Faloutsos (2013) 101

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]

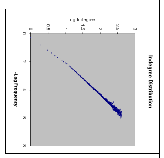
15-826 Copyright: C. Faloutsos (2013) 102

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]

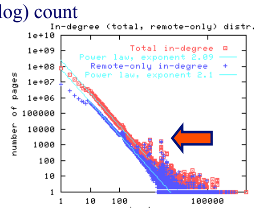
15-826 Copyright: C. Faloutsos (2013) 103

CMU SCS

"Foiled by power law"

- [Broder+, WWW'00]

(log) count



(log) in-degree

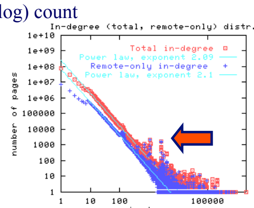
15-826 Copyright: C. Faloutsos (2013) 104

CMU SCS

"Foiled by power law"

- [Broder+, WWW'00]

(log) count



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

(log) in-degree

15-826 Copyright: C. Faloutsos (2013) 105

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 (2013) 106

CMU SCS

Even more power laws:

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

15-826 Copyright: C. Faloutsos (2013) 107

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 (2013) 108

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]
- ...

Settings for fractals:

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

Settings for fractals:

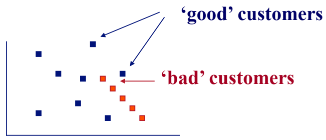
Points; areas, eg:

- 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

CMU SCS

Settings for fractals:

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



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

CMU SCS

Multi-Fractals

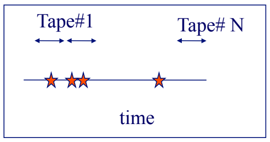
Setting: points or objects, w/ some value, eg:

- 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

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?*



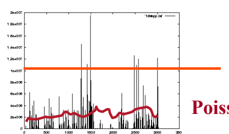
15-826 Copyright: C. Faloutsos (2013) 115

CMU SCS

Use of multifractals

- how often do we exceed the threshold?

#bytes



time

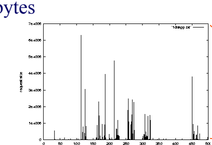
15-826 Copyright: C. Faloutsos (2013) 116

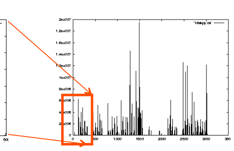
CMU SCS

Use of multifractals cont'd

- Extrapolations for/from samples

#bytes





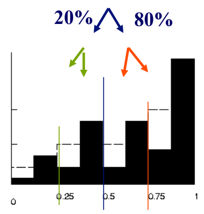
time

15-826 Copyright: C. Faloutsos (2013) 117

CMU SCS

Use of multifractals cont'd

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



15-826 Copyright: C. Faloutsos (2013) 118

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 (2013) 119

CMU SCS

Conclusions

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

15-826 Copyright: C. Faloutsos (2013) 120

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)

CMU SCS

Conclusions - cont'd

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

Bad news:
lead to skewed distributions
(no Gaussian, Poisson,
uniformity, X independence,
mean, variance)

Good news: 

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

CMU SCS

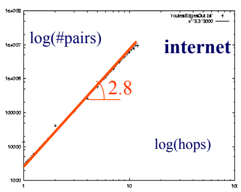
Conclusions

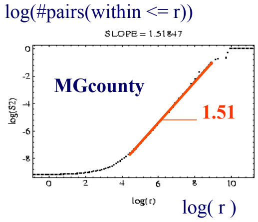
- **tool#1: (for points) 'correlation integral'**: (#pairs within <= r) vs (distance r)
- **tool#2: (for categorical values) rank-frequency** plot (a'la Zipf)
- **tool#3: (for numerical values) CCDF:** Complementary cumulative distr. function (#of elements with value >= a)

CMU SCS

Practitioner’s guide:

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



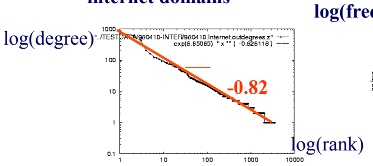


CMU SCS

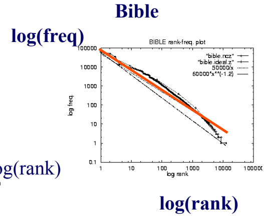
Practitioner’s guide:

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

internet domains



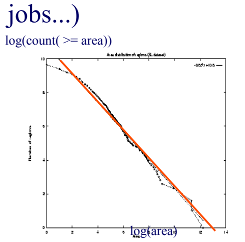
Bible



CMU SCS

Practitioner’s guide:

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



scandinavian lakes

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

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

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.

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

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.

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

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

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

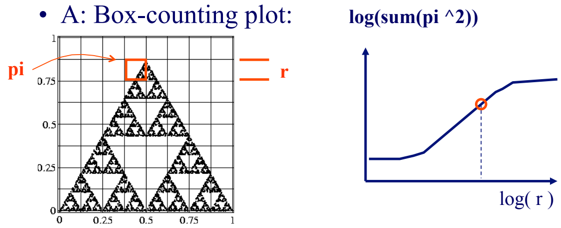
Appendix - Gory details

- 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

CMU SCS

Fast estimation of fd(s):

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



CMU SCS

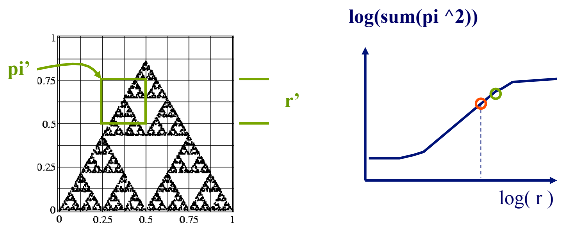
Definitions

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

CMU SCS

Fast estimation of fd(s):

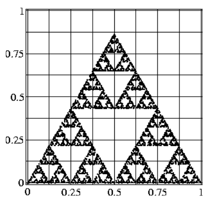
- compute sum(pi^2) for another grid side, r'

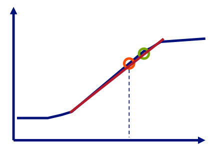


CMU SCS

Fast estimation of fd(s):

- etc; if the resulting plot has a linear part, its slope is the correlation fractal dimension D2





CMU SCS

Definitions (cont'd)

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

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

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

CMU SCS

Hausdorff or box-counting fd:

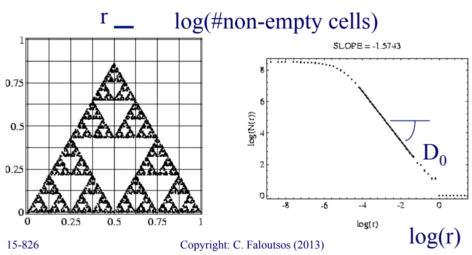
- Box counting plot: Log(N (r)) vs Log (r)
- r: grid side
- N (r): count of non-empty cells
- (Hausdorff) fractal dimension D0:

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

CMU SCS

Definitions (cont'd)

- Hausdorff fd:



15-826 Copyright: C. Faloutsos (2013) 142

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 (2013) 143

CMU SCS

Observations, cont'd

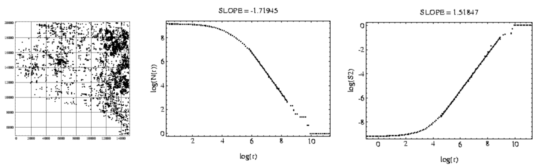
- in general, the Dq's take similar, but not identical, values.
- except for perfectly self-similar point-sets, where Dq=Dq' for any *q, q'*

15-826 Copyright: C. Faloutsos (2013) 144

CMU SCS

Examples:MG county

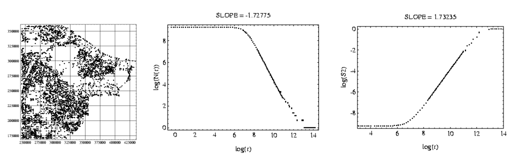
- Montgomery County of MD (road end-points)



CMU SCS

Examples:LB county

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



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)