

15-826: Multimedia (Databases) and Data Mining

Lecture #9: Fractals – examples & algo's

C. Faloutsos

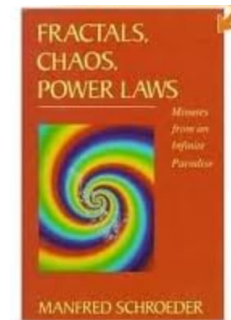
Must-read Material

- 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.

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





Outline

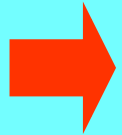
Goal: ‘Find **similar / interesting** things’

- 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





Road map

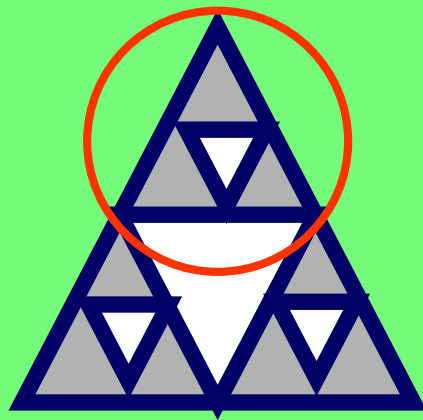
- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems
- ➔ • More tools, **drills**, and examples
- Discussion - putting fractals to work!
- Conclusions – practitioner's guide
- Appendix: gory details - boxcounting plots



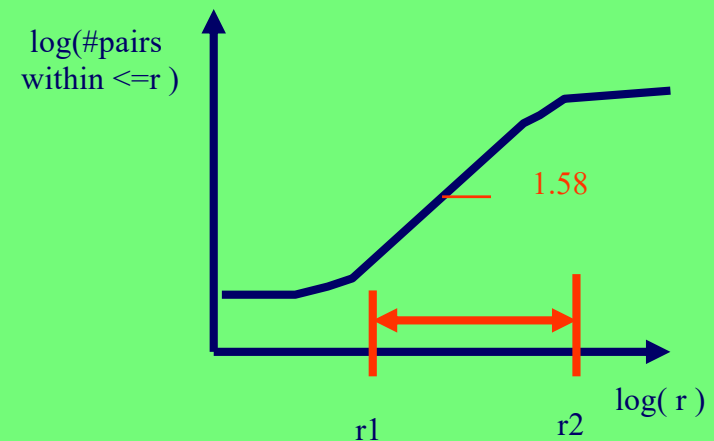
Definitions of f.d.

For mathematical fractal:

$$fd = \frac{\log(n)}{\log(f)}$$



For real set of points:
fd in the **range** (r1, r2):
Slope of corr. integral





Problem

- How to use fractals?



Conclusions

- How to use fractals?
- Tools: Correlation integral; CCDF plot



Road map

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems
- More tools, **drills**, and examples
 - ➔ – Mathematical fractals
 - Corr. integrals
- Discussion - putting fractals to work!





Drill # F1

Q: Give a (mathematical) fractal with $fd = 2$



Drill # F1

Q: Give a (mathematical) fractal with $fd = 2$

A: unit square; circle; surface of a cylinder



Drill # F2

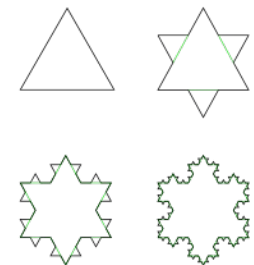
Q: fd of ‘Koch snowflake’?





Drill # F2

Q: fd of ‘Koch snowflake’?



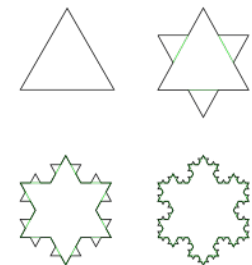
From wikipedia



Drill # F2

Q: fd of 'Koch snowflake'?

A: $\log(4)/\log(3) = 1.26$



From wikipedia

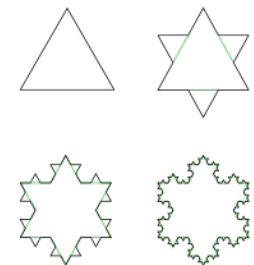


Drill # F2

Q: fd of 'Koch snowflake'?

A: $\log(4)/\log(3) = 1.26$

Q': does that make sense?



From wikipedia



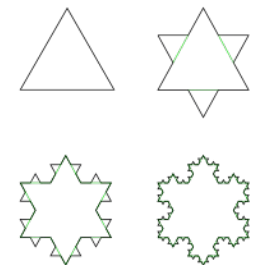
Drill # F2

Q: fd of 'Koch snowflake'?

A: $\log(4)/\log(3) = 1.26$

Q': does that make sense?

A': yes, a bit more complicated
than a line



From wikipedia



Drill # F3

Q: is it possible to have $fd < 1$?



Drill # F3

Q: is it possible to have $fd < 1$?

A: yes – eg., ‘Cantor dust’ (== leave middle third)





Drill # F3'

Q: fd?





Drill # F3'

Q: fd?

A: $\log(2)/\log(3) = 0.63$

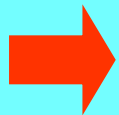
(Q': does it make sense?)





Road map

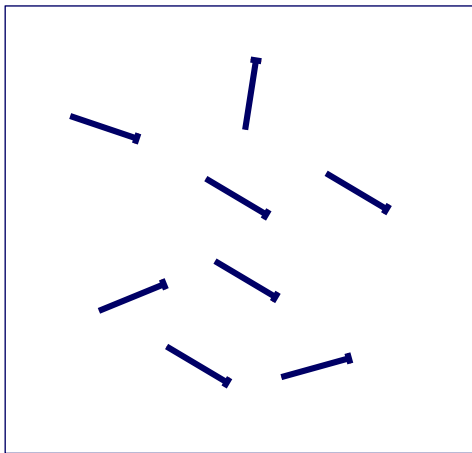
- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems
- More tools, **drills**, and examples
 - Mathematical fractals
 - Corr. integrals
- Discussion - putting fractals to work!





Drill # CI1

Q: points on short line segments, uniformly distributed in the 2-d space – how does the corr. integral look like?

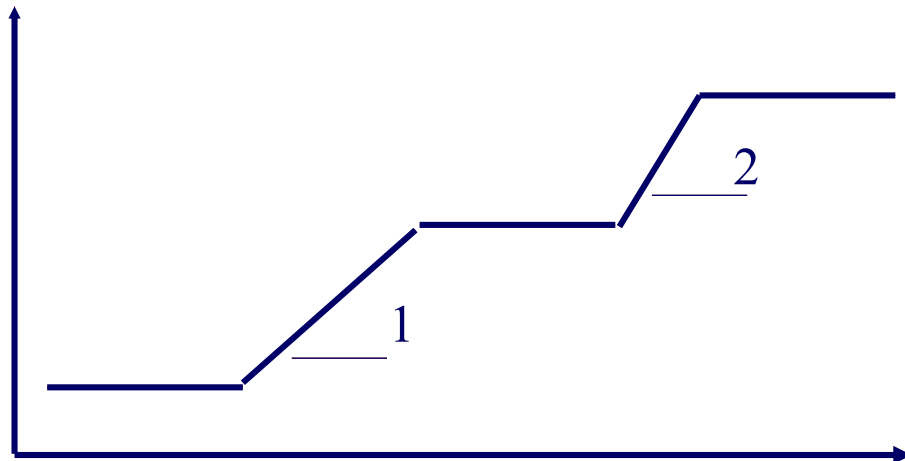
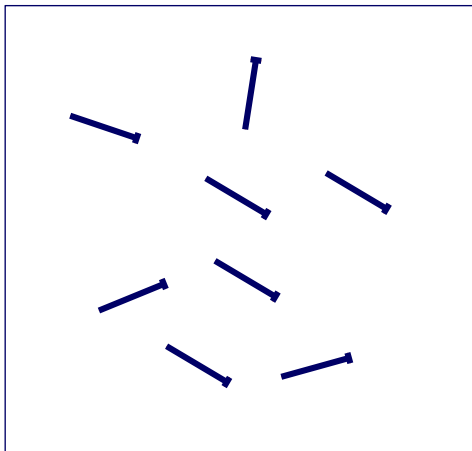




Drill # CI1

Q: points on short line segments, ... CI?

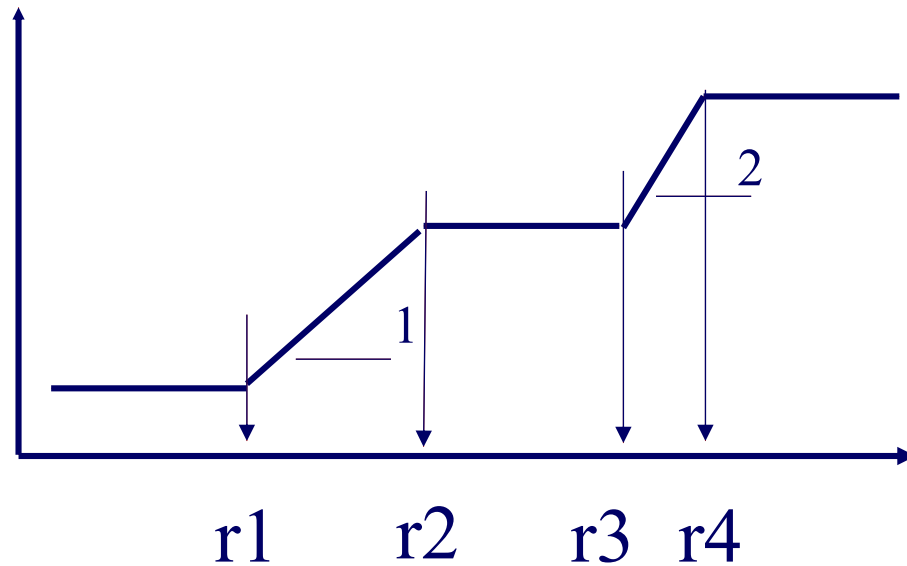
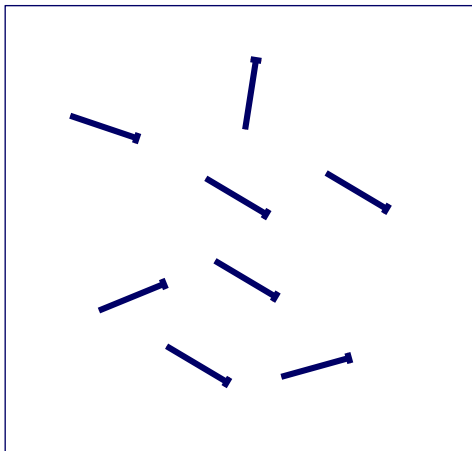
A: left-to-right: slopes $0 - 1 - 0 - 2 - 0$





Drill # CI1'

Q: $r1, r2, r3, r4 = ?$

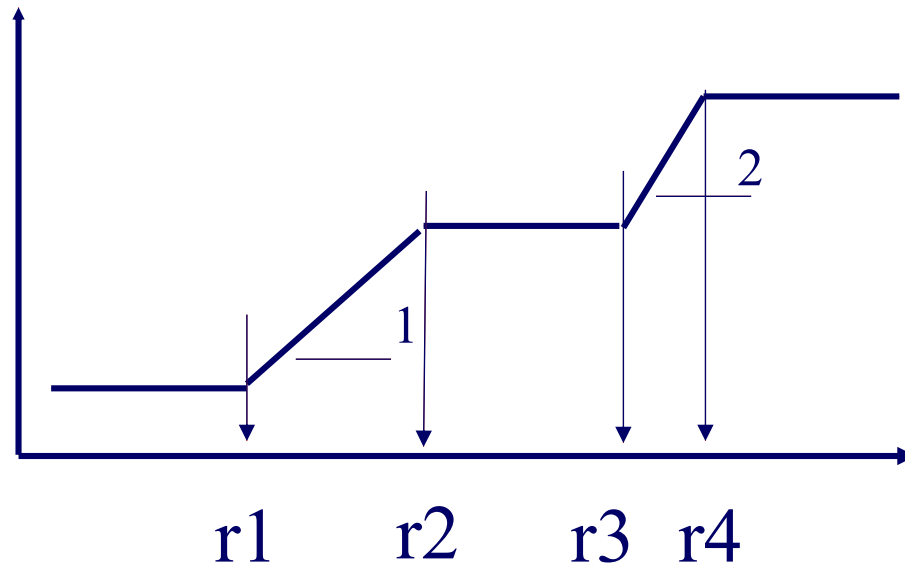
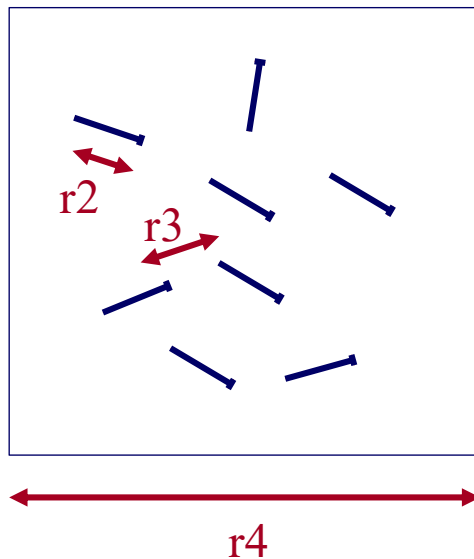




Drill # CI1'

Q: $r1, r2, r3, r4 = ?$

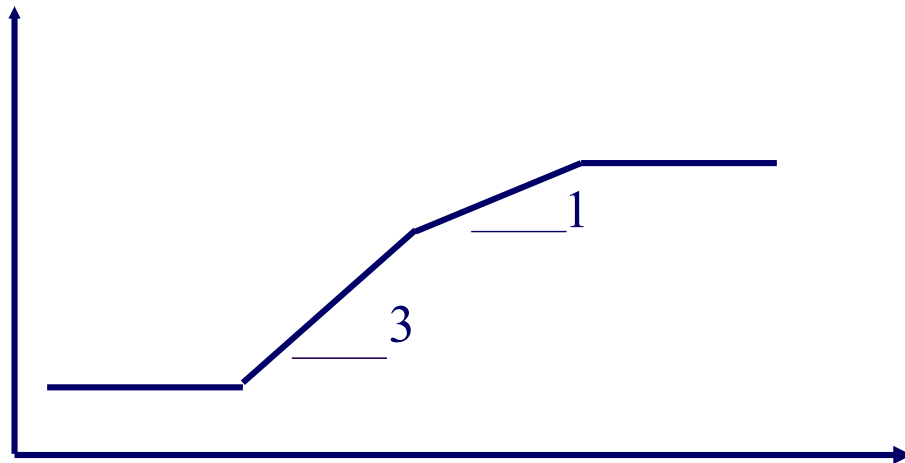
A: min distance; segment length; gap-length; diameter





Drill # CI2

Q: give 1M points in 3-d space, so that the CI has slopes $0 - 3 - 1 - 0$

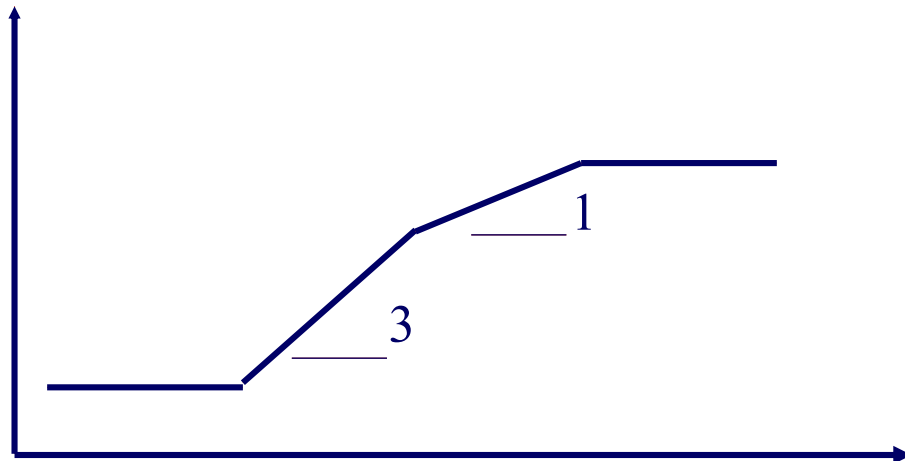




Drill # CI2

Q: give 1M points ...

A: small cubes, along a line, no gaps





Drill # CI3

Q: fd of molecules on a sheet of paper?

Q: ditto, after we crumble the sheet?



Drill # CI3

Q: fd of molecules on a sheet of paper?

A: 2

Q: ditto, after we crumble the sheet?

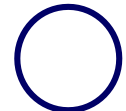
A: $2 < f < 3$



Drill # CI4

Q: guess the fd of each curve / surface:

- a) Piece of (straight) string
- b) Piece of crumbled string / knot
- c) Periphery of a circle
- d) A disk
- e) Surface of a cylinder
- f) Bark of tree





Drill # CI4

Q: guess the fd of each curve / surface:

a) Piece of (straight) string: **1**

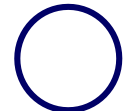
b) Piece of crumbled string / knot: **1-3**

c) Periphery of a circle: **1**

d) A disk: **2**

e) Surface of a cylinder: **2**

f) Bark of tree: **2+**





Drill # CI4

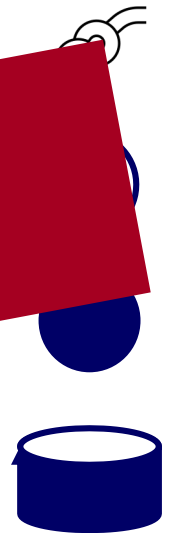
Q: guess the fd of each curve / surface:

- a) Piece of (straight) string: **1**
- b) Piece of crumbled string / knot: **1-3**
- c) Periphery of a circle: **1**

d) A disk:

**Smooth curves / surfaces: 1, 2 resp.
Otherwise, the rougher, the higher**

e) Bark of tree: **2+**





Road map

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

Fractals & power laws:

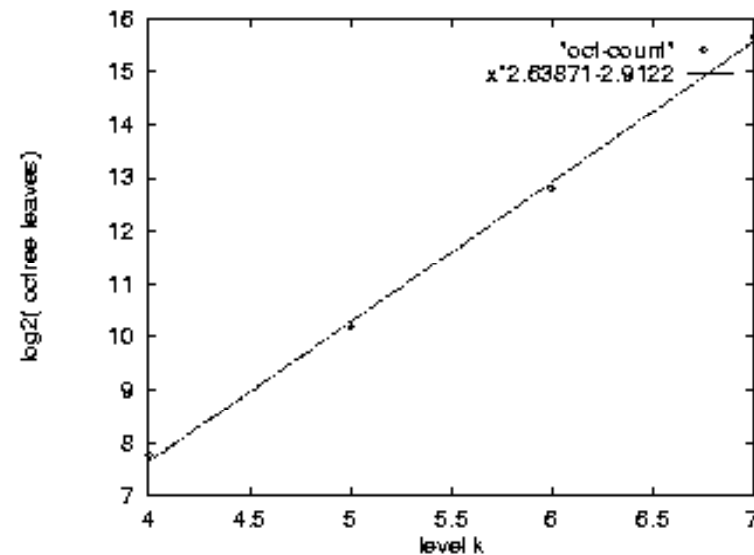
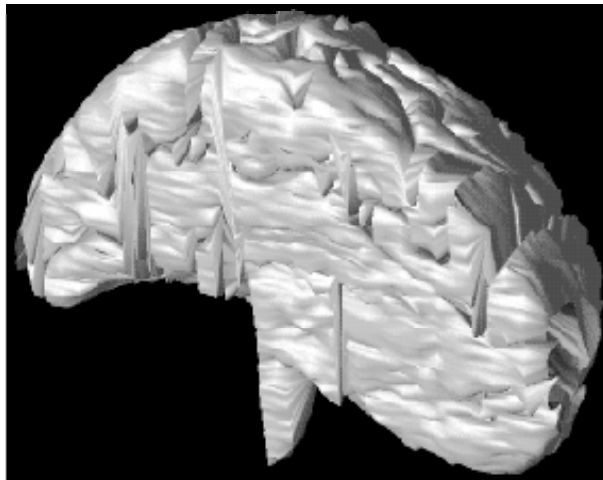
appear in numerous settings:

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

More apps: Brain scans

- Oct-trees; brain-scans

Log(#octants)

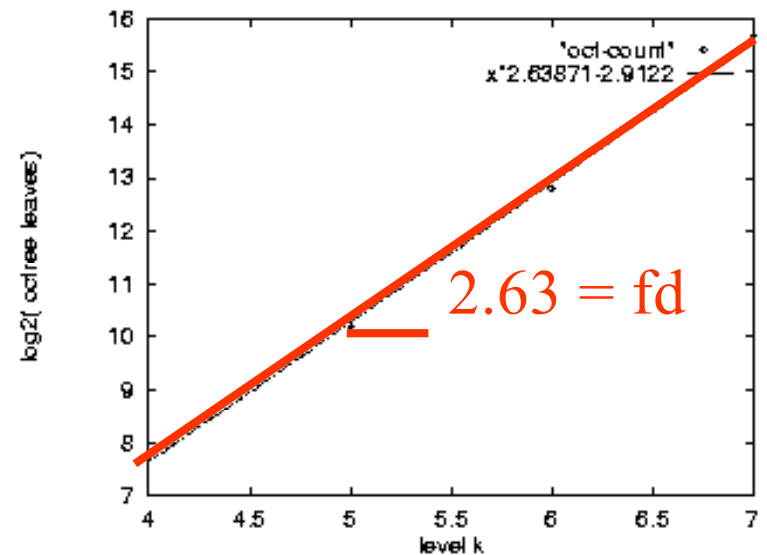
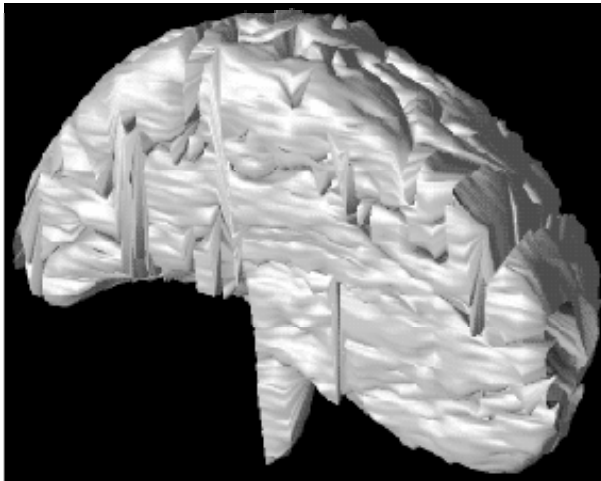


octree levels

More apps: Brain scans

- Oct-trees; brain-scans

Log(#octants)



octree levels

More apps: Medical images

[Burdett et al, SPIE '93]:

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

More fractals:

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



Fractals & power laws:

appear in numerous settings:

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

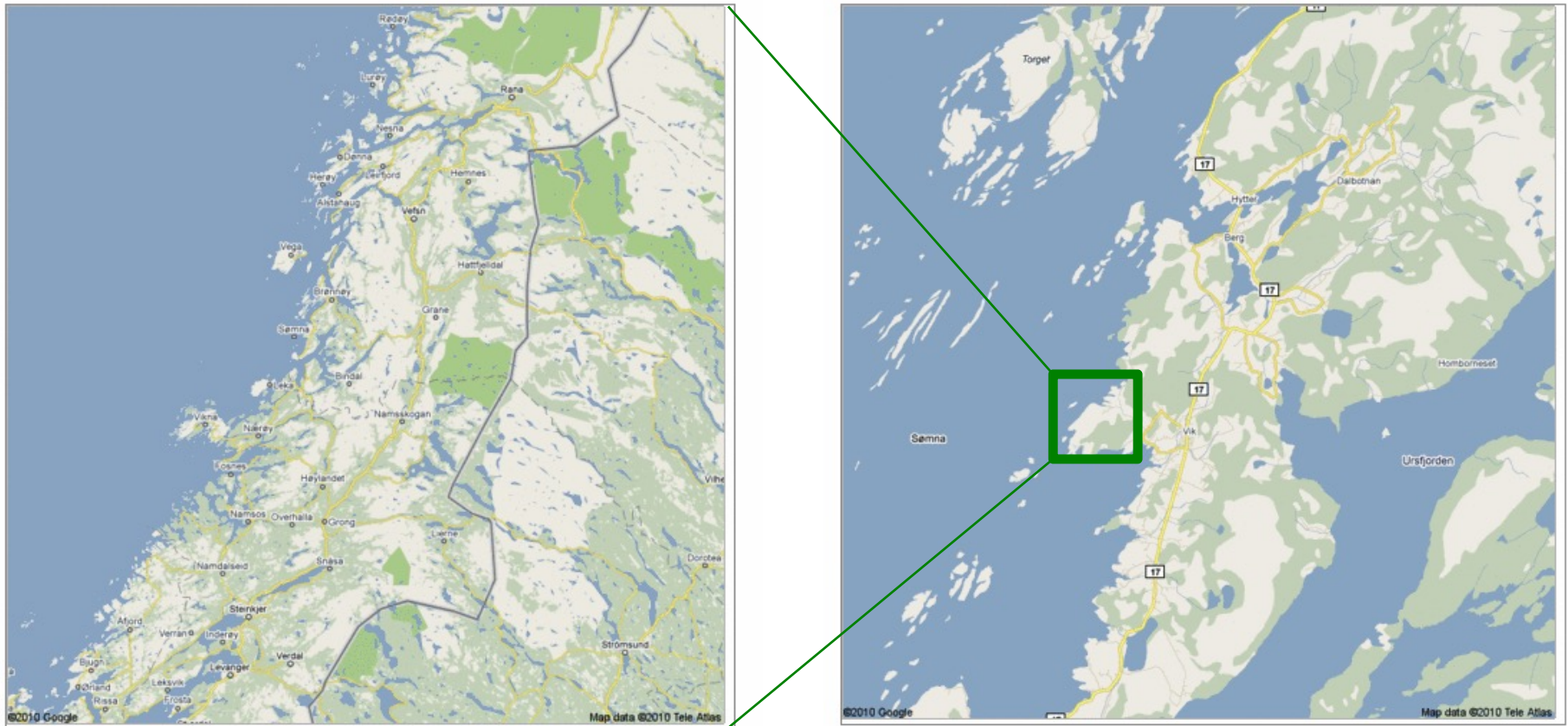
Real, self similar dataset



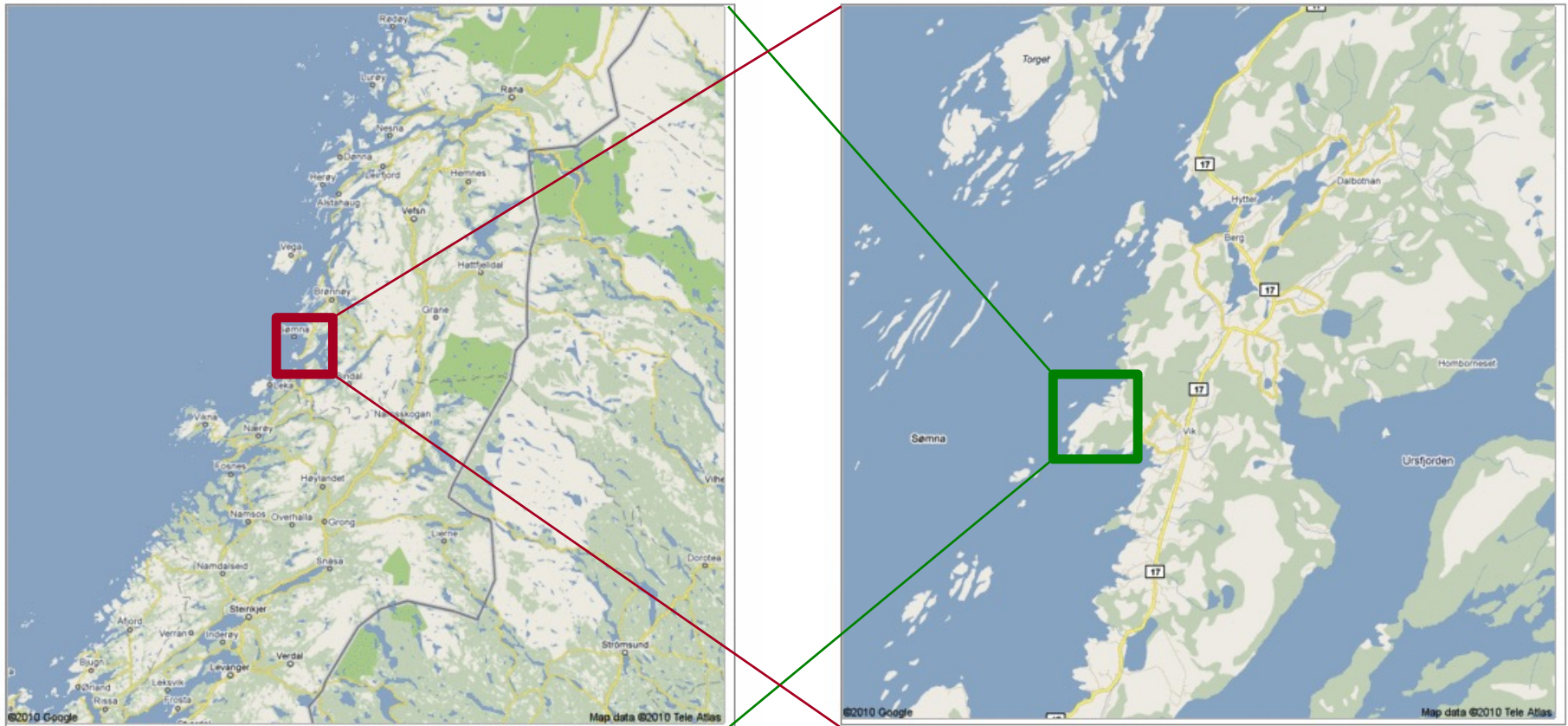
Real, self similar dataset



Real, self similar dataset



Real, self similar dataset



- the red is true
- origin: Norway
- but most other coastlines are ‘self-similar’, too!



More fractals:

- Coastlines: 1.2-1.58





More fractals:

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

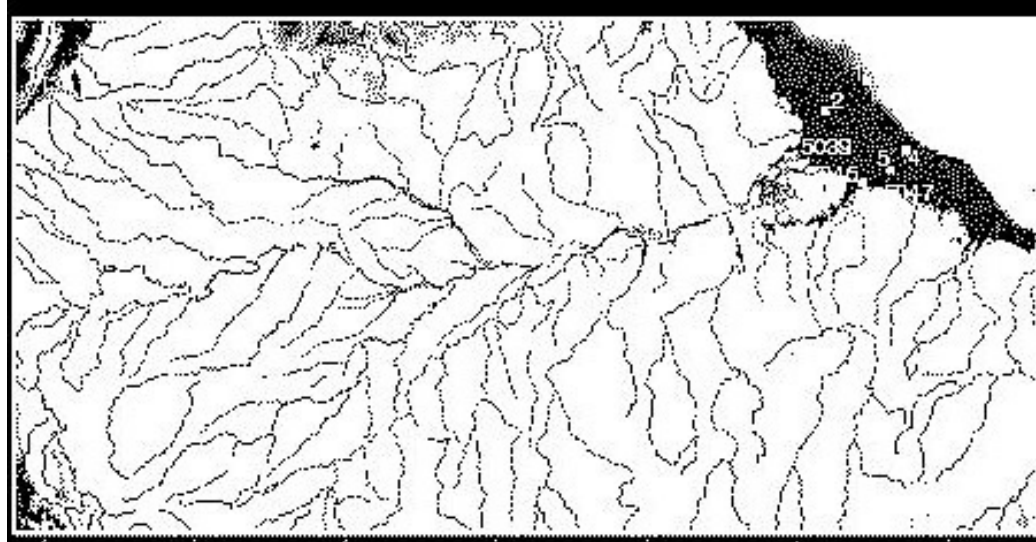
[ems.gphys.unc.edu/nonlinear/fractals/examples.html]



More fractals:

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

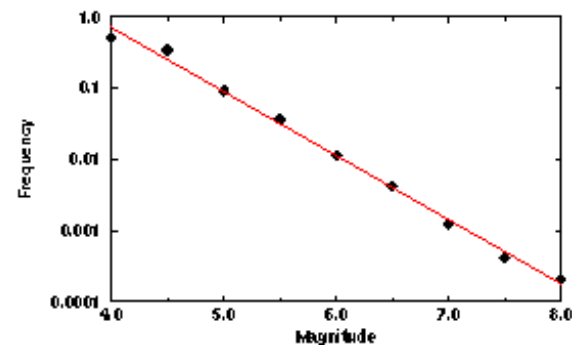
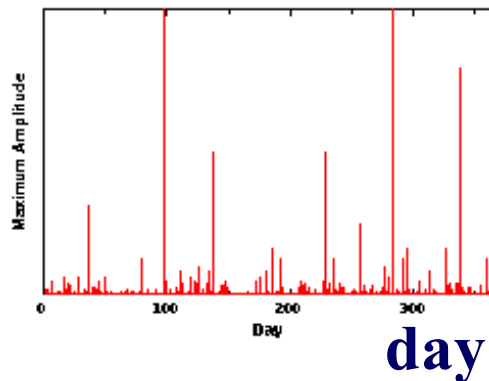
[ems.gphys.unc.edu/nonlinear/fractals/examples.html]



More power laws

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

amplitude



magnitude

Fractals & power laws:

appear in numerous settings:

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

More fractals:

stock prices (LYCOS) - random walks: 1.5

1 year



2 years



Even more power laws:

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

Fractals & power laws:

appear in numerous settings:

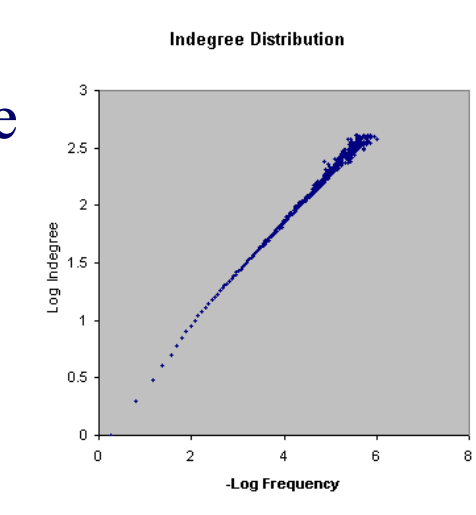
- medical
- geographical / geological
- social
- **computer-system related**

Power laws, cont' d

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

log indegree

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



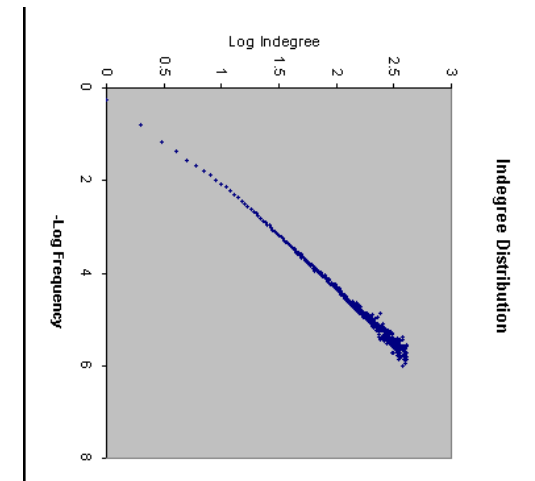
- log(freq)

Power laws, cont' d

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

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

$\log(\text{freq})$

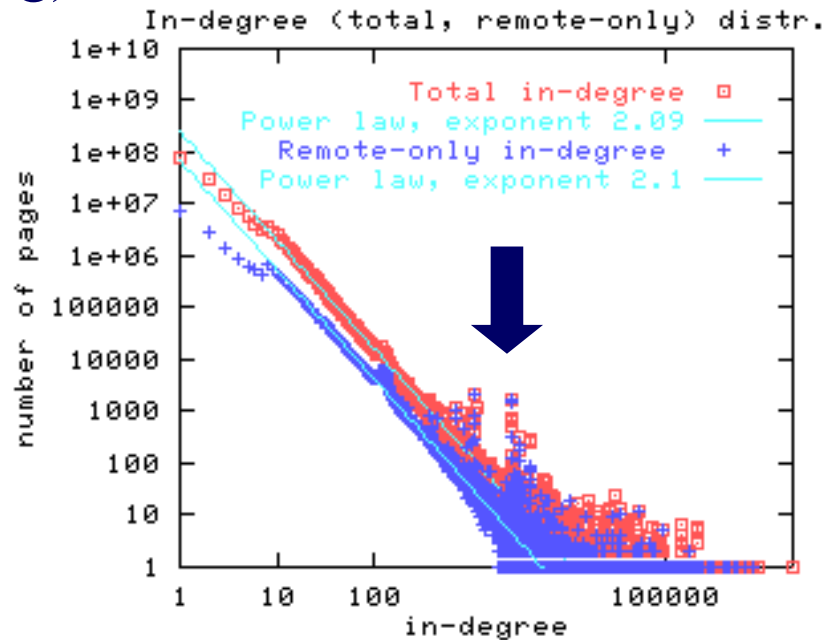


$\log \text{ indegree}$

“Foiled by power law”

- [Broder+, WWW' 00]

(log) count

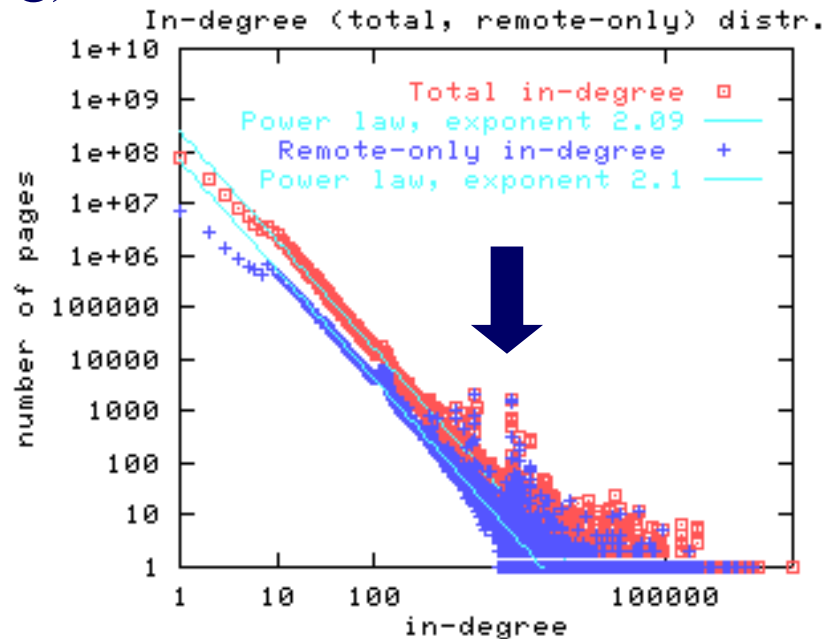


(log) in-degree

“Foiled by power law”

- [Broder+, WWW' 00]

(log) count



(log) in-degree

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

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]

Even more power laws:

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

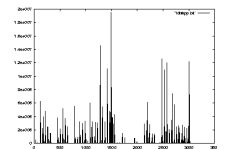
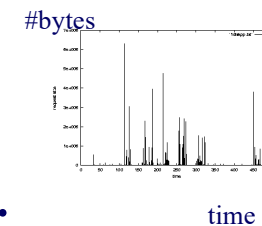
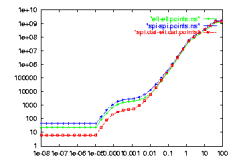
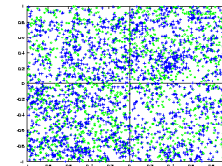


Road map

- 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

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]
 - ...



Conclusions

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

Conclusions

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

Conclusions - cont' d

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

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

Conclusions - cont' d

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

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

Good news:



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



Conclusions

- How to use fractals?
- Tools: Correlation integral; CCDF plot

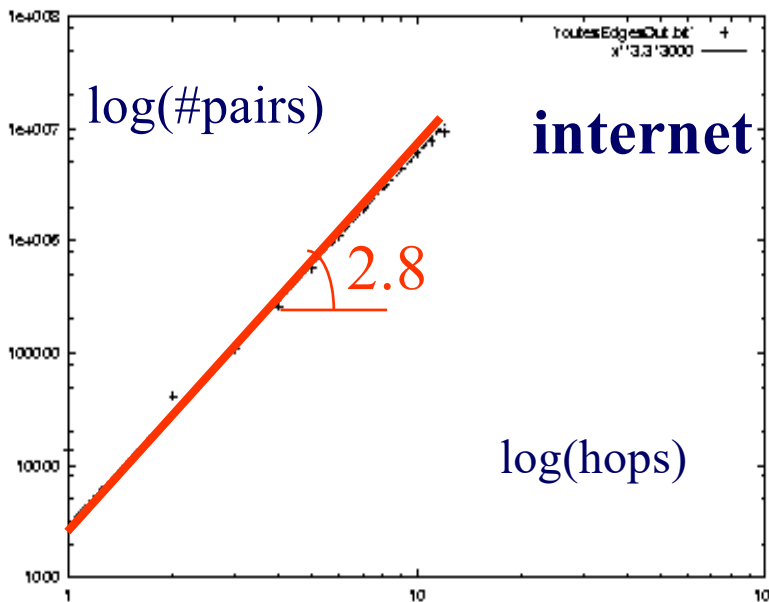


Conclusions - detailed

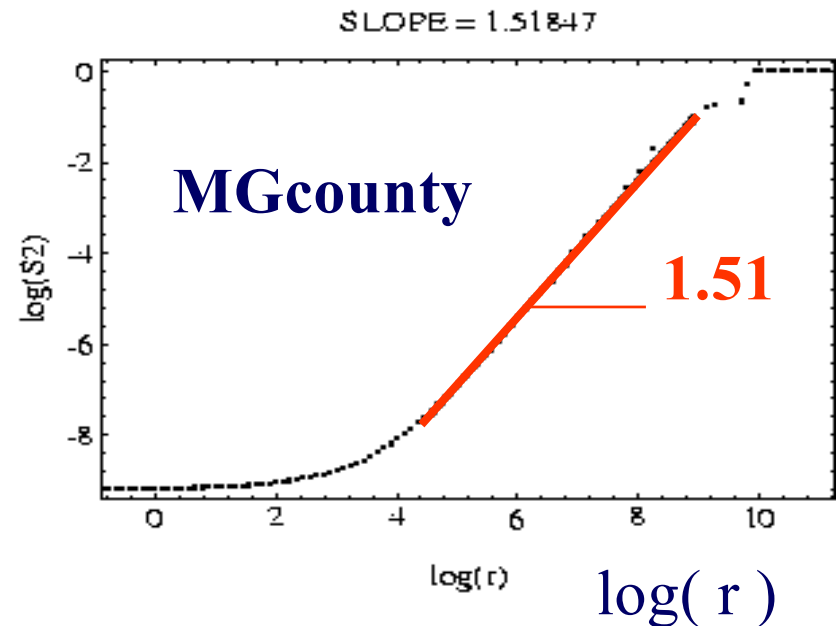
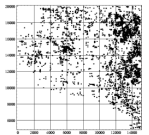
- **tool#1: (for points) ‘correlation integral’** : (#pairs within $\leq r$) vs (distance r)
- **tool#2: (for categorical values) rank-frequency plot** (a’ la Zipf)
- **tool#3: (for numerical values) CCDF:** Complementary cumulative distr. function (#of elements with value $\geq a$)

Practitioner's guide:

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



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

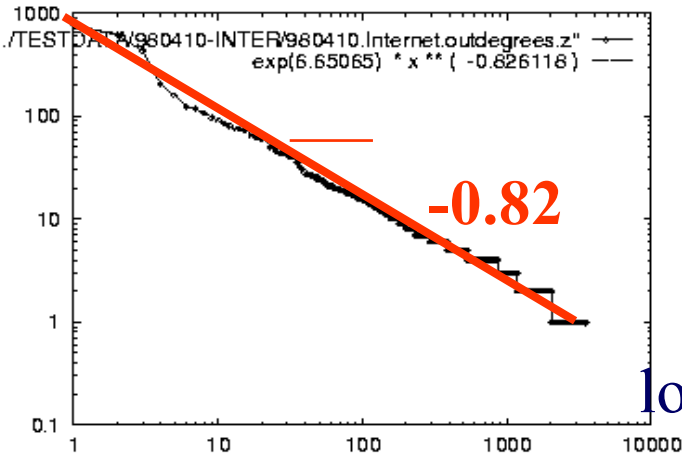


Practitioner's guide:

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

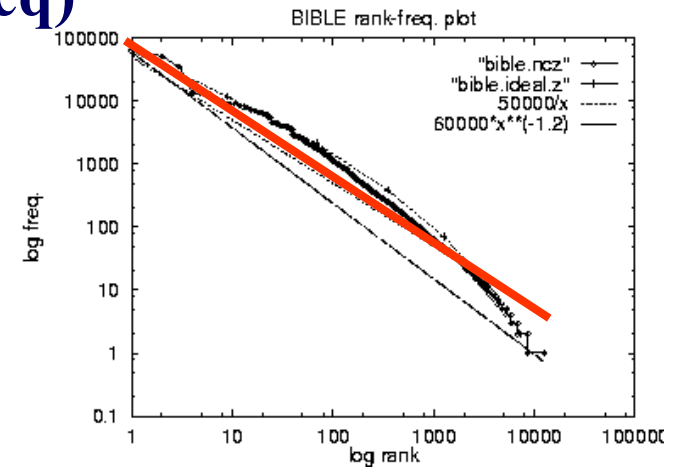
internet domains

log(degree)



log(freq)

Bible

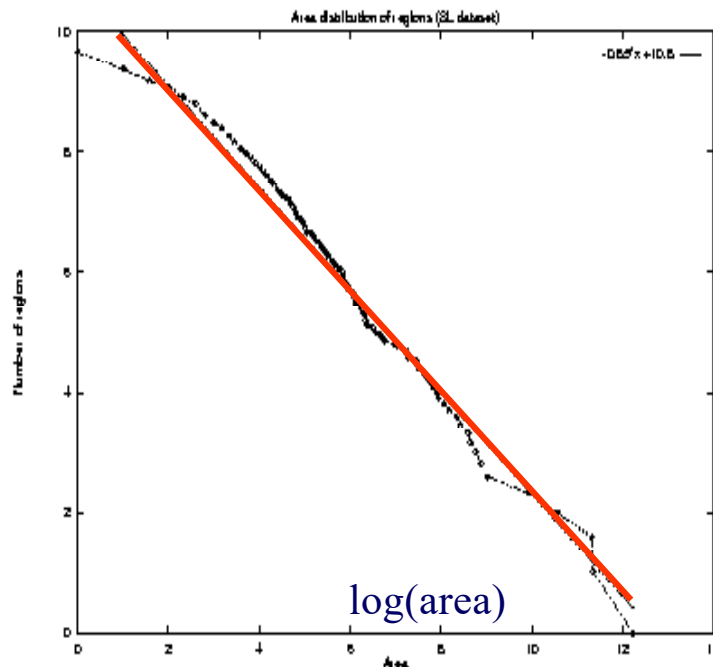


log(rank)

Practitioner's guide:

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

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



scandinavian lakes

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

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

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.

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

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.

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

Fast estimation of $fd(s)$:

- How, for the (correlation) fractal dimension?

Fast estimation of $fd(s)$:

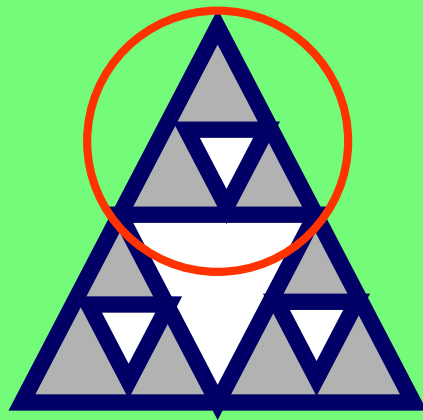
- How, for the (correlation) fractal dimension?
- A1: naïve, $O(N^2)$



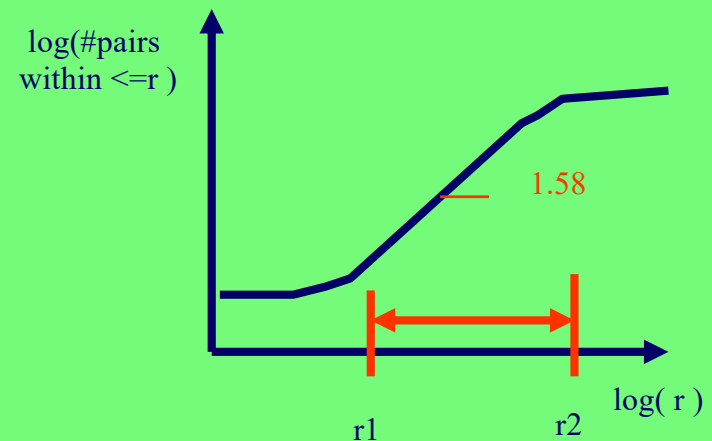
Definitions of f.d.

For mathematical fractal:

$$fd = \frac{\log(n)}{\log(f)}$$



For real set of points:
fd in the **range** (r1, r2):
Slope of corr. integral



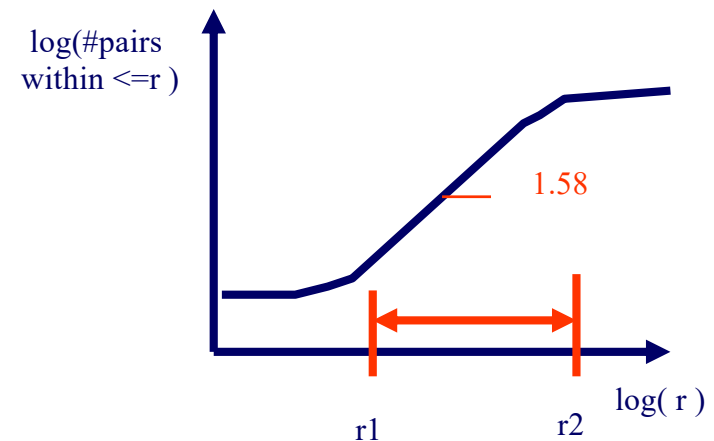


Definitions of f.d.



For real set of points:
fd in the **range** $(r1, r2)$:
Slope of corr. integral

CDF (distances) $\equiv$
[log-log scales]



Fast estimation of $fd(s)$:

CCDF of pairwise distances

- A1) Algo – naïve $O(N^2)$ – anything faster?

Fast estimation of $fd(s)$:

CCDF of pairwise distances

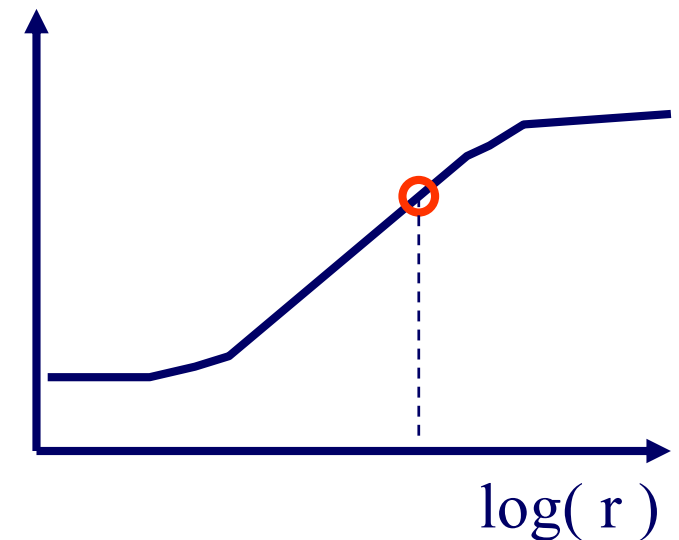
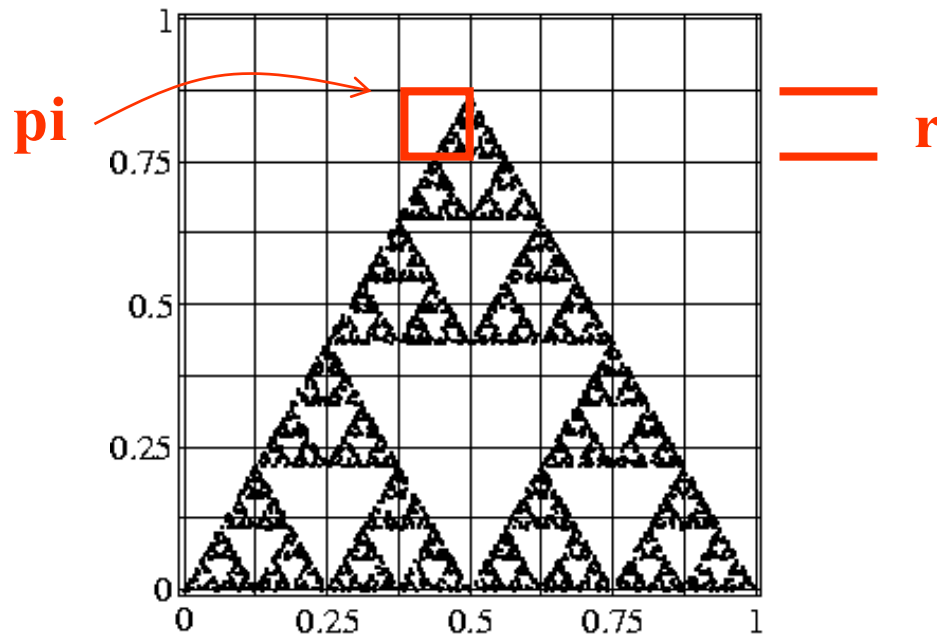
- A1) Algo – naïve $O(N^2)$
- A2) Smarter (sampling - $O(s * N)$)
- A3) Even smarter: next

Faster estimation of $fd(s)$:

- How, for the (correlation) fractal dimension?

Faster estimation of $fd(s)$:

- How, for the (correlation) fractal dimension?
- A3: Box-counting plot: $\log(\sum(\pi^2))$

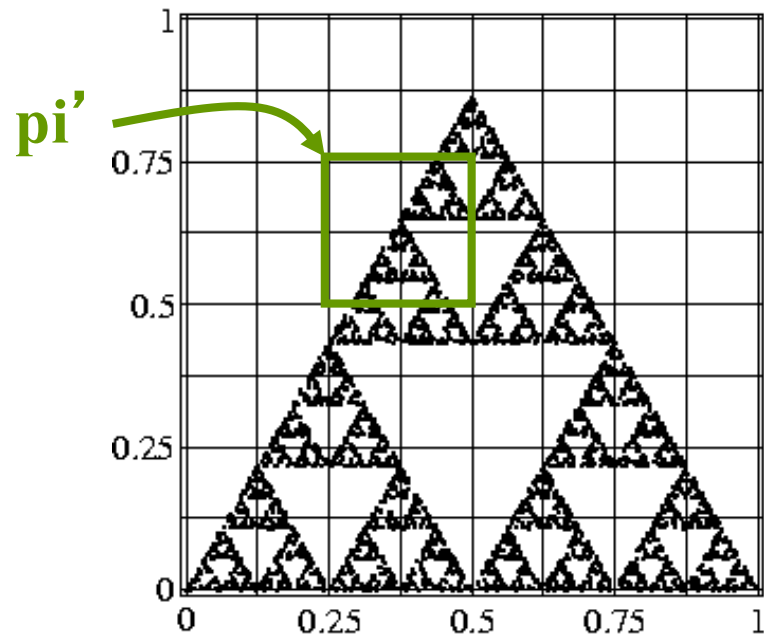


Definitions

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

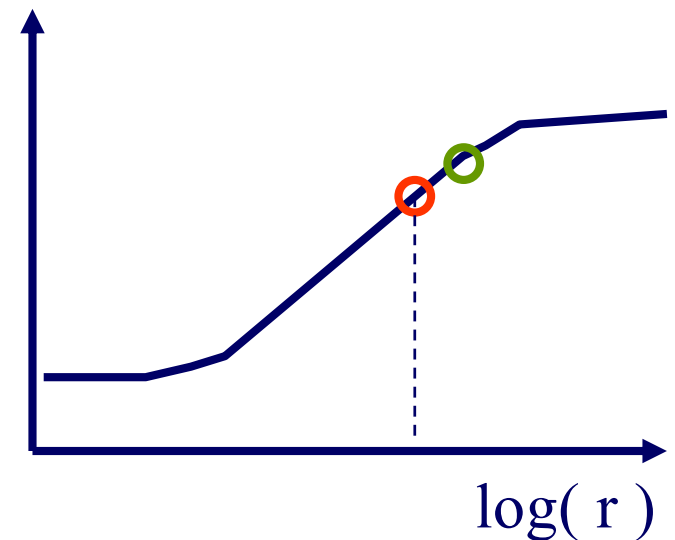
Fast estimation of $fd(s)$:

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



r'

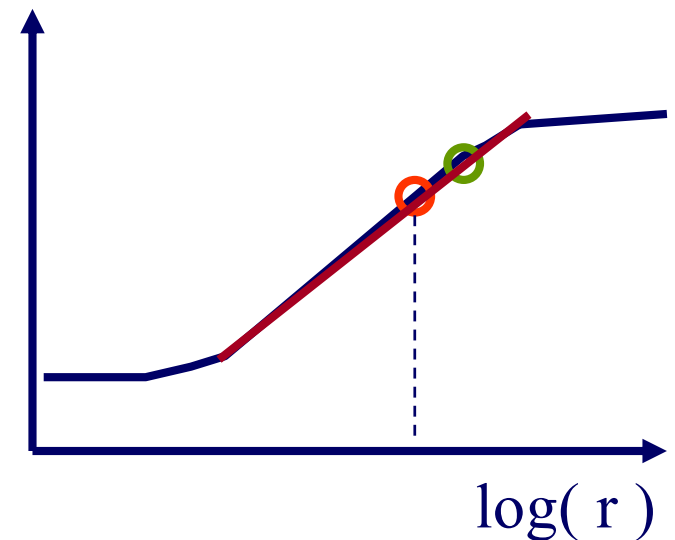
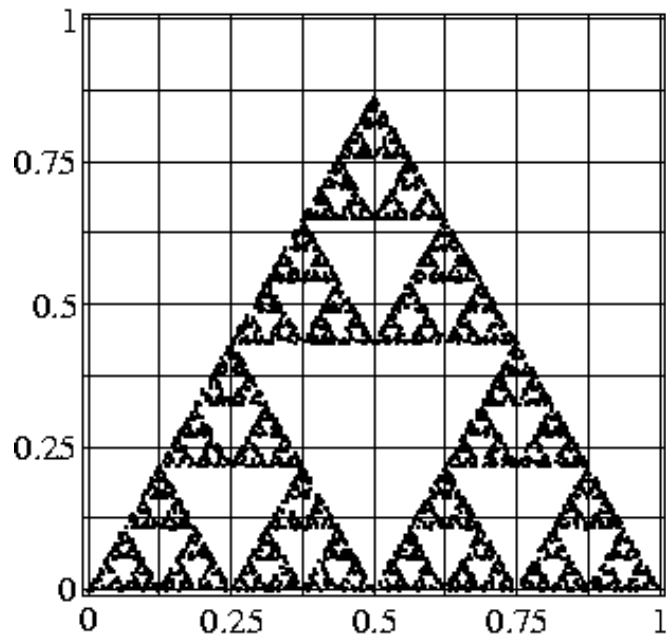
$\log(\sum(\pi^2))$



Fast estimation of $fd(s)$:

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

$\log(\sum(p_i^2))$



Definitions (cont' d)

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

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

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

Hausdorff or box-counting fd:

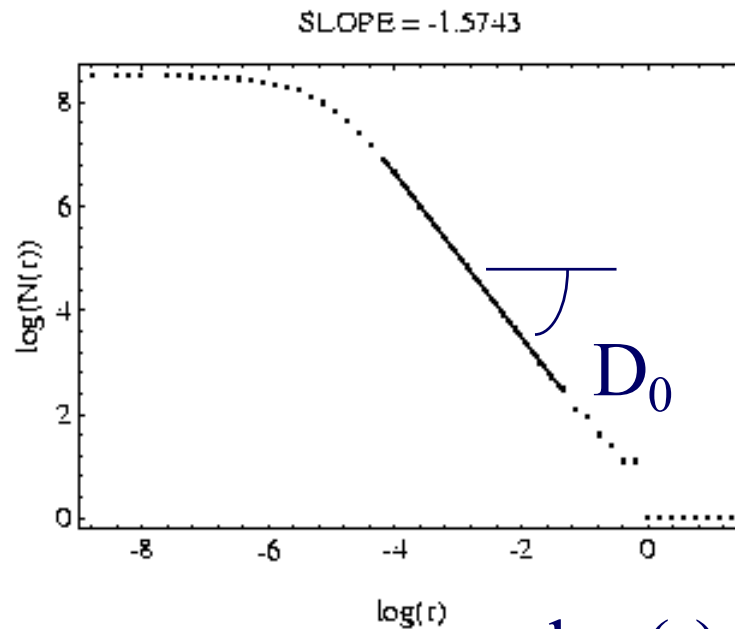
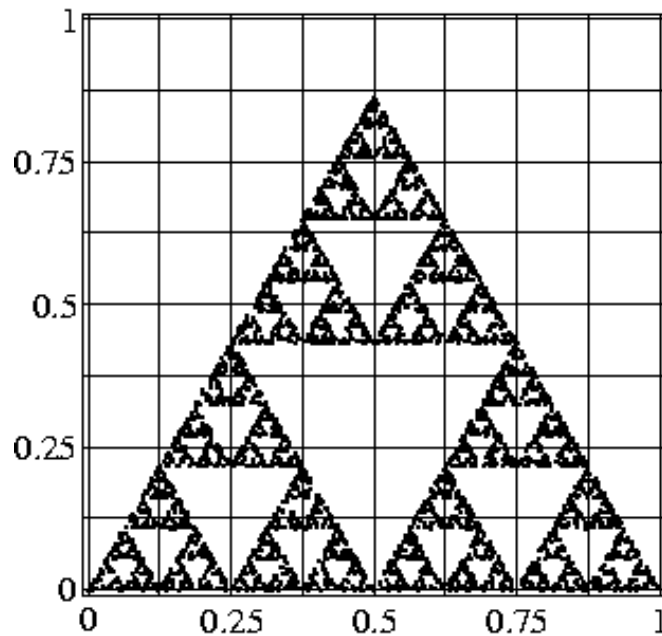
- Box counting plot: $\text{Log}(N(r))$ vs $\text{Log}(r)$
- r : grid side
- $N(r)$: count of non-empty cells
- (Hausdorff) fractal dimension D_0 :

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

Definitions (cont' d)

- Hausdorff fd:

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



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

Observations, cont' d

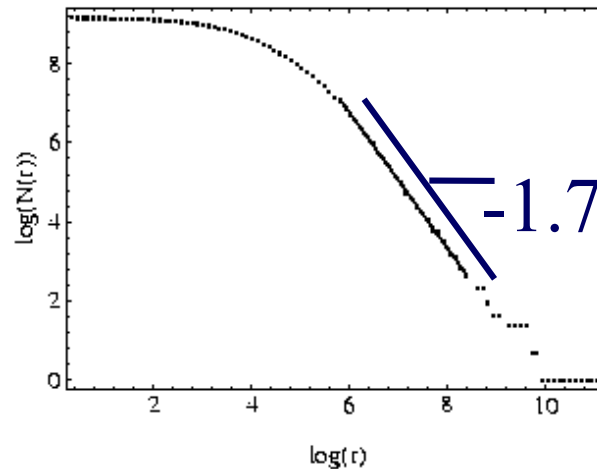
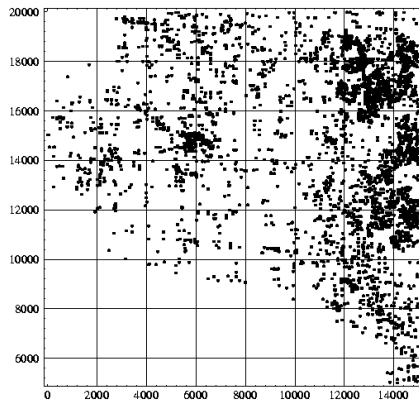
- in general, the Dq 's take similar, but not identical, values.
- except for perfectly self-similar point-sets, where $Dq = Dq'$ for any q, q'

Examples:MG county

- Montgomery County of MD (road end-points)

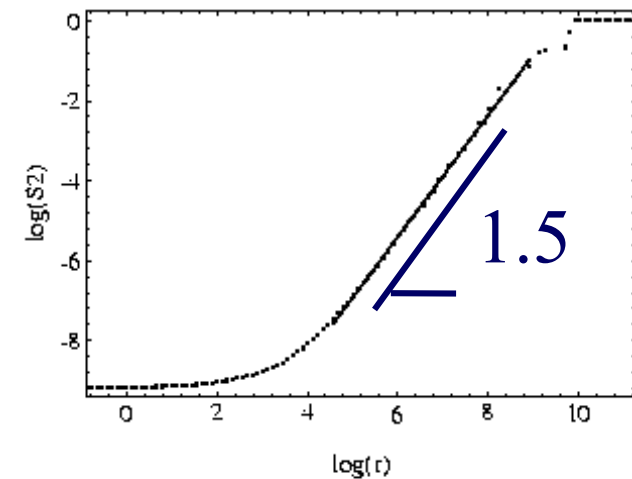
$q=0$

SLOPE = -1.71945



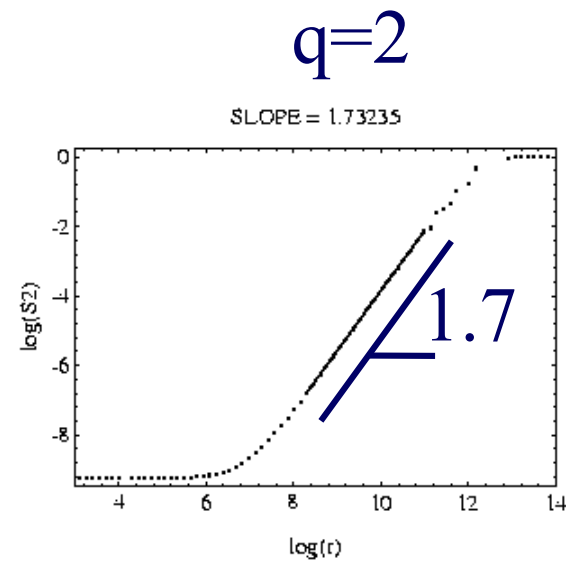
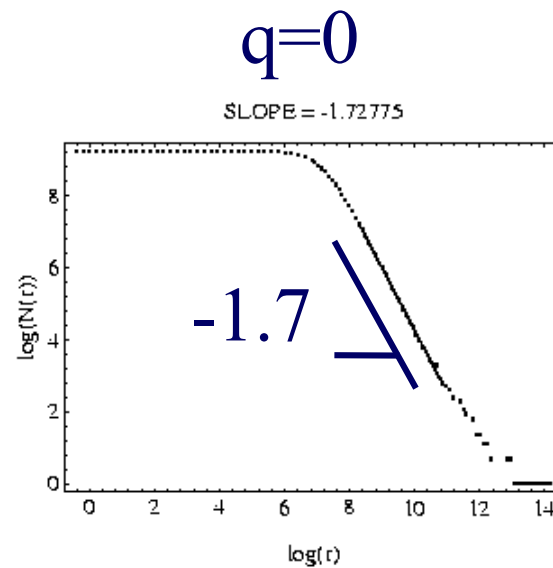
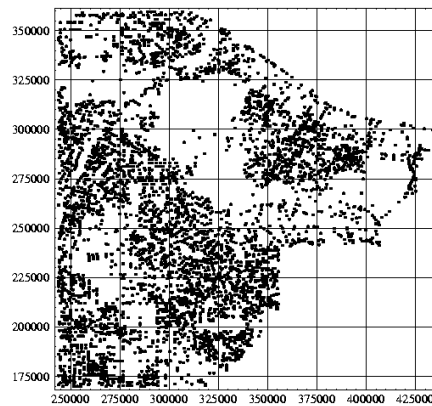
$q=2$

SLOPE = 1.51847



Examples:LB county

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



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)



Conclusions

- How to use fractals?
- Tools: Correlation integral; CCDF plot ($\sim$ Zipf plot)
- Many fractal dimensions – ‘box-counting’ algo

