15-826: Multimedia Databases and Data Mining

Lecture #9: Fractals - introduction

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
  (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
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??

Problem #2 - spatial d.m.

Galaxies (Sloan Digital Sky Survey w/ B. Nichol)
- ‘spiral’ and ‘elliptical’ galaxies
- patterns?
- attraction/repulsion?
- how many ‘spi’ within r from an ‘ell’?

Problem #3: traffic

- disk trace (from HP - J. Wilkes); Web traffic - fit a model
- how many explosions to expect?
- queue length distr.?
Q: Then, how to generate such bursty traffic?
Common answer:

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

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 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: \( \log(3)/\log(2) = 1.58... \)

Dfn of fd:

**ONLY** for a perfectly self-similar point set:

\[
\frac{\log(n)}{\log(f)} = \frac{\log(3)}{\log(2)} = 1.58
\]

Intrinsic (‘fractal’) dimension

- Q: Fractal dimension of a line?
- A: 1 (\( = \log(2)/\log(2) \))
Intrinsic (‘fractal’) dimension

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

Q: fractal dimension of a plane?
- A: nn ( <= r ) ~ r^2
  \[ \text{fd} = \text{slope of } \log(nn) \text{ vs } \log(r) \]

Q: dfn for a given set of points?

\[
\begin{array}{|c|c|}
\hline
x & y \\
5 & 1 \\
4 & 2 \\
3 & 3 \\
2 & 4 \\
\hline
\end{array}
\]

Q: fd of a plane?
- A: nn ( <= r ) ~ r^2
Intrinsic (‘fractal’) dimension

• Algorithm, to estimate it?

Notice

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

  including ‘mirror’ pairs


Sierpinsky triangle

\[ \log(\#\text{pairs within } \leq r) \]
\[ \Rightarrow \text{‘correlation integral’} \]


Observations:

• Euclidean objects have **integer** fractal dimensions
  – point: 0
  – lines and smooth curves: 1
  – smooth surfaces: 2

• fractal dimension \( \Rightarrow \) roughness of the periphery
Important properties

• \( fd = \) embedding dimension \( \rightarrow \) uniform pointset
• a point set may have several \( fd \), depending on scale

2-d

1-d
Important properties

0-d

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

Problem #1: GIS points

Cross-roads of Montgomery county:
• any rules?
### Solution #1

A: self-similarity ->
- \( \leftrightarrow \) fractals
- \( \leftrightarrow \) scale-free
- \( \leftrightarrow \) power-laws
  \( y = x^a, F = C r^{-2} \)
- avg\#neighbors(\( \leq r \)) = \( r^D \log(r) \)

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

\[ \log(r) \]

\[ 1.51 \]

### Solution #1

A: self-similarity

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

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

\[ \log(r) \]

\[ 1.51 \]

### Examples: MG county

- Montgomery County of MD (road endpoints)
**Examples: LB county**

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

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

Galaxies (‘BOPS’ plot - [sigmod2000])

- 1.8 slope
- plateau!
- repulsion!
Spatial d.m.

- 1.8 slope
- plateau!
- repulsion!

Heuristic on choosing # of clusters

ell-ell
spi-spi
spi-ell

log(r)

log(#pairs within <=r )

r1
r2

"ell-ell.points.ns"
"spi-spi.points.ns"
"spi.dat-ell.dat.points"
Spatial d.m.

- 1.8 slope
- plateau!
- repulsion!!
- duplicates

Solution #3: traffic

- disk traces: self-similar:

Solution #3: traffic

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

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

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-comm-7.tar.Z
Web traffic

- [Crovella Bestavros, SIGMETRICS’96]

```
1000 sec; 100sec
10sec; 1sec
```

Tape accesses

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

```
Tape#1  Tape# N
\[ \text{\# tapes retrieved} \]
\[ \text{\# qual. records} \]
```

Tape accesses

```
50-50 = Poisson
```

```
\[ \text{50-50 = Poisson} \]
\[ \text{real} \]
```

```
\[ \text{# tapes retrieved} \]
\[ \text{# qual. records} \]
```
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

A counter-intuitive example

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

A: 1!!

A counter-intuitive example

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

Rank exponent $R$

- Power law in the degree distribution
  [SIGCOMM99]
- internet domains

More tools

- Zipf’s law
- Korcak’s law / “fat fractals”
A famous power law: Zipf’s law

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

freq.

A famous power law: Zipf’s law

log(freq)

- Bible - rank vs frequency (log-log)

log(rank)

A famous power law: Zipf’s law

log(freq)

- Bible - rank vs frequency (log-log)

- similarly, in many other languages; for customers and sales volume; city populations etc etc
A famous power law: Zipf’s law

- Zipf distr:
  \[ \text{freq} = \frac{1}{\text{rank}} \]
- Generalized Zipf:
  \[ \text{freq} = \frac{1}{(\text{rank})^a} \]

Olympic medals (Sidney):

\[ \log(\text{#medals}) \]

\[ y = 0.0045x^{-2.96} \quad R^2 = 0.83 \]

Olympic medals (Sidney’00, Athens’04):

\[ \log(\text{#medals}) \]

\[ \log(\text{rank}) \]
TELCO data

count of customers

# of service units

‘best customer’

Count-frequency plot of real and synthetic data

SALES data – store#96

count of products

# units sold

“aspirin”

Count-frequency plot for store no. 96.

More power laws: areas – Korcak’s law

Scandinavian lakes
Any pattern?
More power laws: areas – Korcak’s law

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

More power laws: Korcak

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

(Korcak’s law: Aegean islands)
Korcak’s law & “fat fractals”

How to generate such regions?

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

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)
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)

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

Other applications: Internet

- How does the internet look like?
Other applications: Internet

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

(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)

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
More power laws on the Internet

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

-0.82

More power laws - internet

• pdf of degrees: (slope: 2.2)

Even more power laws on the Internet

Scree plot for Internet domains (log-log) [sigcomm99]
Fractals & power laws:

appear in numerous settings:

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

More apps: Brain scans

- Oct-trees; brain-scans

![Oct-count](image)

\[ \text{log}_2(\text{octree leaves}) = \text{octree levels} \]

\[ 2.63 = f_d \]

More apps: Brain scans

- Oct-trees; brain-scans

![Oct-count](image)
More apps: Medical images

[Burdett et al, SPIE ‘93]:
• benign tumors: \( fd \approx 2.37 \)
• malignant: \( fd \approx 2.56 \)

More fractals:
• cardiovascular system: 3 (!)
• lungs: 2.9

Fractals & power laws:
appear in numerous settings:
• medical
• geographical / geological
• social
• computer-system related
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]

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)

Even more power laws:

Library science (Lotka’s law of publication count); and citation counts:

(citeseer.nj.nec.com 6/2001)
Even more power laws:

- web hit counts [w/ A. Montgomery]

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]
Power laws, cont’d

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

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

“Foiled by power law”

• [Broder+, WWW’00]

“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]
  • ...

Settings for fractals:

Points; areas (→ fat fractals), 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
Settings for fractals:

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

![Graph showing 'good' and 'bad' customers]

Some uses of fractals:

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

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
Use of multifractals:

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

Use of multifractals

- how often do we exceed the threshold?

Use of multifractals cont’d

- Extrapolations for/from samples
Use of multifractals cont’d

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

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

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)

Good news:
• ‘correlation integral’
  for separability
• rank/frequency plots
• 80-20 (multifractals)
• (Hurst exponent,
  strange attractors,
  renormalization theory,
  ++)
Practitioner’s guide:
- **tool#1**: #pairs vs distance, for a set of objects, with a distance function (slope = intrinsic dimensionality)

- **internet**
  - log(#pairs) vs log(hops)
  - slope = 2.8
- **MCounty**
  - log(#pairs(within <= r)) vs log(r)
  - slope = 1.51

Practitioner’s guide:
- **tool#2**: rank-frequency plot (for categorical attributes)

- **internet domains**
  - log(degree) vs log(rank)
  - slope = -0.82
- **Bible**
  - log(freq) vs log(rank)

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

- **scandinavian lakes**
  - log(count( >= area)) vs log(area)
  - slope = -0.85
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:

• Classic book on fractals:

References

• [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


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

References


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?
- A: Box-counting plot:

\[
\begin{align*}
\text{log}(r) & \quad \text{log}(\text{sum}(\pi^2)) \\
\text{pi} & \quad r
\end{align*}
\]

Definitions

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

Fast estimation of \( fd(s) \):

- compute \( \text{sum}(\pi^2) \) for another grid side, \( r' \)
Fast estimation of $fd(s)$:

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

\[
\frac{\log(\text{sum}(p_i^2))}{\log(r)}
\]

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

\[
D_0 = -\frac{\partial \log(N(r))}{\partial \log(r)}
\]
Definitions (cont’d)

- Hausdorff fd:
  \[ r \rightarrow \log(n) \log(\#\text{non-empty cells}) \]

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 \( D_q \)’s take similar, but not identical, values.
- except for perfectly self-similar point-sets, where \( D_q=D_{q'} \) for any \( q, q' \).
Examples: MG county

- Montgomery County of MD (road endpoints)

Examples: LB county

- Long Beach county of CA (road endpoints)

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)