15-826: Multimedia Databases and Data Mining

Fractals - introduction
C. Faloutsos

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
  – 2-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
  (stores and households ...)
  • 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:

... zero area; infinite length!

Definitions (cont’d)

• Paradox: Infinite perimeter; Zero area!
• ‘dimensionality’: between 1 and 2
• actually: \( \log(3)/\log(2) = 1.58... \)

Dfn of fd:

= \text{log}(n)/\text{log}(f) = \text{log}(3)/\text{log}(2) = 1.58

Intrinsic (‘fractal’) dimension

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

Intrinsic (‘fractal’) dimension

• Q: dfn for a given set of points?

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>
Intrinsic ('fractal') dimension

• Q: fractal dimension of a line?
• A: \( \text{nn} (\leq r) \sim r^d \)
  ('power law': \( y = x^{a} \))

• Q: fd of a plane?
• A: \( \text{nn} (\leq r) \sim r^d \)
  \( d = \text{slope of log(nn) vs log(r)} \)

Sierpinsky triangle

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

\[ \log(r) \]

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

== ‘correlation integral’

Observations:

• Euclidean objects have \textbf{integer} fractal dimensions
  – point: 0
  – lines and smooth curves: 1
  – smooth surfaces: 2
• fractal dimension \( \rightarrow \) roughness of the periphery

Important properties

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

Important properties

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

2-d
Important properties

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

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 ($y=x^a$, $F=C \cdot r^{-b}$)
- $\text{avg}\#\text{neighbors}(\leq r) = r^D$

Solution #1

A: self-similarity
- $\text{avg}\#\text{neighbors}(\leq r) \sim r^{1.51}$
Examples: MG county

- Montgomery County of MD (road end-points)

Examples: LB county

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

Solution#2: spatial d.m.

Galaxies ('BOPS' plot - [sigmod2000])

Spatial d.m.

Heuristic on choosing # of clusters
Spatial d.m.

- 1.8 slope
- plateau!
- repulsion!

Solution #3: traffic

- disk traces: self-similar:
  
  ![Graph showing self-similar disk traces]

80-20 / multifractals

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

Spatial d.m.

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

Solution #3: traffic

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

80-20 / multifractals
More on 80/20: PQRS

- Part of ‘self-* storage’ project

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?)

Web traffic

- [Crovella Bestavros, SIGMETRICS’96]

Tape accesses

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

internet domains

att.com

log(degree) vs log(rank)

ibm.com

-0.82
More tools

- Zipf’s law
- Korčak’s law / “fat fractals”

A famous power law: Zipf’s law

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

A famous power law: Zipf’s law

- 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:
  freq = 1/ rank
- generalized Zipf:
  freq = 1 / (rank)^a

Olympic medals (Sidney):

y = -0.9676x + 2.3054
R^2 = 0.9458

0 0.5 1 1.5 2
0 1 2

Series 1
Linear (Series 1)

log(#medals)

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

TELCO data

More power laws: areas – Korcak’s law

More power laws: areas – Korcak’s law

More power laws: Korcak
Korcak’s law & “fat fractals”

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)

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 - box counting 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 (Korcak’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

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

More power laws - internet

- pdf of degrees: (slope: 2.2)

Log(count)

Log(degree)
Fractals & power laws:
appear in numerous settings:
• medical
• geographical / geological
• social
• computer-system related

Even more power laws on the Internet

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

log(i)
log(i-th eigenvalue)

1
10
100

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

Fractals & power laws:

More apps: Medical images

[Burdeett et al, SPIE ‘93]:
• benign tumors: fd ~ 2.37
• malignant: fd ~ 2.56

More apps: Brain scans

• Oct-trees; brain-scans

Log(#octants)

More fractals:

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

Fractals & power laws:

appear in numerous settings:
• medical
• geographical / geological
• social
• computer-system related

More apps: Brain scans

• Oct-trees; brain-scans

Fractals & power laws:

appear in numerous settings:
• medical
• geographical / geological
• social
• computer-system related

Fractals & power laws:

appear in numerous settings:
• medical
• geographical / geological
• social
• computer-system related

Fractals & power laws:

appear in numerous settings:
• medical
• geographical / geological
• social
• computer-system related

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)

Fractals & power laws:

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

• the fractal dimension for the Amazon river is 1.85 (Nile: 1.4)
  [ems.gphys.unc.edu/nonlinear/fractals/examples.html]

• 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:

(citeenerima.nl, nce.com 6/2001)

log(count)

log(#citations)

Even more power laws:

- Web hit counts [w/ A. Montgomery]

Fractals & power laws:

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

appear in numerous settings:

- Medical
- Geographical / geological
- Social
- Computer-system related

Power laws, cont’d

from [Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins]
Power laws, cont’d

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

![Log-log plot of web site indegree vs frequency](image1)

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

“Foiled by power law”

- [Broder+, WWW’00]

![Graph showing log indegree vs log count](image2)

“The anomalous bump at 120 on the x-axis is due to 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 [PEVA’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

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

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?

#bytes

Poisson

time

Use of multifractals cont’d

• Extrapolations for/from samples

#bytes

Poisson

time

Use of multifractals cont’d

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

20%  80%

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)
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)

**Practitioner’s guide:**

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

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

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

**Resources:**

- Software for fractal dimension
  - http://www.cs.cmu.edu/~christos
  - christos@cs.cmu.edu

**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

- [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.
- [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


References

- [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


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:

\[
\log(\sum(p_i^2))
\]

Definitions

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

Definitions (cont’d)

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

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

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:

\[
\begin{align*}
\tau & \quad \log(\text{#non-empty cells}) \\
\log(r) & \quad \log(\tau)
\end{align*}
\]

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 end-points)

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)