Lecture #8: 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:


- 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
  (stores and households ...)
- patterns?
- attraction/repulsion?
- how many ‘spi’ within r from an ‘ell’?

Problem #3: traffic

# bytes

- disk trace (from HP - J. Wilkes); Web traffic - fit a model
- how many explosions to expect?
- queue length distr.?
Problem #3: traffic

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))

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: \( \text{nn} (\leq r) \sim r^2 \) fd== slope of \( \log(\text{nn}) \) vs log (r)
Intrinsic ('fractal') dimension

- **Local** fractal dimension of point 'P'?
  - **A**: \( \text{nn}_P(\leq r) \sim r^d \)
  - If this equation holds for several values of \( r \),
  - Then, the **local fractal dimension** of point P:
    - Local fd = \( d = 1 \)

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

Intrinsic ('fractal') dimension

- **Global** fractal dimension?
  - **A**: if
    - \( \text{sum}_P \{ \text{nn}_P(\leq r) \} \sim r^d \)
  - Then: \( d = \text{global f.d.} \)
  - Or simply the f.d.
Intrinsic (‘fractal’) dimension

- Algorithm, to estimate it?
- Notice
  - \( \text{Sum}_{all_p} \{ \text{nn}_P(\leq r) \} \) is exactly \( \text{tot#pairs}(\leq r) \)
    including ‘mirror’ pairs

Observations:

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

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

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 ->
  - <= fractals
  - >= scale-free
  - <= power-laws

\[ y = x^a, F = C \cdot r^{(1-2)} \]

- avg#neighbors(<= r) = r^D

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

\[ \log(r) \]

\[ 1.51 \]

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 )
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 accesses

# tapes retrieved

50-50 = Poisson

50-50 = Poisson

real
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!!

avg: 3.3
degree
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

log(rank) vs log(degree)

Correlation: $-0.82$

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?

log(freq) vs log(rank)

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 (Sydney):

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

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

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

Count-frequency plot of real and synthetic data

SALES data – store #96

Count-frequency plot for store no. 96.

More power laws: areas – Korčak’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)
Q: How to generate such regions?  
A: recursively, from a single region

Korcak’s law & “fat fractals”

so far we’ve seen:

• concepts:
  – fractals, multifractals and fat fractals
  – correlation integral (= pair-count plot)
  – CCDF (Korcak’s law)

• tools:
  – rank/frequency plot (Zipf’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 (Korčak’s law)

Next:

- More examples / applications
- Practitioner’s guide
- Box-counting: fast estimation of correlation integral