Data Mining using Fractals and Power laws

Christos Faloutsos
Carnegie Mellon University

THANK YOU!

- Prof. Ed Chan
- Debbie Mustin

Thanks to

- Deepayan Chakrabarti (CMU/Yahoo)

- Michalis Faloutsos (UCR)

- George Siganos (UCR)

Overview

- Goals/ motivation: find patterns in large datasets:
  - (A) Sensor data
  - (B) network/graph data
- Solutions: self-similarity and power laws
- Discussion

Applications of sensors/streams

- 'Smart house': monitoring temperature, humidity etc
- Financial, sales, economic series

Applications of sensors/streams

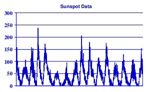
- 'Smart house': monitoring temperature, humidity etc
- Financial, sales, economic series

Tamer; Ihab

School of Computer Science
Carnegie Mellon

Motivation - Applications

- Medical: ECGs +; blood pressure etc monitoring
- Scientific data: seismological; astronomical; environment / anti-pollution; meteorological

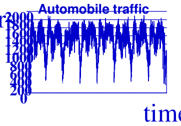


School of Computer Science
Carnegie Mellon

Motivation - Applications (cont'd)

- civil/automobile infrastructure
  - bridge vibrations [Oppenheim+02]
  - road conditions / traffic monitoring

car

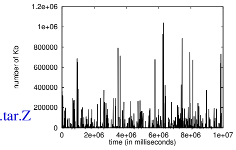


time

School of Computer Science
Carnegie Mellon

Motivation - Applications (cont'd)

- Computer systems
  - web servers (buffering, prefetching)
  - network traffic monitoring
  - ...

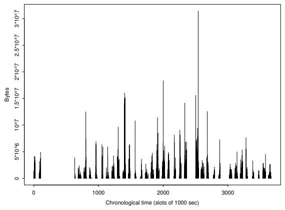


http://repository.cs.vt.edu/lbl-conn-7.tar.Z

School of Computer Science
Carnegie Mellon

Web traffic

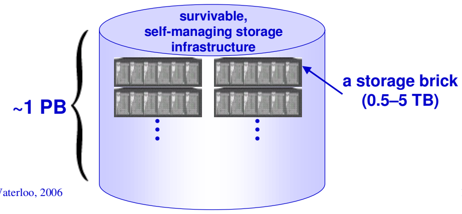
- [Crovella Bestavros, SIGMETRICS'96]



School of Computer Science
Carnegie Mellon

Self-* Storage (Ganger+)

- "self-*" = self-managing, self-tuning, self-healing, …
- Goal: 1 petabyte (PB) for CMU researchers
- www.pdl.cmu.edu/SelfStar

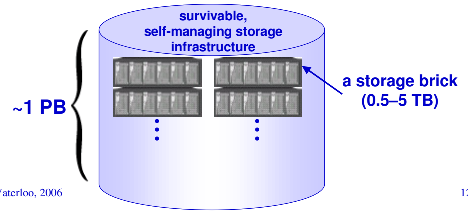


School of Computer Science
Carnegie Mellon

Self-* Storage (Ganger+)

- "self-*" = self-managing, self-tuning, self-healing, …

Ashraf, Ihab, Ken



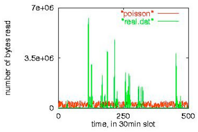
Problem definition

- Given: one or more sequences

  $x_1, x_2, \ldots, x_t, \ldots; (y_1, y_2, \ldots, y_t, \ldots)$
- Find
  - patterns; clusters; outliers; forecasts;

Problem #1

 # bytes

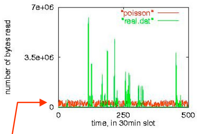


time

- Find patterns, in **large** datasets

Problem #1

 # bytes



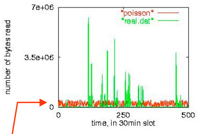
time

Poisson
indep.,
ident. distr

- Find patterns, in **large** datasets

Problem #1

 # bytes



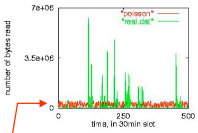
time

~~Poisson~~
indep.,
ident. distr

- Find patterns, in **large** datasets

Problem #1

 # bytes



time

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

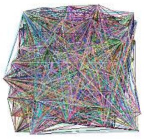
- Find patterns, in **large** datasets

Q: Then, how to generate such bursty traffic?

Overview

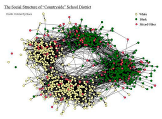
- Goals/ motivation: find patterns in **large** datasets:
  - (A) Sensor data
  - (B) network/graph data
- Solutions: self-similarity and power laws
- Discussion

Problem #2 - network and graph mining

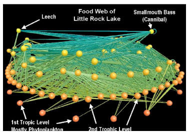


- How does the Internet look like?
- How does the web look like?
- What constitutes a 'normal' social network?
- What is the 'network value' of a customer?
- which gene/species affects the others the most?

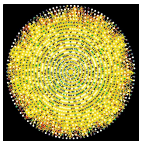
Network and graph mining



Friendship Network
[Moody '01]



Food Web
[Martinez '91]

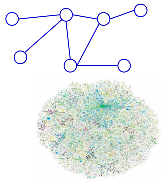


Protein Interactions
[genomebiology.com]

Graphs are everywhere!

Problem#2

Given a graph:



- which node to market-to / defend / immunize first?
- Are there un-natural sub-graphs? (eg., criminals' rings)?

[from Lumeta: ISPs 6/1999]

Solutions

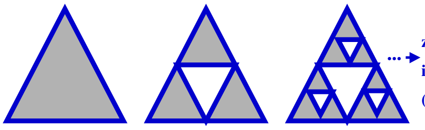
- New tools: power laws, self-similarity and 'fractals' work, where traditional assumptions fail
- Let's see the details:

Overview

- Goals/ motivation: find patterns in **large** datasets:
  - (A) Sensor data
  - (B) network/graph data
- ➡ Solutions: self-similarity and power laws
- Discussion

What is a fractal?

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

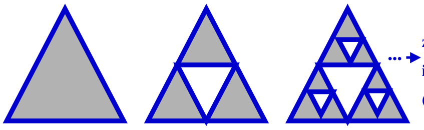


... → **zero area: (3/4)^inf**
infinite length!
(4/3)^inf

Q: What is its dimensionality??

What is a fractal?

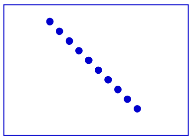
= **self-similar** point set, e.g., Sierpinski triangle:

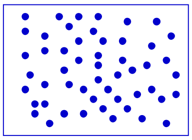


Q: What is its dimensionality??
A: log3 / log2 = 1.58 (!?!)

Intrinsic ('fractal') dimension

- Q: fractal dimension of a line?
- Q: fd of a plane?





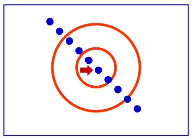
Intrinsic ('fractal') dimension

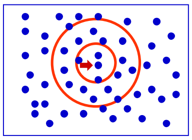
- Q: fractal dimension of a line?
- A: nn (<= r) ~ r^1

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

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

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

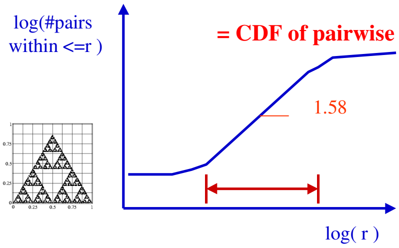




Sierpinsky triangle

== 'correlation integral'

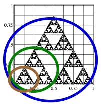
= CDF of pairwise distances

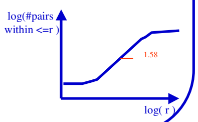


Observations: Fractals <-> power laws

Closely related:

- fractals <=>
- self-similarity <=>
- scale-free <=>
- power laws ($y= x^a$; $F=K\, r^{-2}$)
- (vs $y=e^{-ax}$ or $y=x^a+b$)





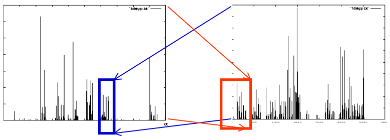
Outline

- Problems
- Self-similarity and power laws
- **Solutions to posed problems**
- Discussion

Solution #1: traffic

- disk traces: self-similar: (also: [Leland+94])
- How to generate such traffic?

#bytes

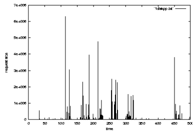


time

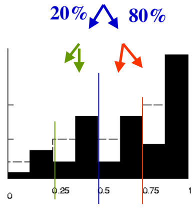
Solution #1: traffic

- disk traces (80-20 ‘law’) – ‘multifractals’

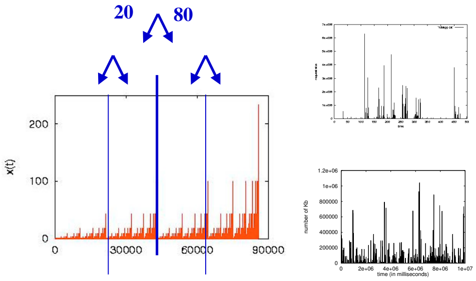
#bytes



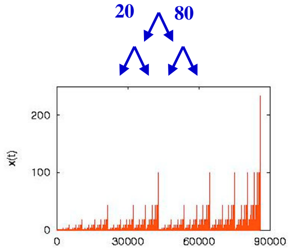
time



80-20 / multifractals



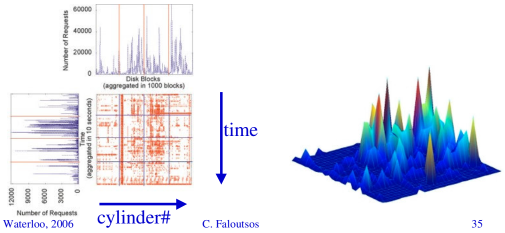
80-20 / multifractals



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

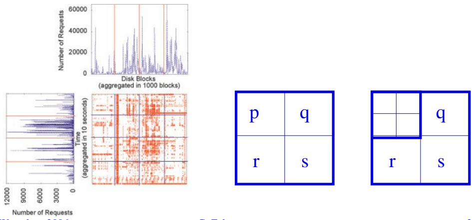
More on 80/20: PQRS

- Part of ‘self-* storage’ project



More on 80/20: PQRS

- Part of ‘self-* storage’ project

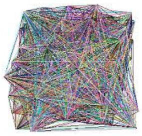


Overview

- Goals/ motivation: find patterns in **large** datasets:
  - (A) Sensor data
  - (B) network/graph data
- Solutions: self-similarity and power laws
  - sensor/traffic data
  - network/graph data
- Discussion

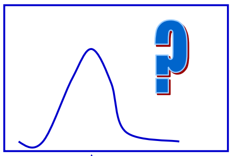
Problem #2 - topology

How does the Internet look like? Any rules?



Patterns?

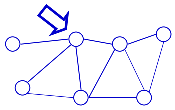
count



avg: 3.3

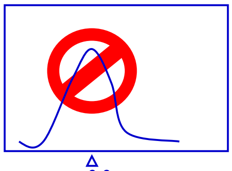
degree

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



Patterns?

count



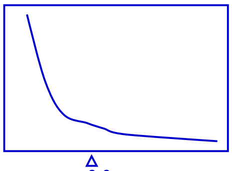
avg: 3.3

degree

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

Patterns?

count



avg: 3.3

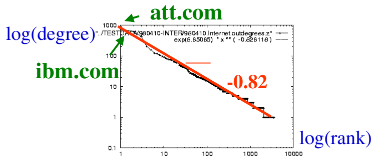
degree

- 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 (!))

Solution#2: Rank exponent *R*

- A1: Power law in the degree distribution [SIGCOMM99]

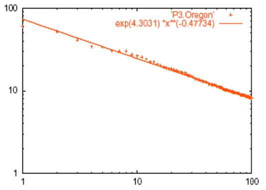
internet domains



School of Computer Science
Carnegie Mellon

Solution#2': Eigen Exponent *E*

Eigenvalue



Exponent = slope

E = -0.48

May 2001

Rank of decreasing eigenvalue

- A2: power law in the eigenvalues of the adjacency matrix

School of Computer Science
Carnegie Mellon

Power laws - discussion

- do they hold, over time?
- do they hold on other graphs/domains?

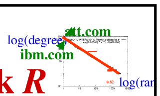
School of Computer Science
Carnegie Mellon

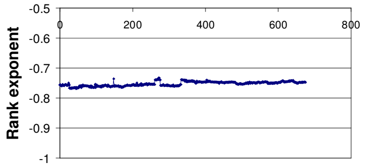
Power laws - discussion

- do they hold, over time?
- Yes! for multiple years [Siganos+]
- do they hold on other graphs/domains?
- Yes!
  - web sites and links [Tomkins+], [Barabasi+]
  - peer-to-peer graphs (gnutella-style)
  - who-trusts-whom (epinions.com)

School of Computer Science
Carnegie Mellon

Time Evolution: rank *R*





Domain level

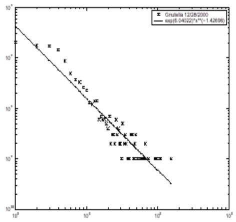
Instances in time: Nov'97 and on

- The rank exponent has not changed! [Siganos+]

School of Computer Science
Carnegie Mellon

The Peer-to-Peer Topology

count



(a) Gnutella snapshot from Dec. 28, 2000 (|r|=0.94)

[Jovanovic+]

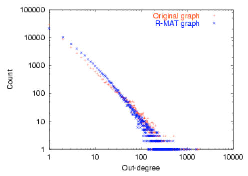
degree

- Number of immediate peers (= degree), follows a power-law

School of Computer Science
Carnegie Mellon

epinions.com

count

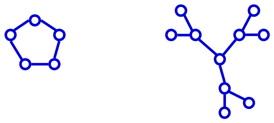


(out) degree

- who-trusts-whom [Richardson + Domingos, KDD 2001]

Why care about these patterns?

- better graph generators [BRITE, INET]
  - for simulations
  - extrapolations
- 'abnormal' graph and subgraph detection



Recent discoveries [KDD'05]

- How do graphs evolve?
- degree-exponent seems constant - anything else?

Evolution of diameter?

- Prior analysis, on power-law-like graphs, hints that

  diameter ~ O(log(N)) or

  diameter ~ O(log(log(N)))
- i.e.., slowly increasing with network size
- Q: What is happening, in reality?

Evolution of diameter?

- Prior analysis, on power-law-like graphs, hints that

  diameter ~ O(log(N)) or

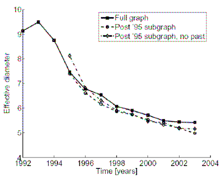
  diameter ~ O(log(log(N)))
- i.e.., slowly increasing with network size
- Q: What is happening, in reality?
- A: It **shrinks**(!!), towards a constant value

Shrinking diameter

[Leskovec+05a]

- Citations among physics papers
- 11yrs; @ 2003:
  - 29,555 papers
  - 352,807 citations
- For each month M, create a graph of all citations up to month M

diameter

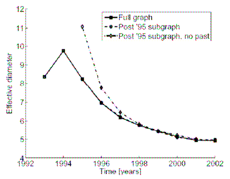


(a) arXiv citation graph

time

Shrinking diameter

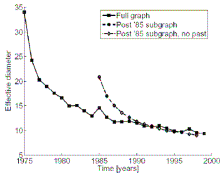
- Authors & publications
- 1992
  - 318 nodes
  - 272 edges
- 2002
  - 60,000 nodes
    - 20,000 authors
    - 38,000 papers
  - 133,000 edges



(b) Affiliation network

Shrinking diameter

- Patents & citations
- 1975
  - 334,000 nodes
  - 676,000 edges
- 1999
  - 2.9 million nodes
  - 16.5 million edges
- Each year is a datapoint

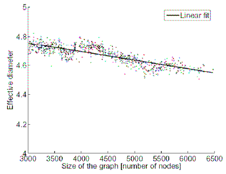


(c) Patents

Shrinking diameter

- Autonomous systems
- 1997
  - 3,000 nodes
  - 10,000 edges
- 2000
  - 6,000 nodes
  - 26,000 edges
- One graph per day

diameter



(d) AS

N

Temporal evolution of graphs

- N(t) nodes; E(t) edges at time t
- suppose that

  N(t+1) = 2 * N(t)
- Q: what is your guess for

  E(t+1) =? 2 * E(t)

School of Computer Science
Carnegie Mellon

Temporal evolution of graphs

- N(t) nodes; E(t) edges at time t
- suppose that

  N(t+1) = 2 * N(t)
- Q: what is your guess for

  E(t+1) =? ~~2~~ * E(t)
- A: over-doubled!

School of Computer Science
Carnegie Mellon

Temporal evolution of graphs

- A: over-doubled - but obeying:

  $\boxed{E(t) \sim N(t)^{a} \quad \text{for all } t}$

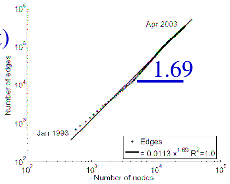
  where $1<a<2$

School of Computer Science
Carnegie Mellon

Densification Power Law

ArXiv: Physics papers and their citations

E(t)

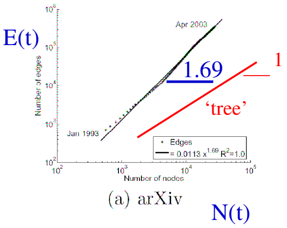


(a) arXiv

N(t)

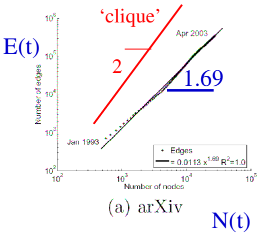
Densification Power Law

ArXiv: Physics papers and their citations



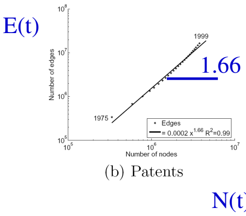
Densification Power Law

ArXiv: Physics papers and their citations



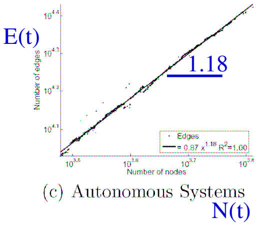
Densification Power Law

U.S. Patents, citing each other



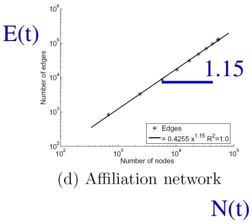
Densification Power Law

Autonomous Systems



Densification Power Law

ArXiv: authors & papers



Outline

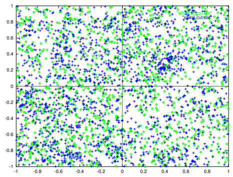
- problems
- Fractals
- Solutions
- **Discussion**
  - **what else can they solve?**
  - **how frequent are fractals?**

What else can they solve?

- separability [KDD'02] ← spatial d.m. Ed, Ihab, Tamer
- 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]
- ...

Problem #3 - spatial d.m.

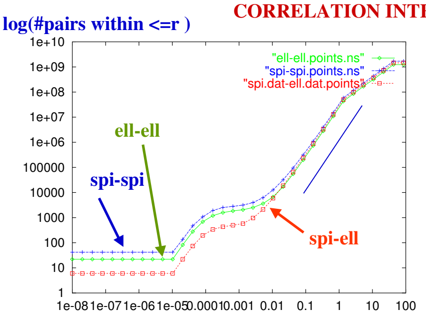
Galaxies (Sloan Digital Sky Survey w/ B. Nichol)



- 'spiral' and 'elliptical' galaxies
- patterns? (not Gaussian; not uniform)
-attraction/repulsion?
- separability??

Solution#3: spatial d.m.

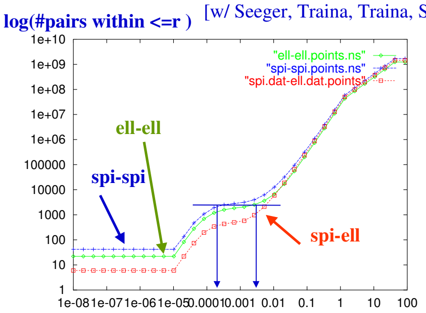
CORRELATION INTEGRAL!



- 1.8 slope
- plateau!
- repulsion!

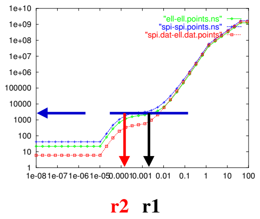
Solution#3: spatial d.m.

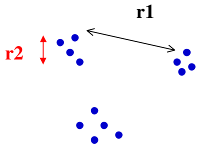
[w/ Seeger, Traina, Traina, SIGMOD00]



- 1.8 slope
- plateau!
- repulsion!

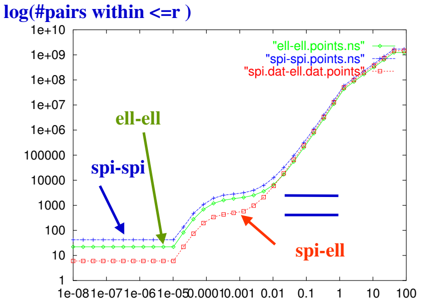
Solution#3: spatial d.m.





Heuristic on choosing # of clusters

Solution#3: spatial d.m.



- 1.8 slope
- plateau!
- repulsion!

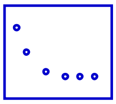
What else can they solve?

- 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]
- ...

Problem#4: dim. reduction

- given attributes x_1, ... x_n
  - possibly, non-linearly correlated
- drop the useless ones

cc

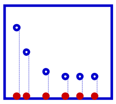


mpg

Problem#4: dim. reduction

- given attributes x_1, ... x_n
  - possibly, non-linearly correlated
- drop the useless ones

cc



mpg

(Q: why?
A: to avoid the 'dimensionality curse')

Solution: keep on dropping attributes, until the f.d. changes! [w/ Traina+, SBBD'00]

Outline

- problems
- Fractals
- Solutions
- **Discussion**
  - **what else can they solve?**
  - **how frequent are fractals?**

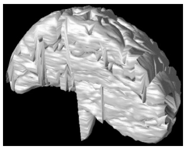
Fractals & power laws:

appear in numerous settings:

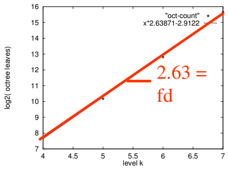
- medical
- geographical / geological
- social
- computer-system related
- <and many-many more! see [Mandelbrot]>

Fractals: Brain scans

- brain-scans



Log(#octants)



octree levels

School of Computer Science
Carnegie Mellon

More fractals

- periphery of malignant tumors: ~1.5
- benign: ~1.3
- [Burdet+]

School of Computer Science
Carnegie Mellon

More fractals:

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

School of Computer Science
Carnegie Mellon

Fractals & power laws:

appear in numerous settings:

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

School of Computer Science
Carnegie Mellon

More fractals:

- Coastlines: 1.2-1.58

1 1.1 1.3

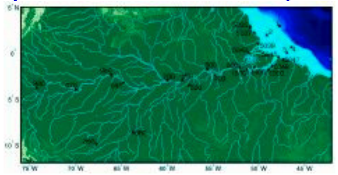
School of Computer Science
Carnegie Mellon

School of Computer Science
Carnegie Mellon

More fractals:

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

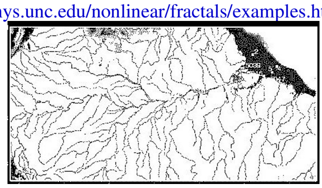
[ems.gphys.unc.edu/nonlinear/fractals/examples.html]



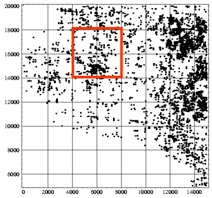
More fractals:

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

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



GIS points

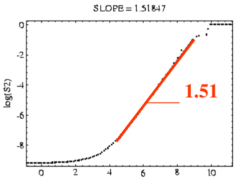


Cross-roads of Montgomery county:

•any rules?

GIS

log(#pairs(within <= r))



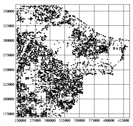
log(r)

A: self-similarity:

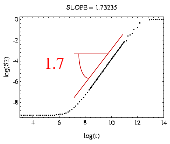
- intrinsic dim. = 1.51

Examples:LB county

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

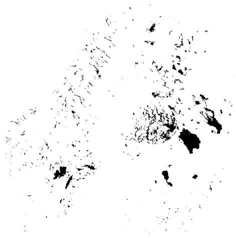


log(#pairs)



log(r)

More power laws: areas – Korcak's law

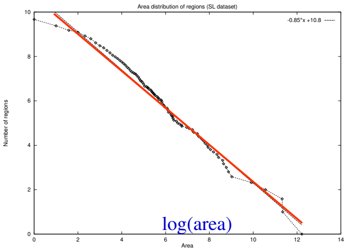


Scandinavian lakes

Any pattern?

More power laws: areas – Korcak's law

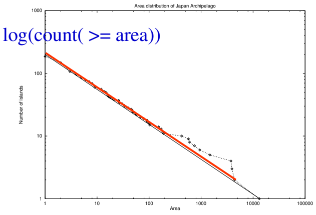
log(count(>= area))



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

School of Computer Science
Carnegie Mellon

More power laws: Korcak



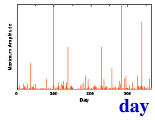
Japan islands;

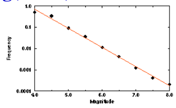
area vs cumulative count (log-log axes)

School of Computer Science
Carnegie Mellon

More power laws

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





School of Computer Science
Carnegie Mellon

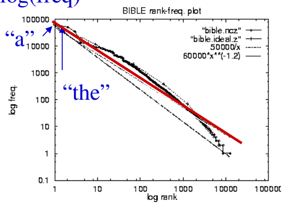
Fractals & power laws:

appear in numerous settings:

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

School of Computer Science
Carnegie Mellon

A famous power law: Zipf's law

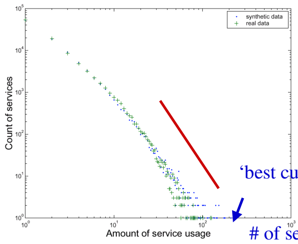


- Bible - **rank** vs. **frequency** (log-log)

"Rank/frequency plot"

School of Computer Science
Carnegie Mellon

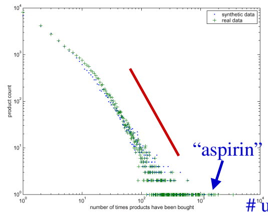
TELCO data



Count-frequency plot of real and synthetic data

School of Computer Science
Carnegie Mellon

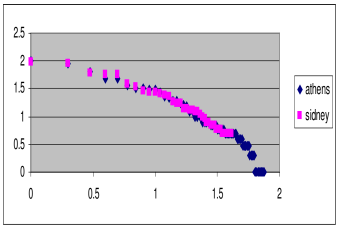
SALES data – store#96



Count-frequency plot for store no. 96.

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

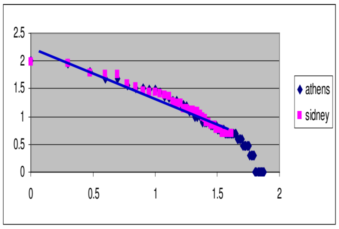
log(#medals)



log(rank)

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

log(#medals)



log(rank)

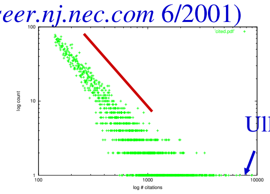
Even more power laws:

- 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)

log(count)



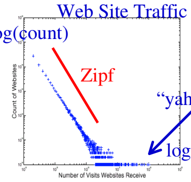
Ullman

log(#citations)

Even more power laws:

- web hit counts [w/ A. Montgomery]

log(count)



"yahoo.com"

log(freq)

Fractals & power laws:

appear in numerous settings:

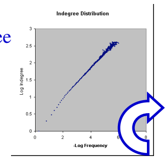
- medical
- geographical / geological
- social
- **computer-system related**

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]



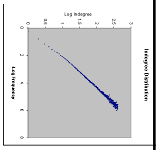
- log(freq)

Power laws, cont'd

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

log(freq)

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



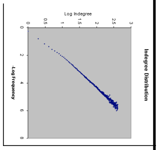
log indegree

Power laws, cont'd

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

log(freq)

Q: 'how can we use these power laws?'

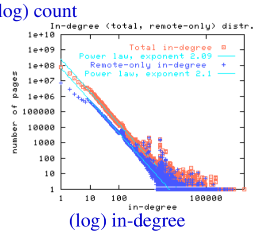


log indegree

"Foiled by power law"

- [Broder+, WWW'00]

(log) count

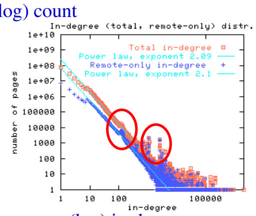


(log) in-degree

"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

Additional projects

- Find anomalies in traffic matrices [under review]
- Find correlations in sensor/stream data [VLDB'05]
  - Chlorine measurements, with Civ. Eng.
  - temperature measurements (INTEL/MIT)
- Virus propagation (SIS, SIR) [Wang+, '03]
- Graph partitioning [Chakrabarti+, KDD'04]

Conclusions

- Fascinating problems in Data Mining: find patterns in
  - sensors/streams
  - graphs/networks

Conclusions - cont'd

New tools for Data Mining: 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,
- ++)

Resources

- Manfred Schroeder "*Chaos, Fractals and Power Laws*", 1991

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

- J. Considine, F. Li, G. Kollios and J. Byers, *Approximate Aggregation Techniques for Sensor Databases* (ICDE'04, best paper award).
- [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

School of Computer Science
Carnegie Mellon

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.
- [sigmod2000] Christos Faloutsos, Bernhard Seeger, Agma J. M. Traina and Caetano Traina Jr., *Spatial Join Selectivity Using Power Laws*, SIGMOD 2000

Waterloo, 2006 C. Faloutsos 115

School of Computer Science
Carnegie Mellon

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

Waterloo, 2006 C. Faloutsos 116

School of Computer Science
Carnegie Mellon

References

- [Leskovec 05] Jure Leskovec, Jon M. Kleinberg, Christos Faloutsos: *Graphs over time: densification laws, shrinking diameters and possible explanations*. KDD 2005: 177-187

Waterloo, 2006 C. Faloutsos 117

School of Computer Science
Carnegie Mellon

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.
- [brite] Alberto Medina, Anukool Lakhina, Ibrahim Matta, and John Byers. *BRITE: An Approach to Universal Topology Generation*. MASCOTS '01

Waterloo, 2006 C. Faloutsos 118

School of Computer Science
Carnegie Mellon

References

- [icde99] Guido Proietti and Christos Faloutsos, *I/O complexity for range queries on region data stored using an R-tree* (ICDE'99)
- Stan Sclaroff, Leonid Taycher and Marco La Cascia , *"ImageRover: A content-based image browser for the world wide web"* Proc. IEEE Workshop on Content-based Access of Image and Video Libraries, pp 2-9, 1997.

Waterloo, 2006 C. Faloutsos 119

School of Computer Science
Carnegie Mellon

References

- [kdd2001] Agma J. M. Traina, Caetano Traina Jr., Spiros Papadimitriou and Christos Faloutsos: *Tri-plots: Scalable Tools for Multidimensional Data Mining*, KDD 2001, San Francisco, CA.

Waterloo, 2006 C. Faloutsos 120

Thank you!

Contact info:

christos <at> cs.cmu.edu

www. cs.cmu.edu /~christos

(w/ papers, datasets, code for fractal dimension estimation, etc)