Next Generation Data Mining Tools: Power Laws and Self-similarity for Graphs, Streams and Traditional Data

Christos Faloutsos

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA
christos@cs.cmu.edu

Abstract. What patterns can we find in a bursty web traffic? On the web or internet graph itself? How about the distributions of galaxies in the sky, or the distribution of a company's customers in geographical space? How long should we expect a nearest-neighbor search to take, when there are 100 attributes per patient or customer record? The traditional assumptions (uniformity, independence, Poisson arrivals, Gaussian distributions), often fail miserably. Should we give up trying to find patterns in such settings?
Self-similarity, fractals and power laws are extremely successful in describing real datasets (coast-lines, rivers basins, stock-prices, brain-surfaces, communication-line noise, to name a few). We show some old and new successes, involving modeling of graph topologies (internet, web and social networks); modeling galaxy and video data; dimensionality reduction; and more.

Introduction – Problem Definition

The goal of data mining is to find patterns; we typically look for the Gaussian patterns that appear often in practice and on which we have all been trained so well. However, here we show that these time-honored concepts (Gaussian, Poisson, uniformity, independence), often fail to model real distributions well. Further more, we show how to fill the gap with the lesser-known, but even more powerful tools of self-similarity and power laws.

We focus on the following applications:

- Given a cloud of points, what patterns can we find in it?
- Given a time sequence, what patterns can we find? How to characterize and anticipate its bursts?
- Given a graph (e.g., social, or computer network), how does it look like? Which is the most important node? Which nodes should we immunize first, to guard against biological or computer viruses?

All three settings appear extremely often, with vital applications. Clouds of points appear in traditional relational databases, where records with k-attributes become points in k-d spaces; e.g. a relation with patient data (age, blood pressure, etc.); in geographical information systems (GIS), where points can be, e.g.,

N. Lavrač et al. (Eds.): ECML 2003, LNAI 2837, pp. 10–15, 2003.

cities on a two-dimensional map; in medical image databases with, for example, three-dimensional brain scans, where we want to find patterns in the brain activation [ACF$^+$93]; in multimedia databases, where objects can be represented as points in feature space [FRM94]. In all these settings, the distribution of k-d points is seldom (if ever) uniform [Chr84], [FK94]. Thus, it is important to characterize the deviation from uniformity in a succinct way (e.g. as a sum of Gaussians, or something even more suitable). Such a description is vital for data mining [AIS93],[AS94], for hypothesis testing and rule discovery. A succinct description of a k-d point-set could help reject quickly some false hypotheses, or could help provide hints about hidden rules.

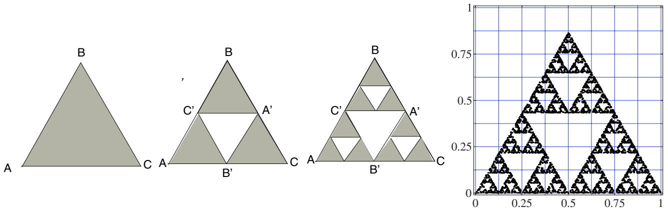
A second, very popular class of applications is time sequences. Time sequences appear extremely often, with a huge literature on linear [BJR94], and non-linear forecasting [CE92], and the recent surge of interest on sensor data [OJW03] [PBF03] [GGR02]

Finally, graphs, networks and their surprising regularities/laws have been attracting significant interest recently. The applications are diverse, and the discoveries are striking. The World Wide Web is probably the most impressive graph, which motivated significant discoveries: the famous Kleinberg algorithm [Kle99] and its closely related PageRank algorithm of Google fame [BP98]; the fact that it obeys a "bow-tie" structure [BKM$^+$00], while still having a surprising small diameter [AJB99]. Similar startling discoveries have been made in parallel for power laws in the Internet topology [FFF99], for Peer-to-Peer (gnutella/Kazaa) overlay graphs [RFI02], and for who-trusts-whom in the epinions.com network [RD02]. Finding patterns, laws and regularities in large real networks has numerous applications, exactly because graphs are so general and ubiquitous: Link analysis, for criminology and law enforcement [CSH$^+$03]; analysis of virus propagation patterns, on both social/e-mail as well as physical-contact networks [WKE00]; networks of regulatory genes; networks of interacting proteins [Bar02]; food webs, to help us understand the importance of an endangered species.

We show that the theory of fractals provide powerful tools to solve the above problems.

Definitions

Intuitively, a set of points is a fractal if it exhibits self-similarity over all scales. This is illustrated by an example: Figure 1(a) shows the first few steps in constructing the so-called *Sierpinski triangle*. Figure 1(b) gives 5,000 points that belong to this triangle. Theoretically, the Sierpinski triangle is derived from an equilateral triangle ABC by excluding its middle (triangle A'B'C') and by recursively repeating this procedure for each of the resulting smaller triangles. The resulting set of points exhibits 'holes' in any scale; moreover, each smaller triangle is a *miniature replica* of the whole triangle. In general, the characteristic of fractals is this *self-similarity* property: parts of the fractal are similar (exactly or statistically) to the whole fractal. For our experiments we use 5,000 sample points from the Sierpinski triangle, using Barnsley's algorithm of Iterated Function Systems [BS88] to generated these points quickly.



(a) construction (b) a finite sample

Fig. 1. Theoretical fractals: the Sierpinski triangle (a) the first 3 steps of its recursive construction (b) a finite sample of it (5K points)

Notice that the resulting point set is neither a 1-dimensional Euclidean object (it has infinite length), nor 2-dimensional (it has zero area). The solution is to consider *fractional* dimensionalities, which are called *fractal dimensions.* Among the many definitions, we describe the *correlation* fractal dimension, D, because it is the easiest to describe and to use.

Let $nb(\epsilon)$ be the average number of neighbors of an arbitrary point, within distance ϵ or less. For a real, finite cloud of E-dimensional points, we follow [Sch91] and say that this data set is *self-similar in the range of scales* r_1, r_2 if

$$nb(\epsilon) \propto \epsilon^D \qquad r_1 \leq \epsilon \leq r_2 \tag{1}$$

The *correlation integral* is defined as the plot of $nb(\epsilon)$ versus ϵ in log-log scales; for self-similar datasets, it is linear with slope D.

Notice that the above definition of fractal dimension D encompasses the traditional Euclidean objects: lines, line segments, circles, and all the standard curves have D=1; planes, disks and standard surfaces have D=2; Euclidean volumes in E-dimensional space have $D = E$.

Discussion – How Frequent Are Self-similar Datasets?

The reader might be wondering whether any real datasets behave like fractals, with linear correlation integrals. *Numerous* the real datasets give linear correlation integrals, including longitude-latitude coordinates of stars in the sky, population-versus-area of the countries of the world [FK94]; several geographic datasets [BF95] [FK94]; medical datasets [FG96]; automobile-part shape datasets [BBB$^+$97,BBKK97].

There is overwhelming evidence from multiple disciplines that fractal datasets appear *surprisingly* often [Man77](p. 447),[Sch91]:

– coast lines and country borders ($D \approx 1.1$ - 1.3);
– the periphery of clouds and rainfall patches ($D \approx 1.35$)[Sch91](p.231);
– the distribution of galaxies in the universe ($D \approx 1.23$);
– stock prices and random walks (D=1.5)
– the brain surface of mammals ($D \approx 2.7$);
– the human vascular system ($D = 3$, because it has to reach every cell in the body!)
– even traditional Euclidean objects have linear box-counting plots, with integer slopes

Discussion – Power Laws

Self-similarity and power laws are closely related. A *power law* is a law of the form

$$y = f(x) = x^a \tag{2}$$

Power laws are the only laws that have no characteristic scales, in the sense that they remain power laws, even if we change the scale: $f(c * x) = c^a * x^a$

Exactly for this reason, power laws and self-similarity appear often together: if a cloud of points is self similar, it has no characteristic scales; any law/pattern it obeys, should also have no characteristic scale, and it should thus be a power law.

Power laws also appear extremely often, in diverse settings: in text, with the famous Zipf law [Zip49]; in distributions of income (the Pareto law); in scientific citation analysis (Lotka law); in distribution of areas of lakes, islands and animal habitats (Korcak's law [Sch91,HS93,PF01]) in earthquake analysis (Gutenberg-Richter law [Bak96]; in LAN traffic [LTWW94]; in web click-streams [MF01]; and countless more settings.

Conclusions

Self-similarity and power laws can solve data mining problems that traditional methods can not. The two major tools that we cover in the talk are: (a) the "correlation integral" [Sch91] for a set of points and (b) the "rank-frequency" plot [Zip49] for categorical data. The former can estimate the intrinsic dimensionality of a cloud of points, and it can help with dimensionality reduction [TTWF00], axis scaling [WF02], and separability [TTPF01]. The rank-frequency plot can spot power laws, like the Zipf's law, and many more.

References

ACF$^+$93. Manish Arya, William Cody, Christos Faloutsos, Joel Richardson, and Arthur Toga. QBISM: A prototype 3-D medical image database system. *IEEE Data Engineering Bulletin*, 16(1):38–42, March 1993.

AIS93. Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. In *Proc. ACM SIGMOD*, pages 207–216, Washington, DC, May 26-28 1993.

AJB99. R. Albert, H. Jeong, and A.-L. Barabasi. Diameter of the world-wide web. *Nature*, 401:130–131, September 1999.

AS94. Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proc. of VLDB Conf.*, pages 487–499, Santiago, Chile, Sept. 12-15 1994.

Bak96. Per Bak. How nature works : The science of self-organized criticality, September 1996.

Bar02. Albert-Laszlo Barabasi. *Linked: The New Science of Networks.* Perseus Publishing, first edition, May 2002.

BBB$^+$97. Stefan Berchtold, Christian Boehm, Bernhard Braunmueller, Daniel A. Keim, and Hans-Peter Kriegel. Fast similarity search in multimedia databases. In *SIGMOD Conference*, pages 1–12, 1997.

BBKK97. Stefan Berchtold, Christian Boehm, Daniel A. Keim, and Hans-Peter Kriegel. A cost model for nearest neighbor search in high-dimensional data space. *PODS*, pages 78–86, 1997.

BF95. Alberto Belussi and Christos Faloutsos. Estimating the selectivity of spatial queries using the 'correlation' fractal dimension. In *Proc. of VLDB*, pages 299–310, Zurich, Switzerland, September 1995.

BJR94. George E.P. Box, Gwilym M. Jenkins, and Gregory C. Reinsel. *Time Series Analysis: Forecasting and Control.* Prentice Hall, Englewood Cliffs, NJ, 3rd edition, 1994.

BKM$^+$00. Andrei Broder, Ravi Kumar, Farzin Maghoul1, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web: experiments and models. In *WWW Conf.*, 2000.

BP98. Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual (web) search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

BS88. M.F. Barnsley and A.D. Sloan. A better way to compress images. *Byte*, pages 215–223, January 1988.

CE92. M. Castagli and S. Eubank. *Nonlinear Modeling and Forecasting.* Addison Wesley, 1992. Proc. Vol. XII.

Chr84. S. Christodoulakis. Implication of certain assumptions in data base performance evaluation. *ACM TODS*, June 1984.

CSH$^+$03. H. Chen, J. Schroeder, R. Hauck, L. Ridgeway, H. Atabaksh, H. Gupta, C. Boarman, K. Rasmussen, and A. Clements. Coplink connect: Information and knowledge management for law enforcement. *CACM*, 46(1):28–34, January 2003.

FFF99. Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262, 1999.

FG96. Christos Faloutsos and Volker Gaede. Analysis of the z-ordering method using the hausdorff fractal dimension. *VLDB*, September 1996.

FK94. Christos Faloutsos and Ibrahim Kamel. Beyond uniformity and independence: Analysis of R-trees using the concept of fractal dimension. In *Proc. ACM SIGACT-SIGMOD-SIGART PODS*, pages 4–13, Minneapolis, MN, May 24-26 1994. Also available as CS-TR-3198, UMIACS-TR-93-130.

FRM94. Christos Faloutsos, M. Ranganathan, and Yannis Manolopoulos. Fast subsequence matching in time-series databases. In *Proc. ACM SIGMOD*, pages 419–429, Minneapolis, MN, May 25-27 1994. 'Best Paper' award; also available as CS-TR-3190, UMIACS-TR-93-131, ISR TR-93-86.

GGR02. Minos N. Garofalakis, Johannes Gehrke, and Rajeev Rastogi. Querying and mining data streams: You only get one look. *ACM SIGMOD*, page 635, June 2002. (tutorial).

HS93. Harold M. Hastings and George Sugihara. *Fractals: A User's Guide for the Natural Sciences.* Oxford University Press, 1993.

Kle99. Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

LTWW94. W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson. On the self-similar nature of ethernet traffic. *IEEE Transactions on Networking*, 2(1):1–15, February 1994. (earlier version in SIGCOMM '93, pp 183-193).

Man77. B. Mandelbrot. *Fractal Geometry of Nature.* W.H. Freeman, New York, 1977.

MF01. Alan L. Montgomery and Christos Faloutsos. Identifying web browsing trends and patterns. *IEEE Computer*, 34(7):94–95, July 2001.

OJW03. C. Olston, J. Jiang, and J. Widom. Adaptive filters for continuous queries over distributed data streams. *ACM SIGMOD*, 2003.

PBF03. Spiros Papadimitriou, Anthony Brockwell, and Christos Faloutsos. Adaptive, hands-off stream mining. *VLDB*, September 2003.

PF01. Guido Proietti and Christos Faloutsos. Accurate modeling of region data. *IEEE TKDE*, 13(6):874–883, November 2001.

RD02. M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *SIGKDD*, pages 61–70, Edmonton, Canada, 2002.

RFI02. M. Ripeanu, I. Foster, and A. Iamnitchi. Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. *IEEE Internet Computing Journal*, 6(1), 2002.

Sch91. Manfred Schroeder. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise.* W.H. Freeman and Company, New York, 1991.

TTPF01. Agma Traina, Caetano Traina, Spiros Papadimitriou, and Christos Faloutsos. Tri-plots: Scalable tools for multidimensional data mining. *KDD*, August 2001.

TTWF00. Caetano Traina, Agma Traina, Leejay Wu, and Christos Faloutsos. Fast feature selection using the fractal dimension,. In *XV Brazilian Symposium on Databases (SBBD)*, Paraiba, Brazil, October 2000.

WF02. Leejay Wu and Christos Faloutsos. Making every bit count: Fast nonlinear axis scaling. *KDD*, July 2002.

WKE00. Chenxi Wang, J. C. Knight, and M. C. Elder. On computer viral infection and the effect of immunization. In *ACSAC*, pages 246–256, 2000.

Zip49. G.K. Zipf. *Human Behavior and Principle of Least Effort: An Introduction to Human Ecology.* Addison Wesley, Cambridge, Massachusetts, 1949.