

Future Directions in Data Mining: Streams, Networks, Self-Similarity and Power Laws

Prof. Christos Faloutsos
Carnegie Mellon University

Abstract

How to spot abnormalities in a stream of temperature data from a sensor? Or from a network of sensors? How does the Internet look like? Are there 'abnormal' sub-graphs in a given social network, possibly indicating, e.g., money-laundering rings?

We present some recent work and list many remaining challenges for these two fascinating issues in data mining, namely, streams and networks. Streams appear in numerous settings, in the form of, e.g., temperature readings, road traffic data, series of video frames for surveillance, patient physiological data. In all these settings, we want to equip the sensors with nimble, but powerful enough algorithms to look for patterns and abnormalities,

- (a) on a semi-infinite stream,
- (b) using finite memory, and
- (c) without human intervention.

For networks, the applications are also numerous: social networks recording who knows/calls/emails whom; the Internet itself, as well as the Web, with routers and links, or pages and hyper-links; the genes and how they are related; customers and products they buy. In fact, any "many-to-many" database relationship eventually leads to a graph/network. In all these settings we want to find patterns and 'abnormalities'; the most central/important

nodes; we also want to predict how the network will evolve; and we want to tackle huge graphs, with millions or billions of nodes and edges.

As a promising direction towards these problems, we present some surprising tools from the theory of fractals, self-similarity and power laws. We show how the 'intrinsic' or 'fractal' dimension can help us find patterns, when traditional tools and assumptions fail. We show that self-similarity and power laws models work well in an impressive variety of settings, including real, bursty disk and web traffic; skewed distributions of click-streams; and multiple, real Internet graphs.

Short Bio

Christos Faloutsos received the B.Sc. degree in Electrical Engineering (1981) from the National Technical University of Athens, Greece and the M.Sc. and Ph.D. degrees in Computer Science from the University of Toronto, Canada. He is currently a Professor at Carnegie Mellon University.

He has received the Presidential Young Investigator Award by the National Science Foundation (1989), two "best paper" awards (SIGMOD 94, VLDB 97), and four teaching awards. He has published over 100 refereed articles, one monograph, and holds four patents. His research interests include data mining, multimedia databases, and database performance.