Large Graph Mining

Christos Faloutsos
CMU

Sept. 12, 2007

ABSTRACT

How do graphs look like? How do they evolve over time? How can we generate realistic-looking graphs?

Graphs appear in a wide variety of settings, including social networks, call graphs, IP traffic matrices, gene regulatory networks and many more. Finding patterns, regularities, and communities are important to help us understand the given graph, so that we can spot anomalies, summarize it, and do forecasting.

In this talk, we review some static and temporal 'laws', and we describe the "Kronecker" graph generator, which naturally matches all of the known properties of real graphs. Moreover, we present tools for discovering anomalies and patterns in two types of graphs, static and time-evolving. For the former, we present the 'CenterPiece' subgraphs (CePS), which expects q query nodes (eg., suspicious people) and finds the node that is best connected to all q of them (eg., the master mind of a criminal group). We also show how to compute CenterPiece subgraphs efficiently. For the time evolving graphs, we present tensor-based methods, and apply them on real data, like the DBLP author-paper dataset, where they are able to find natural research communities, and track their evolution. Finally, we also briefly mention some results on influence and virus propagation on real graphs.

We will also cover some promising future research directions for graph mining. The major challenge for the future is scalability: How can we handle huge graphs, spanning Gigabytes, Terabytes, or even Petabytes, so that we can find patterns and anomalies. We believe that disk-based algorithms, with linear or near-linear response time, are very important for datasets spanning Gigabytes. For even larger ones, an extremely promising direction is to use shared-nothing parallelism, exploiting ideas like the 'map-reduce' of Google and the 'hadoop' of Yahoo.

BIOGRAPHICAL NOTE

Christos Faloutsos is a Professor at Carnegie Mellon University. He has received the Presidential Young Investigator Award by the National Science Foundation

(1989), the Research Contributions Award in ICDM 2006, eleven "best paper" awards, and several teaching awards. He has served as a member of the executive committee of SIGKDD; he has published over 160 refereed articles, 11 book chapters and one monograph. His research interests include data mining for streams and networks, fractals, indexing for multimedia and bio-informatics data, and database performance.