Mining the graph structures of the web

Aristides Gionis

Yahoo! Research, Barcelona, Spain, and
University of Helsinki, Finland
gionis@yahoo-inc.com

Abstract

Graph structures is a general way of modeling entities and their relationships and they are used to describe a wide variety of data including the Internet, the Web, social networks, metabolic networks, protein-interaction networks, food webs, networks of citations among papers, and many more. In the recent years there has been an increasing amount of literature on studying properties, models, and algorithms for graph data. The first part of the seminar gives a brief overview of graph-generation models and graph-mining algorithms. The set of topics includes algorithms for discovering communities, models for characterizing the evolution of graphs over time, as well as discussion on their ubiquitous scale-free properties. In the second part we discuss applications where exploiting the graph structure is beneficial for certain data-mining tasks and we present challenges of graph mining in the context of problems appearing in a search engine.

1 Background

One of the most pervasive properties of real-world graphs is the emergence of power laws that seems to characterize many of their statistical properties [1, 5]. Power laws have intrigued the interest of researchers and many models that attempt to explain their presence in real graphs have been proposed, e.g., see [1, 4, 11]. In the first part of the seminar we discuss properties of power-law distributions and describe underlying processes that generate such distributions [14, 15, 17].

We then discuss the problem of finding communities in graphs, which is related to the problem of graph clustering. We give an overview of objective functions used for the task of finding communities and we review a number of combinatorial and spectral algorithms [7, 8, 16]. Many of the clustering methods are prohibitively expensive for applying them on large-scale graphs, so we also discuss scalable algorithms that have been designed for finding communities on the Web [12].

Finally we review studies on statistical properties of graphs that evolve over time [10, 13].

2 Applications

In the second part of the seminar we present specific applications of graph mining in the context of problems appearing in search engines.

The first application is spam detection [2]. A common approach to detecting spam is to extract a set of content-based and link-based features from Web pages and treat the spam-detection problem as a classification problem. In addition to extracting discriminative features, one can exploit the observation that linked hosts tend to belong to the same class: either both are spam or both are non-spam. We discuss different algorithms that attempt

to leverage this observation and exploit the topology of the web graph in order to improve the accuracy of a baseline feature-based spam-detection system.

Then we discuss the problem of predicting the popularity of items in a dynamic environment in which authors post new items and provide feedback on existing ones [3]. The basic setting can be applied to predict popularity of blog posts, rank photographs in a photo-sharing system, or predict the citations of a scientific article using author information and monitoring the item of interest for a short period of time after its creation. One of the components of the system is the eigenrumor algorithm [6], an adaptation of the HITS algorithm [9].

We conclude by describing complex graph structures that emerge in problems related to search engines and we discuss challenges on mining those graphs.

References

[1] A.-L. Barabasi, R Albert. *Emergence of Scaling in Random Networks.* Science, 286, 1999.

[2] C. Castillo, D. Donato, A. Gionis, V. Murdock, F. Silvestri. *Know your Neighbors: Web Spam Detection using the Web Topology.* 30th Annual International ACM SIGIR Conference, 2007.

[3] C. Castillo, D. Donato, A. Gionis. *Estimating the number of citations of a paper using author reputation.* Submitted for publication.

[4] A. Fabrikant, E. Koutsoupias, C. Papadimitriou. *Heuristically Optimized Trade-offs: A New Paradigm for Power Laws in the Internet.* 29th International Colloquium on Automata, Languages and Programming (ICALP), 2002.

[5] M. Faloutsos, P. Faloutsos, C. Faloutsos. *On Power-Law Relationships of the Internet Topology.* ACM SIGCOMM, 1999.

[6] K. Fujimura, N. Tanimoto. *The EigenRumor algorithm for calculating contributions in cyberspace communities.* Trusting Agents for Trusting Electronic Societies, 2005.

[7] J. Hopcroft, O. Khan, B. Kulis, B. Selman. *Natural communities in large linked networks.* 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.

[8] G. Karypis, V. Kumar. *A fast and high quality multilevel scheme for partitioning irregular graphs.* SIAM Journal on Scientific Computing, 20(1), 1999.

[9] J. Kleinberg. *Authoritative sources in a hyperlinked environment.* Journal of the ACM, 46, 1999.

[10] R. Kumar, J. Novak, P. Raghavan, A. Tomkins. *On the bursty evolution of Blog Space.* 12th International World Wide Web Conference, 2003.

[11] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, E. Upfal. *Stochastic models for the Web graph.* 41th IEEE Symposium on Foundations of Computer Science, 2000.

[12] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins. *Trawling the web for emerging cyber-communities.* 8th International World Wide Web Conference, 1999.

[13] J. Leskovec, J. Kleinberg, C. Faloutsos, *Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations.* International Conference on Knowledge Discovery and Data Mining, 2005.

[14] L. Li, D. Alderson, J. Doyle, W. Willinger. *Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications.* Internet Mathematics, 2006.

[15] M. Mitzenmacher. *A Brief History of Generative Models for Power Law and Lognormal Distributions.* Internet Mathematics, 2004.

[16] M. E. J. Newman. *Power laws, Pareto distributions and Zipf's law.* Contemporary Physics, 46(5), 2005.

[17] M. E. J. Newman, M. Girvan. *Finding and evaluating community structure in networks.* Physical Review E, 2004.