On the Spectral Evolution of Large Networks
PhD Thesis Proposal

Jérôme Kunegis
Institute for Web Science and Technologies
University of Koblenz-Landau
kunegis@uni-koblenz.de

August 19, 2010

Abstract

In my dissertation, I study the spectral evolution of large networks. My main result is an interpretation of the spectrum and eigenvectors of networks in terms of global and local effects. I argue and show that the spectrum describes a network on the global level, whereas eigenvectors describe a network at the local level. I observe this behavior in over one hundred network datasets from social networks, authorship networks, feature networks, rating networks, link networks, folksonomies and other types of networks. Based on the spectral evolution model, I introduce two novel link prediction methods, one based on curve fitting, and one on spectral extrapolation. As special cases I present variants of all methods that apply to bipartite and signed graphs.

1 Introduction

A certain number of machine learning and data mining problems can be formulated as the analysis of large networks: social networks, hyperlink networks, citation graphs, rating graphs, trust networks, communication networks, etc. In these settings, a prominent type of problem is given by link prediction: Learn where new edges will appear. Link prediction can be applied to recommender systems, to collaborative filtering methods, or everytime edges or edge weights are to be predicted. Many different approaches have been proposed to the link prediction problem.

I study the link prediction problem taking an algebraic approach: Given a matrix associated with a network, compute a matrix decomposition, giving eigenvalues (the spectrum) and eigenvectors. Common matrices are the adjacency and Laplacian matrices; common decompositions are the eigenvalue and singular value decompositions. Studying a large collection of network datasets I made the following observations:

- Over time, the eigenvalues increase, while the eigenvectors stay approximately constant.
- The eigenvector components follow certain distributions (e.g. lognormal), and do not change significantly with time.

These observations lead to two novel link prediction algorithms, one based on curve fitting and one based on extrapolation. I show that both algorithms are competitive in many settings.

The methods I present are very general and also apply to networks with special structure and additional features. In bipartite networks, I use the biadjacency matrix instead of the adjacency matrix, leading to odd pseudokernels. In signed graphs, I use the signed Laplacian matrix, giving signed spectral link prediction. As a benefit, the signed Laplacian can be used for signed clustering.

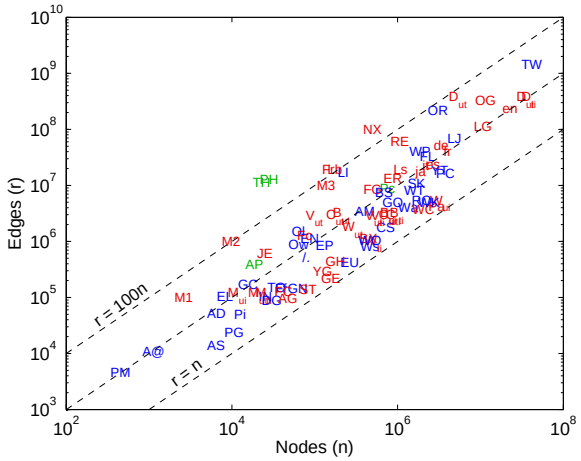


Figure 1: Datasets from my dissertation arranged by total number of nodes and edges. Unipartite networks are shown in red and bipartite networks in blue.

2 Datasets

During my research I have collected network datasets [19]. These datasets are either unipartite (such as social networks) or bipartite (such as rating networks). Some networks have edge weights, e.g. ratings or trust/distrust links, in which case the link prediction problem also consists of predicting the weight of edges. Edge weights admit negative values in some datasets. The networks include social networks, citation graphs, hyperlink graphs, trust networks, rating graphs (i.e. user-item graphs), communication graphs and others. Many of these graphs are used in previous literature in various machine learning and data mining subfields. Edges in some datasets additionally have timestamps. Figure 1 gives an overview of the networks by their size.

I extracted one dataset myself, and studied it separately: the Slashdot Zoo [23]. This dataset has the distinction of containing negative edge weights in a social network.

The following is a sample of datasets used. A longer list is given in [19].

- Unipartite unweighted datasets: hep-th citations [31], the WWW link graph [1], the Advogato trust network [35], Patentcite citations [14], DBLP citations [29], TREC WT10G hyperlinks [2], Citeseer citations [4], Twitter followers [28], English Wikipedia hyperlinks [37], Facebook friends and wall posts [11].
- Unipartite weighted datasets: the Slashdot Zoo [23], LíbímSeTi dating site ratings [6], Epinions [13], Enron e-mails [17].
- Bipartite unweighted networks: DBLP authorship [29], CiteULike tags [9], English Wikipedia categories [37], the edit networks of several language Wikipedias [37], BibSonomy user/item/tag networks [15], Delicious user/item/tag networks [36].
- Bipartite weighted networks: Reuters [32], MovieLens (several sizes) [12], Netflix [3], Book Crossing (BX) [38], Jester [10].

3 Algebraic Graph Theory

My analysis of networks is based on algebraic graph theory [8]. I study the decomposition of certain matrices associated with graphs. The (weighted) edge set of a graph can be represented by a matrix whose characteristics follow those of the graph. Unweighted graphs lead to a 0/1 matrix, undirected graphs to a symmetric matrix, bipartite graphs to a rectangular matrix, and so on. This matrix is called the adjacency matrix and is denoted A. The following decompositions can be computed:

- Singular value decomposition: $A = U\Sigma V^T$
- Eigenvalue decomposition: $A + A^T = U\Lambda U^T$
- Eigenvalue decomposition of the Laplacian matrix: $L = U\Lambda U^T$

The columns U and V are the (Laplacian) eigenvectors of the network. Σ and Λ are the spectra of the network, containing the singular values and eigenvalues.

My thesis consists of studying the behavior of U, V, Σ and Λ of networks as these grow. While studying the Laplacian matrix of weighted networks, I found a way of defining it for networks with negative edge weights [25, 26, 27].

4 Link Prediction

Many machine learning problems applying to networks can be understood as *link prediction*. The two main problems are: predicting where links will appear [30] and predicting the weight of new edges, if edges are weighted (e.g. collaborative filtering [5], predicting trust [13], friendship/enmity prediction [23]). By studying the evolution of networks over time I made the following observations:

- Eigenvectors remain largely constant over time
- Spectra evolve over time

These lead to link prediction algorithms that compute a new spectrum and multiply it with the known eigenvectors to predict links. It turns out that several common link prediction methods are of this form, using various ways of transforming the spectrum [7, 16, 18, 33]. To learn a good spectral transformation, two methods can be used:

- Extrapolate the growth of the eigenvalues over time [20], see Figures 2(a) and 2(b).
- Reduce the problem to a one-dimensional curve fitting problem [21], see Figure 2(c).

I evaluated these two methods on the collection of large network datasets, and found they are both competive in all kinds of link prediction tasks, over all types of network types.

5 Current State

Preliminary results of my work were published as conference papers.

I began my research by studying the collaborative filtering problem. In particular, I applied the Laplacian matrix to networks with negative edge weights [25, 26, 27]. That last and most complete analysis was presented at SIAM SDM 2010.

An analysis of the self-acquired Slashdot Zoo dataset containing positive and negative edges between users ("friends" and "foes") was presented separately at WWW 2009 [23].

My work on graph kernels started by studying their scalability and their application as similarity functions [22, 24]. In a later paper I introduced a method for learning spectral transformations [21] (presented at ICML 2009). An empirical verification of the spectral evolution model, along with the spectral extrapolation algorithm will be presented at CIKM 2010 [20].

A survey of the network datasets I collected was written and submitted to ICDM 2010 [19].

6 Future Work

While these tasks are not part of my dissertation, they follow directly from it.

- Application of spectral learning to other, more complex matrix decompositions: nonnegative matrix factorization, approximations with missing data (e.g. [34]), probabilistic latent semantic analysis, maximum margin matrix factorization.
- Studying link prediction in networks with multiple vertex and node types, leading to new machine learning problems, such as learning relative weights of different edge types.
- Characterizing graphs based on the spectrum and a probability distribution for eigenvectors.

References

[1] ALBERT, R., JEONG, H., AND BARABÁSI, A.-L. The diameter of the World Wide Web. *Nature 401* (1999), 130.

[2] BAILEY, P., CRASWELL, N., AND HAWKING, D. Engineering a multi-purpose test collection for Web retrieval experiments. *Information Processing and Management 39*, 6 (2003), 853–871.

[3] BENNETT, J., AND LANNING, S. The Netflix prize. In *Proc. KDD Cup* (2007), pp. 3–6.

[4] BOLLACKER, K., LAWRENCE, S., AND GILES, C. L. CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications. In *Proc. Int. Conf. on Autonomous Agents* (1998), pp. 116–123.

[5] BREESE, J. S., HECKERMAN, D., AND KADIE, C. Empirical analysis of predictive algorithms for collaborative filtering. In *Proc. Conf. on Uncertainty in Artificial Intelligence* (1998), pp. 43–52.

[6] BROŽOVSKÝ, L., AND PETŘÍČEK, V. Recommender system for online dating service. In *Proc. Znalosti* (2007), pp. 29–40.

[7] CHEBOTAREV, P., AND SHAMIS, E. V. On proximity measures for graph vertices. *Automation and Remote Control 59*, 10 (1998), 1443–1459.

[8] CHUNG, F. *Spectral Graph Theory*. American Mathematical Society, 1997.

[9] EMAMY, K., AND CAMERON, R. CiteULike: A researcher's social bookmarking service. *Ariadne*, 51 (2007).

[10] GOLDBERG, K., ROEDER, T., GUPTA, D., AND PERKINS, C. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval 4*, 2 (2001), 133–151.

[11] GOLDER, S. A., WILKINSON, D. M., AND HUBERMAN, B. A. Rhythms of social interaction: messaging within a massive online network. In *Proc. Int. Conf. on Communities and Technologies* (2007).

[12] GROUPLENS RESEARCH. MovieLens data sets. `http://www.grouplens.org/node/73`, October 2006.

[13] GUHA, R., KUMAR, R., RAGHAVAN, P., AND TOMKINS, A. Propagation of trust and distrust. In *Proc. Int. World Wide Web Conf.* (2004), pp. 403–412.

[14] HALL, B. H., JAFFE, A. B., AND TRAJTENBERG, M. The NBER patent citations data file: Lessons, insights and methodological tools. In *NBER Working Papers 8498, National Bureau of Economic Research, Inc* (2001).

[15] HOTHO, A., JÄSCHKE, R., SCHMITZ, C., AND STUMME, G. BibSonomy: A social bookmark and publication sharing system. In *Proc. Workshop on Conceptual Structure Tool Interoperability* (2006), pp. 87–102.

[16] ITO, T., SHIMBO, M., KUDO, T., AND MATSUMOTO, Y. Application of kernels to link analysis. In *Proc. Int. Conf. on Knowledge Discovery in Data Mining* (2005), pp. 586–592.

[17] KLIMT, B., AND YANG, Y. Introducing the Enron corpus. In *Proc. Conf. on Email and Anti-spam* (2004).

[18] KONDOR, R., AND LAFFERTY, J. Diffusion kernels on graphs and other discrete structures. In *Proc. Int. Conf. on Machine Learning* (2002), pp. 315–322.

[19] KUNEGIS, J., BAUCKHAGE, C., NEUBAUER, N., AND OBERMAYER, K. Empirical verification of four complex network models on seventy-seven large network datasets. Submitted.

[20] KUNEGIS, J., FAY, D., AND BAUCKHAGE, C. Network growth and the spectral evolution model. In *Proc. Int. Conf. on Information and Knowledge Management* (2010).

[21] KUNEGIS, J., AND LOMMATZSCH, A. Learning spectral graph transformations for link prediction. In *Proc. Int. Conf. on Machine Learning* (2009), pp. 561–568.

[22] KUNEGIS, J., LOMMATZSCH, A., AND BAUCKHAGE, C. Alternative similarity functions for graph kernels. In *Proc. Int. Conf. on Pattern Recognition* (2008).

[23] KUNEGIS, J., LOMMATZSCH, A., AND BAUCKHAGE, C. The Slashdot Zoo: Mining a social network with negative edges. In *Proc. Int. World Wide Web Conf.* (2009), pp. 741–750.

[24] KUNEGIS, J., LOMMATZSCH, A., BAUCKHAGE, C., AND ALBAYRAK, S. On the scalability of graph kernels applied to collaborative recommenders. In *Proc. ECAI Workshop on Recommender Systems* (2008), pp. 35–38.

[25] KUNEGIS, J., AND SCHMIDT, S. Collaborative filtering using electrical resistance network models with negative edges. In *Proc. Industrial Conf. on Data Mining* (2007), pp. 269–282.

[26] KUNEGIS, J., SCHMIDT, S., BAUCKHAGE, C., MEHLITZ, M., AND ALBAYRAK, S. Modeling collaborative similarity with the signed resistance distance kernel. In *Proc. European Conf. on Artificial Intelligence* (2008), pp. 261–265.

[27] KUNEGIS, J., SCHMIDT, S., LOMMATZSCH, A., AND LERNER, J. Spectral analysis of signed graphs for clustering, prediction and visualization. In *Proc. SIAM Int. Conf. on Data Mining* (2010), pp. 559–570.

[28] KWAK, H., LEE, C., PARK, H., AND MOON, S.

What is Twitter, a social network or a news media? In *Proc. Int. World Wide Web Conf.* (2010), pp. 591–600.

[29] LEY, M. The DBLP computer science bibliography: Evolution, research issues, perspectives. In *Proc. Int. Symposium on String Processing and Information Retrieval* (2002), pp. 1–10.

[30] LIBEN-NOWELL, D., AND KLEINBERG, J. The link prediction problem for social networks. In *Proc. Int. Conf. on Information and Knowledge Management* (2003), pp. 556–559.

[31] NEWMAN, M. E. J. The structure of scientific collaboration networks. *Proc. National Academy of Sciences 98*, 2 (2001), 404–409.

[32] ROSE, T., STEVENSON, M., AND WHITEHEAD, M. The Reuters corpus volume 1–from yesterday's news to tomorrow's language resources. In *Proc. Int. Conf. on Language Resources and Evaluation* (2002), pp. 29–31.

[33] SMOLA, A., AND KONDOR, R. Kernels and regularization on graphs. In *Proc. Conf. on Learning Theory and Kernel Machines* (2003), pp. 144–158.

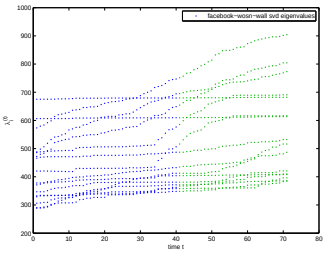
[34] SREBRO, N., AND JAAKKOLA, T. Weighted low-rank approximations. In *Proc. Int. Conf. on Machine Learning* (2003), pp. 720–727.

[35] STEWART, D. Social status in an open-source community. *American Sociological Review 70*, 5 (2005), 823–842.

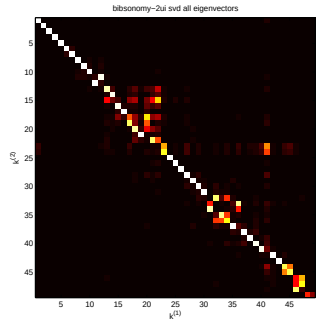
[36] WETZKER, R., ZIMMERMANN, C., AND BAUCKHAGE, C. Analyzing social bookmarking systems: A del.icio.us cookbook. In *Proc. Mining Social Data Workshop* (2008), pp. 26–30.

[37] WIKIMEDIA FOUNDATION. Wikimedia downloads. `http://download.wikimedia.org/`, January 2010.

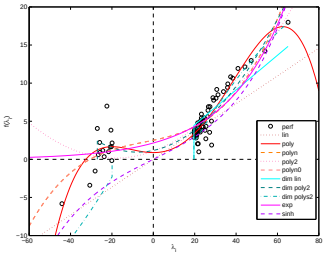
[38] ZIEGLER, C.-N., MCNEE, S. M., KONSTAN, J. A., AND LAUSEN, G. Improving recommendation lists through topic diversification. In *Proc. Int. World Wide Web Conf.* (2005), pp. 22–32.



(a) Spectral growth



(b) Eigenvector similarity



(c) Learning a spectral transformation

Figure 2: Link prediction methods: (a) Evolution of the singular values in the Facebook wall post network. The evolution of the spectrum on the right part (in green) can be extrapolated from the known spectral evolution on the left (in blue). (b) Cosine distance between current and expected eigenvectors in the BibSonomy user-item graph. (white = 1, black = 0) This shows that indeed most eigenvectors remain the same, although a few do not. This plot also shows that some eigenvalue *pass* others. (c) Curve fitting current and expected spectrum in the hep-th citation network, using the method described in [21].