Doctoral Forum and Student Travel Fellowship Application
SIAM Conference on Data Mining (SDM 2010)

Applicant's Name: Jérôme Kunegis

Ph.D. Advisor's Name: Prof. Sahin Albayrak

Department: DAI Laboratory (DAI-Labor)

Institution: Technische Universität Berlin

Contact Address:

Jérôme Kunegis
Technische Universität Berlin / DAI-Labor
Fakultät IV für Elektrotechnik und Informatik
Sekretariat TEL 14
Ernst-Reuter-Platz 7
D-10587 Berlin
Germany

e-mail: jerome.kunegis@dai-labor.de

Expected Date of Ph.D. Thesis Defense[*]**:** 2011-01-15

Ph.D. Thesis Topic: On the Spectral Evolution of Large Networks

Are you a US Citizen or Permanent Resident?[?] No

Gender[‡]: Male

Ethnicity: ----

- Do you have a paper accepted for presentation at SDM 2010? Yes (#116, see [3] below)
- Are you submitting an abstract for the Doctoral Forum (distinct from any paper accepted for inclusion in the SDM proceedings)? Yes

  If so, please be sure to include a 1-2page (11pt Times Roman font, 1 inch margins) abstract summarizing your dissertation research on page 2.

[*] Must be verified by the thesis advisor in the letter of recommendation

[?] This information is used to determine eligibility for specific sources of travel funds

[‡] Special consideration will be given to women and under-represented minorities.

[+] The requirements for advancing to Ph.D. candidacy vary from one institution to another, but typically involves passing an examination of the proposed Ph.D. thesis research proposal.

On the Spectral Evolution of Large Networks

PhD Thesis Abstract

Jérôme Kunegis

Technische Universität Berlin

Abstract. In my dissertation, I study the spectral characteristics 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. This argument is supported by the two following observations: (i) Link prediction in networks can be implemented by learning a transformation of the spectrum, changing the global structure of the network (modeling growth) by keeping the eigenvectors (retaining the identities of nodes). (ii) I introduce a random graph model parameterized by the network's spectrum, with randomly distributed eigenvectors. This random graph model retains the global structure of the original graph and discards any local structure, and can be interpreted as a multidimensional extension of the preferential attachment graph model.

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. Two approaches to studying large networks are link prediction and random graph models. The link prediction problem consists of predicting the location or weight of edges in a network, given past edges. Modeling networks corresponds to finding a random model that matches real-world graphs. These two problems are connected: Given a random graph model, a link prediction algorithm can be derived by assuming the graph will grow but keep its general characteristics, and a link prediction algorithm can lead to a random graph model by assuming that large networks grow from small networks following certain growth patterns.

I study both problems taking a spectral 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 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 the following applications:

- Links can be predicted by predicting spectral growth. I study two different methods for doing this: curve fitting and extrapolation.
- Networks can be characterized by their spectra and eigenvector component distributions, explaining other features such as degree distributions, and resulting in a multidimensional extension of the preferential attachment graph growth model.

Novelty of Work

This dissertation topic builds on a range of previous spectral graph mining and link prediction techniques. The novelty of the dissertation lies in the unified treatment of spectral transformations (making it possible to learn them), and in a justification in terms of local and global network properties. Specific novelties correspond to papers I have written or intend to write:

- A new dataset, the Slashdot Zoo, representing a social network with negative edges [1]
- A method to learn spectral transformations [2]
- A variant of the network Laplacian applicable to signed networks [3]
- An extrapolation method to predict links [4]

Datasets

During the time as a PhD student, I collected over 80 large network datasets, including Advogato, BibSonomy, CiteULike, DBLP, Delicious, Epinions, Facebook, Flickr, MovieLens, Netflix, Twitter, Wikipedia, YouTube and others. I crawled Slashdot.org to collect the "Slashdot Zoo" dataset, and published an analysis of it at WWW 2009 [1].

Methods

Link Prediction by Curve Fitting. In a paper at ICML 2009, I showed how the problem of estimating a parameterized graph kernel can be reduced to a one-dimensional curve fitting problem [2]. This model subsumes the exponential, von Neumann, Laplacian and heat diffusion kernels, and the methods of weighted path counting and rank reduction.

Extrapolation of Spectral Growth. In an unpublished paper submitted to KDD 2010, I developed a link prediction method based on extrapolating the spectral growth of a network over time [4]. This work is done in collaboration with Damien Fay (NUI Galway, Ireland).

Random Graph Models. While the spectra of networks are specific to the network in time, the eigenvectors follow a simpler pattern. Preliminary results suggest the following:

- All eigenvectors of a network follow the same component distribution up to signs.
- This distribution of eigenvector components is lognormal in many cases

Taken together, these lead to a lognormal degree distribution, confirming several previous studies (e.g. the DGX distribution). The overall degree distribution would then follow a mixture of lognormal distributions. Assuming network growth follows a spectral transformation results in a generalization of the preferential attachment model in which each node has a degree specific to a certain number of latent classes.

The work on random graph models is unfinished and unpublished. I intend to discuss the exact scope of this part of my dissertation at the SDM Doctoral Forum.

References

[1] The Slashdot Zoo: Mining a Social Network with Negative Edges, Kunegis et al., WWW 2009.
[2] Learning Spectral Graph Transformations for Link Prediction, Kunegis et al., ICML 2009.
[3] Spectral Analysis of Signed Graphs for Clustering, Prediction and Visualization, Kunegis et al., SDM 2010
[4] Network Growth and the Spectral Evolution Hypothesis, Kunegis et al., submitted to KDD 2010.