

Preface

This edited book contains extended versions of selected papers from ASONAM 2010 which was held at the University of Odense, Denmark, August 9–11, 2010. From the many excellent papers submitted to the conference, 28 were chosen for this volume. The volume explores a number of aspects of social networks, both global and local, and it also shows how social networks analysis and mining may aid web searches, product acceptances and personalized recommendations just to mention a few areas where social networks analysis can improve results in other mostly web-related areas. The application of graph theoretical aspects to social networks analysis is a recurrent theme in many of the chapters, and terminology from graph theory has influenced that of social networks to a large extent.

The theme of the book relates to the influence of technology on social networks and mining. This influence is not new. Technology is the enabling tool for all social networks except for the most trivial. Indeed without technology the only possible social networks would be extremely local and the cohesion of the network would simply have been by oral communication. Wider social networks only became a possibility with the advent of some sort of pictorial representation, for example, the technology of carving on stone. This meant that a message of some form could be read by others when the individual creating the representation was no longer present. Abstractions in the form of pictographs representing ideas and concepts and alphabets improved the technology. The advent of the movable print further sped up the technology. The printing press technology enabled a significant increase in speed for social network communication. These technologies were still limited in what could be disseminated both in time and space, however.

The advent of the electronic means of disseminating ideas and communications together with the development of the Internet opened up the possibility of transmitting ideas and to make connections with an essentially unlimited number of actors (people) with no geographical limitation at very low cost. This technological advance enabled the growth of social networks to sizes that could not be realized with previous technologies. The papers in this volume describe a number of aspects of this new ability to form such networks and they provide new tools and techniques for analyzing these networks effectively.

The first chapter is: *EgoClustering: Overlapping Community Detection via Merged Friendship-Groups* by Bradley S. Rees and Keith B. Gallagher. In this chapter, the authors identify communities through the identification of friendship groups where a friendship-group is a localized community as seen from an individual's perspective that allows him/her to belong to multiple communities. The basic tools of the chapter are those of graph theory. An algorithm has been developed that finds overlapping communities and identifies key members that bind communities together. The algorithm is applied to some standard social networks datasets. Detailed results from the Caveman and Zachary data sets are provided.

The chapter *Evolution of Online Forum Communities* by Mikolaj Morzy is a perfect example of a chapter discussing a theme relating the theme of the volume since the concept of an "online forum" did not exist prior to the current advances in technology. While one can trace the forum idea back to posters on bulletin boards and discussion in the printed literature, the current online forums are highly dependent on the speed and ease of transmission made possible by the Internet. The chapter discusses the evolution of these forums and their social implications. There are large number of forums and that are established that expand, contract, develop, and wither depending on the interest they generate. The paper introduces a micro-community-based model for measuring the evolution of Internet forums. It shows how the simple concept of a micro-community can be used to quantitatively assess the openness and durability of an Internet forum. The authors apply the model to a number of actual forums to experimentally verify the correctness and robustness of the model.

In *Integrating Online Social Network Analysis in Personalized Web Search* by Omair Shafiq, Tamer N. Jarada, Panagiotis Karampelas, Reda Alhajj, and Jon G. Rokne, the authors discuss how a web search experience can be improved through the mining of trusted information sources. From the content of the sources preferences are extracted that reorders the ranking of the results of a search engine. Search results for the same query raised by different users may differ in priority for individual users. For example a search for "The best pizza house" will clearly have a geographical component since the best pizza house in Miami is of no interest to someone searching for the best pizza in New York. It is also assumed that a query posed by a user correlates strongly with information in their social networks. To find the personal interest and social context, the paper therefore considers (1) the activities of users in their social network and (2) relevant information from a user's social networks, based on proposed trust and relevance matrices. The proposed solution has been implemented and tested.

The latent class models (LCMs) used in social science are applied in the context of social networks in *How Latent Class Models Matter to Social Network Analysis and Mining: Exploring the Emergence of Community* by Jaime R. S. Fonseca and Romana Xerez. The chapter discusses the advantages of reducing complex data to a limited number of typologies from a theoretical and empirical perspective. A relatively small dataset was obtained from surveying a community while using the notion of homophile to establish the survey criteria. The methodology is applied in the context of a three-latent class social network and the findings are in terms

of (1) network structure, (2) trust and reciprocity, (3) resources, (4) community engagement, (5) the Internet, and (6) years of residence.

In *Extending Social Network Analysis with Discourse Analysis: Combining Relational with Interpretive Data* by Christine Moser, Peter Groenewegen, and Marleen Huysman the authors investigate social networks that are related to specific interest groups such as Dutch Cake Bakers (DCB). These communities may be quite large (DCB had about 10,000 members at the time of writing the chapter) and they are characterized by a high level of activity; a strong, active, and small core; and an extensive peripheral group. They were able to gather very detailed and massive relational data from their example online communities from which they explored the connections within the communities. The authors then performed a discourse analysis on the content of the gathered messages and by this characterized the interactions in terms of we-them, compliments and empathy, competition and advice, and criticism, thus enabling a deeper understanding of the communities.

Viewing relational databases through their information content for social networks is the topic of the chapter *DB2SNA: An All-in-one Tool for Extraction and Aggregation of Underlying Social Networks from Relational Databases* by Rania Soussi, Etienne Cuvelier, Marie-Aude Aufaure, Amine Louati, and Yves Lechevallier. The authors propose a heterogeneous object graph extraction approach from a relational database which they use to extract a social network. This step is followed by an aggregation step in order to improve the visualization and analysis of the extracted social network. This is followed by an aggregation step using the k-SNAP algorithm which produces a summarized graph in order that the resulting social network graphs can be more easily understood.

The next chapter, *An Adaptive Framework for Discovery and Mining of User Profiles from SocialWeb-Based Interest Communities* by Nima Dokoochaki and Mihhail Matskin, introduces an adaptive framework for semi- to fully automatic discovery, acquisition, and mining of topic style interest profiles from openly accessible social web communities. Their techniques use machine learning tools including clustering and classifying for their algorithms. Three schemes are defined as follows: (1) depth-based, allowing for discovering and crawling of topics on a certain taxonomy tree-depth at each time; (2) n-split, allowing iterative discovery and crawling of all topics while at each iteration gathered data is split for n-times; and finally (3) greedy, which allows for discovery and crawling the network for all topics and processing the cached data. They apply the developed techniques to the social networking site LiveJournal.

The chapter *Enhancing Child Safety in MMOs* by Lyta Penna, Andrew Clark, and George Mohay considers the general issue of how the Internet can be made safe for children, specifically when Massively Multiplayer Online (MMO) games and environments are involved. A particular issue with respect to children and MMOs is the potential for luring a child into an off-line encounter which would in many cases present a hazard to a child. Typical message threads are analyzed for contextual content that might lead to such harmful encounters. The techniques developed to detect potentially unfavorable situations are applied to World of Warcraft as a case study. The chapter extends previous work by the authors.

Virtual communities are studied in *Towards Leader-Based Recommendations* by Ilham Esslimani, Armelle Brun, and Anne Boyer with the aim of discovering community leaders. These leaders influence the opinion and decision making of the rest of the community. Discovering these leaders is important, for example, in the area of marketing, where detecting opinion leaders allows the prediction of future decision making (about products and services), the anticipation of risks (due, e.g., to negative opinions of leaders) and the follow-up of the corporate image (e-reputation) of companies. Their algorithm considers the high connectivity and the potentiality of propagating accurate appreciations so as to detect reliable leaders through these networks. Furthermore, studying leadership is also relevant in other application areas, such as social network analysis and recommender systems.

Name and author disambiguation is an important topic for today's electronic article databases. For example, J. Smith, Jim Smith, J. Peter Smith may be (a) one author using different variations of his name Jim Smith, (b) two authors with variations in the use of their names, or (c) three authors. The chapter *Learning from the Past: An Analysis of Person Name Corrections in the DBLP Collection and Social Network Properties of Affected Entities* by Florian Reitz and Oliver Hoffmann tackles this problem for the DBLP bibliographic database of computer science and related topics. Given the name of an author, the intent is that the DLBP database will provide a list of papers by that author. Although there are a large number of algorithmic approaches to solve this problem, little is known on the properties of inconsistencies in the information in the databases such as variations of names of one individual. The present paper applies a historical and social network approach to the problem. Their algorithms are able to calculate the probability that a name will need correction in the future.

Factors Enabling Information Propagation in a Social Network Site by Matteo Magnani, Danilo Montesi, and Luca Rossi discusses the phenomenon that information propagates efficiently over social networks and that it is much more efficient than traditional media. Many general formal models of network propagation that might be applied to social network information dissemination have been developed in different research fields. This paper presents the result of an empirical study on a Large Social Database (LSD) aimed at measuring specific socio-technical factors enabling information spreading over social network sites.

In the chapter *Detecting Emergent Behavior in a Social Network of Agents* by Mohammad Moshirpour, Shimaa M. El-Sherif, Behrouz H. Far, and Reda Alhajj, the entities of the social networks are agents, that is, computer programs that exchange information with other computer programs and perform specific functions. In this chapter, there are agents handling queries, learning and managing concepts, annotating documents, finding peers, and resolving ties. The agents may work together to achieve certain goals, and certain behavior patterns may develop over time (emergent behavior). The chapter presents a case study of using a social network of a multiagent system for semantic search.

In *Detecting Communities in Massive Networks Efficiently with Flexible Resolution* by Qi Ye, Bin Wu, and Bai Wang the authors are concerned with data analysis on real-world networks. They consider an iterative heuristic approach to extract

the community structure in such networks. The approach is based on local multi-resolution modularity optimization and the time complexity is close to linear and the space complexity is linear. The resulting algorithm is very efficient, and it may enhance the ability to explore massive networks in real time.

The topic of the next chapter *Extraction of Spatio-temporal Data for Social Networks* by Judith Gelernter, Dong Cao, and Kathleen M. Carley is using social networks for the identification of locations and their association with people. This is then used to obtain a better understanding of group changes over time. The authors have therefore developed an algorithm to automatically accomplish the person-to-place mapping. It involves the identification of location and uses syntactic proximity of words in the text to link the location to a person's name. The contributions of this chapter include techniques to mine for location from text and social network edges as well as the use of the mined data to make spatiotemporal maps and to perform social network analysis.

The chapter *Clustering Social Networks Using Distance-Preserving Subgraphs* by Ronald Nussbaum, Abdol-Hossein Esfahanian, and Pang-Ning Tan considers cluster analysis in a social networks setting. The problem of not being able to define what a cluster is causes problems for cluster analysis in general; however, for the data sets representing social networks, there are some criteria that aid the clustering process. The authors use the tools of graph theory and the notion of distance preservation in subgraphs for the clustering process. A heuristic algorithm has been developed that finds distance-preserving subgraphs which are then merged to the best of the abilities of the algorithm. They apply the algorithm to explore the effect of alternative graph invariants on the process of community finding. Two datasets are explored: CiteSeer and Cora.

The chapter *Informative Value of Individual and Relational Data Compared Through Business-Oriented Community Detection* by Vincent Labatut and Jean-Michel Balasque deals with the issue of extracting data from an enterprise database. The chapter uses a small Turkish university as the background test case and develops algorithms dealing with aspects of the data gathered from students at the university. The authors perform group detection on single data items as well as pairs gathered from the student population and estimate groups separately using individual and relational data to obtain sets of clusters and communities. They then measure the overlap between clusters and communities, which turns out to be relatively weak. They also define a predictive model which allows them to identify the most discriminant attributes for the communities, and to reveal the presence of a tenuous link between the relational and individual data.

Considering the data from blogs in a social network context is the topic of *Cross-Domain Analysis of the Blogosphere for Trend Prediction* by Patrick Siehndel, Fabian Abel, Ernesto Diaz-Aviles, Nicola Henze, and Daniel Krause. The authors note first the importance of blogs for communicating information on the web. Blogging over advanced communications devices such as smartphones and other handheld devices has enabled blogging anywhere at any time. Because of this facility, the blogged information is up to date and a valuable source for data, especially for companies. Relevant date, extracted from blogs, can be used to adjust

marketing campaigns and advertisement. The authors have selected the music and movie domains as examples where there is a significant blogging activity and they used these domains to investigate how chatter from the blogosphere can be used to predict the success of products. In particular, they identify typical patterns of blogging behavior around the release of a product by analyzing the terms of posting relevant to the product, point out methods for extracting features from the blogosphere, and show that we can exploit these features to predict the monetary success of movies and music with high accuracy.

Betweenness computation is the topic of *Efficient Extraction of High-Betweenness Vertices from Heterogeneous Networks* by Wen Haw Chong, Wei Shan Belinda Toh, and Loo Nin Teow. The efficient computation of betweenness in a network is computationally expensive, yet it is often the set of vertices with high betweenness that is of key interest in a graph. The authors have developed a novel algorithm that efficiently returns the set of vertices with the highest betweenness. The convergence criterion for the algorithm is based on the membership stability of the high-betweenness set. They also show experimentally that the algorithm tends to perform better on networks with heterogeneous betweenness distributions. The authors have applied the algorithm developed to the real-world cases of Protein, Enron, Ticker, AS, and DBLP data.

Engagingness and Responsiveness Behavior Models on the Enron E-mail Network and their Application to E-mail Reply Order Prediction deals with user interactions in e-mail systems. The authors note that user behaviors affect the way e-mails are sent and replied. They therefore investigate user engagingness and responsiveness as two interaction behaviors that give us useful insights into how users e-mail one another. They classify e-mail users in two categories: engaging users and responsive users. They propose four model types based on e-mail, e-mail thread, e-mail sequence, and social cognitively. These models are used to quantify the engagingness and responsiveness of users, and the behaviors can be used as features in the e-mail reply order prediction task which predicts the e-mail reply order given an e-mail pair. Experiments show that engagingness and responsiveness behavior features are more useful than other non-behavior features in building a classifier for the e-mail reply order prediction task. An Enron data set is used to test the models developed.

In the chapter *Comparing and Visualizing the Social Spreading of Products on a Large Social Network* by Pål Roe Sundsøy, Johannes Bjelland, Geoffrey Canright, Kenth Engø-Monsen, and Rich Ling, the authors investigate how products and services adoption is propagated. By combining mobile traffic data and product adoption history from one of the markets for the telecom provider Telenor the social network among adopters is derived. They study and compare the evolution of adoption networks over time for several products: the iPhone handset, the Doro handset, the iPad 3G, and video telephony. It is shown how the structure of the adoption network changes over time and how it can be used to study the social effects of product diffusion. Supporting this, they find that the adoption probability increases with the number of adopting friends for all the products in the study. It is postulated that the strongest spreading of adoption takes place in the dense core

of the underlying network, and gives rise to a dominant LCC (largest connected component) in the adoption network, which they call the social network monster. This is supported by measuring the eigenvector centrality of the adopters. They postulate that the size of the monster is a good indicator for whether or not a product is going to “take off.”

The next chapter is *Virus Propagation Modeling in Facebook* by W. Fan and K. H. Yeung, where the authors model virus propagation in social networks using Facebook as a model. It is argued that the virus propagation models used for e-mail, IM, and P2P are not suitable for social networks services (SNS). Facebook provides an experimental platform for application developers and it also provides an opportunity for studying the spreading of viruses. The authors find that a virus will spread faster in the Facebook network if Facebook users spend more time on it. The simulations in the chapter are generated with the Barabasi-Albert (BA) scale-free model. This model is compared with some sampled Facebook networks. The results show that applying BA model in simulations will overestimate the number of infected users a little while still reflecting the trend of virus spreading.

The chapter *A Local Structure-Based Method for Nodes Clustering. Application to a Large Mobile Phone Social Network* by Alina Stoica and Zbigniew Smoreda and Christophe Prieur presents a method for describing how a node of a given graph is connected to a network. They also propose a method for grouping nodes into clusters based on the structure of the network in which they are embedded using the tools of graph theory and data mining. These methods are applied to a mobile phone communications network. The paper concludes with a typology of mobile phone users based on social network cluster, communication intensity, and age.

In the chapter *Building Expert Recommenders from E-mail-Based Personal Social Networks* by Veronica Rivera-Pelayo, Simone Braun, Uwe V. Riss, Hans Friedrich Witschel, and Bo Hu, the authors investigate how to identify knowledgeable individuals in organizations. In such organizations, it is generally necessary to collaborate with people in any organization, to establish interpersonal relationships, and to establish sources for knowledge about the organization and its activities. Contacting the right person is crucial for successfully accessing this knowledge. The authors use personal e-mail corpora as a source of information of a user since it contains rich information about all the people the user knows and their activities. Thus, an analysis of a person’s e-mails allows automatically constructing a realistic image of the surroundings of that person. They develop ExpertSN, a personalized Expert Recommender tool based on e-mail Data Mining and Social Network Analysis. ExpertSN constructs a personal social network from the e-mail corpus of a person by computing profiles including topics represented by keywords and other attributes.

The most common way of visualizing networks is by depicting the networks as graphs. In *Pixel-Oriented Network Visualization: Static Visualization of Change in Social Networks* by Klaus Stein, René Wegener, and Christoph Schlieder, the networks are described in a matrix form using pixels. They claim that their approach is more suitable for social networks than graph drawing since graph drawing results in a very cluttered image even for moderately sized social networks. Their technique

implements activity timelines that are folded to inner glyphs within each matrix cell. Users are ordered by similarity which allows to uncover interesting patterns. The visualization is exemplified using social networks based on corporate wikis.

The chapter *TweCoM: Topic and Context Mining from Twitter* by Luca Cagliero and Alessandro Fiori is concerned with knowledge discovery from user-generated content from social networks and online communities. Many different approaches have been devoted to addressing this issue. This chapter proposes the TweCoM (Tweet Context Miner) framework which entails the mining of relevant recurrences from the content and the context in which Twitter messages (i.e., tweets) are posted. The framework combines two main efforts: (1) the automatic generation of taxonomies from both post content and contextual features and (2) the extraction of hidden correlations by means of generalized association rule mining. In particular, relationships holding in context data provided by Twitter are exploited to automatically construct aggregation hierarchies over contextual features, while a hierarchical clustering algorithm is exploited to build a taxonomy over most relevant tweet content keywords. To counteract the excessive level of detail of the extracted information, conceptual aggregations (i.e., generalizations) of concepts hidden in the analyzed data are exploited in the association rule mining process. The extraction of generalized association rules allows discovering high-level recurrences by evaluating the extracted taxonomies. Experiments performed on real Twitter posts show the effectiveness and the efficiency of the proposed technique.

In the chapter *Application of Social Network Metrics to a Trust-Aware Collaborative Model for Generating Personalized User Recommendations* by Iraklis Varlamis, Magdalini Eirinaki, and Malamati Louta, the authors discuss trustworthiness of recommendations in social networks which discuss product placement and promotion. The authors note that community-based reputation can aid in assessing the trustworthiness of individual network participants. In order to better understand the properties of links, and the dynamics of social networks, they distinguish between permanent and transient links and in the latter case, they consider the link freshness. Moreover, they distinguish between the propagation of trust in a local level and the effect of global influence and compare suggestions provided by locally trusted or globally influential users. The dataset extended Epinions is used as a testbed to evaluate the techniques developed.

Optimization Techniques for Multiple Centrality Computations by Christian von der Weth, Klemens Böhm, and Christian Hütter applies optimization techniques to identify important nodes in a social network. The authors note that many types of data have a graph structure and that, in this context, by identifying central nodes, users can derive important information about the data. In the social network context, it can be used to find influential users and in a reputation system it can identify trustworthy users. Since centrality computation is expensive, performance is crucial. Optimization techniques for single centrality computations exist, but little attention so far has gone into the computation of several centrality measures in combination. In this chapter, the authors investigate how to efficiently compute several centrality measures at a time. They propose two new optimization techniques and demonstrate

their usefulness both theoretically as well as experimentally on synthetic and on real-world data sets.

Movie Rating Prediction with Matrix Factorization Algorithm by Ozan B. Fikir, İlker O. Yaz, and Tansel Özyer discusses a movie rating recommendation system. Recommendation systems is one of the research areas studied intensively in the last decades and several solutions have been elicited for problems in different recommendation domains. Recommendations may differ by content, collaborative filtering, or both. In this chapter, the authors propose an approach which utilizes matrix value factorization for predicting rating i by user j with the sub matrix as k -most similar items specific to user i for all users who rate all items. Previously predicted values are used for subsequent predictions and they investigate the accuracy of neighborhood methods by applying the method to the prizing of Netflix. They have considered both items and users relationships on Netflix dataset for predicting ratings. Here, they have followed different ordering strategies for predicting a sequence of unknown movie ratings and conducted several experiments.

Finally, we would like to mention the hard work of the individuals who have made this valuable edited volume possible. We also thank the authors who submitted revised chapters and the reviewers who produced detailed constructive reports which improved the quality of the papers. Various people from Springer as well deserve much credit for their help and support in all the issues related to publishing this book. In particular, we would like to thank Stephen Soehnen for his dedication, seriousness, and generous support in terms of time and effort. He answered our e-mails on time despite his busy schedule, even when he was traveling.

A number of organizations supported the project in various ways. We would like to mention the University of Odense, which hosted ASONAM 2010; the National Sciences and Reserch Council of Canada, which supported several of the editors financially through its granting program; the Joint Research Centre (JRC) of European Commission, which supported one of the editors from its Global Security and Crisis Management Unit.

Sogutozu Ankara, Turkey
Calgary, AB, Canada
Ispra, Italy
Leiden, The Netherlands

Tansel Özyer
Jon Rokne
Gerhard Wagner
Arno Reuser

The Influence of Technology on Social Network Analysis
and Mining

Özyer, T.; Rokne, J.; Wagner, G.; Reuser, A.H.P. (Eds.)

2013, XXIII, 643 p., Hardcover

ISBN: 978-3-7091-1345-5