Charu C. Aggarwal *Editor*

Social Network Data Analytics

Springer

Social Network Data Analytics

Charu C. Aggarwal
Editor

Social Network Data Analytics

Springer

Editor
Charu C. Aggarwal
IBM Thomas J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532, USA
charu@us.ibm.com

ISBN 978-1-4419-8461-6 e-ISBN 978-1-4419-8462-3
DOI 10.1007/978-1-4419-8462-3
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011922836

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Preface

Social networks have been studied fairly extensively over two decades in the general context of analyzing interactions between people, and determining the important structural patterns in such interactions. The trends in recent years have focussed on *online* social networks, in which the social network is enabled as an *internet application*. Some examples of such networks are *Facebook*, *LinkedIn* and *MySpace*. Such social networks have rapidly grown in popularity, because they are no longer constrained by the geographical limitations of a conventional social network in which interactions are defined in more conventional way such as face-to-face meetings, or personal friendships.

The infrastructure which is built around social networks can support a rich variety of data analytic applications such as search, text analysis, image analysis, and sensor applications. Furthermore, the analysis and evolution of the structure of the social network is also an interesting problem in of itself. While some of these problems are also encountered in the more conventional notion of social networks, many of the problems which relate to the data-analytic aspects of social networks are relevant only in the context of online social networks. Furthermore, online social networks allow for more efficient data collection on a large scale, and therefore, the computational challenges are far more significant.

A number of books have been written in recent years on the topic of social networks, though most of these books focus on the non-technological aspect, and consider social networks more generally in the context of relationships between individuals. Therefore, these books mostly focus on the social, structural, and cognitive aspects of the social network, and do not focus on the unique issues which arise in the context of the interplay between the structural and data-centric aspects of the network. For example, an online social network may contain various kinds of contents or media such as text, images, blogs or web pages. The ability to mine these rich sources of information in the context of a social network provides an unprecedented challenge and also an opportunity to determine useful and actionable information in a wide variety of fields such as marketing, social sciences, and defense. The volume of the data available is also a challenge in many cases because of storage and efficiency

constraints. This book provides a first comprehensive compendium on recent research on the *data-centric aspect* of social networks.

Research in the field of online social networks has seen a revival in the last ten years. The research in the field is now reaching a level of maturity where it is useful to create an organized set of chapters which describe the recent advancements in this field. This book contains a set of survey chapters on the different data analytic issues in online social networks. The chapters describe the different facets of the field in a comprehensive way. This creates an organized description of the significant body of research in the important and emerging field of online social networks.

Chapter 1

AN INTRODUCTION TO SOCIAL NETWORK DATA ANALYTICS

Charu C. Aggarwal
IBM T. J. Watson Research Center
Hawthorne, NY 10532
charu@us.ibm.com

Abstract The advent of online social networks has been one of the most exciting events in this decade. Many popular online social networks such as *Twitter*, *LinkedIn*, and *Facebook* have become increasingly popular. In addition, a number of multimedia networks such as *Flickr* have also seen an increasing level of popularity in recent years. Many such social networks are extremely rich in content, and they typically contain a tremendous amount of *content* and *linkage* data which can be leveraged for analysis. The linkage data is essentially the graph structure of the social network and the communications between entities; whereas the content data contains the text, images and other multimedia data in the network. The richness of this network provides unprecedented opportunities for data analytics in the context of social networks. This book provides a *data-centric view* of *online* social networks; a topic which has been missing from much of the literature. This chapter provides an overview of the key topics in this field, and their coverage in this book.

Keywords: Social Networks, Data Mining

1. Introduction

This chapter will provide an introduction of the topic of social networks, and the broad organization of this book. Social networks have become very popular in recent years because of the increasing proliferation and affordability of internet enabled devices such as personal computers, mobile devices and other more recent hardware innovations such as internet tablets. This is evidenced by the burgeoning popularity of many online social networks such as *Twitter*,

C. C. Aggarwal (ed.), *Social Network Data Analytics,*
DOI 10.1007/978-1-4419-8462-3_1,

Facebook and *LinkedIn*. Such social networks have lead to a tremendous explosion of network-centric data in a wide variety of scenarios. Social networks can be defined either in the context of systems such as *Facebook* which are explicitly designed for social interactions, or in terms of other sites such as *Flickr* which are designed for a different service such as content sharing, but which also allow an extensive level of social interaction.

In general, a social network is defined as a network of *interactions* or *relationships*, where the nodes consist of actors, and the edges consist of the relationships or interactions between these actors. A generalization of the idea of social networks is that of *information networks*, in which the nodes could comprise *either* actors or *entities*, and the edges denote the relationships between them. Clearly, the concept of social networks is not restricted to the specific case of an internet-based social network such as *Facebook*; the problem of social networking has been studied often in the field of sociology in terms of generic interactions between any group of actors. Such interactions may be in any conventional or non-conventional form, whether they be face-to-face interactions, telecommunication interactions, email interactions or postal mail interactions.

The conventional studies on social network analysis have generally not focussed on *online interactions*, and have historically preceded the advent and popularity of computers or the internet. A classic example of this is the study of Milgram [18] in the sixties (well before the invention of the internet), who hypothesized the likelihood that any pair of actors on the planet are separated by at most six degrees of separation. While such hypotheses have largely remained conjectures over the last few decades, the development of online social networks has made it possible to test such hypotheses at least in an online setting. This is also referred to as the *small world* phenomenon. This phenomenon was tested in the context of MSN messenger data, and it was shown in [16] that the average path length between two MSN messenger users is 6.6. This can be considered a verification of the widely known rule of "six degrees of separation" in (generic) social networks. Such examples are by no means unique; a wide variety of online data is now available which has been used to verify the truth of a host of other conjectures such[1] as that of *shrinking diameters*[15] or *preferential attachment*. In general, *the availability of massive amounts of data in an online setting has given a new impetus towards a scientific and statistically robust study of the field of social networks.*

[1]The shrinking diameter conjecture hypothesizes that the diameters of social networks shrink in spite of the addition of new nodes, because of an increase in the density of the underlying edges. The preferential attachment conjecture hypothesizes that new nodes and edges in the social networks are more likely to be attached to the dense regions of the network.

This data-centric impetus has lead to a significant amount of research, which has been unique in its statistical and computational focus in analyzing large amounts of online social network data. In many cases, the underlying insights are applicable to the conventional social network setting as well. Before discussing the research topics in more detail, we will briefly enumerate the different settings for social network analysis, and specifically distinguish between the conventional and non-conventional scenarios. Specifically, these different settings are as follows:

- The most classical definition of a social network is one which is based purely on human interactions. This is the classical study of social networks in the field of sociology. These studies have traditionally been conducted with painstaking and laborious methods for measuring interactions between entities by collecting the actual data about human interactions manually. An example is the six-degrees-of-separation experiment by Milgram [18], who used postal mail between participants in order to test whether two arbitrary actors could be connected by a chain of 6 edges with the use of locally chosen forwards of the mail. Such experiments are often hard to conduct in a completely satisfactory way, because the actors in such experiments may have response rates which cannot be cleanly modeled in terms of social interaction behavior. An example is the case of the Milgram experiment, in which the results have often been questioned [14] because of the low forward rate of the letters which never reached the target. Furthermore, such social experiments are often biased towards high status targets in order to ensure likelihood of logical forwards. However, these results have eventually been accepted at least from a qualitative perspective, even though the rule of six degrees may not be precisely correct, depending upon the nature of the network which is being studied. Nevertheless, the "small world phenomenon" definitely seems to be correct, since the diameters of most such networks are relatively small.

  The social analysis of such networks has also been modeled in the field of cognitive science, where the cognitive aspects of such interactions are utilized for analytical purposes. Much of the research in the traditional field of social networks has been conducted from this perspective. A number of books [7, 24, 25] provide an understanding of this perspective. However, this work does not discuss the data-centric issues which are common to *online and internet-enabled* social networks.

- A number of technological enablers such as telecommunications, electronic mail, and electronic chat messengers (such as Skype, Google Talk or MSN Messenger), can be considered an indirect form of social networks, because they are naturally modeled as communications between

different actors. One advantage of such applications is that the traces of the communication data are often available (subject to some privacy controls). This data can be used for extensive analysis of such social networks. Some examples are the extensive analysis on the Enron email data set [28], or the recent verification of the six degrees of separation conjecture in the context of the MSN messenger data in [16].

- In recent years, a number of sites have arisen explicitly in order to model the interactions between different actors. Some examples of such social networks are *Facebook*, *MySpace*, or *LinkedIn*. In addition, sites which are used for sharing online media content, such as *Flickr*, *Youtube* or *delicious*, can also be considered indirect forms of social networks, because they allow an extensive level of user interaction. In these cases, the interaction is centered around a specific service such as content-sharing; yet many fundamental principles of social networking apply. We note that such social networks are extremely *rich*, in that they contain a tremendous amount of content such as text, images, audio or video. Such content can be leveraged for a wide variety of purposes. In particular, the interaction between the links and content has provided impetus to a wide variety of mining applications. In addition, social media outlets provide a number of unique ways for users to interact with one another such as posting blogs, or tagging each other's images. Which such forms of interaction are indirect, they provide rich *content-based* knowledge which can be exploited for mining purposes. In recent years, it has even become possible to integrate *real-time sensor-based content* into dynamic social networks. This is because of the development of sensors, accelerometers, mobile devices and other GPS-enabled devices, which can be used in a social setting for providing a dynamic and interactive experience.

- Finally, a number of social networks can also be constructed from specific kinds of interactions in different communities. A classical example would be the scientific community in which bibliographic networks can be constructed from either co-authorship or citation data. These can be used in conjunction with the content of the publications in order to derive interesting trends and patterns about the underlying papers. We note that much of the analysis for the first case above applies to this as well, though a lot of data and content is available because of the way in which such documents networks are archived. A number of document collections and bibliographic networks are archived explicitly, and they can be used in conjunction with more principled data-centric techniques, because of the content which is available along with such networks.

While the results of this book may be applicable to all the different kinds of social networks, the specific focus is on the data-centric issues which arise in the context of online social networks. It is also important to understand that an online social network can be defined much more generally than an online site such as *Facebook*, *Twitter* or *LinkedIn* which are formally advertised as social networking sites. In fact, *any web-site or application which provides a social experience in the form of user-interactions* can be considered to be a form of social network. For example, media-sharing sites such as *Flickr*, *Youtube*, or *delicious* are formally not considered social networks; yet they allow for social interactions in the context of information exchange about the content being shared. Similarly, many mobile, web, and internet-driven chat applications also have social aspects embedded in them. Furthermore, many mobile applications such as *Google Latitude* allow the implicit embedding of sensor or GPS information, and this is used in order to enable user interactions. These are all novel forms of social networks, each of which brings with it a set of unique challenges for the purpose of analysis. Therefore, our definition of social networks is fairly broad, and many chapters will study aspects which are relevant to these alternative forms of social networks.

This chapter is organized as follows. In the next section, we discuss the main thrusts in the field of social networks. In section 3, we discuss the organization of the book and their relationships to these different thrusts. In section 4, we present the conclusions and related directions in the field.

2. Online Social Networks: Research Issues

The field of online social networks has seen a rapid revival in recent years. A key aspect of many of the online social networks is that they are *rich in data*, and provide unprecedented challenges and opportunities from the perspective of knowledge discovery and data mining. There are two primary kinds of data which are often analyzed in the context of social networks:

- **Linkage-based and Structural Analysis:** In linkage-based and structural analysis, we construct an analysis of the linkage behavior of the network in order to determine important nodes, communities, links, and evolving regions of the network. Such analysis provides a good overview of the global evolution behavior of the underlying network.

- **Adding Content-based Analysis:** Many social networks such as *Flickr*, *Message Networks*, and *Youtube* contain a tremendous amount of content which can be leveraged in order to improve the quality of the analysis. For example, a photograph sharing site such as *Flickr* contains a tremendous amount of text and image information in the form of user-tags and images. Similarly, blog networks, email networks and message boards contain text content which are linked to one another. It has been ob-

served that combining content-based analysis with linkage-based analysis provides more effective results in a wide variety of applications. For example, communities which are designed with text-content are much richer in terms of carrying information about the topical expertise of the underlying community.

The other main differences which arises in the context of social network algorithms is that between *dynamic analysis* and *static analysis*. In the case of static analysis, we assume that the social network changes slowly over time, and we perform an analysis of the whole network in batch mode over particular snapshots. Such is often the case for many networks such as bibliographic networks in which the new events in the network may happen only slowly over time. On the other hand, in the case of many networks such as instant messaging networks, interactions are continuously received over time at very large rate, which may lead to network streams. The analysis of such networks is much more challenging, and is a subject of recent research [2–5]. The temporal aspect of networks often arises in the context of dynamic and evolving scenarios. Many interesting temporal characteristics of networks can be determined such as evolving communities, interactions between entities and temporal events in the underlying network.

Dynamic networks also arise in the context of *mobile applications*, in which moving entities constantly interact with one another. Such dynamic networks arise in the context of moving entities which interact with one another over time. For example, many mobile phones are equipped with GPS receivers, which are exploited by the applications on these phones. A classical example of such an application is the *Google Latitude* application which is capable of keeping track of the locations of different users, and issuing alerts when a given user is in the vicinity. Such dynamic social networks can be modeled as dynamic graphs for which the edges change constantly over time. Such dynamic graphs lead to massive challenges in processing because of the extremely large number of connections between the entities which may need to be tracked simultaneously. In such cases, graph stream mining applications are required in order to perform effective online analysis. Such applications are typically required to be able to summarize the network structure of the data in real time and utilize it for a variety of mining applications. Some recent advances in this direction are discussed in [1, 3].

A number of important problems arise in the context of structural analysis of social networks. An important line of research is to try to understand and model the nature of very large online networks. For example, the verification of the *small world phenomenon*, *preferential attachment*, and other general structural dynamics has been a topic of great interest in recent years. Since a significantly larger amount of data is available for the case of online social networks, the verification of such structural properties is much more robust

in terms of statistical significance. For the first time, it has actually become possible to study these classical conjectures in the context of massive amounts of data.

The most well known structural problem in the context of social networks is that of *community detection*. The problem of community detection is closely related to the problem of finding structurally related groups in the network. These structurally related groups are referred to as *communities*. Some well known methods for community detection are discussed in [8, 9, 11, 19]. The problem of community detection arises both in a static setting in which the network changes slowly over time, as well as a dynamic setting in which the network structure evolves rapidly. While these problems have been studied in the traditional literature in the context of the problem of graph partitioning [11], social networks are significantly larger in size. Furthermore, a significant amount of content may be available for improving the effectiveness of community discovery. Such challenges are unique to the online scenario, and has lead to the development of a significant number of interesting algorithms.

Social networks can be viewed as a structure which enables the dissemination of information. This is direct corollary of its enabling of social interactions among individuals. The *analysis of the dynamics of such interaction* is a challenging problem in the field of social networks. For example, important news is propagated through the network with the use of the interactions between the different entities. A well known model for influence propagation may be found in [20]. The problem of influence analysis is very relevant in the context of social networks, especially in the context of determining the most influential members of the social network, who are most likely to propagate their influence to other entities in the social network [12]. The most influential members in the social network may be determined using flow models as in [12], or by using page rank style methods which determine the most well connected entities in the social network.

Finally, an important class of techniques is that of inferring links which are not yet known in the social networks. This problem is referred to as that of *link inference* [17]. The link prediction problem is useful for determining important *future* linkages in the underlying social network. Such future linkages provide an idea of the future relationships or missing relationships in the social network. Link prediction is also useful in a number of*adversarial applications* in which one does not fully know the linkages in an enemy or terrorist network, and uses automated data mining techniques in order to estimate the underlying links.

Many of the above mentioned applications can be greatly improved with the use of content information. For example, content can be associated with nodes in the community, which has been shown to greatly improve the quality of the clusters in the underlying network [26]. This is because the content informa-

tion in different parts of the network is often closely related to its structure; the combination of the two can provide useful information which cannot be obtained from either as a single entity. It has also been observed [10] that the use of content information can also improve the qualitative results on problems such as link inference. In general, the incorporation of content can improve the end result of a wide variety of inference problems in social and information networks. In the case of photograph and video sharing web sites such as *Flickr* and *YouTube*, the content can be very rich and can contain data of different types, such as text, audio or video. Such heterogeneous data requires the design of methods which can learn and analyze data with heterogeneous feature spaces. In some cases, it is also useful to design methods which can transfer knowledge from one space to another. This has lead to an increasing interest in the field of *transfer learning* which uses the implicit links creates by users (such as tags) in order to transfer knowledge [27] from one space to another. Such methods can be particularly useful when a significant amount of content is available for learning in some spaces, but only a scan amount of content is available for learning in others. In the next section, we will discuss how the different chapters of this book are organized in the context of the afore-mentioned topics.

3. Research Topics in Social Networks

This book is organized into several chapters based on the topics discussed above. This chapter will discuss the different topics in detail and their relationship to the corresponding chapters. The discussion of these topics in this section is organized in approximate order of the corresponding chapters. The broad organization of the chapters is as follows:

- The first set of chapters are based on structural analysis of social networks. These include methods for statistical analysis of networks, community detection, classification, evolution analysis, privacy-preserving data mining, link inference and visualization.

- The second set of chapter are focussed on content-based mining issues in social networks. We have included chapters on several different kinds of content: (a) General data mining with arbitrary kinds of data (b) Text mining in social networks (b) Multimedia mining in social networks (d) Sensor and stream mining in social networks.

  We discuss how these different kinds of content can be leveraged in order to make interesting and valuable inferences in social networks. The richness of the underlying content results in a number of interesting inferences which are not possible with the use of purely structural methods.

Next, we will discuss the individual chapters in the book in detail, and how they relate to the above themes:

Statistical Analysis of Social Networks: The work of Milgram [18] laid the foundation for a more extensive analysis of the structural properties of large scale networks. In chapter 2, we study the important statistical properties of "typical" social networks. Some interesting questions which are examined in the chapter are to explore how typical social networks look like, on a large scale. The connectivity behavior of the nodes is examined to see if most nodes have few connections, with several "hubs" or whether the degrees are more evenly distributed. The clustering behavior of the nodes in typical social networks is examined. Another issue which is examined are the typical temporal characteristics of social networks. For example, it is examined how the structure varies as the network grows. As the network evolves over time, new entities may be added to the network, though certain graph properties may continue to be retained in spite of this. The behavior and distribution of the connected components of the graph is also examined.

Random Walks and their Applications in Social Networks: Ranking is one of the most well known methods in web search. Starting with the well known page-rank algorithm [6] for ranking web documents, the broad principle can also be applied for searching and ranking entities and actors in social networks. The page-rank algorithm uses random walk techniques for the ranking process. The idea is that a random walk approach is used on the network in order to estimate the probability of visiting each node. This probability is estimated as the *page rank*. Clearly, nodes which are structurally well connected have a higher page-rank, and are also naturally of greater importance. Random walk techniques can also be used in order to personalize the page-rank computation process, by biasing the ranking towards particular kinds of nodes. In chapter 3, we present methods for leveraging random walk techniques for a variety of ranking applications in social networks.

Community Detection in Social Networks: One of the most important problems in the context of social network analysis is that of community detection. The community detection problem is closely related to that of clustering, and it attempts to determine regions of the network, which are dense in terms of the linkage behavior. The topic is related to the generic problem of graph-partitioning [13] which partitions the network into dense regions based on the linkage behavior. However, social networks are usually dynamic, and this leads to some unique issues from a community detection point of view. In some cases, it is also possible to integrate the content behavior into the community detection process. In such cases, the content may be leveraged in order to de-

termine groups of actors with similar interests. Chapter 4 provides an overview of some of the important algorithms on the problem of community detection.

Node Classification in Social Networks: In many applications, some of the nodes in the social network may be labeled, and it may be desirable to use the attribute and structural information in the social network in order to propagate these labels. For example, in in a marketing application, certain nodes may be known to be interested in a particular product, and it may be desirable to use the attribute and structural information in the network in order to learn other nodes which may also be interested in the same product. Social networks also contain rich information about the content and structure of the network, which may be leveraged for this purpose. For example, when two nodes in a social network are linked together, it is likely that the node labels are correlated as well. Therefore, the linkage structure can be used in order to *propagate* the labels among the different nodes. Content and attributes can be used in order to further improve the quality of classification. Chapter 5 discusses a variety of methods for link-based node classification in social networks.

Evolution in Dynamic Social Networks: Social Networks are inherently dynamic entities; new members join them, old members stop participating, new links emerge as new contacts are built, and old links become obsolete as the members stop interacting with one other. This leads to changes in the structure of the social network as a whole and of the communities in it. Two important questions arise in this context: (a) What are the laws which govern long term changes in the social network over time, which are frequently observed over large classes of social networks? (b) How does a community inside a social platform evolve over time? What changes can occur, and how do we capture and present them?
Chapter 6 elaborates on these questions in more detail. Advances associated with the first set of questions are studied in Chapter 2, and to a lesser extent in Chapter 6. Advances on the second question are studied in Chapter 6, where the main focus is on evolution in social networks.

Social Influence Analysis: Since social networks are primarily designed on the basis of the interactions between the different participants, it is natural that such interactions may lead to the different actors influencing one another in terms of their behavior. A classic example of this would be a viral marketing application in which we utilize the messages between interconnected participants in a social network in order to propagate the information across different parts of the network. A number of natural questions arise in this context:
(a) How do we model the nature of the influence across actors?
(b) How do we model the spread of influence?

(c) Who are the most influential actors for influence spread?
Chapter 7 studies these issues in considerable depth and provides a deep understanding of the nature of influence analysis in social networks.

Expert Discovery in Networks: Social networks can be used as a tool in order to identify experts for a particular task. For example, given the activities of candidates within a context (*e.g.*, authoring a document, answering a question), we first describe methods for evaluating the level of expertise for each of them. Often, experts are organized in networks that correspond to social networks or organizational structures of companies. Many complex tasks often require the collective expertise of more than one expert. In such cases, it is more realistic to require a team of experts that can collaborate towards a common goal. Chapter 8 discusses methods for determining teams of experts which can perform particular tasks.

Link Prediction in Social Networks: Much of the research in mining social networks is focussed on *using the links* in order to derive interesting information about the social network such as the underlying communities, or labeling the nodes with class labels. However, in most social networking applications, the links are dynamic, and may change considerably over time. For example, in a social network, friendship links are continuously created over time. Therefore, a natural question is to determine or predict future links in the social network. The prediction process may use either the structure of the network or the attribute-information at the different nodes. A variety of structural and relational models have been proposed in the literature for link prediction [17, 21–23]. Chapter 9 provides a detailed survey of such methods.

Privacy in Social Networks: Social networks contain a tremendous information about the individual in terms of their interests, demographic information, friendship link information, and other attributes. This can lead to disclosure of different kinds of information in the social network, such as identity disclosure, attribute disclosure, and linkage information disclosure. Chapter 10 discusses a detailed survey of privacy mechanisms in social networks in context of different kinds of models and different kinds of information which can be disclosed.

Visualizing Social Networks: As social networks become larger and more complex, reasoning about social dynamics via simple statistics is cumbersome, and not very intuitive. Visualization provides a natural way to summarize the information in order to make it much easier to understand. Recent years have witnessed a convergence of social network analytics and visualization, coupled with interaction, that is changing the way analysts understand and character-

ize social networks. In chapter 11, the main goal of visualization is discussed in the context of user understanding and interaction. The chapter also examines how different metaphors are aimed towards elucidating different aspects of social networks, such as structure and semantics. A number of methods are described, where analytics and visualization are interwoven towards providing a better comprehension of social structure and dynamics.

Data Mining in Social Media: Social Media provides a wealth of social network data, which can be mined in order to discover useful business applications. Data mining techniques provide researchers and practitioners the tools needed to analyze large, complex, and frequently changing social media data. An overview on the topic of data mining in social media is provided in Chapter 12. This chapter introduces the basics of data mining in the context of social media, and discusses how to mine social media data. The chapter also highlights a number of illustrative examples with an emphasis on social networking sites and blogs.

Text Mining in Social Networks: Social networks contain a lot of text in the nodes in various forms. For example, social networks may contain links to posts, blogs or other news articles. In some cases, users may tag one another, which is also a form of text data on the links. The use of content can greatly enhance the quality of the inferences which may be made in the context of graphs and social networks. In chapter 13, we present methods for using text mining techniques in social networks in the context of a variety of problems such as clustering and classification.

Integrating Sensors and Social Networks: Many mobile phones provide the ability for actors to interact with one another dynamically, and in real time depending upon their location and status. Such applications also result in the generation of massive streams in real time, which can be used to make inferences about one another, or about the aggregate properties of the objects which are being tracked. Since location information is private, this also naturally leads to a number of privacy concerns from a processing perspective. Chapter 14 discusses such methods for incorporating sensor data as an integral part of social network data analytics.

Multimedia Information Network Analysis in Social Media: Many forms of media sharing sites such as *Flickr* and *Youtube* provide the ability to share media. Such shared media are often used in conjunction with the interactions of different users, such as the placing of tags or comments on the different images. Such rich context-based information networks can be mined for a wide variety of applications by leveraging the combination of user tags and image

data in the mining and retrieval process. Chapter 15 discusses methods for mining multimedia information networks with social media.

Social Tagging: Much of the interaction between users and social networks occurs in the form of *tagging*, in which users attach short descriptions to different objects in the social network, such as images, text, video or other multimedia data. Chapter 16 provides a detailed survey of various aspects of tagging. The chapter discusses properties of tag streams, tagging models, tag semantics, generating recommendations using tags, visualizations of tags, applications of tags, integration of different tagging systems and problems associated with tagging usage. Many interesting issues are discussed, such as the reason why people tag, what influences the choice of tags, how to model the tagging process, kinds of tags, different power laws observed in tagging domain, how tags are created and how to choose the right tags for recommendation.

4. Conclusions and Future Directions

This book is primarily focussed on providing readers with an introduction to the area of social networks. The broad area is so vast, that it is probably not possible to cover it comprehensively in a single book. The problem of social network data analytics is still in its infancy; there is a tremendous amount of work to be done, particularly in the area of content-based and temporal social networks. Some key research directions for the future are as follows:

- **Content-based Analysis:** Much of the past research in this area has been based on *structural analysis* of social networks. Such analysis primarily uses linkage structure only in order to infer interesting characteristics of the underlying network. Some recent research [26] has shown that the inclusion of content information can yield valuable insights about the underlying social network. For example, the content at a given node may provide more information about the expertise and interests of the corresponding actor.

- **Temporal Analysis:** Most of the research in social networks is based on static networks. However, a number of recent studies [8, 9, 11] have shown that the incorporation of temporal evolution into network analysis significantly improves the quality of the results. Therefore, a significant amount of work remains to be done on dynamic analysis of social networks which evolve rapidly over time.

- **Adversarial Networks:** In adversarial networks, it is desirable to determine the analytical structure of a network in which the actors in the network are adversaries, and the relationships among the different adversaries may not be fully known. For example, terrorist networks would

be a typical adversarial network to a law enforcement agency. Such networks are far more challenging because the links may not be known a-priori, but may need to be inferred in many cases. Such inferred links may need to be used for analytical purposes.

In addition, we expect that it will be increasingly important to analyze networks in the context of heterogeneous data, in which the links are of different types and correspond to different kinds of relationships between the actors. A generalization of the concept of social networks is that of *information networks*, in which the nodes could be either actors of entities, and the edges correspond to logical relations among these entities. Such networks are also heterogeneous, and therefore it is increasingly important to design tools and techniques which can effectively analyze heterogeneous networks.

References

[1] C. C. Aggarwal, H. Wang. Managing and Mining Graph Data, *Springer*, 2010.

[2] C. C. Aggarwal. P. Yu. Online Analysis of Community Evolution over Data Streams, *SIAM Conference on Data Mining*, 2005.

[3] C. C. Aggarwal, Y. Zhao, P. Yu. On Clustering Graph Streams, *SIAM Conference on Data Mining*, 2010.

[4] C. C. Aggarwal, Y. Li, P. Yu, R. Jin. On Dense Pattern Mining in Graph Streams, *VLDB Conference*, 2010.

[5] C. C. Aggarwal, Y. Zhao, P. Yu. Outlier Detection in Graph Streams, *ICDE Conference*, 2011.

[6] S. Brin, L. Page. The Anatomy of a Large Scale Hypertextual Search Engine, *WWW Conference*, 1998.

[7] P. J. Carrington, J. Scott, S. Wasserman. *Models and Methods in Social Network Analysis (Structural Analysis in the Social Sciences)*, Cambridge University Press, 2005.

[8] D. Chakrabarti, R. Kumar, A. Tomkins. Evolutionary Clustering, *ACM KDD Conference*, 2000.

[9] Y. Chi, X. Song, D. Zhou, K. Hino, B. L. Tseng. Evolutionary spectral clustering by incorporating temporal smoothness. *KDD Conference*, 2007.

[10] L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models of relational structure. In *ICML*, pages 170–177, 2001.

[11] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, B. L. Tseng. FacetNet: A framework for analyzing communities and their evolutions in dynamic networks. *WWW Conference*, 2008.

[12] D. Kempe, J. Kleinberg, E. Tardos. Maximizing the Spread of Infuence in a Social Network, *ACM KDD Conference*, 2003.

[13] B. W. Kernighan, S. Lin, An efficient heuristic procedure for partitioning graphs, *Bell System Technical Journal*, 1970.

[14] J. Kleinberg. Complex Networks and Decentralized Search Algorithms, *Proceedings of the International Congress on Mathematics*, 2006.

[15] J. Leskovec, J. Kleinberg, C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. *ACM KDD Conference*, 2005.

[16] J. Leskovec, E. Horvitz. Planetary-Scale Views on a Large Instant-Messaging Network, *WWW Conference*, 2008.

[17] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *LinkKDD*, 2004.

[18] S. Milgram. The Small World Problem, *Psychology Today*, 1967.

[19] M. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 2006.

[20] M. E. J. Newman. The spread of epidemic disease on networks, *Phys. Rev*. E 66, 016128, 2002.

[21] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. *UAI*, pages 485–492, 2002.

[22] B. Taskar, M. F. Wong, P. Abbeel, and D. Koller. Link prediction in relational data. *NIPS*, 2003.

[23] C. Wang, V. Satuluri, and S. Parthasarathy. Local probabilistic models for link prediction. *ICDM*, pages 322–331, 2007.

[24] S. Wasserman, K. Faust. *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*, Cambridge University Press, 1994.

[25] D. J. Watts. Six Degrees: The Science of a Connected Age, *W. W. Norton and Company*, 2004.

[26] Y. Zhou, H. Cheng, and J. X. Yu. Graph clustering based on structural/attribute similarities. *PVLDB*, 2(1):718–729, 2009.

[27] Y. Zhu, S. J. Pan, Y. Chen, G.-R. Xue, Q. Yang, Y. Yu. Heterogeneous Transfer Learning for Image Classification. *AAAI*, 2010.

[28] http://www.cs.cmu.edu/~enron

Chapter 2

STATISTICAL PROPERTIES OF SOCIAL NETWORKS

Mary McGlohon
School of Computer Science
Carnegie Mellon University
mmcgloho@cs.cmu.edu

Leman Akoglu
School of Computer Science
Carnegie Mellon University
lakoglu@cs.cmu.edu

Christos Faloutsos
School of Computer Science
Carnegie Mellon University
christos@cs.cmu.edu

Abstract In this chapter we describe patterns that occur in the structure of social networks, represented as graphs. We describe two main classes of properties, static properties, or properties describing the structure of snapshots of graphs; and dynamic properties, properties describing how the structure evolves over time. These properties may be for unweighted or weighted graphs, where weights may represent multi-edges (e.g. multiple phone calls from one person to another), or edge weights (e.g. monetary amounts between a donor and a recipient in a political donation network).

Keywords: Power laws, network structure, weighted graphs

What do social networks look like on a global scale? How do they evolve over time? How do the different components of an entire network form? What

C. C. Aggarwal (ed.), *Social Network Data Analytics,*
DOI 10.1007/978-1-4419-8462-3_2,