Green Computing Towards Green IT
Mrs. I. Faritha Beevi,
Assistant Professor
Sadakathullah Appa College
Palayamkottai – 627 011 India.

A SURVEY OF GRAPH MINING IN SOCIAL NETWORK ANALYSIS

Abstract

Graph mining is an important research area within the domain of data mining. Graph mining has become a popular area of research in recent years because of its numerous applications in a wide variety of practical fields, including computational biology, sociology, software bug localization, keyword search, and computer networking. Graphs have been used for modeling entities and their relationships such as the Internet, the web, social networks, metabolic networks, protein interaction networks, food webs, citation networks, and many more. In this paper we present a survey of Social Network applications in Graph Mining. These are used to extract patterns, trends, classes, and clusters from graphs.

Keywords: Data Mining, Graph Mining, Social Networks

1. INTRODUCTION

Graph Mining

Graphs become increasingly important in modeling complicated structures, such as circuits, images, chemical compounds, protein structures, biological networks, social networks, the Web, workflows, and XML documents. Many graph search algorithms have been developed in chemical informatics, computer vision, video indexing, and text retrieval. With the increasing demand on the analysis of large amounts of structured data, graph mining has become an active and important theme in data mining. Among the various kinds of graph patterns, frequent substructures are the very basic patterns that can be discovered in a collection of graphs. They are useful for characterizing graph sets, discriminating different groups of graphs, classifying and clustering graphs, building graph indices, and facilitating similarity search in graph databases. Recent studies have developed several graph mining methods and applied them to the discovery of interesting patterns in various applications. For example, there have been reports on the

discovery of active chemical structures in HIV-screening datasets by contrasting the support of frequent graphs between different classes. There have been studies on the use of frequent structures as features to classify chemical compounds, on the frequent graph mining technique to study protein structural families, on the detection of considerably large frequent subpathways in metabolic networks, and on the use of frequent graph patterns for graph indexing and similarity search in graph databases. Although graph mining may include mining frequent subgraph patterns, graph classification, clustering, and other analysis tasks, in this section we focus on mining frequent subgraphs. We look at various methods, their extensions, and applications.

2. Literature Review

2.1 Social Network Analysis

The rise of social network analysis in the field of computer science owes much to the spread of the Internet and the increasing reliance upon this medium for communication. Email alone has increased in importance from being a feature to a necessity to a critical business tool over the last 15 years.Against this background it has become increasingly obvious that the use of email networks on the one hand and the use of similarly linked structures in the Internet can be analyzed with the help of social network analysis. Among the number of applicable data available one must count

- Email networks
- The World Wide Web link structure
- Peer-to-Peer networks
- Information flow networks
- Recommender networks
- Client maintained content networks, User Groups

These are just a few of the structured networks found today. Another impulse has been given, after which research has been geared towards the specific problem of detecting small groups within large networks, such as detecting terrorist cells and classification by association as well as link prediction. The spread of malware of viral attacks in computer networks has also been a point of research in recent

years. The field of virtual epidemiology takes some pointers from biology and treats the spread of viral attacks like a biological contagion. Interestingly, the spread of a computer virus often coincides with the spread of information in social networks. Generally the robustness and strength of networks with respect to electronic attack has been an interest in recent years. This ranges as far as the formal treatment of information warfare and the use of network analysis in intrusion detection.

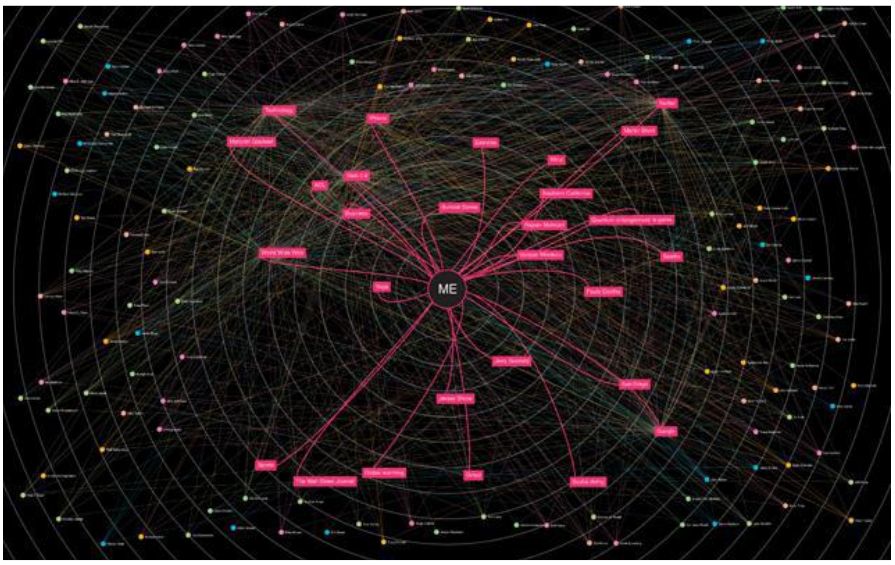

Fig: Social Network Analysis

2.1.1 Properties

A social network has a few central properties revealing information about the structure of the network itself, and the members of the network. Properties of complex networks and random networks have great impact upon the analysis thereof. Especially a search performed in a complex network is greatly dependant on the completeness of the data at hand.

Actor	Cent	. Position
Andy Zipper	0.088335	Vice President Enron Online
Louise Kitchen	0.078280	Founder Enron Online
John Lavorato	0.076380	CEO of Enron
Barry Tycholiz	0.071972	Vice President

Tana Jones	0.055695	N/A
Mike Grigsby	0.047822	director of corporate strategy and Development.
Richard Sanders	0.047530	Vice President and Assistant. General Counsel for Enron Wholesale Services,
Geoff Storey	0.045056	N/A
David Delainey	0.042674	CEO, Enron Energy Services
Michelle Cash	0.037957	Assistant

Table 2.1: Centrality in the Enron email network

Centrality

The centrality of an actor within a network attempts to capture the importance with respect to how much this actor is involved in the communications activity. This centrality can be defined in various ways. Table shows a sample listing of the centrality values of actors in the Enron network

• In/Out link

Basically a measure based on the number of links going in as well as out from an actor. The more links he has, the more connected and embedded within the network he is. While it works well on small networks, it is intrinsically local and takes no account of transitivity in the graph.

• Random Walk

An idea put forward to alleviate the necessity of combing through the entire graph, it proposes to measure centrality as a function of how many random walks through the graph hit a given actor.

• Between ness

Probably the most common measure of centrality implemented deems an actor to be central to the network in direct relation to how many paths it lies in between all other pairs of nodes.

• Rank

It can be taken as an extension of the in/out link centrality idea, only encompassing transitivity. Thus indirect communication to nodes further away in the network is also considered to add to the centrality of an actor.

• Eigenvalue

As a graph can be represented by an adjacency matrix, and an adjacency matrix can yield real Eigenvectors. This is the case when the communication is encoded as imaginary values in the matrix, as well as being hermitian. The resulting (real) Eigenvectors can be interpreted as the main conversations, with actors ranked by importance within a conversation.

Prestige

Another important aspect has been the recognition of the level of influence a member of the group has on the whole, designated as prestige. While centrality can be seen as a measure capturing how involved an actor is in a network, prestige aims to capture his importance.

• In-Link

Contrary to the measure of centrality mentioned above, this measure aim to capture how many actors refer to a given actor as a measure to how much he is sought after. The more in-links, the more prestige (or perhaps expertise) he has.

Similarly, the determination of prestige under consideration of transitivity yields the Rank prestige measure. As such it is more comprehensive and sensitive than the localized version above. Considerations of prestige have been given increased thought in recent years, especially when considering how to increase influence through the maximization of rank in a network .This basically represents the drive for importance found in much blogging and Internet based communication behavior today.

Graph Size

The intention under which social network analysis has been developed has been the use and observation of a distinct group of limited size. This has several implications which will become central

to this thesis. Especially the detection of denser sub-graphs within a network has gained relevance in recent years.

Cliques

The definition and subsequent detection of subgroups within a group of people has become known as clique detection. The current methodologies focus on discovering complete or nearly complete sub-graphs within a network. The definition off a clique originates from the observation of small subgroups within a population in which every member is in contact with every other member. These complete sub-graphs form a core group of social interaction of interest to the research done. Today the intention behind the definition of a clique can be seen somewhat more relaxed, as the completeness of a sub-graph is often more a hindrance than a help. The broadening of clique to include actors in contact with each other on a more intense level than the graph at large falls more into line with communication patterns observed today. Another point is the fact that the size of typical networks has increased to the point of diluting the original intention of social network analysis. It has become necessary to either change network analysis, or preprocess the data to restore social network analysis meaning and functionality.

2.2 An Algorithmic Perspective

So what does this mean from an algorithmic perspective? Generally all operations are performed on adjacency matrices, thus framing the dimension to be considered in the efficiency calculations. A number of approaches have been put forward to deal with adjacency matrices efficiently and effectively, such as the detection of networks in blogging data, or the treatment of social networks in conjunction with content.

2.2.1 Graph Mining

The subject of graph mining, while relevant in this context, will be treated in. Nonetheless it has to be mentioned that there is a great deal of research done to extract sub-graphs from massive graphs. These approaches all bear on the subsequent analysis using social network analysis.

SNA by Social Scientists

The approaches used by social scientists are based on methodologies developed as far back as the mid- to late seventies, and as such are poorly adapted to the computational power available today.

• Block Models are used in the analysis of social networks when the individual actors are combined into discrete subgroups. The subgroups are then linked with one or more link types, each expressing a different relationship. This approach aims to provide a slightly more generalized form of social network analysis based on groups rather than individual actors.

• Positional Analysis can be said to relate to the previously mentioned Block Models. In it the position of an actor within a network as defined by the in- and/or out-degree describes a role within the network.

• Relational Equivalence

It aims to group actors by their role within a network. Actors with equivalent

Structural topologies are assumed to posse's similar role within a social network. For the course of this thesis we will focus on the aspects of centrality and prestige in social networks. The extension to other aspects of social network analysis is an area to be explored by future work.

Centrality/Prestige algorithms

Eigenvector Approaches

The eigenvector approach is an effort to find the most central actors (i.e. those with the smallest farness from others) in terms of the "global" or "overall" structure of the network, and to pay less attention to patterns that are more "local." The method used to do this (factor analysis) is beyond the scope of the current text. In a general way, what factor analysis does is to identify "dimensions" of the distances among actors. The location of each actor with respect to each dimension is called an "eigenvalue," and the collection of such values is called the "eigenvector." Usually, the first dimension captures the "global" aspects of distances among actors; second and further dimensions capture more specific and local sub-structures.

3. Limits of Social Network Analysis

This thesis has focused on the problems posed by social network analysis of massive communications networks. Not only does the expressiveness of social network analysis decrease with

increasing graph size, but the larger and more varied the graph the more ambiguous are the conclusions drawn from analysis of important actors within them. This thesis touches upon several disparate fields of computer science, on the one hand using graph theory. By partitioning massive communication graphs, we are able to instill a more focused meaning upon social network analysis of relevant sub-graphs. This has been followed up in two main directions, on the one hand by extracting sub-networks determined by common topics and possibly tracking such topic evolution over time, and on the other hand by focusing on the structure of the network and extracting network leaders or dense communication sub-networks.

3.1 Integrating Content into Social Network Analysis

We have taken the inherent assumption of social network analysis, namely the disregard of content when judging centrality and prestige of actors in a social network, into consideration when analyzing massive graphs. The idea of calculating the centrality of a large, semantically not homogeneous network dilutes the expressiveness of social network analysis to the degree of generality. This has lead us to take content into account when performing social network analysis. By focusing on the topics contained in communication networks, we can pull topic based sub-networks out of the massive graph. Such a topic based network, when treated to social network analysis, now describes centrality and prestige in a much more precise setting than the entire graph. Whereas the result is not as pronounced as in the Enron data set, we have found relevant cohesive sub networks in both cases. The resulting increased expressiveness of topic based centrality has consequently yielded a better understanding of the communications taking place within the complete communication networks.

3.1.1 Interpreting Sub-graphs

Furthermore the use of Eigenvector based social network analysis can also lead to the discovery of relevant topics and central actors. These dynamics would have applications for example in detecting shifts of interest or the rise of new ’buzz words’ in a given community. Thus the ’spin-doctors’ and the ’pet-topics’ would become visible. This could be relevant for example for the development of innovative products and services if such an analysis could be performed to ’keep an eye’ on topics in R&D communities. As a conclusion it can be stated that the method proposed can yield interesting

insights into the 'mechanics' and 'dynamics' of communicating groups as could be shown in this feasibility study. To gain valid insights from the combination of the analysis of the communication behavior with content analysis it is necessary to perform a considerable effort in data preprocessing such as natural language processing within a concept tailored to the goals of the analysis and the linguistic specifics of the text corpus. Especially the dynamics of actor-to-actor Communication and the dynamics of use-of-words make this necessary but worthwhile.

3.1.2 Analyzing Real World Data

The implementation of the algorithms and user interfaces which have made the analysis possible have been to a large degree performed with an eye to better understand the details of the data mining and social network analysis process. To this end it is expected for the framework to be expanded and amended in the future, always being guided by the future work and research to be done in this area of research. But precisely because the goal of this implementation has been the academic understanding, it is not planned to introduce this software suite into commercial use. While parts of the analysis framework presented in this thesis will continue to fulfill an important role in preliminary research, especially when considering social network analysis, the functionality with regards to other areas of research will most likely be included from additional sources. The data sets in question were quite large and complex, but every set of real world data presents its own caveats. We ran into numerous instances which necessitated particular tuning of our approach. This indicates a great deal of exploration of the quirks must be made if a large-scale application should be developed from these approaches.

4. Conclusion

The application of social network analysis to graphs found in the World Wide Web and the Internet has received increasing attention in recent years. Networks as diverse as those generated by e-mail communication, instant messaging, and link structure in the Internet as well as citation and collaboration networks have all been treated with this method. So far these analyses solely utilize graph structure.. The more people partake in email exchanges, and the larger the Internet communities grow, the more traffic is generated. Networks under scrutiny today are increasing in size and

complexity, being often an amalgamation and super-positioning of multiple networks. These massive communications corpora can be captured by communications graphs, which are increasingly difficult to the generalized application of social network analysis.

We propose to apply the field of content analysis to the process of social network analysis. By extracting relevant and cohesive sub-networks from massive graphs, we obtain information on the actors contained in such sub-networks to a much firmer degree than before. While social network analysis can extract the relevance and impact of actors within a network from structural considerations, it does not regard content of the network under scrutiny. Thus the consistency of the network has a great impact upon the interpretability of the results of social network analysis. But as networks get larger, the more general and imprecise are the conclusions drawn from social network analysis. Google will score a very high centrality, based solely on the fact it links to almost everything. Therefore the intention of the analysis methods does not fall into line with the expected data found in real world problems. We believe the combination of content with social network analysis will greatly improve our understanding of real-world communication.

5. Future Work

The idea of topic centrality and topic prestige we propose in this chapter can go a step further in order to analyze the implications of social network analysis when actors have centrality values in multiple topic networks. In this case an actor can act as a connector between various topics, by constituting a knowledge or information bridge between discrete topics and their associated actors.

References:

[1] vorgelegt von Maximilian Viermetz “Partitioning Massive Graphs for Content Oriented Social Network Analysis” aus London, U.K.2008.

[2] Vijender Singh, Deepak Garg “Survey of Finding Frequent Patterns in Graph Mining: Algorithms and Techniques” International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-3, July 2011.

[3] K.Lakshmi1 and Dr. T. Meyyappan2 “FREQUENT SUBGRAPH MINING ALGORITHMS - A SURVEY AND FRAMEWORK FOR CLASSIFICATION “Natarajan Meghanathan, et al. (Eds): ITCS, SIP, JSE-2012, CS & IT 04, pp. 189–202, 2012. © CS & IT-CSCP 2012 DOI: 10.5121/csit.2012.2117.

[4]*Christos Faloutsos* CMU “Mining Graphs and Tensors “www. cs.cmu.edu /~christos.

[5] DEEPAYAN CHAKRABARTI AND CHRISTOS FALOUTSOS “Graph Mining: Laws, Generators, and Algorithms “ *Yahoo! Research and Carnegie Mellon University.*

[6] Hassan Sayyadi, Shanchan “Graph Mining “Department of computer Science University of Maryland-College Park, CMSC 828G Survey.

[7] Robert A. Hanneman “Introduction to social network methods “ (Department of Sociology, University of California, Riverside) and Mark Riddle (Department of Sociology, University of Northern Colorado)

[8] D. Kavitha, B.V. Manikyala Rao, V.Kishore Babu “ A Survey on Assorted Approaches to Graph Data Mining “*International Journal of Computer Applications (0975 – 8887) Volume 14– No.1, January 2011.*

[9]Reywemlinger Slides.pdf,” Introduction to Graph Mining “ Washington state University.

[10] Sangameshwar Patil “Graph mining and other research at TRDDC,TCS “

[11] Takashi Washio, Hiroshi Motoda, “State of the Art of Graphbased Data Mining “The Institute of Scienti£c and Industrial Research, Osaka University 81,Mihogaoka, Ibarakishi Osaka, Japan .

[12] Yi Jia · Jintao Zhang · Jun Huan “An efficient graph-mining method for complicated and noisy data with real-world applications **“**Received: 29 January 2010 / Revised: 10 November 2010 / Accepted: 20 November 2010 © Springer-Verlag London Limited 2011.

[13]”Graph Mining, Social Network Analysis, and Multirelational Data Mining**”** at *www.nd. edu/_networks/publications.htm#talks0001*.

[14] Daniele Loiacono ” Graph Mining and Social Network Analysis Data Mining and Text Mining” (UIC 583 @ Politecnico di Milano).

[15] “Graph Mining” at www.dataminingtools.net

[16] Hossein Rahmani, "An Introduction to Graph Mining "hrahmani@liacs.nl 8 December 2009.

[17] Harsh J. Patel1, Rakesh Prajapati2, Prof. Mahesh Panchal3, Dr. Monal J. Patel4 "A Survey of Graph Pattern Mining Algorithm and Techniques **" ,** *International Journal of Application or Innovation in Engineering & Management (IJAIEM)* Volume 2, Issue 1, January 2013

[18] Chuntao Jiang, Frans Coenen and Michele Zito," A Survey of Frequent Subgraph Mining Algorithms "*The Knowledge Engineering Review*, Vol. 00:0, 1–31.c 2004, Cambridge University Press.

[19] Prof. Michalis Vazirgiannis LIX," Web graph mining " Ecole Polytechnique, & INFRES , Telecom Paristech http://www.lix.polytechnique.fr/~mvazirg/ March 2012.

[20] Yiping Zhan "Tools for graph mining "yzhan@cs.cmu.edu 12th June 2003