Current Issues in IT-Management

Applicability of Graph Metrics when Analyzing Online Social Networks

Robert Hilbrich
Robert@Hilbri.ch

January 6, 2008

Tutor: Seda Gürses

Abstract

Online Social Networks are special social networks with a limited user interface: the internet browser. As internet access is available from almost everywhere today by using a variety of new technologies, the *interface barrier* is effectively eliminated. Online social networks seem to leave the stage of being an *extension* to the *real* social network and start to become a *social-network-superset* - at least for some people. This allows the following question: can we use the same approach to analyze *real* social networks with these new *virtual* communities?

Therefore, an introduction to graphs and graph metrics in general is given, followed by the description of the modelling specifics for online social networks. A case-study dealing with the search for groups conducting illicit activities is used to show the advantages, disadvantages and unresolved issues regarding the application of graph metrics for analyzing social networks and online social networks in particular.

Contents

List of Figures

1 Introduction

During the last years the role and the importance of online social network services, such as `facebook.com`, `xing.com` or `studivz.de`, has changed from a niche to a widely used and well accepted way to keep contact to friends or business contacts. The role of these emerging virtual communities seem to have changed from a small *extension* to a *reflection* or even a *superset* of the *real* social network.

It is therefore reasonable to examine the similarities and differences between *online* and *real* social networks. This will help to answer the following question: can we use the same approach to analyze real social networks when analyzing online social networks?

Graph theory is a mathematical concept to express relations and for that reason it is also a common and promising methodology to analyze social networks. Is graph theory also applicable to analyze virtual communities?

Before this question can be examined more closely, the concept of a *graph model* needs to be properly introduced. Therefore, section 2 contains a brief introduction to graph theory. The following section examines the differences between online and real social networks and presents a way to transform structural information, relationships for example, from these networks into a graph model.

Section 4 covers the analytical process of the graph model in general by introducing special graph properties. The case study presented in section 5 serves as an example application and provides a solid basis for understanding and assessment of the analytical process itself. Finally, section 6 presents the results and draws a conclusion.

2 Graphs and Metrics

Graphs and their *metrics* are the two basic tools that will be referred to throughout the whole paper. Although most people seem to have an intuitive idea about the meaning of these terms, it is still necessary to present a formalized definition. This section is supposed to deliver a common understanding about the approaches taken and conclusion drawn in the paper.

2.1 What are Graphs?

Graphs are a general *modelling technique* used in many research disciplines to represent *pairwise relations* among a set of objects. Graphs can be simplified as a set of linked nodes and imply an abstraction of reality. An example for a graph can be seen in figure 1.

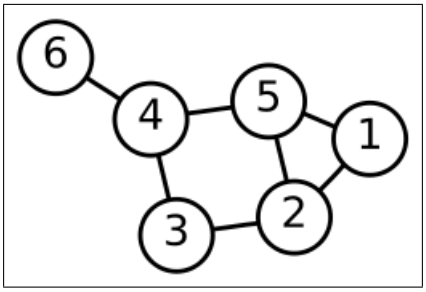

Figure 1: An example for a graph with six nodes

In *graph theory* a graph G is formally defined as a set of *vertices* (nodes) v which is connected by *edges* (links) e. A common notation is:

$$G = (v, e)$$

Edges can also be either *bi-directional* or *uni-directional* to represent the characteristics of the underlying relation. Uni-directional edges are represented as arrows and edges

without arrows are assumed to be bi-directional. Consequently a graph containing uni-directional edges is referred to as a *directed graph*, it is otherwise referred to as an *undirected graph*. Examples for both graphs can be seen in figure 2.

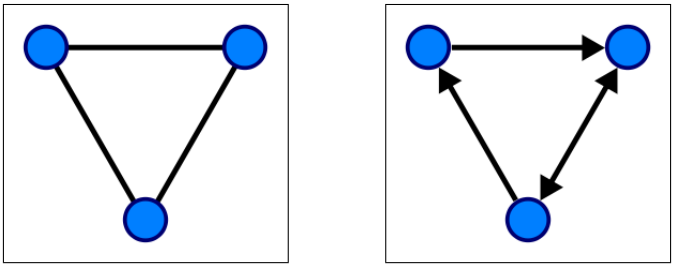

Figure 2: An undirected graph (left) and a directed graph (right)

Additionally, graphs can also represent the *intensities* of the underlying relations. This is done by adding *weights* to the edges, so that the resulting graph is called a *weighted graph*. These weights are used to model distances, costs or durations. An example for a weighted graph can be seen in figure 3.

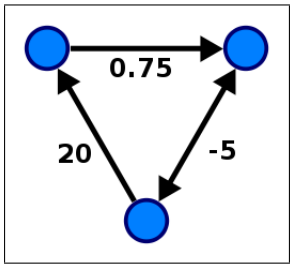

Figure 3: A weighted graph

Although, graphs itself are a *time-static* structure to represent relations, they can be used to model *time-dynamic* models, such as *evolutions of relations*, in contrast to the traditional *snapshot of relations*. An approach for computation-friendly dynamic graph model can be found in [CPV01] or in [TBWK07].

When referring to links and their structure, the notion of a *path* and a *cycle* is common, so it is necessary to give a definition for these terms as well. A *cycle* refers to a chain where the initial and terminal node is the same and that does not use the same link more than once.

A *path* on the other hand is a sequence of links that are traveled in the same direction. For a path to exist between two nodes, it must be possible to travel an uninterrupted sequence of links. Finding all the possible paths in a graph is a fundamental attribute in measuring accessibility and flows.

2.2 What are Metrics?

Graphs are an essential tool to build a model of the research phenomena at hand, but graphs itself do not contain sophisticated techniques for a detailed analysis. Actually, it is difficult to decide whether two given graphs are *similar*. Which criteria should be applied for this decision? Is the number of nodes sufficient to identify similar graphs or should the number of edges also be taken into account?

For a detailed graph analysis, special *measures*, that will be used for *benchmarking* or proper *interpretation* of graph properties, need to be defined. These measures are called *metrics*. Although metrics in mathematics usually refer to an abstraction of distances, metrics in general can be almost arbitrarily defined with regard to the research area.

Graph metrics can be understood as a mapping function from a set of graph properties to a rational number. This approach simplifies graph analysis as the complexity level to work with abstract graph properties is now reduced to simple algebra. These metrics can be used to either compare (sub-) graphs with other (sub-) graphs or nodes with other nodes. It is common to use several metrics for a detailed graph analysis.

A simple metric to compare two graphs is the number of nodes. It could be used as a measure for the number of participants of a network, but does not allow further analysis of the *overall relatedness*. This could be accomplished by using an additional metric reflecting the link structure, such as the *maximum path length*.

As there is a variety of metrics reflecting several distinct graph properties, the choice

which metrics should be used is not an easy one. However, only a proper choice allows justified interpretations later in the analytical process. Some metrics may also be difficult to compute when the graph reaches a certain size. This should be kept in mind during the planning phase of the analytical process.

3 Modelling Online Social Networks

In the previous section, the tools to build and analyze network-like structures have been introduced. In this section a general approach to modelling *online social networks* in particular will be presented. Therefore the notion of an online social network will be explained at first, followed by the description of the modelling approach. Finally, critical remarks will be presented that should lead to a *reasonable* interpretation of the results from the following analysis.

3.1 What are Online Social Networks?

In this paper, a *social network* refers to a social structure consisting of *individuals*. They are tied together by a certain mutuality, such as friends, interests, family relations, social status or common values. A social network should not be confused with *social network services*, such as `StudiVZ.de` or `XING.com`.

However, an *online social network* is a special *social network* which can only be accessed via internet. The web-browser's window becomes the main interface for the users to interact. Thus it is also referred to as a *virtual community*. Social network services provide the technical infrastructure, but they do not constitute a social network.

A *social network* is the core concept of *social network analysis* (SNA), which is a key technique in many modern research areas, such as modern sociology, anthropology, sociolinguistics, geography, social psychology, communication studies, information science,

organizational studies, economics, and biology. In [Kre06] social network analysis is referred to as a

> [...] mapping and measuring of relationships and flows between people, groups, organizations, computers, web sites, and other information/knowledge processing entities. The nodes in the network are the people and groups while the links show relationships or flows between the nodes. SNA provides both a visual and a mathematical analysis of human relationships. [...]

The history and development of SNA is best described in [Fre04] or in [Sco00] pages 15-31.

Probably the first online social network was the world-wide collaboration of researchers using the world wide web. In [Kle99] J. Kleinberg models web pages and their references to each other as a directed graph and examines the question: "What are the recurring patterns of linkage that occur across the Web as a whole?" ([Kle99] page 2).

Using a graph theoretical approach, he was finally able to *cluster* web pages into three main categories - *authorities*, *hubs* and *communities* - just by evaluating their link structure. A similar approach can be used to locate and group individuals based on a certain criteria just by analyzing the *network structure* of an online social network.

3.2 General Modelling Approach

We will now clarify the question of what can be used as a *link structure* in an online social network. One of the core ideas of social network services, is the possibility to *promote* other individuals to being your friends or business contacts. This is done by sending special invitations which the recipient can either confirm or deny. It can be understood as a *two-way-handshake*. After completion it gives both individuals the right to access personal information on each other's profile page that was not declared

to be public. This *handshake process* and the resulting relation can be used as the *link structure* for building a graph model as it reflects the intuitive idea of a *link* in a social network.

Both ends of a link need to confirm that they share the same underlying relation with each other, such as being friends or business contacts. This approach also allows for comparisons between real social networks and virtual ones by applying a common idea of *connectedness* to the virtual world.

The simple approach to modelling an online social network is straight forward. Every member of the virtual community becomes a node in the graph. Additionally, every two members from the community who completed the *two-way-handshake* share an undirected link. The result is an undirected graph representing the community structure.

While the simple approach can be used as a first approach for an analysis, it is a good idea to take other features of online social networks into account. The first addition comes from the communication style within these virtual communities. Every exchange of information can be understood as an *exchange of messages* between two people. These can either be personal messages which suits the intuitive understanding of a message, but it also extends to *wallpaper posts*. Every message for a wallpaper is just a message made public by the recipient. Even the two-way handshake can be seen as an exchange of two messages.

This communication style demands the graph to be *directed* and we can use the exchange of a message to add a directed link from the sender to the recipient. Assuming, that people who are exchanging a lot of messages also share a deeper *bond* between them, weights can be added to the links to indicate the intensity of the relationship. The weights may correspond to the number of messages exchanged in the past. Further studies exploring the *bonds* between people and their communication style should be conducted in order to gain a proper graph model that corresponds as much as possible to real social networks.

The second feature that can be added to the modelling approach, is the possibility in re-

cent social network services to *join a group*. All group members share a common interest in a certain topic and are able to exchange ideas or *meet* like-minded people. Representing a group as a special node the *association* process of joining can be easily understood as a directed link from the individual to the group node.

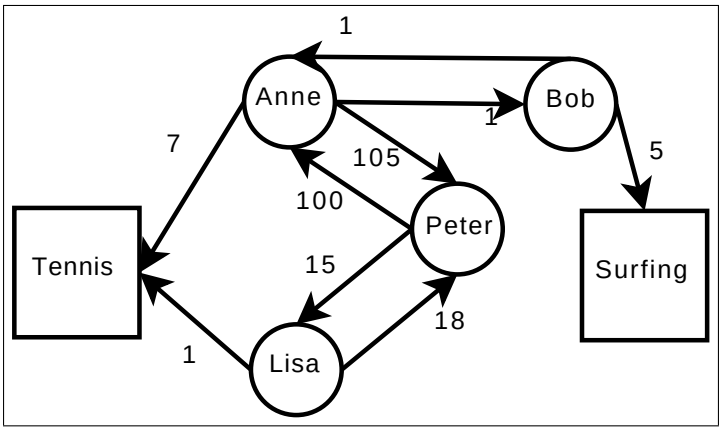

Figure 4: An example for the advanced modelling approach

An example for the advanced modelling approach can be seen in figure 4. In this example there are four individuals modelled as nodes: *Anne*, *Bob*, *Peter* and *Lisa*. There are also two groups called *Tennis* and *Surfing*. The corresponding nodes are squared to denote their speciality.

Additionally, there are also weights given for each link to indicate the intensity of the relation. For example, the relation between *Anne* and *Peter* is stronger compared to the relation between *Anne* and *Bob*. Please note that weights for links can also be used to express the extent of interest between an individual and each of its groups.

3.3 Critical Remarks

Although online social networks seem to be quite similar to non-virtual communities, there are still some fundamental differences that should be paid attention to when analyzing the resulting graph. This section will illuminate the differences and reveal

problems when comparing virtual to non-virtual communities using a graph theoretical approach.

Consider a given graph for an online social network and this intuitive metric:

> The popularity of an individual corresponds to the number of people in their direct neighbourhood.

Translating this informal description into a graph metric leads to:

> The extent of popularity of a node corresponds to its degree.

This metric is intuitive and suits the common understanding of popularity and seems to fit for non-virtual communities.

However, in virtual communities it is possible to encounter individuals having more than 100000 *friends*. How does that reflect their popularity? Is it possible to have that many *friends* in a non-virtual community?

Social network services offer possibilities that are not existing in non-virtual communities. They radically simplify the effort to *meet* new people and also provide simple tools to exchange messages and *manage* the relationships. This is one aspect where virtual communities differ substantially from non-virtual communities.

Over the last years, many social network services emerged and to differentiate themselves, each focused on a certain *customer base*. There is StudiVZ for students, XING for the business *elite*, Facebook for young people in general and many more. It is also quite common to have an account for several - not only one - social network services.

Consequently one might be a member of *several* online social networks at the same time. How does that compare to non-virtual communities? How to reason about an individual when just referring to a facet of its *whole* social network?

With regard to this variety of social network services each of them is build to suit the needs for a certain clientele. This leads to a *label* that is virtually attached to the members of that community. They can be the *business elite* or just *fun-loving students*. Is there something similar in non-virtual communities?

While it is difficult to change ones social network in the *off-line world*, it is very easy to join or leave a community in the *online world* as accounts are usually offered for free. It is now possible for people to freely choose which *label* they want to attach their online presence. For this reason the question whether there might be a hidden agenda or a purpose behind the structure of an online social network needs to be looked at.

Needless to say that non-virtual social networks are also the result of personal preferences and individual objectives, but the speed and simplicity for a change of a social network in the online world is radically higher.

All these remarks should be considered thoroughly when working with models from online social networks. Although virtual and non-virtual communities share the same nature, they do still come with fundamental differences.

4 Analyzing Virtual Communities

After having introduced the modelling approach for virtual communities, this section focuses on the analysis of the graph models. The first part overviews the general approach of a graph analysis and is followed by a review of common graph metrics in the second part. A case study of an interesting research topic using a graph theoretical approach constitutes the third part. With regard to this case study, the last part focuses on implicit assumptions made and dangers of false interpretations encountered throughout the process.

4.1 General Approach

This section introduces a generic graph theoretical approach to analyzing social networks. Itemising each step is necessary to understand the research groundwork of the case study and estimate the dangers of false interpretations described in the last part of the current section.

When analysing a certain phenomenon in social networks usually the very first step is to describe it informally. Separating it from other non-related aspects of social interaction leads to a *model* which is an abstraction of the *real* phenomenon. The next step is a very important and difficult one. It deals with the question of how the research phenomenon is represented or identifiable in the structure of a social network.

The answer to this question leads to several characteristics of its structural context capturing the main aspects of the model. These characteristics can be expressed using *common* social network analysis metrics - as introduced in the next section - or *custom* metrics can be defined, maybe even based on common metrics, to reflect the model as precisely as possible.

Describing the initial phenomenon with graph metrics leads to a formal description that will be referred to as a *footprint* or *pattern* in the graph structure. This pattern may be very simple and allow a basic *pattern search* in the graph model.

However most interesting research phenomena do not come with a simple and unambiguous reflection in the structure of the underlying social network. In this case it is not possible to deduce the combination of metrics manually that would capture the phenomenon in an accurate and concise way.

Machine learning algorithms are the solution for this problem. Please refer to [Nil05] for a detailed introduction to machine learning.

The core idea is to assume the existence of a footprint in the structure and use a large

set of data to *train* the search algorithms. This training is done by feeding both, the training data and a manual decision - what is considered to be part of the footprint and what is not considered to be part of it - into the machine learning algorithms. After substantial training, these algorithms are then able to search for similar patterns in another set of data - the *verification set*.

The last steps are usually a rather *iterative process*, where each step is supposed to refine the footprint in order to get better trained decision algorithms leading to a higher accuracy rate.

4.2 Specifying the Footprint - SNA Metrics

While the last section introduced the analytical process as a whole, this section focuses on the graph-theoretical *footprint* of the phenomenon in the social network.

In section 2.2 the notion of a *metric* was introduced and explained as a measure for structural characteristics of a given graph. Therefore these metrics can be used to describe the *footprint* mentioned above.

As the simple metrics shown in section 2.2 are not sufficient for a detailed analysis, more specialized metrics are needed for social network analysis to properly reflect social interactions within a graph. These metrics can be found in [Kre06], in [Wat99], in [RAH05] in chapter 10 and in [CGM04b]. [Soc07] also provides an overview about special metrics used in social networks analysis.

Itemizing all common social network analysis metrics and explaining their specifics is beyond the scope of this paper. In the following, two metrics, that will be referred to in section 5, will be introduced.

Throughout this paper the following definitions will be used:

- G - a graph

- $N(G)$ - the set of *nodes* in the graph G
- $E(G)$ - the set of *edges* in the graph G
- n_i, n_j, - individual nodes
- $dist(u, v)$ - the distance (length of the shortest path) between the nodes u and v

The first metric to be introduced is the *density*. The density measures the ratio of existing edges to potentially existing edges. *Undirected density* can be formally expressed as (cf. [CGM04b] page 3):

$$Den_U(G) = \frac{2|E(G)|}{|N(G)| \cdot (|N(G)| - 1)}$$

The second metric that will be used later on is the *characteristic path length* (CPL). The CPL is the average distance between any two nodes in a graph. There are several definitions available. In conjunction with [CGM04b], the following formal definition is used:

$$CPL(G) = \frac{1}{|N(G)| \cdot (|N(G)| - 1)} \cdot \sum_{i \neq j} \frac{1}{dist(n_i, n_j)}$$

Using the metrics *density* and *CPL*, it is now possible to model certain characteristics of a social network in which links represent the existence of communication paths between individuals. An example application will be presented in the next section.

5 Case-Study

The company 21st Century Technologies inc. is a team of American researches who specialized in "[...] graph pattern matching, image processing, genetic algorithms, natural language processing, machine learning, complex network modeling, data mining, data fusion, game theory, and group detection.[...]" (cf. `http://www.21technologies.com`).

They conducted a study on the applicability of graph metrics in social networks regarding the classification of terrorist activity and published their findings in [CGM04a], [CM04a] and [CGM04b]. According to their publications, they achieved a 93% classification accuracy on synthetic data regarding the "two-class classification problem (terrorist or non-terrorist)" (cf. [CM04a] page 1).

In the following, their approach will presented briefly, so that it is possible to evaluate the application of graph metrics in the next section by looking at the implicit assumptions made and the risks of false interpretations allowed.

5.1 Graph theoretical Approach

Based on the assumption that groups conducting illicit activities often organize and communicate in a special form, the *Leninist cell structure*, the approach taken is to use pattern classification techniques applied to evolutions of SNA metrics to identify these groups. This is supposed to "[. . .] help identify members of terrorist groups before they act" (cf. [CM04b]. page 1).

The evolution of the communication structure of the Leninist cell is assumed to be a distinctive result of the efforts to hide illicit activities from the public. Therefore a distinction is drawn between two communication patterns: the *hub-and-spoke*-evolution and the *small-world* evolution (cf. [Wat99]). While the latter represents *typical* interpersonal communication network, the first refers to the Leninist network.

According to the model, a Leninist cell can be in one of two states: *sleeping* or *active*. The active state results in a change of the communication patterns from *mediated* to *non-mediated*. An example for these different communication patterns can be seen in figure 5.

The specific pattern for a Leninist network is therefore a period of mediated communication followed by non-mediated communication. During the sleeping state, two members of a group will not communicate directly for security reasons. All communication

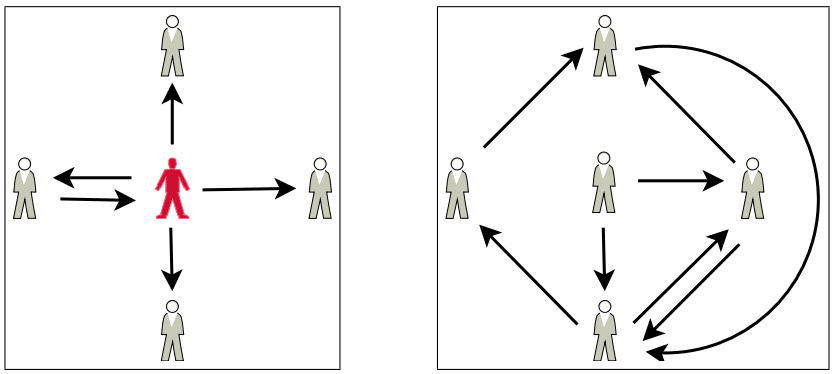

Figure 5: Mediated communication (left) and non-mediated communication (right)

will go through the group leader acting as a mediator. However, in the active state with non-mediated communication, efficiency is supposed to prevail the need for secrecy.

On the other hand, the small world evolution is considered to be characteristic of everyday social interaction and features generally *non-mediated* communication during its evolution. Also, direct communication is preferred over communication through an intermediary which is referred to as the *geodesic assumption*. It means that the members of this network prefer the shortest path for their interaction. The tendency to develop redundant communication paths is also supposed to be a characteristic of everyday social interaction.

Modelling the *footprint* was mainly done with two SNA metrics: characteristic path length (CPL) and density. Both metrics have been introduced in section 4.2. As a result of the mediated communication, the CPL will remain higher compared to the non-mediated communication. In order to reduce the number of possible leaks, the density is expected to be lower during the sleeping state. During the transitioning process, the CPL is supposed to drop to a level similar to the small world evolution.

A labeled and synthetic dataset - a *phi graph* with several parameters - was used for the classifier training and the observability level was set to 100% for both the training and the testing dataset. The group classification was implemented with hidden Markov models

(HMM). The classification performance was measured by computing the accuracy rate, which is "the percentage of groups within the test dataset that were assigned the correct classification" (cf. [CM04a] page 6).

5.2 Applicability Evaluation

Although the approach presented in the previous section focuses on social networks in general, it could also be applied to data gathered from online social networks. The effort to gather the required data may even be smaller as there are already central databases from these community services keeping track of all required data. Therefore this section is geared to evaluate the applicability of graph metrics when analyzing online social networks.

With the astonishing accuracy rate in mind graph metrics appear to be very efficient. It now seems possible to group and locate individuals based on their goals or agenda just by analyzing their communication structure. Is this the research groundwork for an *early warning system for terrorist attacks*?

With regard to the simplicity of the data collection in online social networks as members of online communities reveal their friends, their messages and other personal information at least to the provider of these services (ref. [War08]), graph analysis appears to very cost-effective. Expensive wiretapping may have been rendered unnecessary.

The methodology to apply pattern matching algorithms is also a very convenient way to analyze the data collected. This approach can be automated to a high degree and seems to promise good results. It can also be easily scaled to work on large graphs with millions of nodes.

Additionally, companies like *AT&T* (ref. [CPV01]) sponsor research in the area of dynamic graphs and graph metrics to develop techniques to detect fraud. This indicates that graph analysis left the niche of computer sciences and shows the importance of

graph analysis in a wide range of applications.

Although this methodology is appraised by companies, like *AT&T*, or government agencies, like the US *Defense Advanced Research Projects Agency (DARPA)*, it is still necessary to look into arguments against the application of graph metrics.

At first, it is necessary to check whether the graph analysis meets the regulatory framework, such as civil and criminal law. At least in Germany, it is not allowed by law to conduct a graph analysis as described in the case study. The problem lies in the dataset. Searching for terror cells in a graph containing the communication structure of innocent German citizens is not allowed as everyone automatically becomes a suspect in an investigation for no apparent reason.

However, different conditions, such as a limited graph of a customer base, or other goals, such as fraud detection instead of detection secretive activity, may allow the application of graph metrics.

The collection of a proper dataset for analysis may constitute another challenge depending on the application. The example previously presented depends on the collection of an almost complete graph of the communication structure containing all media possible. Unless the graph covers all means of communication and interaction, the analysis results can only be used as a rough approximation and should be considered highly error-prone. Other application areas may be limited to a small number of ways for individuals to interact, so that proper data collection can be achieved.

In order to conduct a proper analysis of a phenomenon, it is necessary to build a model and determine its footprint in the underlying graph dataset (ref. section 4.2). The example presented reduced the general idea of *secretive activity* to the *Leninist cell*-model and determined the corresponding footprint in order to detect terrorist groups. There are two important assumptions made with this approach:

1. All terrorist groups follow the Leninist cell model.
2. Only terrorist groups lead to the characteristic footprint.

With regard to the first assumption, it is easy to imagine other forms of collaboration to have similar results, so that the model of the phenomenon *secretive activity* is incomplete. Due to the variety of organizational forms it may be difficult to create a *complete model* to properly represent *secretive activity*.

The second assumption is a very grave one and shows that graph metrics and graph analysis should be applied in a sensible way. Consider the communication structure of an organizer of a conference. During the planning of the conference, every participant only communicates to the organizer which results in a *mediated communication*. However, during the conference the participants will most likely interact with each other which leads to a *non-mediated communication*. A similar footprint does not necessarily indicate the same phenomenon.

Assuming there is a phenomenon and its footprint allowing proper identification, so that all assumptions above are reasonable, there are still some challenges to master. Usually, just the existence of a footprint is assumed, but is not characterized down to graph properties. The characterization is a very difficult and complex task and it is mostly impossible to be done manually - at least for non-trivial problems. That is the reason why machine learning algorithms are applied to conduct the pattern search (ref. section 4.1).

These algorithms need training in order to tune their search parameters according to the footprint. This is an iterative process and the resulting search quality is highly dependent on the training data sets. The training is the critical part. Over-trained algorithms tend to find too little, under-trained algorithms tend to find too much.

However there is no general way to determine the perfect level of training. Another training dataset may lead to other search parameters in the algorithms. Many papers have been published dealing with this topic, especially the training of neuronal networks. This does not only lead to footprints not being found (*false-negatives*), but also to false classifications (*false-positives*)!

The example presented used a synthetic graph model, taken from Duncan J. Watts (ref. [Wat99]), a *phi graph*. The relatively high accuracy rate was achieved for this specific

synthetic graph. How does the result relate to the same approach based on *real* social networks? There is more research needed in order to properly estimate the accuracy rate when using a real, non-synthetic graph model.

Even if it was possible to reach an accuracy rate of 96%, there are still challenges left. As graph analysis can be easily scaled to large graphs, the remaining 4% can still be a lot of false-positives or false-negatives. A company with a customer base of about one million customers is not uncommon and a likely adopter of graph analysis methodologies due to their large customer base. However 4% of *falsely accused* customers are still 400000 people!

6 Conclusion

Graphs provide an efficient way to store and represent data regarding relations among objects or individuals. They focus on the structure, indicated by nodes and edges, and intensities, indicated by weights. Relations can be properly represented by using directed or undirected graphs. The environment and context of the relations is not relevant and therefore not part of the graph model. While time-static graph models contain a snapshot, time-dynamic graphs are used to depict evolutions and developments of relations.

Graph metrics are measures capturing certain graph properties. They can be understood as the footprint of a given *phenomenon* in the graph model. As it is difficult to characterize the footprint manually, machine-learning algorithms are applied to allow an iterative approximation of the footprint and its representing metrics. Once a footprint is properly outlined, it can be used to search for new occurrences of the same phenomenon.

Recently graph analysis with dynamic graphs and certain graph metrics received a lot of research attention from social network analysts. Social network analysis (SNA) is an intermediary research discipline and tries to infer knowledge about individuals or groups of individuals by analysing their social structure.

SNA applications making use of graph analysis typically share a common goal. They search for individuals or groups of people with similar characteristics being expressed in graph metrics. This way of analysing a graph is referred to as the search for *communities of interest*. Depending on the application, these communities may be a group of people conducting secretive activities or customers of a company being victims of identity theft.

Online social networks or virtual communities are special social networks. Although both networks are similar in most aspects, there are still some fundamental differences. They mainly derive from the new technical platform offering new possibilities, such as having more than 1000 *friends*.

Since members of virtual communities share information about their social network voluntarily - at least from the provider perspective - and since the importance of online social networks rose dramatically over the last years, it is reasonable to look into the applicability of SNA by using data from virtual communities.

Modifications to the graph model have been presented to describe an efficient way to represent data from virtual communities. As the result is still a specialized graph, the classical search application for communities of interest (COI) is still applicable. In order to depict the chances and risks by using graph metrics when analyzing online social networks, a case study for a typical search for COI is presented and followed by a critical evaluation.

Although graph metrics offer a fast and very efficient way for clustering a social network and searching for individuals sharing certain characteristics, they are still a limited and abstract view of the *real* network environment. The variety of difficulties encountered and implicit assumptions made when trying to capture a complex phenomenon such as *secretive activities* or *terror cells* in graph metrics, clearly illuminated the limitations of their application in social network analysis.

References

[CGM04a] Coffman, Thayne; Greenblatt, Seth; Marcus, Sherry: Graph-based technologies for intelligence analysis. In: *Commun. ACM*, volume 47(3):pp. 45–47, 2004. ISSN 0001-0782.

[CGM04b] Coffman, Thayne; Greenblatt, Seth; Marcus, Sherry: Sensitivity of social network analysis metrics to observation noise. In: *Proc. IEEE Aerospace Conf.* 2004.

[CM04a] Coffman, Thayne; Marcus, Sherry: Dynamic classification of groups using social network analysis and hmms. In: *Proc. IEEE Aerospace Conf.* 2004.

[CM04b] Coffman, Thayne; Marcus, Sherry: Pattern classification in social network analysis: A case study. In: *Proc. IEEE Aerospace Conf.* 2004.

[CPV01] Cortes, Corinna; Pregibon, Daryl; Volinsky, Chris: Communities of interest. In: *Lecture Notes in Computer Science*, volume 2189:pp. 105–??, 2001.

[Fre04] Freeman, Linton C: *The development of social network analysis. A study in the sociology of science*. Emperical Press, Vancouver, 2004.

[Kle99] Kleinberg, Jon M.: Hubs, authorities, and communities. In: *ACM Comput. Surv.*, volume 31(4es), 1999. ISSN 0360-0300. doi:10.1145/345966.345982. URL `http://portal.acm.org/citation.cfm?id=345982`.

[Kre06] Krebs, Valdis: Social Network Analysis, A Brief Introduction. `http://www.orgnet.com/sna.html`, 2006.

[Nil05] Nilsson, Nils J.: Introduction to Machine Learning. `http://robotics.stanford.edu/people/nilsson/mlbook.html`, 2005.

[RAH05] Robert A. Hanneman, Mark Riddle: Introduction to social network methods. `http://faculty.ucr.edu/~hanneman/nettext/`, 2005.

[Sco00] Scott, John: *Social Network Analysis, A Handbook, Second Edition*. SAGE Publications, 2000.

[Soc07] Wikipedia. `http://en.wikipedia.org/wiki/Social_network_analysis`, 2007.

[TBWK07] Tantipathananandh, Chayant; Berger-Wolf, Tanya; Kempe, David: A framework for community identification in dynamic social networks. In: *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 717–726. ACM, New York, NY, USA, 2007. ISBN 978-1-59593-609-7. doi:http://doi.acm.org/10.1145/1281192.1281269.

[War08] Ward, Mark: Cyber thieves target social sites. `http://news.bbc.co.uk/1/hi/technology/7156541.stm`, 2008.

[Wat99] Watts, Duncan J.: *Small worlds: the dynamics of networks between order and randomness*. Princeton University Press, Princeton, NJ, USA, 1999. ISBN 0-691-00541-9.