Thesis for the Degree of Doctor of Philosophy

Improving Community Detection Methods for Network Data Analysis

Farnaz Moradi

Division of Networks and Systems
Department of Computer Science and Engineering
Chalmers University of Technology
Göteborg, Sweden 2014

Improving Community Detection Methods for Network Data Analysis
Farnaz Moradi
ISBN: 978-91-7597-041-7

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 3722
ISSN: 0346-718X

Technical report 112D
Department of Computer Science and Engineering

Division of Networks and Systems
Chalmers University of Technology
SE-412 96 GÖTEBORG, Sweden
Phone: +46 (0)31-772 10 00

Author e-mail: `moradi@chalmers.se`

Printed by Chalmers Reproservice
Göteborg, Sweden 2014

ABSTRACT

Empirical analysis of network data has been widely conducted for understanding and predicting the structure and function of real systems and identifying interesting patterns and anomalies. One of the most widely studied structural properties of networks is their community structure. In this thesis we investigate some of the challenges and applications of community detection for analysis of network data and propose different approaches for improving community detection methods.

One of the challenges in using community detection for network data analysis is that there is no consensus on a definition for a community despite excessive studies which have been performed on the community structure of real networks. Therefore, evaluating the quality of the communities identified by different community detection algorithms is problematic. In this thesis, we perform an empirical comparison and evaluation of the quality of the communities identified by a variety of community detection algorithms which use different definitions for communities for different applications of network data analysis. Another challenge in using community detection for analysis of network data is the scalability of the existing algorithms. Parallelizing community detection algorithms is one way to improve the scalability of community detection. Local community detection algorithms are by nature suitable for parallelization. One of the most successful approaches to local community detection is local expansion of seed nodes into overlapping communities. However, the communities identified by a local algorithm might cover only a subset of the nodes in a network if the seeds are not selected carefully. The selection of good seeds that are well distributed over a network using only the local structure of a network is therefore crucial. In this thesis, we propose a novel local seeding algorithm, which is based on link prediction and graph coloring, for selecting good seeds for local community detection in large-scale networks.

Overall, mining network data has many applications. The focus of this thesis is on analyzing network data obtained from backbone Internet traffic, social networks, and search query log files. We show that mining the structural and temporal properties of email networks generated from Internet backbone traffic can be used to identify unsolicited email from the mixture of email traffic. We also show that a link based community detection algorithm can separate legitimate and unsolicited email into distinct communities. Moreover, we show that, in contrast to previous studies, community detection algorithms can be used for network anomaly detection. We also propose a method for enhancing community detection algorithms and present a framework for using community detection as a basis for network misbehavior detection. Finally, we show that network analysis of query log files obtained from a health care portal can complement the existing methods for semantic analysis of health related queries.

Keywords: Networks, Community Detection Algorithms, Overlapping Communities, Seed Selection, Misbehavior Detection, Spam, Medical Query Logs

Preface

This thesis is based on the work contained in the following publications:

- Farnaz Moradi, Tomas Olovsson, Philippas Tsigas, "*Towards Modeling Legitimate and Unsolicited Email Traffic Using Social Network Properties*," in *Proceedings of the 5th Workshop on Social Network Systems (SNS'12)*, pp. 9:1 - 9:6, ACM, Bern, Switzerland, April, 2012.
- Farnaz Moradi, Tomas Olovsson, Philippas Tsigas, "*An Evaluation of Community Detection Algorithms on Large-Scale Email Traffic,*" in *Proceedings of the 11th International Conference on Experimental Algorithms (SEA'12)*, Lecture Notes in Computer Science Vol.: 7276, pp. 283 - 294, Springer-Verlag, Bordeaux, France, June, 2012.
- Farnaz Moradi, Tomas Olovsson, Philippas Tsigas, "*Overlapping Communities for Identifying Misbehavior in Network Communications,*" in *Proceedings of the 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'14),* Lecture Notes in Computer Science Vol.: 8443, pp. 398-409, Springer-Verlag, Tainan, Taiwan, May, 2014.
- Farnaz Moradi, Tomas Olovsson, Philippas Tsigas, "*A Local Seed Selection Algorithm for Overlapping Community Detection,*" in *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'14),* Beijing, China, August, 2014.
- Farnaz Moradi, Ann-Marie Eklund, Dimitrios Kokkinakis, Philippas Tsigas, Tomas Olovsson, "*A Graph-Based Analysis of Medical Queries of a Swedish Health Care Portal,*" in *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi'14),* pp. 2–10, Gothenburg, Sweden, April, 2014.

Acknowledgments

First and foremost, I would like to express my profoundest gratitude to my supervisors, Prof. Philippas Tsigas and Associate Prof. Tomas Olovsson, for their constant guidance and support. They have always inspired me by showing excitement for any result I have presented during our meetings and cheering me up anytime I was disappointed. I am also very much in their intellectual debt.

I extend my sincere gratitude to Associate Prof. Dimitrios Kokkinakis for the excellent collaboration we had. I also thank Prof. Per-Larsson Endefors for his invaluable suggestions during my PhD follow up meetings.

I am also grateful to my colleagues in the Networks and Systems division who have contributed immensely to a friendly and productive working environment. I thank Magnus for being supportive, friendly, and fun and for all the advice he has given me. I also thank Marina and Ali for always being helpful and supportive. I would also like to give my appreciation to all the current and former members of the division. Many thanks to Andreas, Bapi, Daniel, Elad, Erland, Georgios, Iosif, Laleh, Nhan, Olaf, Oscar, Pierre, Thomas, Valentin, Vilhelm, Vincenzo, Wolfgang, Yiannis, Zhang, and all the other new members of the division. I am also thankful to all my colleagues in the department for an excellent working environment. I would especially like to express my gratitude to Peter, Eva, Tiina, and Marianne. I also thank my friends Negin, Fatemeh, and Behrooz for the good times we spent in the department.

Finally, my deepest appreciation goes to my family and friends. I am especially grateful to my parents for their unwavering love, selfless support, and encouragement over the years. I would also like to thank my wonderful husband, Mohammad Reza, who has supported me at each step of the way with his love and patience. You are the best and I am really grateful to everything you have done for me and I am proud of everything we have achieved together.

Farnaz Moradi
Göteborg, 2014

Contents

List of Figures

Part I

INTRODUCTION

1 INTRODUCTION

Advances in technology and computation have provided the possibility of collecting and mining a massive amount of real-world data. Mining such "big data" allows us to understand the structure and the function of real systems and to find unknown and interesting patterns.

Many types of real-world datasets can be modeled with *networks*. A network provides a powerful mathematical tool to represent the relations in the data. Networks generated from real-world data are often divided into four categories, social, information, technological, and biological networks [1]. A *social network* is a network connecting the people who contact or interact with each other. Social networks are not limited to "online social networks" such as Facebook, Twitter, or LinkedIn. Other examples of social networks are the network of people collaboration, co-authorships, and co-appearance, as well as networks of communication between people such as telephone calls and emails. An *information network* is a network of entities containing information such as World Wide Web, network of citations, and word co-occurrence networks. A *technical network* refers to a man-made network such as the Internet, the electric power grid, networks of roads, railways, and airline routes. A *biological network* represents a biological system such as a network of metabolic pathways, protein-protein interactions, the food web, and the network of blood vessels.

In this thesis we consider networks from two categories, i.e., social networks and information networks. The focus of the thesis is on the structural properties of these networks and the algorithms which exist for study of these properties, particularly their community structure.

This thesis is organized into two parts. The first part is an introduction to the thesis and the second part consists of a collection of papers. The remainder of the introduction is organized as follows. In Section 1.1 we briefly summarize the structural properties of social and information networks. In Section 1.2 we focus on the community structure of networks and existing algorithms for identifying network communities and investigate a number of challenges in community detection, namely quality evaluation, scalability, and seed selection. In Section 1.3 we look into a number of applications of mining real network data for identifying

interesting patterns and anomalies. In particular we look into identifying sources of unsolicited email traffic based on the communication patterns observed on an Internet backbone link. We also study the application of intrusion detection using network flow data, scalable identification of communities in social networks, and analysis of large query log files by identifying communities of related words from a word co-occurrence network. In Section 1.4 we present the real datasets which we have used in this thesis for generating different networks and analyzing their structural properties. More specifically we describe the collection process of email and flow data from an Internet backbone link, as well as the data which was obtained from different social networks and the query logs of a health care portal. In Section 1.5 our approaches towards analysis of network data and a brief description of the appended papers are presented. Section 1.6 summarizes our contributions in the thesis and, finally, Section 1.7 concludes the thesis and present possible future research directions.

1.1 Structural Properties of Networks

A great deal of work has been devoted to study the structure and dynamics of networks generated from real-world data. These networks are not random networks and the nodes in these networks are organized into specific structures. A wide variety of network mining methods and algorithms exists which can be used to uncover the structure of such networks.

Traditionally, network data was modeled as random graphs [2]. However, empirical studies on different types of real network data have revealed interesting properties such as the "small-world effect" [3], also known as "six degrees of separation" [4], and the scale-free behavior of networks [5, 6]. These properties show that social and information networks are fundamentally different from other types of networks such as random networks [1]. A review of the structural properties of these networks can be found in [7].

Many real networks have been modeled as *small-world* networks. A *small-world* network has a small *effective diameter* and the distance between any pair of nodes in the network is relatively short. The distance between two nodes is measured as the number of edges in the shortest path connecting them. In addition to small effective diameters or short average path lengths, small-world networks tend to be highly clustered which can be quantified using the average *clustering coefficient* of the networks [3].

Another robust measure of the structure of networks is their *degree distribution* which characterizes the spread in the node degrees. It has been shown that for social and information networks the degree distribution has a power law tail. This means that in these networks most of the nodes have a very low degree while a few of the nodes have very high degrees. Such networks are also known as *scale-free* networks [5, 6].

Numerous attempts to model the structure of social networks have also taken other structural properties into account: the distribution of the size of the connected components of the network, the presence of a giant connected component (GCC), and the community structure of the networks. The studies of the changes of structural properties of networks over time have also revealed interesting properties of network evolution. As the networks grow over time, they become more dense (*densification power law*) and the average distance between their nodes shrinks (*shrinking diameter*) [9]. There are many other patterns which have been observed in real world networks. A summary of different patterns, particularly the patterns observed in weighted networks can be found in [8].

1.2 Community Detection

An excessively studied structural property of real-world networks is their community structure. The community structure captures the tendency of nodes in the network to group together with other similar nodes into communities. This property has been observed in many real-world networks. Despite excessive studies of the community structure of networks, there is no consensus on a single quantitative definition for the concept of *community* and different studies have used different definitions. A community, also known as a *cluster*, is usually thought of as a group of nodes that have many connections to each other and few connections to the rest of the network. Identifying communities in a network can provide valuable information about the structural properties of the network, the interactions among nodes in the communities, and the role of the nodes in each community.

1.2.1 Algorithms

A wide variety of *community detection* algorithms, also known as *clustering* algorithms, have been proposed to identify the communities in a network. Since different community detection algorithms use different definitions of a community, they yield different communities. Figure 1.1 shows an example of the communities identified by two fundamentally different community detection algorithms on a real network (Zachary's network of karate club members [10]).

Many traditional community detection methods are borrowed or inspired from graph clustering algorithms. *Partitioning* the nodes in a network into a predetermined number of disjoint communities is one of the traditional methods for identifying communities. However, since the community structure of real-world networks are not usually known, making assumptions about the number of communities or the size of the communities are not realistic. Moreover, many real-world networks have a hierarchical structure where meaningful communities at different scales can exist and such community structures cannot be captured by partitioning algorithms. Therefore, another group of community detection algorithms have been introduced which can identify hierarchical communities. *Hierarchical* clus-

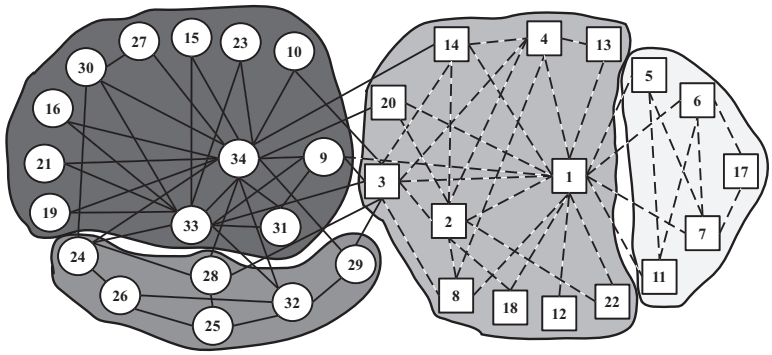


Figure 1.1: *The square and round nodes show the two groups of the members in the Zachary karate club network. The four grey communities are found by applying a node-based modularity optimization algorithm [11]. The solid and dashed edges show the two communities identified by a link-based community detection algorithm [12].*

tering techniques can be divided into *agglomerative* and *divisive* methods [13]. Agglomerative algorithms use a bottom-up approach where clusters are iteratively merged. Divisive algorithms use a top-down approach where the clusters are iteratively split. Overall, using hierarchical algorithms allow us to choose the suitable level of hierarchy and study the communities at that level of hierarchy.

In many real-world networks, nodes can naturally belong to multiple communities, therefore the communities can overlap. In social networks, an individual can belong to a community of family members, to a community of friends, and to a community of colleagues. In an information network, a web page can cover topics that are associated with different communities. Traditional community detection algorithms fail to uncover the community overlaps. Not being able to identify community overlaps in networks with naturally overlapping communities means missing valuable information about the structure of the network [14]. Therefore, *overlapping community detection algorithms* have gained a lot of attention. Overlapping communities can be identified using different approaches. One of these approaches is based on partitioning the edges of a network into communities rather than partitioning the nodes [12, 15]. A thorough review and comparison of different types of overlapping community detection algorithms can be found in [16].

The majority of existing community detection algorithms implicitly assume that the entire structure of the network is known and is available. We refer to these types of algorithms as *global algorithms*, since they require a global knowledge of the whole network in order to uncover all the communities in that network. Since such knowledge might not be available for large networks, *local algorithms* are gaining more popularity [23, 27–29]. Local algorithms typically start from a number of given *seed* nodes and expand them into possibly overlapping communities by examining only a small part of the network. Since it is possible to find local com-

Table 1.1: *Community Detection algorithms.*

	Algorithm	Type	Description	Complexity
Non-Overlapping	Blondel [11]	G,H	*Fast modularity maximization (Louvain)* is a greedy approach to modularity maximization and unfolds a hierarchical community structure.	$O(m)$
	Infomap [17], InfoH [18]	G,H	*Maps of random walks* finds communities based on the compression of the description length of the average path of a random walker over the network. *Multilevel compression of random walks* is the hierarchical version of infomap which minimizes a hierarchical map equation to find the shortest multilevel description length.	$O(m)$
	RN [19]	G,H	*Potts model community detection* minimizes the Hamiltonian of a local objective function (the absolute Potts model).	$O(m^{1.3})$
	MCL [20]	G,NH	*Markov Clustering* is based on the probability of random walks remaining for a long time in a dense community before moving to another community.	$O(nK^2)$
Overlapping	LC [15]	G,H	*Link Community detection* uses the similarity of the edges to identify hierarchical communities of edges rather than communities of nodes.	$O(nK^2)$
	LG [12]	G,H	*Line Graph and graph partitioning* runs a non-overlapping node-based algorithm on a line graph induced from the original graph to identify overlapping link-based communities.	$O(nm^2)$
	SLPA [21]	G,H	*Speaker listener Label Propagation* is an extension to the label propagation algorithm where nodes adopt multiple labels based on the majority labels in their neighborhood.	$O(tm)$
	OSLOM [22]	L,H	*Order Statistics Local Optimization Method* identifies significant communities with respect to a Null model similar to modularity.	$O(n^2)$
	DEMON [23]	L,NH	*Democratic Estimate of the Modular Organization of a Network* is a local algorithm which uses the label propagation algorithm to find communities in the egonet of each node and then merges them into larger communities.	$O(nK^{3-\alpha})$
	PPR [24]	L,NH	*Personalized PageRank-based*, is a local algorithm which uses the *PageRank-Nibble* algorithm [25] to approximate a personalized PageRank vector from a given *seed* node and then uses the method in [26] to create the communities based on a scoring function.	$O(\sum_{C \in \mathcal{C}} vol(C))$

In the "Type" column, L and G denote local and global, and H and NH denote hierarchical and non-hierarchical, respectively. The LG algorithm can find hierarchical communities if the node-based algorithm is hierarchical.
In the "Complexity" column, n denotes the number of nodes, m denotes the number of edges, K is the maximum node degree, t is the number of algorithm iterations selected, α is the power-law exponent, $vol(C)$ is the sum of the degree of all the nodes in a community C, and $\mathcal{C}$ is the set of all the identified communities.

munities from each seed independently, they are very suitable for being parallelized and therefore can scale well. The local communities identified from each seed can be aggregated in order to uncover the global community structure of the network. However, if the local community detection algorithm is naively started from each node in a network, it can lead to many redundant communities and therefore is computationally expensive. Therefore, it is important to identify a number of good seeds which are well distributed over the network by using a *seeding algorithm* before running the local community detection. On the other hand, if the seeding algorithm does not select enough seeds, the communities might only cover a subset of the nodes in a network and therefore, the problem of selecting a reasonable number of seeds which are well-distributed over the network is challenging. These challenges are further investigated in Section 1.2.4.

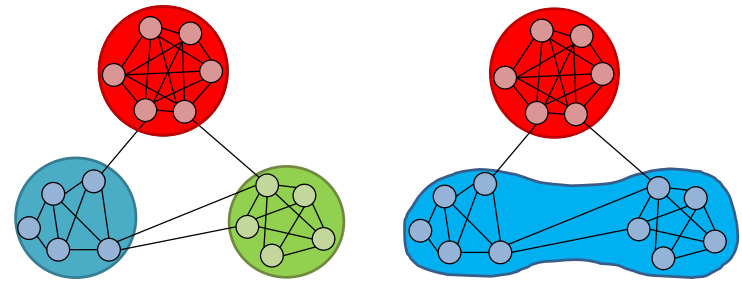
In addition to different types of community detection algorithms, recently, a number of studies have focused on proposing methods for improving the quality of the existing community detection algorithms. Ciglan et al. [30] introduced a method for adding edge weights to unweighted networks as a pre-processing step to improve the quality of the identified communities with respect to ground truth data. Soundarajan et al. [28] introduced a template for using existing community detection algorithms for identifying more realistic communities. Another approach for improving community detection is to use ensemble clustering, which is inspired by ensemble learning, where multiple community detection algorithms run as an ensemble and the identified communities are combined to improve the community qualities. Staudt et al. [31] showed that ensemble clustering can be used to achieve the best trade-off between quality of the communities and the speed of community detection.

Thorough reviews of different types of community detection algorithms can be found in [13, 16, 32]. Table 1.1 summarizes the algorithms which are used throughout this thesis.

1.2.2 Quality Evaluation

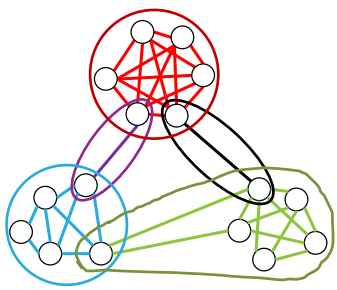
Given the diverse nature of real-world networks and the high diversity of community detection algorithms, it is necessary to perform experimental evaluation of the algorithms to find the most suitable method for each type of network. However, due to the ambiguity in the definition of a community, extracting communities and evaluating their quality is proven to be very difficult.

Figure 1.2 shows the communities identified by different community detection algorithms (see Table 1.1) in a toy network. It can be seen that different types of algorithms identify different communities in the network since they use different definitions for communities and take different approaches for identifying these communities. In order to find out which algorithm yields the best set of communities, it is necessary to use a quantitative measure to evaluate the quality of the communities identified by each algorithm.

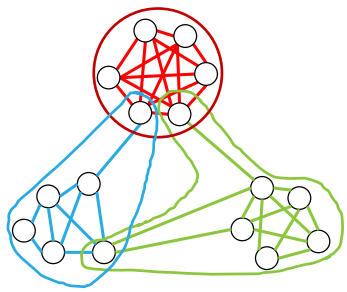


(a) Blondel, Infomap, RN, MCL, PPR

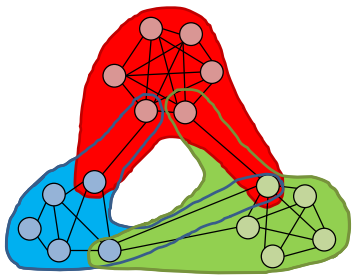
(b) OSLOM



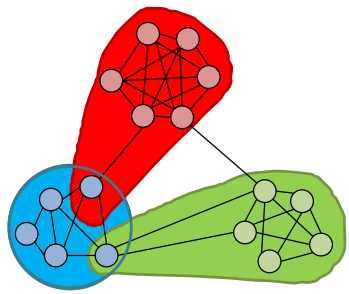
(c) LC



(d) LG



(e) DEMON



(f) SLPA

Figure 1.2: *A comparison of the communities yield by different community detection algorithms on a toy example network.*

The most widely used structural *quality function* is *modularity* [33] which is also widely used as an *objective function* or *scoring function* to be optimized by community detection algorithms. In addition to modularity, many other quality functions have been used and proposed in the literature. However, it has been shown that there is no single perfect quality function for comparison of the quality of the communities identified by different algorithms [34]. Moreover, many of the existing quality functions are designed for evaluating disjoint communities and extending them for evaluation of overlapping communities is not straightforward [16].

One of the methods which is widely used for evaluating and comparing the identified communities by different algorithms is to use synthetic networks from different *benchmarks*. In the GN benchmark [35], communities of the same size are embedded into a network for a given expected degree and a given ratio of internal to external connections between the communities. Other benchmarks have been proposed to improve and complement GN for example for overlapping communities. One such widely used benchmark is the LFR benchmark [36] which introduces heterogeneity into degree and community size distributions of a network.

The main reason for using benchmark graphs for evaluating community detection algorithms, is the lack of *ground truth* information about the communities in real-world networks. Recently, more studies have used ground truth data. Ground truth data is usually obtained from meta data or explicit group memberships of the nodes. Ahn et al. [15] used meta data, e.g., tags assigned by users to annotate the items in a co-purchase network, to define a number of quality functions based on the purity of the attributes of nodes in communities and to assess how well the identified communities reflect the meta data. Abrahao et al. [37] identified ground truth communities from annotations, e.g., product categories and groups of protein functions, and compared the structural properties of the communities detected by different algorithms with ground truth communities. Yang and Leskovec [24] have studied a large number of social, collaboration, and information networks to define ground truth communities based on the explicit declaration of group membership by the nodes. Their comparison of the ground truth communities with different definitions of communities have shown that *conductance* is the best scoring function for networks with well-separated and non-overlapping communities, while the *triad-participation ratio* is the best scoring function for networks with densely overlapping communities.

In this thesis, in addition to the above methods for evaluating community quality, we also propose to evaluate the *logical quality* of the communities identified by different algorithms. The logical quality is defined based on the type of the edges inside communities and how homogeneous these edges are. In other words, the communities in which all of the edges are homogeneous, i.e., are of the same type, are considered to have perfect logical quality (see Section 1.5.2).

1.2.3 Scalability

Identifying high quality communities from large-scale real-world networks is typically computationally expensive and does not scale well. One approach for improving the scalability of community detection is to use parallelism. Parallelism can significantly speed up the community detection and is also necessary for coping with the massive volume of real-world datasets.

Recently, a number of studies have proposed parallel community detection algorithms. Yang and Leskovec [42] proposed BigClam which is a model-based parallel algorithm for community detection. Prat-perez [43] proposed SCD which is a parallel scalable algorithm which identifies disjoint communities.

In addition to designing new parallel algorithms, there has been a number of attempts to parallelize conventional community detection algorithms in order to improve their scalability. Staudt et al. [31] provided the parallel implementation of the Louvain algorithm by Blondel et al. [11] and the label propagation algorithm [38]. Cheong et al. [39] proposed a hierarchical parallel algorithm based on the Louvain algorithm implemented on single- and multi-GPU (Graphics Processing Unit). Soman et al. [40] proposed a community detection algorithm based on label propagation optimized for GPU architectures. Kuzmin et al. [41] proposed a parallel version of the SLPA [21] algorithm for shared and distributed memory machines.

Another fast and scalable approach to community detection is to use local community detection algorithms. In local algorithms, the computations can be done in parallel starting from seed nodes and expanding them into communities by only investigating the neighborhood of the seed nodes in the network. A naive approach to local community detection is to expand every node in the network into a community. However, this approach is computationally expensive and will generate many duplicate communities. Therefore, the challenge is to select an optimal number of seeds to be expanded into communities which can cover the majority of the nodes in a network.

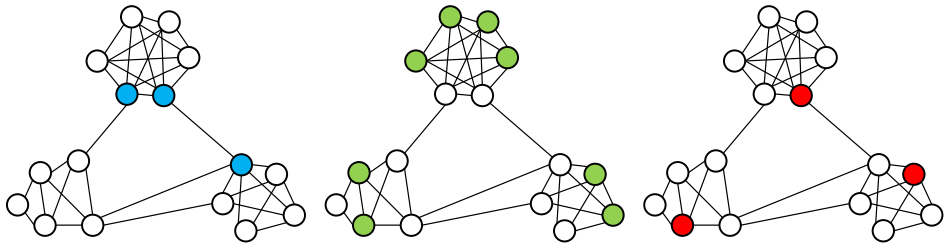
1.2.4 Seed Selection

One of the most successful community detection methods is local seed expansion which is, as mentioned earlier, also very scalable since it is parallelizable by nature. However, the problem of selecting *good seeds* to be expanded into high quality overlapping communities is far from trivial and is not widely studied.

A good seed is usually assumed to have many neighbors inside the target community. Andersen et al. [25] theoretically showed that a seed set that is "nearly contained" in a target community is a good seed set for that community. They also showed that a randomly selected seed set from a target community can also be a good choice for identifying that community. However, Whang et al. [29] showed that careful selection of seeds leads to better results compared to a simple random selection.

One approach for selecting good seeds in a network is to use non-structural knowledge of the network if such information exists. As an example, Gargi et al. [14] have considered non-structural properties of the Youtube video network and have selected the nodes which correspond to videos with the highest view count as the seeds. Unfortunately, such non-structural information might not be available for many types of networks particularly when no global knowledge about the network exists.

In other studies, the structural properties of the networks have been used for seed selection. Shen et al. [44] proposed to use maximal cliques as seeds since they form the core of the communities. However, this approach is computationally expensive. It was shown by Gleich et al. [45] that the *egonets* with low conductance



(a) SH (k=3), MD (b) EC (c) CN+coloring (our algorithm)

Figure 1.3: *A comparison of the seeds yield by different seed selection algorithms on a toy example network.*

(EC) are good seeds for finding the best communities of a network with respect to conductance. However, Whang et al. [29] showed that the communities expanded from these egonets do not achieve a good coverage of the network. Chen et al. [46] proposed an algorithm for selecting the nodes with local maximal degree (MD) as seeds and suggested to repeatedly remove the identified communities expanded from the selected seeds from the network and find new seeds in the remaining parts of the network to improve the coverage.

Whang et al. [29] have proposed two seeding algorithms which can achieve good coverage: Graclus centers and Spread hub. In the *Graclus centers*, first a partitioning algorithm is used in order to find k partitions, where k is pre-determined, and then the nodes in the center of these partitions are selected as seeds. In the *spread hub* algorithm (SH), first the nodes in the network are sorted based on their degree, then the nodes with the highest degree are selected as seeds until at least k nodes are selected. These seeding methods are both shown to perform well in large real-world networks. However, these methods require that the number of seeds to be selected is known in advance. Unfortunately, making assumptions about the number of communities in a network is not realistic since the community structure of real-world networks is normally unknown to us.

Figure 1.3 shows the seed nodes which are selected by different seeding methods. It can be seen that different algorithms pick different nodes as seeds since they take different structural properties of the nodes into account. In this thesis, we propose a new seed selection algorithm which does not require global information about the network nor the number of seeds to be picked, and still is able to select a reasonably small number of good seeds which are well distributed over the network (see Section 1.5.4).

1.2.5 Other Challenges

Despite the excessive number of community detection algorithms proposed in the literature, identifying communities in real-world networks is still a challenge. The

challenges are not limited to quality evaluation of the identified communities and the scalability of the algorithms. Some other challenges, which are not covered in the thesis, but are very important to be studied are as follows.

- Identifying communities in dynamic networks, where new nodes can join, existing nodes can leave the network and new edges can be formed and existing edges can break.
- Studying the stability of communities identified by different algorithms, particularly in evolving networks.
- Combining structural and non-structural information, where such knowledge exists, for identifying more realistic communities.
- Interpreting what the identified communities show about the function of the system and how the output of a community detection algorithm can be used for different applications.

1.3 Applications

Mining large-scale real-world network data has many different applications such as understanding the function of a system, modeling and predicting its behavior, and identifying outliers and anomalies. In this section we present three network data analysis applications which are the focus of this thesis.

1.3.1 Unsolicited Email Detection

Email is one of the most common services on the Internet with everyday business and personal communications depending on it. Unfortunately, the vast amount of unsolicited email (*spam*) consumes network and mail server resources, imposes security threats, and costs businesses significant amounts of money. Spam can also be exploited for phishing and scam and it can carry Trojans, worms, or viruses, making email unreliable.

It is known that a large fraction of spam originates from *botnets* [47, 48]. A botnet is a collection of compromised hosts (*bots*) where each bot contributes to conducting malicious activities or attacks such as distributed denial of service (DDoS), scanning, click frauds, and sending spam. Therefore, identifying the source of spam can lead to the detection of the source of other malicious activities on the Internet.

Numerous attempts to fight spam have led to implementation of anti-spam tools that are quite successful in hiding the spam from users' mailboxes. Most of the conventional approaches inspect email contents at the receiving mail servers, and are very resource-intensive. Although such *content-based filters* are effective in learning what the content of spam looks like, the spammers are very agile in obfuscating email contents and encapsulating their messages in other formats such as images to bypass these filters.

As a complement to content-based filters, *pre-filtering* strategies are widely used to stop spam before the email content is received and examined by the mail servers. A commonly used pre-filtering method is *IP blacklisting*. The receiving mail servers can consult IP blacklists to decide whether to accept or reject an incoming email. However, IP addresses are not persistent, they can be obtained from dynamic pools of addresses and they can be stolen [47, 49]. In addition, bots usually send spam at a low rate to each individual domain and do not reuse IP addresses that have become blacklisted.

In addition to the above mentioned anti-spam strategies, numerous other spam detection and prevention techniques have been introduced. Approaches such as enforcing laws and regulations, requesting proof-of-work (e.g., processing time) [50], mail quota enforcement [51], port blocking, and user monitoring are proposed to stop spam at the sender side. Greylisting [52], reputation-based approaches, sender authentication, and domain verification are approaches that can be used on the receiver side before accepting email contents. Replacing SMTP with a new protocol or deploying overlay authentication protocols, are some other ideas proposed to stop spam during transit.

Recently, approaches that focus on the network-level behavior of spam have gained attention. These approaches are concerned about email sending behavior of the spammers, which is expected to be more difficult for them to change than the content of the email [53–55]. In order to improve and come up with more such methods, there is a need to understand the network-level characteristics of spam and how it differs from legitimate email (*ham*) traffic.

It is known that spam is sent automatically, therefore it is expected that it does not exhibit the *social* properties of human-generated communications [56–59]. The social properties of email communications can be studied by analyzing the structure of *email networks* generated from email traffic. An email network is an implicit social network in which each node represents an email address and each edge represents an email. It has been shown that email networks have the same structural properties that other social and interaction networks have [60–62]. Our intuition is that the structural properties of email networks containing unsolicited email are not similar to the structure of email networks containing only legitimate email. Therefore, analysis of email networks generated from a mixture of email communications can be used for identifying the distinguishing properties of ham and spam which can potentially be used for detecting the botnets based on their anti-social behavior rather than on the content of what they send.

1.3.2 Network Intrusion Detection

Networked systems are continuously under attack causing considerable damages, therefore, network intrusion detection systems are widely deployed. Network intrusions can be identified using two different approaches, i.e., misuse detection and anomaly detection. Techniques for misuse detection rely on the signatures of attacks, and search for patterns of well-known attacks to identify intrusions, there-

fore, they lack the ability to detect new intrusions or zero-day attacks. Anomaly detection techniques, on the other hand, do not require prior knowledge of an attack signature. However, they might have a high false positive rate.

In this thesis, we focus on anomaly detection-based intrusion detection systems. Anomaly detection has been extensively studied in the context of different application domains and a variety of techniques have been proposed. An overview of anomaly detection methods can be found in [63].

Anomalies are patterns in network traffic that do not conform to normal behavior. Any change in the network usage behavior or malicious activities such as DoS attacks, port scanning, unsolicited traffic, and worm outbreaks, can be seen as anomalies in the traffic.

The main challenge in using anomaly detection for identifying misbehaving hosts is to define normal behavior and draw boundaries between normal and abnormal communication patterns. One approach to defining normality is to look into the social behavior of normal nodes. Since many types of intrusions are automatically generated, it is expected that they do not conform to the expected normal social behavior. Therefore, a number of features that are representative of (anti)social communication patterns can be extracted for identification of misbehaving nodes.

Recently, it was shown that network intrusions can successfully be detected by examining the network communications that do not respect the community boundaries [64]. In such an approach, normality is defined with respect to social behavior of nodes concerning the communities to which they belong and intrusion is defined as "*entering* communities to which one does not belong". In this thesis we propose an alternative definition for anomaly/intrusion and study how the network structure and the community structure of graphs generated from network traffic can be used for network misbehavior detection (see Section 1.5.3).

1.3.3 Query Analysis

Logs of search engines contain a wealth of information from the queries submitted by users. Query logs have been widely studied and analyzed in order to improve the service provided to the users and to better understand their behavior and needs. Analysis of web query logs can provide useful information regarding the use of a site considering when and how users seek information for topics covered by the site [65]. Extracting information from query logs can also be useful for different types of users such as terminologists, infodemiologists, and web analysts, as well as specialists in Natural Language Processing (NLP) technologies such as information retrieval and text mining.

Medical and health information seeking on the Internet is quite common. Mining query logs of medical search can be beneficial to public officials in health and safety organizations, epidemiologists, and medical data analysts. Information extracted from large-scale logs can be used both for a general understanding of public health awareness and the information seeking patterns of users, and for optimizing

search indexing, recommendations, query completion and presentation of results for improved public health information.

In order to study query logs, several graph-based relations among queries can be used [66]. A co-occurrence network for the words which co-occurred in different queries is an information network which we use to capture the relations between the words. We further study the structural and temporal properties of the co-occurrence network and show that it is similar to other information and social networks. We also look into the community structure of the network and how the identified communities can potentially be used for improving our understanding of the language used by users of the health care portal and improving their search experience (Section 1.5.5).

1.4 Data Collection

Getting access to and performing analysis of large-scale real-world datasets is crucial for many different applications. Collection and processing of real data is far from trivial. The challenges involved are both of general and technical nature. Getting access to the data, privacy and ethical concerns, pre-processing and analysis of the dataset are just a number of challenges that need to be addressed before the data can be used for an application. The main challenge, however, is handling the massive amount of data. The data collection process has to keep up with the speed in which the data is being produced or received. It is usually inevitable to sample the data, to process summaries of the data or to only focus on analyzing snapshots of data obtained during limited time windows. In some cases such as Internet traffic collection, special measurement equipments which can cope with full link-speed or allow high sampling frequencies are required. After the collection, the data also needs to be parsed or pre-processed before it is possible to extract relevant information for example to create a network from the relations observed in the datasets. In many cases, obtaining ground-truth data for evaluating the results of the data analysis can also be impossible or non-trivial. In this thesis we have collected and obtained different types of real data including data captured from a high speed Internet backbone link, data from social and information networks, and query log files from a health care portal.

1.4.1 Email Dataset

One of the datasets which is collected by us is an email dataset which is used for understanding the characteristics of legitimate and unsolicited email. The study of the characteristics of email and spam can be conducted using different types of email data. A number of studies have used SMTP log files from mail servers [49, 57, 59, 67–69]. Although such datasets are limited to communications to/from a single domain, they contain detailed information about each email and the statistical summaries of accepted and rejected email communications, which

allows comparison of the behavior of spam, ham, and the rejected traffic. The spam captured in honeypots or relay sinkholes have also been used to study the characteristics of spam [53, 70]. The honeypots only attract spammers, therefore they do not allow the comparison of different characteristics and communication patterns of spam and ham. Flow-level data collected on access routers have also been used to study the properties of spam and rejected traffic [71]. These flows only contain packet headers, and although they are not limited to a single domain, they do not carry enough information to allow distinguishing spam from ham to study their distinct characteristics. Another type of data that has been used to understand the sending behavior of spam was collected from inside spam campaigns [48, 72, 73]. The data collected at these campaigns has the view point of spammers and makes it possible to closely investigate how spam is sent.

In our studies, we have used yet another type of email data. Our dataset enables us to study the behavior of legitimate and unsolicited traffic from the perspective of a network device which monitors the traffic traversing a backbone link. The collected email traffic is not limited to a single organization or domain and allows us to classify the observed email into ham, spam, and rejected communications to compare their characteristics.

Collection of large datasets from backbone Internet traffic can face several challenges [74]. Not only is mere physical access to optical Internet backbone links needed, but also rather expensive equipment in order to deal with the large data volumes arriving at high speeds. Adding to the complexity, the collected data traces must be desensitized since they may contain privacy-sensitive data. Packets also need to be reassembled into application level "conversations" so that, finally and maybe the most challenging part, methods and algorithms suitable for analysis of massive data volumes can be run [75].

Our datasets were generated passively capturing traffic on a 10 Gbps backbone link of SUNET (the Swedish University Network) [76]. The collection location is shown in Figure 1.4. Each dataset was collected over 14 consecutive days with roughly a year time span between them.

The process of collecting data and generating the first dataset is described in more detail in the following. Table 1.2 summarizes the collected data during 14 consecutive days in March 2010. The second dataset was also collected similarly during 14 consecutive days in spring 2011.

We used a hardware filter to only capture traffic to and from *port 25* which resulted in more than 183 GB of SMTP data. The captured packets belonging to a single flow were then aggregated to allow the analysis of complete SMTP sessions.

The collected data contained both *SMTP requests* and *SMTP replies*. As each SMTP request flow corresponds to an SMTP session, it can carry one or more emails, thus we had to extract each email from the flows by examining the SMTP commands. The resulting extracted email transaction contained the SMTP commands including the email addresses of the sender and the receiver(s), email headers, and the email content.

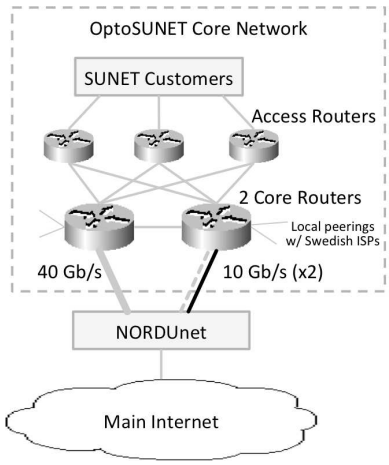


Figure 1.4: *OptoSUNET core topology. All SUNET customers are via access routers connected to two core routers. The SUNET core routers have local peering with Swedish ISPs, and are connected to the international commodity Internet via NORDUnet. SUNET is connected to NORDUnet via three links: a 40 Gbps link and two 10 Gbps links. Our measurement equipment collects data on the first of the two 10 Gbps links (black) between SUNET and NORDUnet.*

After the collection phase, first the dataset was pruned of all unusable email traces. For example, flows with no payload are mainly scanning attempts and should not be considered in the classification. Also, SMTP flows missing the proper commands were excluded from the dataset as they most likely belong to other applications using port 25. Encrypted email communications cannot be analyzed, and were also eliminated.[1] Any email with an empty sender address is a notification message, such as a non-delivery message [77]; it does not represent a real email transmission and was also excluded. Finally, any email transaction that was missing either the proper starting/ending or any intermediate packet was considered as incomplete. Possible reasons for having incomplete flows include transmission errors and measurement hardware limitations caused by a framing synchronization problem.

The remaining email transactions were then classified as *accepted*, i.e. those emails that are delivered by the mail servers, or *rejected*. An email transaction can fail at any time before the transmission of the email data (header and content) due to rejection by the receiving mail server. Therefore, *rejected* emails are those that do not finish the SMTP command exchange phase and consequently do not send any email data. The rejections are mostly because of spam pre-filtering strategies

[1] Around 3.8% of the flows carried encrypted SMTP sessions.

Table 1.2: *Email dataset statistics (2010).*

	Incoming ($/10^6$)	Outgoing ($/10^6$)
Packets	626.9	170.1
Flows	34.9	11.9
Distinct source IPs	2.30	0.01
Distinct destination IPs	0.57	1.94
SMTP Replies	2.84	9.14
Email:	19.3	0.73
Ham email	1.32	0.21
Spam email	1.66	0.20
Rejected email	16.3	0.31

deployed by mail servers including blacklisting, greylisting, DNS lookups, and user database checks.

Finally, we discriminated between *spam* and *ham* in our dataset. As we have captured the complete SMTP flows, including IP addresses, SMTP commands, and email contents, we can establish a ground truth for further analysis of *only* the spam traffic properties and a comparison with the corresponding legitimate email traffic. We deployed the widely-used spam detection tool called SpamAssassin[2] to mark emails as spam and ham. SpamAssassin uses a variety of techniques for its classification, such as header and content analysis, Bayesian filtering, DNS blocklists, and collaborative filtering databases.[3]

The final pre-processing step of the dataset was to desensitize any user data. Immediately after the classification of emails into ham and spam, we discard the content of the emails and anonymized the email and IP addresses in the headers [75]. Once the sensitive data was discarded, the resulting anonymized dataset had a size of 37 GB.

The second dataset from 2011 was collected and pre-processed similarly to the first dataset. The infrastructure and the data collection equipment was updated during the one year time span between the collections. Although, the changes have caused differences in the collected data, these differences are in our favor since they allow us to compare our observations over time and verify that our findings are not limited to a single vantage point.

[2] `http://spamassassin.apache.org`

[3] The well-trained SpamAssassin applied to our dataset was in use for a long time at our university, incurring an approximate false positive rate of less than 0.1%, and an detection rate of 91.3% after around 94% of the spam being rejected by blacklists.

Table 1.3: *Unique hosts during the data collection 2010-04-01.*

	Inside SUNET		Outside SUNET	
Incoming Link	Destination IPs	970,149	Source IPs	24,587,096
Outgoing Link	Source IPs	23,600	Destination IPs	18,780,894

1.4.2 Flow Dataset

In order to study other types of misbehavior in network traffic such as network intrusions, we have used network flow data collected from the backbone link of SUNET. The network flow data was collected from the same location as the email dataset (see Figure 1.4).

For a period of more than six months, a 24 hour snapshot of all flows was regularly collected once a week. The dataset contains a total of 12 billion flows in both directions. Table 1.3, summarizes all unique IP addresses found during a single collection day to give an idea of the scale of the traffic passing by the measuring point.

This dataset also contains metadata, including, for example, hosts known to aggressively spread malware at the time of the collected snapshots. The source addresses of these malicious sources in the dataset were defined by using the lists reported by DShield and SRI Malware Threat Center during the data collection period [78, 79]. By using the flow data together with this information, we can then make more targeted types of analysis of hosts, despite their addresses being anonymized.

We have used flow data from seven days in the dataset in order to study a community-based network intrusion detection method (Section 1.5.3). More details about the collection of the dataset and other analysis performed on the data can be found in [80].

1.4.3 Social and Information Network Datasets

In addition to data from real network traffic, we have used data from other types of social and information networks. We have used publicly available datasets provided by the Stanford Large Network Collection [81] including a product network from Amazon, a collaboration network from DBLP computer science bibliography, and the social networks of users in Youtube and Livejournal. These datasets also include the information about the ground truth communities.

In the *Amazon* network, nodes are products in the Amazon website and two products have an edge if they were co-purchased frequently. The ground truth is based on the product categories defined by Amazon. In the *DBLP* network, nodes are authors and two authors are connected with an edge if they have co-authored at least one paper, and the ground truth is obtained based on the publication venues. In the *Youtube* and *LiveJournal* networks the nodes are the users of the

video sharing and online blogging websites, respectively, and the edges correspond to friendships and the ground truth is based on user-defined groups.

In addition to above datasets, we have collected a dataset from the *SoundCloud* sound sharing site (`http://soundcloud.com/`). In SoundCloud, similar to Twitter, users can follow each other, and popular artists tend to attract a large number of followers. For the collection of Soundcloud data, we alternated between random sampling and breadth-first-search, so that we could capture local neighborhood information while covering different parts of the network [82]. After data collection, we generated a network of "follow" relations, where the nodes are the users, and an edge (u, v) exists if the user u follows the user v.

The data collection from SoundCloud is an ongoing process and by the time this thesis is being written, we have collected data from more than 39 million users with more than 642 million follows and around 76 thousand groups. We are going to publish a more complete version of the datasets after finishing the collection process. By the time we started to use the SoundCloud dataset, we had around 5 million users in the dataset. Even though our work is focused on a small subset of the whole user base, this network has been the largest social network which we used in our studies. In this thesis, we have used the datasets presented in this section for evaluating our proposed local seed selection algorithm. Our algorithm selects seeds by merely investigating the direct neighborhood of each node in the network and therefore does not require the global structure of the network to be accessible, so our analysis is not affected from the lack of global data.

1.4.4 Medical Query Logs

The last dataset which we used was obtained from the query logs of a Swedish health care portal. We obtained 67 million queries for the period October 2010 to the end of September 2013. The data was provided by vardguiden.se through an agreement with the company Euroling AB which provides indexing and searching functionality to vardguiden.se. 27 million of the queries are unique before any kind of normalization, and 2.2 million after case folding.

The obtained queries are then automatically annotated with semantic labels using two medically-oriented semantic resources, i.e., the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) and the National Repository for Medicinal Products (NPL), as well as a named entity (including the ontological categories location, organization, person, time, and measure entities) recognizer. We used these labels to identify semantic communities based on the co-occurrence of words in the queries.

Moreover, from each query which contained more than one word/term, we extracted the words and created a network of word co-occurrences. We are interested in analyzing the relations between the words and the language being used in the queries, so the single-word queries were not of interest to us. This network was used for structural analysis and identification of graph communities.

Overall, the semantic and graph analysis of query logs can be of great interest for different types of studies and can reveal important information about the usage patterns, information needs, and the language of the users of the website (Section 1.5.5).

1.5 Our Approach

As presented in the previous section we have collected and obtained large volumes of real-world data and constructed different networks from the datasets and studied their structural properties. In this section we summarize our approaches towards the different applications which we had at hand. The details of our approaches are covered in the appended papers.

1.5.1 Structural and Temporal Analysis of Email Networks

In order to understand the characteristics of unsolicited email traffic and how they differ from legitimate traffic, we have performed a *social network analysis* of real email traffic (Section 1.4.1). Our hypothesis is that social network analysis of email traffic can reveal the differences in the communication patterns of legitimate and unsolicited email traffic and can be used for identifying the sources of spam.

In order to verify our hypothesis, we have generated *email networks* from the observed email communications in which each node represents an email address and each edge represents an observed email communication between a pair of nodes. The generated email network from the larger dataset contains 10,544,647 nodes and 21,562,306 edges, and the email network from the smaller dataset contains 4,525,687 nodes and 8,709,216 edges. Based on our ground truth, we have generated a number of ham, spam, rejected, and complete email networks, and have studied and compared their structural and temporal properties. We have looked into the (in-/out-)degree distribution, average shortest path length, average clustering coefficient, distribution of the size of the connected components, the percentage of total nodes in the giant connected component, as well as how these properties change over time as the networks grow.

Our study reveals that the legitimate email traffic exhibit similar structural properties as other social and interaction networks, and therefore a ham network can be modeled as a scale-free small-world network. We also show the similarities and differences in the structural and temporal properties of email networks of ham and spam, and show that the anti-social behavior of spam and rejected traffic is not hidden in a mixture of email traffic and causes anomalies (outliers) in the structural properties of email networks. We also propose a method for identifying spamming nodes by finding the outliers in the structural properties of email networks which mainly are caused by the spammers.

1.5.2 Evaluation of Community Detection Algorithms

Despite the excessive number of studies on community detection there is no consensus on a definition for a community and different community detection algorithms have been proposed in the literature based on the different definitions. Therefore, it is not clear how to evaluate which algorithm is most suitable to be used for different types of networks. Moreover, due to the ambiguity in the definitions for community, assessing the quality of the communities identified by different algorithms can be challenging.

In this thesis, we have conducted an empirical study to compare and evaluate a variety of community detection algorithms based on a set of structural and logical quality functions on our email networks. We have evaluated the structural quality of the communities using different well-known and widely-used quality functions, namely modularity, coverage, and conductance. We have also proposed to use the *logical quality* of the communities based on how homogeneous the edges inside the communities are. A community which only contains the same type of edges is considered to have a perfect logical quality. Our aim is to find the most suitable approach that can separate ham and spam emails from the mixture of traffic into distinct communities.

Our study shows that both ham and spam networks, as well as networks containing a mixture of both, exhibit a community structure, and that different community detection algorithms can be used to unfold the communities of these networks. However, we also show that there is a trade-off in creating high structural quality and high logical quality communities. We reveal that although different community detection algorithms use different approaches to define and extract the communities of a network, algorithms that create communities with similar granularity and size distribution also achieve similar structural and logical qualities. We confirm that community detection algorithms which find coarse-grained communities achieve high structural quality. However, we reveal that they fail to find communities with high logical quality since they tend to combine smaller homogeneous communities into mixed communities in favor of better structural quality. We also show that an edge-based community detection algorithm can achieve a high logical quality since it can separate ham and spam emails into distinct communities.

1.5.3 Identifying Misbehavior Using Community Detection Algorithms

Recently, it was shown that the community structure of a flow network can be used for successful intrusion detection [64]. In a community-based anomaly detection method, normality is defined with respect to the social behavior of nodes concerning the communities to which they belong. Nodes that participate in anti-social communications and disrespect community boundaries by "*entering* communities to which they do not belong" can be identified as anomalous by a community-based anomaly detection method. Despite the fact that these methods use a notion of

community, Ding et al. [64] showed that a traditional modularity maximizing community detection algorithm is not suitable for intrusion detection in network flow data since the majority of intruders end up inside a large community and do not enter other communities.

Our intuition is that, in contrast to Ding et al. [64], community detection algorithm can be used for successful network anomaly/intrusion detection. In order to verify this, we look into communities identified by different types of community detection algorithms to extend and complement the work in [64]. Our hypothesis is that misbehaving nodes tend to *belong to multiple communities*. However, a vast variety of community detection algorithms partition network nodes into disjoint communities where each node only belongs to a single community, therefore they cannot be directly used for verifying our hypothesis. Therefore, we introduce *auxiliary communities* to enhance non-overlapping community detection algorithms. This enhancement is achieved by adding a layer of auxiliary communities over the boundary nodes of neighboring communities, allowing nodes to be members of several communities. Therefore, this enhancement enables us to show that, in contrary to [64], it is possible to use community detection algorithms for identifying anomalies in network traffic.

In addition to traditional community detection algorithms, numerous overlapping algorithms exist which allow a node to belong to several overlapping communities [16]. We also compare our proposed enhancement method for non-overlapping community detection algorithms with a number of overlapping algorithms for network anomaly detection, and show that they have comparable performance.

Finally, we propose a framework for network misbehavior detection. The framework allows us to incorporate a community detection algorithm for identifying anomalous nodes that belong to multiple communities. However, since legitimate nodes can also belong to several communities [24], we also introduce a number of application-specific filters based on different graph properties to be used for discriminating the legitimate nodes from the anti-social nodes in the community overlaps, thus reducing the induced false positives. Our experiments show that our framework is suitable for identifying intruders and the sources of scanning attacks from flow networks, and the sources of spam from email networks.

1.5.4 Local Seed Selection for Overlapping Community Detection Algorithms

Local community detection algorithms are gaining more attention than global algorithms which require the structure of the whole network to be known. In local algorithms, first local communities are identified independently of each other only based on local knowledge of the network, then they are combined to provide the global community structure of the network. Local algorithms are easy to parallelize and therefore can scale well. However, the selection of good seeds to be expanded into communities that achieve good coverage of the network is challenging. Our

aim is to design a local *seeding algorithms* which can select a reasonable number of seeds which are well-distributed over the network and therefore can lead to communities covering the majority of the nodes.

Existing seeding algorithms either require a global knowledge of the entire network to be available or they will fail to pick an adequate number of seeds which can lead to incomplete coverage of the network. Therefore, in this thesis we further study the problem of local seed selection for finding a reasonably small number of seeds. The seeds identified by such a seeding algorithm can then be expanded into high quality overlapping communities using high quality local community detection algorithms such as the Personalized PageRank-based algorithm (PPR) [24, 83].

We propose a novel seed selection algorithms for local overlapping community detection. First, we define a *similarity score* which is calculated as the sum of the similarity of a node with all of its connected neighbors by adopting the *similarity indices* from *link prediction* techniques. In link prediction, the aim is to estimate connections that are very likely to be formed between nodes in a network, therefore link prediction methods typically use a similarity index to calculate the similarity of the nodes which are not directly connected. If two nodes have a high similarity, it is predicted that an edge will be formed between them. However, in our algorithm, we use similarity indices to calculate the similarity of the nodes which already share an edge. Our intuition is that a node that has a high aggregated similarity with its neighbors is expected to belong to the same community as its neighbors. Therefore, we propose to select the node with the highest score in its neighborhood as a seed and expand it into a community. We have compared a number of different widely used similarity indices for our seeding algorithm and have also compared our seeding algorithm with a number of existing local seeding algorithms.

Although we show that by using similarity scores we can identify a small number of very good seeds, we can also show that similar to other local seeding algorithms, the expanded communities from these seeds do not achieve a high coverage of the network. Therefore, we propose to use distributed random graph coloring for enhancing our local seed selection algorithm. In order to combine similarity scores with graph coloring for seed selection, we propose a *biased graph coloring algorithm* in which the nodes with high similarity score are assigned a specific color and color conflicts between neighbors are resolved at random. This enhancement of our seeding algorithm makes sure that good seeds which have received the specific color are well distributed over the network. Our biased coloring algorithm can also be used for enhancing and improving other existing local seeding methods.

Our novel local seeding algorithms is parameter free, finds seeds that are well distributed over the network, and does not pick neighboring nodes as seeds and therefore does not lead to many duplicate communities. We empirically evaluate the execution time of local community detection when seeding is used as the first step of community detection and compare the quality and the coverage of the communities expanded from the selected seeds using large-scale real-world networks. Our experiments show that by using seeding, the execution time of community

detection is dramatically reduced and the average quality of the communities is preserved and a high coverage is achieved.

1.5.5 Graph-based Analysis of Medical Queries

Large search query logs carry a wealth of information about the behavior of the users in information seeking and the language they use. Similar to many other types of data, query log files can also be modeled as networks.

Our hypothesis is that graph-based analysis of words which have co-occurred in different queries can provide a better understanding of the relations of words and terms in different domains and in different languages. In order to verify our hypothesis, we have generated a *word co-occurrence network* from the query logs of a Swedish health care website. We study the structural and temporal properties of the generated network and show that it is similar to other existing information and social networks. We also look into the community structure of the word co-occurrence network in order to understand the relation between the words in a medical domain.

Moreover, we have introduced *semantic communities* which are communities of words which have co-occurred with a semantic label. These labels are added to the queries using medically-oriented semantic resources. We also apply a personalized PageRank-based community detection algorithm to the generated word co-occurrence network and compare the identified *graph communities* with the semantic communities. Our experiments show that while semantic communities can cover only a small percentage of all the words in the logs, the graph communities can cover the vast majority of the words. Therefore, the graph-based analysis can capture more relations among the words which have been used in the queries. Moreover, the graph and semantic analysis capture different relations between the words and identify communities which are only partially similar and therefore can be used to complement each other. Overall, our graph-based approach can be used as the first step towards a better understanding of the language usage in medical domain as well as for providing better services and recommendations to the users of the health care portal.

1.6 Summary of Contributions

This section summarizes the contributions of the papers included in this thesis.

1.6.1 PAPER I

In this paper, we show that an email network generated from legitimate email traffic collected on an Internet backbone link (a ham network) can be modeled as a scale-free small-world network similar to other social and interaction networks. We also show the similarities and the differences in the structure of ham and spam

networks and how they change over time. We reveal that the anti-social behavior of spam is not hidden in a mixture of email traffic and causes anomalies (outliers) in the structural properties of email networks. Moreover, we propose a simple method for identifying the nodes that correspond to outliers in the degree distribution of email networks and show that they are mainly sending spam.

1.6.2 PAPER II

In this paper, we study the community structure of ham, spam, and email networks generated from real email traffic and compare a number of well-known community detection algorithms for identifying the communities of these networks. Our experiments reveal that there is a trade-off in creating high structural quality and high logical quality communities. We propose to evaluate the logical quality of the communities based on the homogeneity of the edges inside each community, and show that regardless of the approaches used to define and extract communities, the algorithms that create communities with similar granularity and size distribution also achieve similar structural and logical qualities. We also show that the most successful community detection algorithm for achieving high logical quality (i.e., clustering ham and spam emails into distinct communities), finds overlapping communities by partitioning the edges of the network instead of the nodes.

1.6.3 PAPER III

In this paper, we extend and complement the previous work on community-based intrusion detection. We hypothesize that misbehaving nodes tend to *belong to multiple communities*. To investigate our hypothesis, we consider different definitions for communities, and propose a framework in which different types of community detection algorithms can be used as the basis for network anomaly and intrusion detection. We propose two enhancement methods for adding auxiliary communities over the disjoint communities identified by non-overlapping community detection algorithms. We show that by using our enhancement methods, it is possible to use traditional community detection algorithms for identifying anomalies in network traffic which is in contrast to the observations in [64].

Moreover, we propose a framework that allows us to incorporate communities identified by overlapping algorithms for identifying anomalous nodes that belong to multiple communities. We show that the algorithms which tend to identify coarse-grained communities are not suitable for network misbehavior detection. We also propose to use application-specific filters to filter out legitimate nodes which can naturally belong to several communities. Our experiments reveal that our framework is suitable for identifying scanning nodes from network flow traffic as well as spammers from email traffic.

1.6.4 PAPER IV

In this paper, we propose a novel distributed seed selection algorithm for local overlapping community detection. We define a similarity score using the similarity indices from link prediction techniques and propose an algorithm in which each node compares its similarity score with all its neighbors, and the nodes which have the highest score in their neighborhood are selected as seeds. We show that this algorithm succeeds in selecting a small number of very good seeds which are expanded into high quality communities but cannot cover the whole network. We also propose to use graph coloring for enhancing our local seed selection algorithm in order to improve the coverage. We propose a *biased graph coloring* algorithm in which the nodes with high similarity score are assigned a specific color and color conflicts between neighbors are resolved at random. Our experiments using large-scale real-world social networks show that our seeding algorithm is fast, and leads to high quality communities with a good coverage of the networks.

1.6.5 PAPER V

In this paper, we create a word co-occurrence network from query log files obtained from a medical and health care portal. We show that this network has the same structural and temporal properties that other information networks exhibit. We use a local overlapping community detection algorithm to identify the communities in the co-occurrence network. We also use the semantic labels assigned to the queries in the log files and define semantic communities which are communities of words which have co-occurred with a semantic label. We compare the graph communities with the semantic communities and show that our graph-based analysis of queries can improve and complement the semantic analysis. We also study how the length of the time window in which queries are observed can affect our graph-based analysis.

1.7 Conclusions and Future Work

In this thesis, we have looked into algorithms and methods for analyzing networks generated from large-scale real-world datasets. Particularly, we have focused on the community structure of networks and have looked into the challenges and the applications of community detection algorithms.

One of the challenges in identifying communities in a network is the selection of the most suitable algorithm for the network, since different algorithms use different definitions for communities and use different methods for identifying the communities. In this thesis we have performed an empirical comparison and evaluation of a number of different community detection algorithms and show that there is a trade-off between the structural and the logical quality of the communities identified by different algorithms. Therefore, an algorithm which can create communities

with very high structural quality might not be the most suitable algorithm for the application at hand, for example, separating different types of edges into distinct communities.

Another challenge in using community detection algorithms for analysis of large datasets is scalability. It has been shown that local seed expansion algorithms are very successful in fast and scalable detection of high quality communities. In this thesis, we have proposed a fast local seed selection algorithm which can be used as a pre-processing step for local community detection using seed expansion. Our algorithm can dramatically reduce the execution time of community detection while preserving the quality of the identified communities and achieving a good coverage of the network. Moreover, there are many interesting trade-offs between the number of selected seeds, the quality, and the coverage of communities which can be further studied. Another property which can further be taken into account for seed selection is to reduce the number of duplicate communities.

In addition to investigating and addressing some of the challenges of community detection, we have also looked into some of the applications of network analysis and community detection. One of the applications which has been considered in this thesis is identifying the source of unsolicited email. Our goal has been to reveal the differences and similarities in the communication patterns of legitimate and unsolicited email by mining email networks generated from traffic seen on an Internet backbone link. To pursue this goal, we have taken a social network analysis approach and show that the behavior of spam senders causes anomalies in the structural properties of email networks, and these anomalies can be detected using an outlier detection approach. We can also show that spam and ham, which are mixed in the observed traffic, can be separated into distinct communities by deploying a link community detection algorithm. Moreover, we have proposed a framework for network misbehavior detection which takes advantage of overlapping communities for identifying sources of spam as well as sources of other types of malicious traffic such as scanning. We are able to show that misbehaving nodes belong to multiple communities and they can be identified by either using overlapping community detection algorithms or by enhancing non-overlapping algorithms with auxiliary communities.

The proposed approaches in this thesis for identifying sources of misbehavior are promising and can potentially be used to complement existing anti-spam and intrusion detection methods. The advantage of deploying our approaches is that they provide us with the possibility of stopping unwanted traffic closer to its source by merely observing the communication patterns of network traffic, for example email communications. However, there is more work to be done before our findings can be deployed practically as part of a working anti-spam or intrusion detection tool. One desirable future direction is to investigate how our methods can be combined with each other to be used as a stand-alone detection system or in cooperation with existing tools. One possibility is to deploy a network device that monitors the traffic on a link and that is able to tag suspicious traffic or populate

a blacklist. Moreover, a study of the robustness of our findings in order to see how easy it is for the spammers or intruders to change their sending behavior and how easy it is to evade detection is another future research area.

Another goal of this thesis has been to improve community detection algorithms so that they could be used for different applications. We have introduced auxiliary communities to enhance existing non-overlapping community detection algorithms in order to identify sources of misbehavior from real network traffic. However, our approach can potentially be extended for converting disjoint communities into overlapping communities which will allow the use of existing non-overlapping community detection algorithms for identifying overlapping communities.

In this thesis, we have also shown how to use network mining and community detection methods to analyze other types of large datasets such as the query logs obtained from a Swedish health care portal. A future direction is to improve our graph-based query analysis by improving the pre-processing of the data, for example by representing different variations of words with a single node in the word co-occurrence network, filtering out non-medical related words, and introducing edge weights based on the frequency of word co-occurrences. Moreover, other information from the logs can be deployed to better understand the language used by users and to be able to improve the search experience of the users by providing better suggestions and recommendations to them.

Overall, with advances in technology and computation and proliferation of smart and mobile devices, new opportunities for collecting and analyzing big data emerge and more and more applications can benefit from the extracted knowledge from the data. Therefore, there is an increasing need for fast, dynamic, and scalable solutions which also open more research questions. One of the challenges is to design new network mining algorithms and to improve existing ones to run in parallel and in distributed settings. The designed parallel and distributed algorithms also need to cope with the lack of global knowledge of the networks, as well as the dynamically changing structure of networks. Moreover, there is also a need for improving the quality of the network mining algorithms, particularly community detection algorithms. Recent studies using ground truth data have revealed that existing community detection algorithms are not very successful in identifying the real communities in large networks, therefore new approaches to community detection which for example take non-structural properties of the networks into account, are desirable. Another interesting future research direction is to develop efficient methods, such as visualization, for interpreting the output of different graph algorithms, to allow better understanding of the structure of networks and identifying interesting patterns and anomalies. Finally, extending the applicability of network mining algorithms to more real-time domains and applications is another challenging future direction.

Bibliography

[1] MEJ Newman and Juyong Park, "Why social networks are different from other types of networks," *Physical Review E*, vol. 68, no. 3, Sept. 2003.

[2] Paul Erdos and Alfred Renyi, "On The Evolution of Random Graphs," *Publication of the Mathematical Institute of the Hungarian Academy of Science*, vol. 5, pp. 17–61, 1960.

[3] D J Watts and S H Strogatz, "Collective Dynamics of 'Small-World' Networks.," *Nature*, vol. 393, no. 6684, pp. 440–2, June 1998.

[4] S Milgram, "The Small World Problem," *Psychology today*, vol. 2, pp. 60–67, 1967.

[5] A.L. Barabási and R. Albert, "Emergence of Scaling in Random Networks," *Science*, vol. 286, no. 5439, pp. 509, 1999.

[6] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On Power-Law Relationships of the Internet Topology," *ACM SIGCOMM Computer Communication Review*, vol. 29, no. 4, pp. 251–262, 1999.

[7] Reka Zsuzsanna Albert, *Statistical Mechanics of Complex Networks*, Ph.D. thesis, University of Notre Dame, 2001.

[8] Leman Akoglu, *Mining and Modeling Real-world Networks : Patterns , Anomalies , and Tools*, Ph.D. thesis, Carnegie Mellon University, 2012.

[9] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos, "Graphs Over Time: Densification Laws, Shrinking Diameters and Possible Explanations," in *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining - KDD '05*. 2005, p. 177, ACM Press.

[10] W.W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, pp. 452–473, 1977.

[11] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre, "Fast Unfolding of Communities in Large Networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, pp. P10008, Oct. 2008.

[12] T. Evans and R. Lambiotte, "Line graphs, link partitions, and overlapping communities," *Physical Review E*, vol. 80, no. 1, pp. 1–8, July 2009.

[13] Santo Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, Feb. 2010.

[14] Ullas Gargi and Wenjun Lu, "Large-Scale Community Detection on YouTube for Topic Discovery and Exploration," in *Proceedings of the Fifth International Conference on Weblogs and Social Media*. 2011, The AAAI Press.

[15] Yong-Yeol Ahn, James P Bagrow, and Sune Lehmann, "Link communities reveal multiscale complexity in networks.," *Nature*, vol. 466, no. 7307, pp. 761–4, Aug. 2010.

[16] Jierui Xie, S Kelley, and BK Szymanski, "Overlapping community detection in networks: the state of the art and comparative study," *ACM Computing Surveys*, vol. 45, no. 4, 2013.

[17] Martin Rosvall and Carl T Bergstrom, "Maps of random walks on complex networks reveal community structure.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 4, pp. 1118–23, Jan. 2008.

[18] Martin Rosvall and Carl T Bergstrom, "Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems.," *PloS one*, vol. 6, no. 4, pp. e18209, Jan. 2011.

[19] Peter Ronhovde and Zohar Nussinov, "Multiresolution community detection for megascale networks by information-based replica correlations," *Physical Review E*, vol. 80, no. 1, pp. 1–18, July 2009.

[20] Stijn VAN Dongen, *Graph Clustering by Flow Simulation*, Ph.D. thesis, University of Utrecht, The Netherlands, 2000.

[21] Jierui Xie and BK Szymanski, "Towards Linear Time Overlapping Community Detection in Social Networks," in *the 16th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining*. 2012, pp. 25–36, Springer-Verlag.

[22] Andrea Lancichinetti, Filippo Radicchi, José J Ramasco, and Santo Fortunato, "Finding statistically significant communities in networks.," *PloS one*, vol. 6, no. 4, pp. e18961, Jan. 2011.

[23] Michele Coscia, Giulio Rossetti, Fosca Giannotti, and Dino Pedreschi, "DEMON: a local-first discovery method for overlapping communities," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*. 2012, p. 615, ACM Press.

[24] Jaewon Yang and Jure Leskovec, "Defining and evaluating network communities based on ground-truth," in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2012, pp. 1–8.

[25] Reid Andersen and Kevin Lang, "Communities from seed sets," in *Proceedings of the 15th international conference on World Wide Web - WWW '06*. 2006, p. 223, ACM Press.

[26] Daniel A Spielman and Shang-Hua Teng, "Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems," in *Proceedings of the 36th annual ACM symposium on Theory of computing - STOC '04*. 2004, p. 81, ACM Press.

[27] Aaron Clauset, "Finding local community structure in networks," *Physical Review E*, vol. 72, no. 2, pp. 026132, Aug. 2005.

[28] Sucheta Soundarajan and John E Hopcroft, "Use of Local Group Information to Identify Communities in Networks," *ACM Transactions on Knowledge Discovery from Data (to appear)*, 2014.

[29] Joyce Jiyoung Whang, David F Gleich, and Inderjit S Dhillon, "Overlapping community detection using seed set expansion," in *Proceedings of the 22nd ACM international Conference on information & knowledge management - CIKM '13*. 2013, pp. 2099–2108, ACM Press.

[30] Marek Ciglan, Michal Laclavík, and Kjetil Nø rvåg, "On community detection in real-world networks and the importance of degree assortativity," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13*, New York, New York, USA, 2013, p. 1007, ACM Press.

[31] Christian L Staudt and Henning Meyerhenke, "Engineering High-Performance Community Detection Heuristics for Massive Graphs," in *International Conference on Parallel Processing*, 2013.

[32] Satu Elisa Schaeffer, "Graph Clustering," *Computer Science Review*, vol. 1, no. 1, pp. 27–64, Aug. 2007.

[33] M E J Newman, "Modularity and community structure in networks.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 23, pp. 8577–82, June 2006.

[34] Helio Almeida, Dorgival Guedes, Wagner Meira Jr., and Mohammad J. Zaki, "Is There a Best Quality Metric for Graph Clusters?," in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD*, Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, Eds. 2011, pp. 44–59, Springer-Verlag.

[35] M Girvan and M E J Newman, "Community structure in social and biological networks.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–6, June 2002.

[36] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi, "Benchmark graphs for testing community detection algorithms," *Physical Review E*, vol. 78, no. 4, pp. 1–5, Oct. 2008.

[37] Bruno Abrahao, Sucheta Soundarajan, John Hopcroft, and Robert Kleinberg, "On the separability of structural classes of communities," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*. 2012, p. 624, ACM Press.

[38] Usha Raghavan, Réka Albert, and Soundar Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical Review E*, vol. 76, no. 3, pp. 036106, Sept. 2007.

[39] Chun Yew Cheong, Huynh Phung Huynh, David Lo, Rick Siow, and Mong Goh, "Hierarchical Parallel Algorithm for Modularity-Based Community Detection Using GPUs," in *Proceedings of the 19th international conference on Parallel Processing*, 2013, pp. 775–787.

[40] Jyothish Soman and Ankur Narang, "Fast Community Detection Algorithm with GPUs and Multicore Architectures," *2011 IEEE International Parallel & Distributed Processing Symposium*, pp. 568–579, May 2011.

[41] Konstantin Kuzmin, S. Yousaf Shah, and Boleslaw K. Szymanski, "Parallel Overlapping Community Detection with SLPA," *2013 International Conference on Social Computing*, pp. 204–212, Sept. 2013.

[42] Jaewon Yang and Jure Leskovec, "Overlapping community detection at scale," in *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*, New York, New York, USA, 2013, p. 587, ACM Press.

[43] Arnau Prat-pérez and David Dominguez-sal, "High Quality , Scalable and Parallel Community Detection for Large Real Graphs Categories and Subject Descriptors," in *WWW*, 2014.

[44] Huawei Shen, Xueqi Cheng, Kai Cai, and Mao-Bin Hu, "Detect overlapping and hierarchical community structure in networks," *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 8, pp. 1706–1712, Apr. 2009.

[45] David F. Gleich and C Seshadhri, "Vertex neighborhoods, low conductance cuts, and good seeds for local community methods," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*. 2012, pp. 597–605, ACM Press.

[46] Qiong Chen and Ming Fang, "An Efficient Algorithm for Community Detection in Complex Networks," in *the 6th Workshop on Social Network Mining and Analysis*, 2012.

[47] Anirudh Ramachandran and Nick Feamster, "Understanding the Network-Level Behavior of Spammers," in *Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications - SIGCOMM '06*, New York, New York, USA, 2006, p. 291, ACM Press.

[48] Christian Kreibich, Chris Kanich, Kirill Levchenko, Brandon Enright, Geoffrey M. Voelker, Vern Paxson, and Stefan Savage, "On the Spam Campaign Trail," in *Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, June 2008, vol. 453, pp. 697–8.

[49] Zhenhai Duan, K. Gopalan, and X. Yuan, "Behavioral Characteristics of Spammers and Their Network Reachability Properties," in *2007 IEEE International Conference on Communications*. June 2007, pp. 164–171, IEEE.

[50] Martin Abadi, Mike Burrows, Mark Manasse, and Ted Wobber, "Moderately hard, memory-bound functions," *ACM Transactions on Internet Technology*, vol. 5, no. 2, pp. 299–327, May 2005.

[51] Michael Walfish, J D Zamfirescu, Hari Balakrishnan, David Karger, and Scott Shenker, "Distributed Quota Enforcement for Spam Control," in *Proceedings of the 3rd conference on Networked Systems Design & Implementation*. 2006, USENIX Association.

[52] Evan Harris, "The Next Step in the Spam Control War: Greylisting," *http://projects.puremagic.com/greylisting/whitepaper.html*, 2003.

[53] Anirudh Ramachandran, Nick Feamster, and Santosh Vempala, "Filtering Spam with Behavioral Blacklisting," in *Proceedings of the 14th ACM conference on Computer and communications security - CCS '07*, New York, New York, USA, 2007, p. 342, ACM Press.

[54] Guofei Gu, Roberto Perdisci, Junjie Zhang, and Wenke Lee, "BotMiner: Clustering Analysis of Network Traffic for Protocol- and Structure-Independent Botnet Detection," in *the 17th conference on Security symposium*. 2008, pp. 139–154, USENIX Association.

[55] Robert Beverly, "Exploiting Transport-Level Characteristics of Spam," *5th Conference on Email and Anti-Spam (CEAS)*, 2008.

[56] P.O. Boykin and V.P. Roychowdhury, "Leveraging social networks to fight spam," *Computer*, vol. 38, no. 4, pp. 61–68, Apr. 2005.

[57] Luiz H Gomes, Rodrigo B Almeida, and Luis M A Bettencourt, "Comparative Graph Theoretical Characterization of Networks of Spam and Legitimate Email," in *Conference on Email and Anti-Spam (CEAS)*, 2005.

[58] Ho-yu Lam and Dit-yan Yeung, "A Learning Approach to Spam Detection based on Social Networks," in *Conference on Email and Anti-Spam (CEAS)*, 2007.

[59] Chi-Yao Tseng and Ming-Syan Chen, "Incremental SVM Model for Spam Detection on Dynamic Email Social Networks," *2009 International Conference on Computational Science and Engineering*, pp. 128–135, 2009.

[60] Holger Ebel, Lutz-Ingo Mielsch, and Stefan Bornholdt, "Scale-free topology of e-mail networks," *Physical Review E*, vol. 66, no. 3, pp. 1–4, Sept. 2002.

[61] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos, "Graph Evolution: Densification and Shrinking Diameters," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, pp. 2–es, Mar. 2007.

[62] Gueorgi Kossinets and Duncan J Watts, "Empirical Analysis of an Evolving Social Network," *Science (New York, N.Y.)*, vol. 311, no. 5757, pp. 88–90, Jan. 2006.

[63] Varun Chandola, Arindam Banerjee, and Vipin Kumar, "Anomaly Detection: A Survey," *ACM Computing Surveys*, vol. 41, no. Sep, pp. 1–72, 2009.

[64] Qi Ding, Natallia Katenka, Paul Barford, Eric Kolaczyk, and Mark Crovella, "Intrusion as (anti)social communication," in *Proceedings of the 18th ACM SIGKDD conference on Knowledge discovery and data mining - KDD '12*, 2012, p. 886.

[65] Judit Bar-Ilan, Zheng Zhu, and Mark Levene, "Topic-specific analysis of search queries," in *Proceedings of the 2009 workshop on Web Search Click Data - WSCD '09*. 2009, pp. 35–42, ACM Press.

[66] Ricardo Baeza-Yates, "Graphs from Search Engine Queries," in *Theory and Practice of Computer Science*. 2007, vol. 4362, pp. 1–8, Springer.

[67] Luiz Henrique Gomes, Cristiano Cazita, Jussara M. Almeida, Virgílio Almeida, and Wagner Meira, "Workload models of spam and legitimate e-mails," *Performance Evaluation*, vol. 64, no. 7-8, pp. 690–714, Aug. 2007.

[68] Yinglian Xie, Fang Yu, Kannan Achan, Eliot Gillum, Moises Goldszmidt, Ted Wobber, C Computer Communication, and Networks Network, "How Dynamic are IP Addresses?," in *Proceedings of the 2007 conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM'07)*. 2007, pp. 301–312, ACM.

[69] Yehonatan Cohen, Daniel Gordon, and Danny Hendler, "Early Detection of Outgoing Spammers in Large-Scale Service Provider Networks," in *Detection of Intrusions and Malware, and Vulnerability Assessment*. 2013, pp. 83–101, Springer Berlin Heidelberg.

[70] Abhinav Pathak, Y Charlie Hu, and Z Morley Mao, "Peeking into Spammer Behavior from a Unique Vantage Point," in *Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*. 2008, pp. 3:1—-3:9, USENIX Association.

[71] Dominik Schatzmann, Martin Burkhart, and Thrasyvoulos Spyropoulos, "Inferring Spammers in the Network Core," in *Proceedings of the 10th International Conference on Passive and Active Network Measurement*. 2009, pp. 229–238, Springer-Verlag.

[72] Chris Kanich, Christian Kreibich, Kirill Levchenko, Brandon Enright, Geoffrey M Voelker, Vern Paxson, and Stefan Savage, "Spamalytics: An Empirical Analysis of Spam Marketing Conversion," in *Proceedings of the 15th ACM conference on Computer and communications security - CCS '08*, New York, New York, USA, 2008, p. 3, ACM Press.

[73] Christian Kreibich, C Kanich, and K Levchenko, "Spamcraft: An inside look at spam campaign orchestration," in *the 2nd USENIX conference on Large-scale exploits and emergent threats: botnets, spyware, worms, and more*. 2009, USENIX Association.

[74] Wolfgang John, Sven Tafvelin, and Tomas Olovsson, "Passive internet measurement: Overview and guidelines based on experiences," *Computer Communications*, vol. 33, no. 5, pp. 533–550, Mar. 2010.

[75] Farnaz Moradi, Magnus Almgren, Wolfgang John, Tomas Olovsson, and Philippas Tsigas, "On Collection of Large-Scale Multi-Purpose Datasets on Internet Backbone Links," in *Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, 2011.

[76] "SUNET (Swedish University Network), http://www.sunet.se/," .

[77] J. Klensin, "Simple Mail Transfer Protocol, Request for Comments, RFC 5321 (Draft Standard)," Oct. 2008.

[78] DShield, "Recommended block list," 2010.

[79] SRI International Malware Threat Center, "Most aggressive malware attack source and filters," 2010.

[80] Magnus Almgren and Wolfgang John, "Tracking Malicious Hosts on a 10Gbps Backbone Link," in *15th Nordic Conference in Secure IT Systems*, 2010.

[81] "Stanford Large Network Dataset Collection, http://snap.stanford.edu/data/index.html," .

[82] Ömer Yüksel, "Local Community Detection in Complex Networks," *Master thesis, Chalmers University of Technology*, 2013.

[83] Reid Andersen, Fan Chung, and Kevin Lang, "Local Graph Partitioning using PageRank Vectors," in *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*. 2006, pp. 475–486, IEEE.

Part II

PAPERS