The Identification of Like-minded Communities on Online Social Networks

THIS THESIS IS

PRESENTED TO THE

SCHOOL OF COMPUTER SCIENCE & SOFTWARE ENGINEERING

FOR THE DEGREE OF

MASTER OF SCIENCE (RESEARCH)

OF

THE UNIVERSITY OF WESTERN AUSTRALIA

By

Kwan Hui Lim

Jan 2013

Abstract

The efficient identification of communities with common interests is a key consideration in applying targeted advertising and viral marketing to online social networking sites. Existing approaches involve large-scale community detection on the entire social network before determining the interests of individuals within these communities. These approaches are both computationally intensive and may result in communities without a common interest. We propose two methods for detecting these like-minded communities using either topological or interaction links. Both methods are based on our selection algorithm for identifying users with common interests based on their following of celebrities that represent various interest categories. After identifying these users with common interests, we detect communities among them using either topological links (based on mutual friendship relationship) or interaction links (based on frequency and patterns of direct communication). Our evaluation on Twitter shows that both methods are able to detect communities comprising members that are well-connected and cohesive, and these communities become more connected and cohesive with the deepening or specialization of interest. Our proposed methods also result in communities that interact actively about their common interests (based on #hashtags and @mentions), with interaction-based method performing better than its topological-based counterpart.

Equally important in targeted advertising and viral marketing is the efficient detection of a community that is centered at an individual of interest (i.e. an influential individual). Most community detection algorithms are designed to detect all communities in the entire network graph. As such, it would be computationally intensive to first detect all communities followed by identifying communities where the individual of interest belongs to, especially for large-scale networks. We propose a community detection algorithm that directly detects the community centered at an individual of interest, without the need to first detect all communities. Our proposed algorithm utilizes an expanding ring search starting from the individual of interest as the seed user. Following which, we iteratively include users at increasing number of hops from the seed user, based on our definition of a community. This iterative step continues until no further users can be added, thus resulting in the detected community comprising the list of added users. We evaluate our algorithm on three real-life social networks and the YouTube online social network, and show that our algorithm is able to detect communities that strongly resemble the corresponding real-life communities.

Acknowledgements

First and foremost, I would like to thank my main supervisor, Professor Amitava Datta and co-supervisor, Winthrop Professor Mohammed Bennamoun for their supervision and guidance. The countless advice they have given have been most useful throughout the course of my MSc candidature. Apart from academic advice, they have provided great insights for developing my future research career. I also thank Professor Amitava Datta and Associate Professor Chris McDonald for the opportunity and enriching experience of teaching in the unit, Programming and Systems.

In particular, I am most grateful to Amitava for introducing me to research life in 2006 as an Honours candidate under his supervision. The great experience and his outstanding supervision prompted me to seriously consider further studies and eventually pursue a career in research. Subsequently, I had the privilege to work with Amitava again in 2011 as part of my MSc research degree. In addition, I also extend my deepest appreciation to his wife, Lakshmi for the many delicious meals and constant encouragement she provided.

My candidature has also been enriched with the great company of fellow students: Alvaro Monsalve, Amardeep Kaur, Deepak Garg, Maria Bravo Rojas, Matthew Heinsen Egan, Nasrin Moradmand, and Xiaohang Ma. In particular, I thank Alvaro for the many research related (and unrelated) discussion and also Amardeep and Maria for participating in our Data Mining reading group. In addition, my candidature have been greatly facilitated by the administrative and technical assistance provided by staff from the UWA Graduate Research School and the School of Computer Science and Software Engineering.

I also gratefully acknowledge the financial support of the Australian Government, University of Western Australia (UWA) and School of Computer Science and Software Engineering (CSSE) in the form of an International Postgraduate Research Scholarship, Australian Postgraduate Award, UWA CSSE Ad-hoc Top-up Scholarship and UWA Safety Net Top-Up Scholarship. In particular, I am also most grateful to Amitava for setting up and awarding the UWA CSSE Ad-hoc Top-up Scholarship.

During my candidature, I also had the opportunity to attend and present at four different conferences due to the financial support provided by the UWA Graduate Research School (GRS), UWA Postgraduate Students' Society (PSA), Institute of Electrical and Electronics Engineers (IEEE) and Association for Computing Machinery (ACM). This support was in the form of a UWA Graduate Research Travel Award, UWA PSA Conference Travel Award, ACM SIGWEB Student Travel Award and IEEE ComSoc Student Travel Grant.

This thesis has also benefited greatly from the valuable and detailed comments of my examiners, Lexing Xie from the Australian National University and Aixin Sun from the Nanyang Technological University. In addition, this thesis comprises a series of publications that has also benefited greatly from the useful comments of its anonymous reviewers. I sincerely thank both my thesis examiners and anonymous reviewers for their time and efforts in examining/reviewing my work.

Last but not least, the support and understanding of my family and relatives has been most crucial in my initial decision to commence on further studies and the eventual completion of this MSc degree. I am also most fortunate and thankful for an understanding and patient girlfriend who endures all my long hours and late-nights spent in the lab. For you, I completed my Master's degree nine months ahead of the normal duration.

Publications

This thesis contains published work and/or work prepared for publication. The bibliographical details of the work are outlined below.

Fully Refereed Publications

1. Kwan Hui Lim and Amitava Datta. A Seed-Centric Community Detection Algorithm based on an Expanding Ring Search. *Proceedings of the 1st Australasian Web Conference (AWC'13)*. Pages 21-26. Adelaide, Australia. Jan 2013.

2. Kwan Hui Lim and Amitava Datta. Tweets Beget Propinquity: Detecting Highly Interactive Communities on Twitter using Tweeting Links. *Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence (WI'12)*. Pages 214-221. Macau, China. Dec 2012.

3. Kwan Hui Lim and Amitava Datta. Following the Follower: Detecting Communities with Common Interests on Twitter. *Proceedings of the 23rd ACM Conference on Hypertext and Social Media (HT'12)*. Pages 317-318. Milwaukee, WI, USA. Jun 2012.

4. Kwan Hui Lim and Amitava Datta. Finding Twitter Communities with Common Interests using Following Links of Celebrities. *Proceedings of the 3rd International Workshop on Modeling Social Media (MSM'12), in-conjunction with HT'12*. Pages 25-32. Milwaukee, WI, USA. Jun 2012.

Non/Lightly Refereed Publications

5. Kwan Hui Lim. Detecting Like-minded Communities with Common Interests on Twitter. *Proceedings of the 18th UWA School of Computer Science and Software Engineering Research Conference (CSSE'12).* Pages 56-59. Pinjarra, Australia. Nov 2012.

My MSc candidature also resulted in work that was published during my candidature but do not contribute directly to this thesis. The bibliographical details of the work are outlined below.

Fully Refereed Publications

6. Kwan Hui Lim and Amitava Datta. Enhancing the TORA Protocol using Network Localization and Selective Node Participation. *Proceedings of the 23rd IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'12).* Pg 1503-1508. Sydney, Australia. Sep 2012.
7. Kwan Hui Lim and Amitava Datta. An In-depth Analysis of the Effects of IMEP on TORA Protocol. *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC'12).* Pg 3051-3056. Paris, France. Apr 2012

My contribution in all the above publications was at least 85%. I developed and implemented the proposed algorithms/approaches, performed the experiments and wrote the papers. My supervisor, Professor Amitava Datta reviewed the papers and provided useful feedback to improve the quality and readability of the papers. The publication of these papers was possible due to the administrative and financial support of UWA, CSSE, GRS, PSA, IEEE and ACM.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In recent years, Online Social Networks (OSN) such as Twitter and Facebook have gained immense popularity and rapid growth. The prevalence of OSNs is further supported by studies showing that “social networking sites now reach 82% of the world’s online population” and “nearly 1 in every 5 minutes spent online is now spent on social networking sites” [17]. Many popular OSNs are also characterized by a large user base and high user activity such as Twitter that comprises 500 million users who produce 2,200 tweets per second [58, 8]. This large user base and high user activity provide tremendous opportunities for companies to effectively reach out to a large audience group (of potential consumers) on similar OSNs. In turn, this audience group may further propagate information about the products/services provided by these companies.

One important problem in the application of targeted advertising and viral marketing to OSNs is the efficient identification of like-minded communities with common interests in large social networks [26, 29]. Such communities should ideally comprise users of the right demographics who are also well-connected among themselves. The identification of the right demographic group is important to ensure the right product-audience matching [26] and the connectedness of this group facilitates the subsequent word-of-mouth advertising [29].

Another important problem in advertising and marketing is the detection of a community surrounding a particular individual of interest. The interest in this individual is because he/she is determined to be influential in the spread of product/service information [29]. Examples of such individuals could be celebrities or politicians due to their ability to reach out to many people (i.e. high degree of links). In addition to applications in advertising and marketing, such communities are also of interest for counter-terrorism and epidemic modeling purposes. For these purposes, the individuals of interest could correspond to individuals who are at the heart of a terrorist organization (counter-terrorism) or a high-risk individual for an infectious disease (epidemic modeling) [59, 11].

1.1 Motivations and Proposed Solutions

For the purpose of detecting like-minded communities comprising users with common interests, most of the current approaches involve first detecting all communities, followed by determining the interests of these communities [27, 37]. These approaches involve a lengthy and intensive process of detecting communities for the entire social network, which is both large and growing daily. In addition to the lengthy and intensive community detection process, many of the detected communities may not share the interest we are looking for.

Our proposed methods involve first identifying celebrities that are representative of an interest category, before detecting communities based on linkages and communication patterns among followers of these celebrities. Our methods directly identify communities comprising like-minded individuals with common interests, and differ from existing ones that first detect all communities, followed by identifying the topics they are interested in [27, 37]. Also, our methods do not unnecessarily detect communities that do not share any specific interest. Instead, our methods allow for the efficient detection of only communities sharing a common interest and can be applied to targeted advertising and viral marketing. In addition, our methods are able to detect communities at different levels of interest.

Similarly for the purpose of detecting communities centered at individuals of interest, most community detection algorithms aim to detect all communities in the entire network graph [20]. This process is both tedious and computationally intensive due to the large-scale of current social networks, ranging from scientific collaboration networks to online social networking sites. As such, it would be more efficient to focus directly on a community that is centered at this influential individual, compared to first detecting all communities followed by identifying the communities that this individual belongs to.

Hence, we propose a community detection algorithm that directly detects a community centered at an individual of interest. Our proposed algorithm starts from a seed user (i.e. the individual of interest) and performs an expanding ring search to iteratively include users into that community. Users are included into the community based on a metric of their number of links to other users in the community. This iterative adding of users continues until no further users satisfying the metric could be added. While there exists seed-based community detection algorithms, they either require a set of seed users [3, 2] or could potentially exclude the seed user in the detected community [42]. In addition, our algorithm includes a strength factor that allows us to adjust the strength and size of the detected communities.

1.2 Contributions

Our main contributions in this thesis are as follows:

- A selection algorithm for identifying users with common interest based on their following of celebrities (Chapter 3).
- A topology-based approach for detecting social network communities that share common interest, termed as the Common Interest Community Detection (CICD) method (Chapter 3).
- An interaction-based approach for detecting highly interactive communities that frequently communicate about their common interests, termed as the Highly Interactive Community Detection (HICD) method (Chapter 3).

- A seed-centric community detection algorithm for finding the community centered at an individual of interest, using an expanding ring search (Chapter 4).
- An evaluation of our CICD method on the Twitter social network, including a study of the characteristics of Twitter communities that share common interests, and an investigation into the effects of deepening or specialization of interest on these communities (Chapter 5).
- An evaluation of our HICD method on the Twitter social network, including a study into the communication behaviour and patterns of these communities, and a preliminary study into the evolution of links among these communities over time. (Chapter 6).
- An evaluation of our seed-centric community detection algorithm on three real-life social networks and a large-scale YouTube[1] social network (Chapter 7).

1.3 Organization of the Thesis

This thesis is organized and structured in the following chapters. In Chapter 2, we first present a literature review which examines the overall development in the general area of community detection algorithms. Following which, we discuss some related work in the specific areas of community detection of like-minded communities and seed-based community detection. In particular, we highlight the differences between these related work and our proposed algorithms. Also, we provide some background information on the Twitter and YouTube social networks, which are used for experimentation in our thesis.

In Chapter 3, we present the CICD and HICD methods for detecting like-minded communities which comprise users sharing common interests. The CICD and HICD methods respectively utilize topological and interaction information for detecting such communities. Both methods first build upon our proposed selection algorithm for identifying a set

[1] http://www.youtube.com/

of users with common interest, based on their following of celebrities that are representative of interest categories.

In Chapter 4, we present another community detection algorithm, with the main purpose of (directly) finding the community centered at an individual of interest. Our proposed algorithm utilizes an expanding ring search starting from the individual of interest as the seed node and iteratively includes users at increasing number of hops. We also introduce our definition of a community with a strength factor that allows us to adjust the size and strength of the detected communities.

In Chapter 5, we evaluate our CICD method on the Twitter social network using topological measures such as community size, average clustering coefficient, average path length, average degree and diameter. Specifically, we study the topological structure of communities with common interest and investigate how these topological structures change as users (in these communities) develop deeper or specialized interests.

In Chapter 6, we next evaluate our HICD method, also using topological measures on the Twitter social network. In addition, we further evaluate this method based on the communication behaviour and patterns among users in the detected communities, specifically the frequency and content of messages (tweets) sent. We also perform a preliminary analysis on the temporal evolution of topological links among these users and study trends in the creation and deletion of topological links.

In Chapter 7, we evaluate our seed-centric community detection algorithm on three real-life social networks and the large-scale YouTube social network. The three real-life social networks allow us to evaluate the effectiveness of our algorithm in detecting communities that strongly correspond to their real-life counter parts. We also perform a similar evaluation on the YouTube social network using an approximation of real-life communities based on the YouTube groups that an user belongs to.

In Chapter 8, we conclude this thesis by summarizing our contributions and discussing some limitations of our work. In addition, we also propose some future research directions in the areas of community detection and social network analysis.

Chapter 2

Literature Review and Background Information[1]

In this chapter, we present some background information on the Twitter and YouTube social networks, and a literature review of previous work that are closely related to our research. Specifically, we give a broad overview of the development of community detection algorithms with an emphasis on the algorithms which are related to our proposed algorithms. In particular, we discuss about various seed-based community detection algorithms and the differences between these algorithms and our proposed seed-centric community detection algorithm. Also, we study some related work regarding the detection of like-minded communities and the study of interaction among users. Similarly, we discuss how our proposed CICD and HICD methods are different from these related work.

[1] This chapter is based on the following publications by Kwan Hui Lim and Amitava Datta: "A Seed-Centric Community Detection Algorithm based on an Expanding Ring Search", *in Proceedings of the 1st Australasian Web Conference (AWC'13)* [41]; "Tweets Beget Propinquity: Detecting Highly Interactive Communities on Twitter using Tweeting Links", *in Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence (WI'12)* [40]; "Following the Follower: Detecting Communities with Common Interests on Twitter", *in Proceedings of the 23rd ACM Conference on Hypertext and Social Media (HT'12)* [39]; and "Finding Twitter Communities with Common Interests using Following Links of Celebrities", *in Proceedings of the 3rd International Workshop on Modeling Social Media (MSM'12), in-conjunction with HT'12* [38].

2.1 Background Information

As our research is concerned with the detection of like-minded communities on OSNs, we begin by describing two popular OSNs, namely the Twitter and YouTube social networks. These OSNs are also used in our subsequent experiments to evaluate the effectiveness of our proposed algorithms.

2.1.1 Twitter Online Social Network

Twitter is a popular micro-blogging service that allows messages of up to 140 characters to be posted and received by registered users. These messages are called tweets and they can be posted via the Twitter website, short messaging services or third party applications. Tweets form the basis of social interactions in Twitter where a user is kept updated of the tweets of someone he/she is following. A user can also forward the tweets of others to all users following him/her, which is called retweeting. In addition, users can @mention each other in their tweets (via @username) or #hashtag keywords or topics for easy search by others (via #topic). The popularity of Twitter is seen from its large user base of 500 million users who produce 2,200 tweets per second [58, 8]. The popularity of Twitter and availability of data have created plenty of interest in its academic study in recent years [13, 33, 52].

Twitter also provides an Application Programming Interface (API) with the functionality to collect data such as user profiles, linkages among users, tweets, retweets and @mentions [25]. This API allows developers to create applications for Twitter and researchers to study the characteristics of an online social network from individual to community levels. Currently, there is an hourly rate limit on the number of API calls that can be executed.

2.1.2 YouTube Online Social Network

YouTube is an online video sharing service that allows short video clips to be uploaded and shared with the general public. In addition, a YouTube user is able to upload a video clip as a private video and allocate permissions to a specific list of users who can view these videos. The popularity of YouTube can be seen from its high volume of video upload (60 minutes of video uploaded per second) and high number of video views (four billion videos viewed per day) [9].

Apart from its video sharing functionality, YouTube also provides the services of a typical OSN. For example, users are able to comment on each other's videos, privately message other users and even add other users as friends. Of particular interest is the existence of YouTube groups that users are able to join.[2] These YouTube groups comprise a set of users who share similar interest in producing or uploading videos related to that YouTube group. Similar to Twitter, YouTube also provides an API that allows developers to upload and download videos, retrieve video comments and collect data such as user profiles.

2.2 Detecting Like-minded Communities

More closely related to our proposed CICD method are studies that focused on detecting communities in OSNs, particularly like-minded communities where its members share common interests. One such study resulted in the LikeMiner system which identifies popular topics on OSNs based on the explicit "likes" indicated by users [28]. In turn, these topics can be based on textual or graphical information that are determined from comments/messages and pictures/videos respectively. LikeMiner is then able to predict the interests of a user based on the interests of his/her friends.

Similarly, the Friendship and Interest Propagation (FIP) model identifies interests of an

[2] There has since been changes to some YouTube functionalities such as the availability of YouTube groups. As the dataset used for our experiments are based on these old functionalities, we shall describe the YouTube OSN based on these functionalities.

individual and potential friendship links with other users [62]. The FIP model determines the interests of an individual user based on the interests of his/her friends and recommends friends based on those sharing similar interests. This model builds upon the concept of homophily which states that users with similar interests are more likely to be mutual friends compared to users with dissimilar interests. Specifically, the FIP model presents a unified framework to simultaneously identify interests and predict potential friendship links.

In their study of Twitter, Java et al. used the Hyperlink-Induced Topic Search algorithm to detect communities based on a set of hubs and authority, and the CPM algorithm to detect overlapping communities on the Twitter OSN [27]. After detecting all communities, they studied the key terms used by the users (in their tweets) among these communities. Through this tweet analysis, they found that such communities share common interest, which are further divided into formal and informal ones. In addition, Java et al. also noticed that the probability of two persons being connected is negatively correlated with their geographic distance.

Li et al. proposed the TTR-LDA community detection algorithm using the Latent Dirichlet Allocation model and GN algorithm with an inference mechanism for topic distribution [37]. They used the TTR-LDA algorithm to first detect all communities among the top 50,000 taggers in Delicious[3], followed by determining the interest topics of these communities. Next, they modeled the temporal evolution of these interest topics among the detected communities. In particular, they observed that communities share common interests which divide into defined sub-categories over time.

Using BibSonomy[4], Atzmueller and Mitzlaff demonstrated an approach for mining communities with common descriptive features [5]. This approach integrates a database (of user attributes) and topological graph (of user links) into a dataset comprising only links connecting two users with the same attribute. Communities are then detected based on the desired attribute using this new collection of links. This approach could potentially be

[3] http://delicious.com/

[4] http://www.bibsonomy.org/

used to detect like-minded communities with common interests by modeling the database of user attributes as potential interests based on explicit tags on BibSonomy.

2.2.1 Discussion

Our proposed CICD method is designed to detect like-minded communities that comprise users with common interest. One main difference between our CICD method with the LikeMiner system [28] and FIP model [62] is that both the LikeMiner system and FIP model detect individuals (instead of communities) with common interests. On the other hand, our CICD method detects communities (instead of only individuals) where all of its members share a common interest. In addition, the LikeMiner system and FIP model are based on explicitly declared interests ("likes" on Facebook and application items on Yahoo! Pulse[5] for the LikeMiner system and FIP model respectively), whereas our CICD method implicitly infer interests based on a user's followings.

As for the studies by Java et al. [27] and Li et al. [37], they first detect all communities in their network datasets (Twitter and Delicious respectively), followed by determining the interests of these communities. The main difference with our CICD method is that we do not detect all communities then determine their interests but rather, focus directly only on communities sharing specific interests that we are interested in. Also, Li et al. based their experiments on only the top users of Delicious whereas ours is based on the full dataset of the Twitter social network.

While the approach by Atzmueller and Mitzlaff [5] can be applied to detect like-minded communities with common interest, our method is able to detect communities with varying levels of interest. We determine the interest level of users in these communities based on the number of celebrities (of a representative interest category) that these users follow. Furthermore, our method implicitly infer a user's interests based on his/her followings while the approach by Atzmueller and Mitzlaff needs to build user attributes using explicit tags on BibSonomy.

[5] http://pulse.yahoo.com

2.3 Studying Interaction Among Communities

As our proposed HICD method uses communication links to detect highly interactive communities, we now examine related work that study communication patterns on OSNs. Various models have been proposed for studying and predicting general information diffusion on Twitter based on a combination of message content, user profiles and tweeting timings [21, 45, 63]. Romero et al. [54] and Huang et al. [23] studied the diffusion of #hashtags on Twitter and investigated the factors behind the mass adoption of #hashtags and their subsequent dying off.

In addition, tweets have been analyzed to determine their credibility, sentiments and relation to real-life events. Using the tweeting patterns of a user, tweet content and external references, Castillo et al. [12] proposed a method to determine the credibility of tweets. Similarly, Becker et al. [7] presented a real-time system to detect tweets that describe real-life events. Also, Kouloumpis et al. [31] studied the sentiment of tweets based on the usage of #hashtags, emoticons, caps and punctuations. While these studies analyze tweeting patterns and contents, they do not use tweeting links to detect communities with common interests.

Many authors have also used the interaction frequency among users of OSNs to study information diffusion and the topological characteristics of entire OSNs. Various authors constructed interaction graphs to study the general structure and behaviour of users on OSNs such as Cyworld and Facebook [15, 60, 49]. Similarly, the interaction activity between users has also been used to construct networks for the purpose of studying information diffusion on Twitter and Flickr [54, 14, 61].

2.3.1 Discussion

Our proposed HICD method aims to detect highly interactive communities comprising users that frequently communicate about their common interests. This method uses communication data such as the interaction frequency and patterns among users to detect these highly interactive communities. The main difference of our method (from these related work) is that we use interaction frequency to detect highly-interactive communities with common interests, while these authors use it only for studying information diffusion on the overall structure of OSNs. Furthermore, our proposed method imposes a set of criteria for selecting users (with common interest) before constructing a network based on their direct communication with each other.

Community detection is also a common research problem on other real-life social networks, such as scientific collaboration networks [6, 18]. However, these methods consider only topological links to detect community structures, which does not translate to interactive communities [15, 60]. Our proposed study differs from these earlier work as we examine the existence of a highly interactive community with common interests, based on direct communication among the users (instead of only topological links). In addition, we study their communication patterns by examining content such as keywords, #hashtags, URLs and @mentions, and how users follow or unfollow each other (in Chapter 6), instead of using only certain aspects of communication (e.g. only #hashtags).

2.4 Community Detection Algorithms

In the last decade, one important and intensively studied problem in social networks is the effective detection of community structures from an underlying network graph [20]. As such, there exists an extensive literature on community detection algorithms and we give a broad history of the development of these algorithms, coupled with elaborations on a selection of popular algorithms. In particular, we describe a series of seed-based community detection algorithms and highlight the key differences between these algorithms and our proposed seed-centric community detection algorithm.

2.4.1 Hierarchical Clustering Algorithms

While there are many categories and types of community detection algorithms, most of the early methods fall under the hierarchical clustering category. In turn, the hierarchical clustering category is further divided into the agglomerative and divisive sub-categories. Agglomerative methods consider individual nodes and group them together into larger communities if they satisfy a similarity measure. On the contrary, divisive methods consider the entire network graph and iteratively remove edges that are below the similarity measure. This iterative removal of edges continues until distinct communities remain.

One of the first and most popular hierarchical clustering algorithms is the Girvan-Newman (GN) algorithm that utilizes a divisive approach to community detection [22]. The GN algorithm uses a similarity measure of edge-betweenness, defined as the number of shortest paths that passes between two adjacent nodes (i.e. an edge). As a high edge-betweenness value indicates that the two nodes sharing that edge belong to different communities, the edges with the highest edge-betweenness are iteratively removed to split the entire network into distinct communities. However, the calculation of edge-betweenness is computationally expensive as it is a global parameter that requires topological information of the entire network. For a sparse network with N nodes, the GN algorithm has a complexity of $O(N^3)$ which makes it unfeasible for large networks due to the potentially long computation time (from calculating the edge-betweenness for all edges in the entire network).

The Racicchi et al. algorithm builds upon the GN algorithm and uses an edge-clustering coefficient (instead of edge-betweenness) to detect communities [53]. The edge-clustering coefficient is defined as the number of actual triangles (which include this edge) out of the total possible number of such triangles. An edge (pair of nodes) connecting two different communities is likely to have a low edge-clustering coefficient and would be recursively removed until separate communities are left. On the other hand, two nodes within the same community share an edge with a high edge-clustering coefficient as they would be connected to many other nodes within the community. Since the edge-clustering coefficient is a local parameter, this algorithm has a complexity of $O(N^2)$, improving upon that of the GN algorithm.

One key characteristic of hierarchical clustering algorithms is their inability to classify a node into multiple communities (i.e. detect overlapping communities). This functionality is important as overlapping communities are a common characteristic of social networks and a user could be a member of many communities [51, 35, 20].

2.4.2 Overlapping and Compression-based Community Detection

Algorithms such as the Clique Percolation Method (CPM) were then developed to overcome this problem of detecting overlapping communities [51]. The CPM algorithm defines a community as one with a series of adjacent k-cliques, where a k-clique comprises k nodes that are interconnected. CPM first identify all k-cliques in the network and connect them if they are adjacent. Two k-cliques are adjacent if they share $(k-1)$ common nodes. This procedure of connecting k-cliques continues iteratively until no adjacent k-cliques can be found. The result is a set of communities formed by the series of connected k-cliques based on the adjacency criterion.

Rosvall and Bergstrom developed the Infomap algorithm which detects communities using random walkers based on the principle that a random walker is likely to spend more time traversing nodes within a community [55]. Infomap approaches community detection as a coding or compression problem where the network graph can be compressed to retain its key structures. These key structures represent communities or clusters that are found within the network graph. Infomap uses random walkers on the network graph to analyze information flow where the random walker is more likely to traverse within a cluster of nodes belonging to the same community. In addition, Lancichinetti and Fortunato conducted a comprehensive evaluation of numerous community detection algorithms and found that Infomap performs the best in terms of execution speed and correctness [34].

2.4.3 Seed-based Community Detection

More closely related to our proposed algorithm (for detecting seed-centric communities centered at an individual of interest) are community detection algorithms that are based on a set of seed nodes. Andersen and Lang proposed an algorithm based on a series of random walkers, each traversing a limited number of steps starting from a set of seed nodes [3]. This algorithm then uses network flow to clean up the results before returning the detected community based on a selection of nodes that the random walkers have traversed through. Also based on a set of seed nodes, Andersen et al. proposed a local community detection algorithm using a modified version of the PageRank algorithm [2]. A series of random walkers start from this set of seed nodes and each node they traverse is considered for inclusion into the community based on the value of their resulting PageRank vector.

Similarly, there are various algorithms for detecting communities using a single seed node. Clauset introduced the local modularity R which measures how much a node is on the boundary of the community [16]. This local modularity R is calculated based on the number of edges within the community divided by the number of edges linking the community to an outside node. Clauset then starts from a seed node and iteratively adds neighbouring nodes into the community if these nodes maximize the modularity R. This iterative adding of nodes continues until all nodes have been considered, thus resulting in the detection of a local community.

In the same spirit as Clauset (i.e. the local maximization of modularity), Luo et al. proposed an algorithm that starts from a seed node and uses iterative steps of adding and deleting nodes based on the local maximization of modularity [42]. At each addition step, neighbouring nodes (from the seed node) are added if they maximize the local modularity. Similarly, each deletion step iteratively removes nodes from the list of added nodes (to maximize the local modularity), provided that the removal of this node does not separate the detected community into disconnected components. Of note, this deletion step could also potentially remove the seed node. These addition and deletion steps continue iteratively until no further nodes can be added to the community (due to the local maximization of modularity).

2.4.4 Discussion

Our proposed seed-centric community detection algorithm aims to find communities centered at an individual of interest. The main difference between our algorithm and the hierarchical clustering algorithms (GN algorithm [22] and Racicchi et al. algorithm [53]) is that our algorithm directly detects a community centered at this individual and is able to detect overlapping communities based on a selection of seed users. On the other hand, the hierarchical clustering algorithms are designed to detect all communities in a network (a computationally intensive task) and is unable to classify a node into multiple communities (i.e. detect overlapping communities).

Similarly, the overlapping and compression-based community detection algorithms (CPM [51] and Infomap [55]) are also designed to detect all communities in a network. While the ability to detect overlapping communities is useful for applications in social networks, the need to detect all communities is a computationally expensive task. The main difference of our proposed algorithm with these earlier ones is that we focus directly on the community centered at this individual, instead of having to first detect all communities. Since we only detect the community centered at this individual, we need not perform the lengthy and computationally intensive task of first detecting all communities followed by finding the communities where this individual belongs to.

Our proposed algorithm also differs from the seed-based community detection algorithms by Andersen and Lang [3], and Andersen et al. [2] in that we detect communities surrounding a single seed node whereas they require a set of seed nodes. Also, our method differs in the definition of the metric that is used to determine whether a node should be included in a community. Specifically, our metric is a local parameter that is based on the number of internal and external links of a user (to his/her community), coupled with an adjustable strength factor. As such, we are able to tailor our algorithm to detect communities of different strength and size.

While the algorithms by Clauset [16] and Luo et al. [42] are able to detect communities using a single seed user, the difference between these algorithms and our proposed algorithm is in the definition of the metric for including nodes into the community. Specifically, our proposed metric includes an adjustable strength factor that allows us to detect communities of different strength and size (whereas these earlier algorithms are unable to do so). In addition, the algorithm by Luo et al. could potentially exclude the seed node that may result in a detected community without the seed node. This potential exclusion of the seed node is another key difference with our algorithm, which ensures that the seed node is still included in the detected community.

Chapter 3

Approaches for Detecting Like-minded Communities[1]

In this chapter, we propose two approaches for detecting like-minded communities comprising users with common interests. Both approaches are based on a selection algorithm for identifying users with common interests using their following of celebrities. After identifying users with common interest, we then detect communities among these users using either topological or interaction links. The consideration of either topological or interaction links is the main difference between our two proposed approaches. We shall first introduce the notations and definitions used, then describe our proposed selection algorithm for identifying users with common interest, followed by our two approaches for detecting like-minded communities.

[1] This chapter is based on the following publications by Kwan Hui Lim and Amitava Datta: "Tweets Beget Propinquity: Detecting Highly Interactive Communities on Twitter using Tweeting Links", *in Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence (WI'12)* [40]; "Following the Follower: Detecting Communities with Common Interests on Twitter", *in Proceedings of the 23rd ACM Conference on Hypertext and Social Media (HT'12)* [39]; and "Finding Twitter Communities with Common Interests using Following Links of Celebrities", *in Proceedings of the 3rd International Workshop on Modeling Social Media (MSM'12), in-conjunction with HT'12* [38].

3.1 Notations and Definitions

We model the Twitter social network as a directed graph, $G = (U, L)$ where U refers to the set of users/nodes and L refers to the set of links/edges.[2] A followership link $(i, j) \in L$ indicates that user $i \in U$ is a follower of user $j \in U$, while a friendship link $Fr_{i,j}$ indicates $(i, j) = (j, i)$. We classify a Twitter user as a celebrity if he/she has more than 10,000 followers. As required, the definition of a celebrity can also be made more or less stringent by respectively increasing or decreasing the required number of followers.

The interest of a user in a category cat, Int_{cat} is inferred by the number of celebrities (of category cat) that the user follows. Although Int_{cat} represents the interest level of a user in a category, this metric is subjective due to the celebrities selected. The accuracy of Int_{cat} is dependent on the correct classification of celebrities into their respective categories, which is subjective as some celebrities loosely belong to multiple categories (e.g. a singer that has starred in some movies). We minimize this subjective judgment by using information on Wikipedia[3] to classify these celebrities into their respective categories. On the Wikipedia page of a celebrity, there is an "occupation" field which we use to determine the categories this celebrity belong to. Thus, this process minimizes the chances of classifying celebrities into the wrong category. While this classification is currently performed manually, we are working on automating this process using a library of keywords-to-category mapping.

Twitter users directly communicate with other users by posting a tweet containing @username of the other user, along with the actual message. This direct communication is also called the @mentioning of other users. We define $M_{i,j}$ as a tweet posted by user i that contains a @mention of user j (i.e. the @mentioning process). Next, we also model the communication intensity $I_{i,j}$ of user i to j as the number of @mentions user i makes of user j. Table 1 lists a summary of the notations and definitions used in our CICD and HICD methods.

[2] The terms nodes and users are used interchangeably but they refer to the same entity. Similarly, the terms edges and links are used interchangeably.

[3] http://en.wikipedia.org/

Table 1: Notations and Definitions - CICD and HICD Methods

Notations	**Definitions**
(i, j)	A followership (uni-directional) link from user i to user j
$Fr_{i,j}$	A friendship (bi-directional) link between users i and j
Celebrity	A user with more than 10,000 followers
Int_{cat}	The interest level of a user in category cat
$M_{i,j}$	A tweet containing a @mention of user j by user i
$I_{i,j}$	The number of times user i @mentions user j

While we describe our notations and definitions on the basis of Twitter, these notations and definitions can also be applied to other OSNs. Followership and friendship links respectively correspond to uni-directional and bi-directional links (or their appropriate representation) on other OSNs. As such, the definitions of a celebrity and Int_{cat} remains unchanged (as earlier described). Similarly on other OSNs, $M_{i,j}$ represents the private messages (or wall messages) that user i sends (or posts) to user j, and $I_{i,j}$ is the frequency of this messaging process.

3.2 Overall Framework

The overall framework for detecting like-minded communities involves first identifying a set of users with common interest, followed by detecting communities among this set of users based on either topological or interaction links. The type of links used for detecting like-minded communities divides our framework into two different methods, namely the CICD method using topological links and the HICD method using interaction links. Both the CICD and HICD methods are based on the set of users with common interest which are identified based on their followings of representative celebrities. We shall first begin by describing our proposed selection approach for identifying users with common interest.

3.2.1 Identifying Users with Common Interest

The first step in selecting users with common interest is to identify a set of n celebrities $c_1, c_2, ..., c_n$, that belongs to a common interest category.[4] As mentioned in Section 3.1, the classification of celebrities into their respective interest categories could be subjective. Therefore, we utilize information on Wikipedia (the "occupation" field) to ensure the accurate classification of these celebrities.

Our next step is to retrieve the set of Twitter users who follow all celebrities in a given category. Given a pre-identified set of n celebrities $c_1, c_2, ..., c_n$, we next identify all the followership links for each individual celebrity in this set. Consider celebrity $c_j, 1 \leqslant j \leqslant n$, and all the followership links for this celebrity $\bigcup_i link(i, c_j)$. We construct the set:

$$\mathcal{P} = \bigcap (\bigcup_i link(i, c_j)), for\ 1 \leqslant j \leqslant n$$

$\mathcal{P}$ is the set of fans who follow all the n celebrities in the set $\bigcup c_j, for\ 1 \leqslant j \leqslant n$. Fig. 1 shows an example of the representation of Set $\mathcal{P}$, which (in this case) are fans who follow all three sports celebrities. Thus, Set $\mathcal{P}$ is also the intersection set among the fans of each celebrity, as listed in Fig. 1.

The criteria for constructing Set $\mathcal{P}$ can also be relaxed such that we select users who follow x out of n celebrities, where $0 < x < n$. In doing so, the value of x would determine the interest level Int_{cat} of the resulting Set $\mathcal{P}$. As the value of Int_{cat} is inversely correlated to the size of Set $\mathcal{P}$, a user has the flexibility to construct a larger Set $\mathcal{P}$ with the trade-off in a lower level of interest.

After constructing Set $\mathcal{P}$ (comprising users with common interest), we next determine the common links among these users, where such links can be either topological or interaction links. Topological links correspond to followership or friendship links while interaction links correspond to communication (@mentioning) links. The choice of link

[4] Instead of using a random selection of n celebrities (to overcome any sampling bias), we will choose celebrities based on popularity (i.e. number of followers) to maximize the size of the subsequently detected communities.

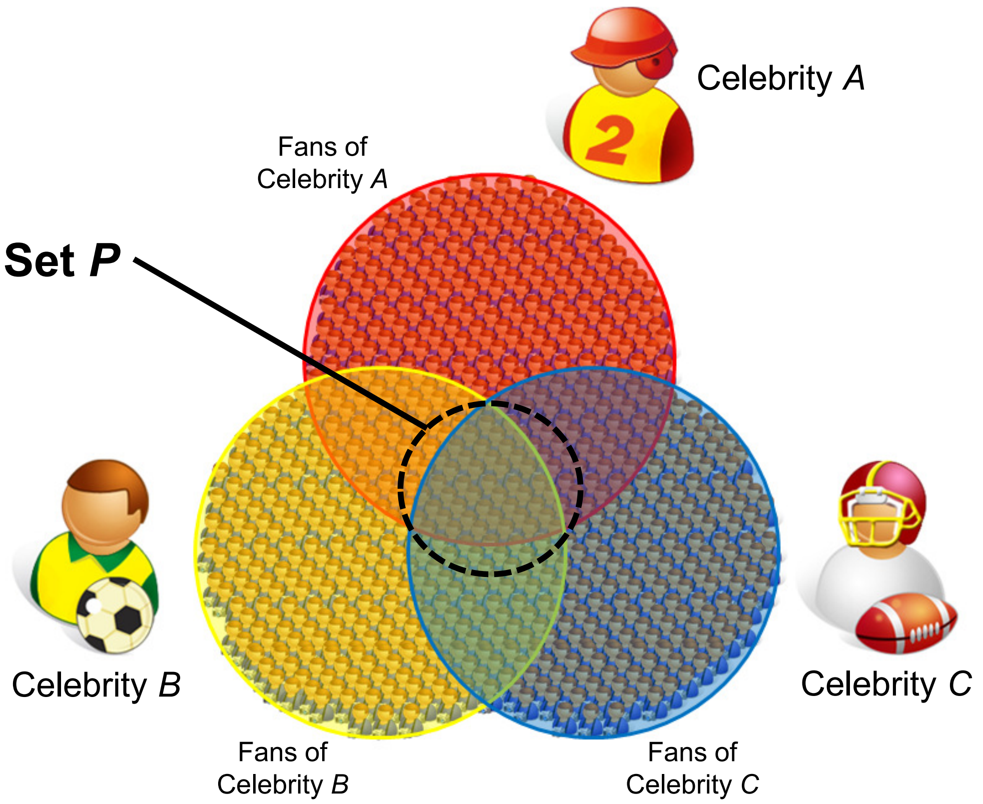


Figure 1: Illustration of Set P

type results in two different approaches to detecting like-minded communities, namely the topological-based CICD method and the interaction-based HICD method.

3.2.2 Topological Approach: Common Interest Community Detection

Our first proposed approach, the CICD method aims to detect like-minded communities comprising users with common interest, using only topological links. Topological links can be either followership or friendship links as introduced in Table 1. In this approach, we consider only friendship links (among Set $\mathcal{P}$) for community detection as friendship links are stronger and more reflective of real-life interactions. Using this set of friendship links (which corresponds to an undirected graph), we try to detect communities among the members of Set $\mathcal{P}$ next using the CPM algorithm developed by Palla et al. [51]. For our experiments, we use CPM with a k-value of 3 as this produces the best results in detecting communities compared to other k-values.

Similarly, we also detect communities among the members of $\mathcal{P}$ next using the Infomap algorithm by Rosvall and Bergstrom [55]. Using both CPM and Infomap show that our proposed method produces results that are independent of the chosen community detection algorithm and their unique characteristics. CPM was chosen due to its ability to detect overlapping communities (which reflects real-life social communities) while Infomap was selected due to its superior performance compared to other algorithms [20]. The communities detected by the CICD method shall be known as the link-based communities, denoted Com_{CICD}.

3.2.3 Interaction Approach: Highly Interactive Community Detection

Our previous approach, the CICD method considers only topological information (such as follower/following links) but not user activity (such as communication/tweeting patterns and frequency). In a community where its users share common interests and are well-connected, the tweeting frequency and content of tweets are other factors that determine the speed of information diffusion. Many studies also support this observation, noting that only a small subset of users (among those connected by topological links) frequently interact with each other [15, 60]. Thus, it is necessary to consider user activity (communication/tweeting patterns and frequency) in addition to topological information for community detection, especially for advertising and marketing purposes. We now present the HICD method for identifying communities where its members not only share common interests but actively and frequently communicate about the common interests.

Our proposed HICD method detects a highly interactive community using the communication information (tweeting pattern and frequency) among the users. This approach involves identifying community members based on their frequency of direct communication with other users in the community. As previously defined in Table 1, $I_{i,j}$ is the number of times user i @mentions user j in his/her tweets. Using this definition of $I_{i,j}$, we next build a list of weighted edges between two users i and j as a tuple $(i, j, I_{i,j})$ where $i, j \in \mathcal{P}$, and user j could be either an ordinary user or celebrity. Using a predetermined intensity threshold T, we remove all tuples $(i, j, I_{i,j})$ if $I_{i,j} < T$ or $I_{j,i} < T$. In short, we are building a new set of users $\mathcal{Q}$ comprising only edges that exceed the threshold T.

Similar to the CICD method, we now detect communities among this set $\mathcal{Q}$ of users using CPM and Infomap where the detected communities shall be referred to as the tweet-based community, Com_{HICD}. These stringent requirements for constructing Set $\mathcal{Q}$ ensures that the resulting Com_{HICD} is well-connected, cohesive and communicate frequently about their common interest.

3.3 Summary

In this chapter, we first introduced a selection algorithm for identifying users with common interest. The interests of users are inferred based on the celebrities they follow and we first identify a set of celebrities that represents a certain interest category. Using this set of celebrities, we retrieve the followers of these celebrities and select users who follow all celebrities in this set. This set of selected users are deemed to have a common interest based on their following of these representative celebrities. In addition, we also proposed the CICD and HICD methods for detecting like-minded communities (using the common links among this set of users with common interest). Both methods aim to directly detect like-minded communities comprising users with common interest, without the need to first detect all communities followed by selecting the communities with common interests.

Our two proposed methods differ mainly in the usage of links for community detection. The CICD method detects communities using only topological information such as explicit bi-directional links. These bi-directional links are reflected in Twitter as a pair of users with mutual follower/following links (i.e. friendship links), which are more representative of real-life social relationships. On the other hand, the HICD method uses implicit link information that is derived from communication links. These communication links are based on users @mentioning each other and result in communities that are more interactive, especially about the common interest. Due to this different usage of links, the communities detected by the CICD and HICD methods may overlap but are unlikely to be a subset of one another (as users may @mention each other even when they are not topologically connected).

There are distinct advantages and disadvantages to both methods. The CICD method is able to detect like-minded communities using a single snapshot of the topological structure of the OSN (i.e. the topological links among users). However, using only topological links for detecting communities may not necessarily correspond to communities that are highly interactive. On the other hand, the HICD method is able to detect such highly interactive communities using communication (@mentioning) links among these users. However, such communication links cannot be retrieved in a single snapshot (unlike topological links) and instead have to be periodically retrieved at specific time intervals. Thus, the trade-off between the CICD and HICD methods are with the ease of links retrieval and the interactivity of the detected communities.

However, both the CICD and HICD methods share the same ability to effectively detect like-minded communities, which are important for applications in targeted advertising and viral marketing. Equally important to advertising and marketing efforts are key individuals whose influence and connections could facilitate the rapid diffusion of product/service information. More precisely, the communities centered at such individuals would be crucial for advertising and marketing purposes. We shall present an algorithm for detecting such communities (centered at these influential individuals) in the next chapter.

Chapter 4

An Approach for Detecting Seed-centric Communities using an Expanding Ring Search[1]

In the previous chapter, we introduced the CICD and HICD methods for detecting like-minded communities, with the purpose of applications in advertising and marketing. Similarly, we are also interested in the communities centered at certain individuals of interest, which are important not only for advertising and marketing but also counter-terrorism and epidemic modeling. Such individuals are deemed to be influential in the spread of product information, at the heart of a terrorist organization, or a high-risk individual for an infectious disease, thus our interest in the communities centered at these individuals. As such, we now propose a seed-centric community detection algorithm for finding the community centered at an individual of interest, using an expanding ring search and our definition of a community.

[1] This chapter is based on the following publication by Kwan Hui Lim and Amitava Datta: "A Seed-Centric Community Detection Algorithm based on an Expanding Ring Search", *in Proceedings of the 1st Australasian Web Conference (AWC'13)* [41].

4.1 Notations and Definitions

While there exist many definitions of a community within a larger set of users, most of these definitions are generally based on the concept that the community comprises individuals who are more densely connected to each other in the community than to those outside the community. Specifically, Radicchi et al. introduced the concept of strong and weak communities where strong communities comprise *individual* users each of whom has more links within this community than outside, while weak communities comprise users who *collectively* have more links within this community than outside [53]. In particular, we implemented a modified version of Radicchi et al.'s definition of a strong community by introducing a community strength factor for adjusting the strength of the community detected.

We first model the social network as an undirected, unweighted graph, $G = (N, E)$ where N and E respectively refer to the set of nodes/users and edges/links in the graph.[2] Undirected links correspond to social links that are reciprocal and reflective of real-life friendships, thus our choice of undirected links for the algorithm. While our description uses unweighted links, the algorithm could cater for weighted links by implementing a simple filtering scheme based on the weights of links. This filtering scheme would work in such a way that links below a certain threshold weight are excluded for consideration as part of the graph.

Each user $i \in N$ has k_i links, each link pointing to another user either within or outside the community. The number of links pointing to users within the community is denoted as k_i^{in} and those outside the community as k_i^{out}. In addition, we introduce a community strength factor f that allows us to adjust the size and strength of the detected communities. Our definition of a community is as denoted:

$$k_i^{in} > k_i^{out} \times f \tag{1}$$

[2] The terms nodes and users are used interchangeably but they refer to the same entity. Similarly, the terms edges and links are used interchangeably.

Our proposed algorithm differs from that of Radicchi et al. in two ways. Firstly, we introduce a community strength factor f to their original definition of a strong community. This community strength factor f allows us to adjust the strength and size of the community detected. Secondly, the method proposed by Radicchi et al. takes an entire graph and iteratively divides it until the separate communities emerge, whereas our algorithm starts from a single seed user and gradually builds up the community surrounding this user. As future work, we intend to enhance our proposed algorithm by providing an automated way to determine the community strength factor f based on certain desired community characteristics.

We also use the term n-hop or nth degree neighbours of a node i to refer to the set of nodes who can be reached by node i through n number of intermediate nodes. Table 2 lists a summary of the notations and definitions used in our algorithm.

Table 2: Notations and Definitions - Seed-centric Community Detection

Notations	**Definitions**
k_i	The total number of links of node i
k_i^{in}	The number of internal (within the community) links of node i
k_i^{out}	The number of external (outside the community) links of node i
f	The community strength factor

4.2 Overall Framework

The overall framework of our algorithm can be divided into two main stages, namely the identification of a seed node (i.e. the user of interest) followed by an iterative expanding ring search to include users into the community (based on our definition of a community as stated in Equation 1). We shall now further elaborate on our algorithm which is presented in Algorithm 1.

Algorithm 1 can be broadly divided into the following specific steps:

1. Identify a user of interest as the seed node and include this user as part of the community. (Fig. 2a)

2. Retrieve all neighbouring nodes of the seed node. Include these 1st degree (one-hop) neighbours as part of the community. (Fig. 2b)

3. Retrieve all the 2nd degree (two-hops) neighbours of the seed node (i.e. neighbours of the neighbours of the seed node). Include them as part of the community if they fulfill our definition of a community as stated in Equation 1. (Fig. 2c)

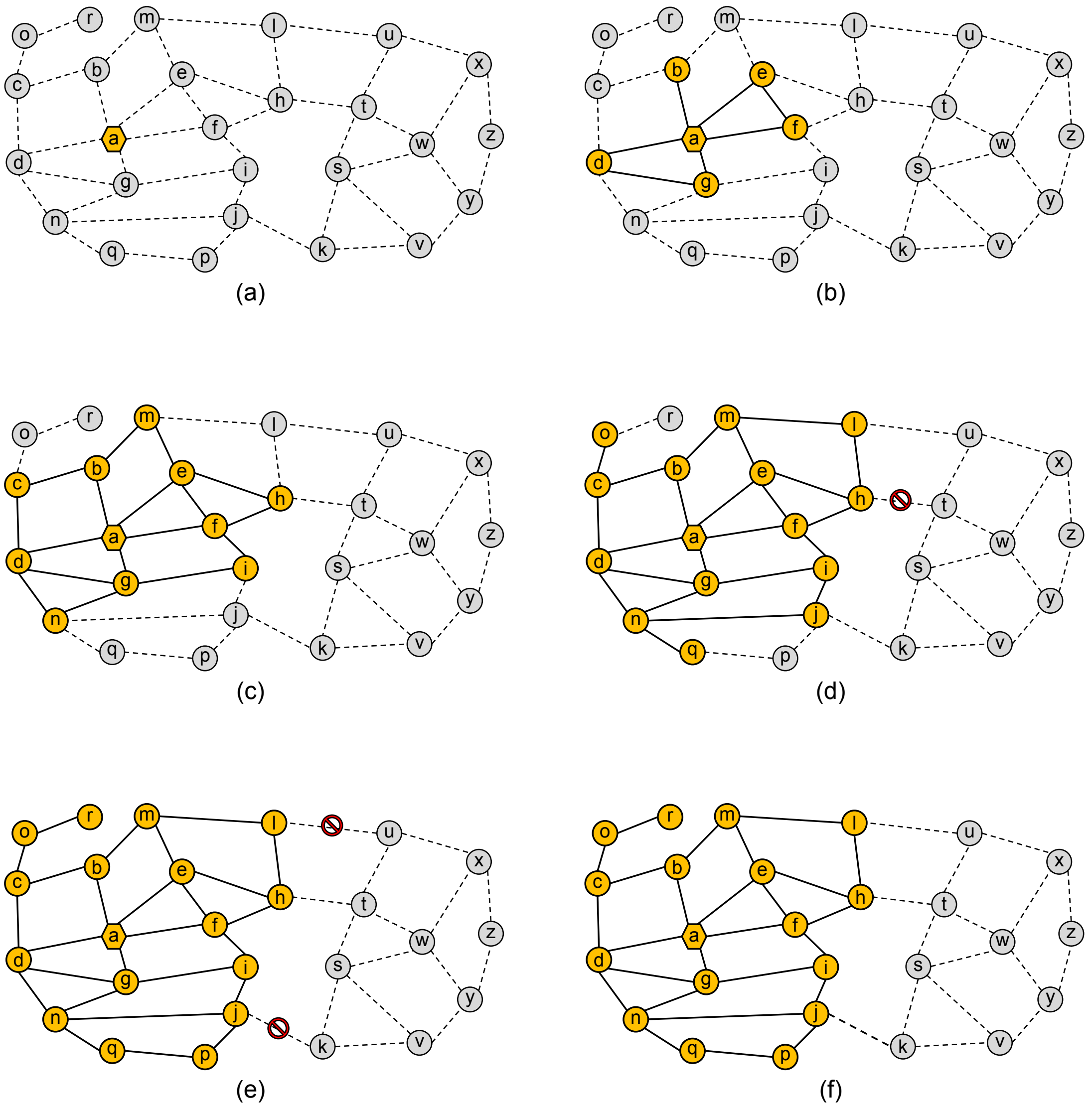
4. Repeat Step 3 for the 3rd, 4th, nth degree neighbours until no further nodes can be added to the community. (Fig. 2d and 2e)

5. The eventual list of included nodes would be the community centered at the seed node. (Fig. 2f)

4.2.1 Identifying a Seed Node

Our algorithm is primarily tailored for detecting a community that is centered at an individual of interest. As such, Step 1 of Algorithm 1 is to identify such a user as the seed node s. In real-life, this seed node s can correspond to an individual with a large number of links to other users, or a person in a particularly influential position. Examples of such an influential person could be the CEO of a company or the director of a research institute.

While our algorithm is designed to detect communities centered at individuals of interest, it is possible to detect all communities for the entire social network. We could achieve the detection of all communities by a selection of seed nodes that is an approximation of the set of influential users throughout the entire social network. Kewalramani [30] also observed that the entire community structure for a network could be determined based on the selection of an appropriate set of seed nodes.



(a) (b)

(c) (d)

(e) (f)

Figure 2: An illustration of our seed-centric community detection algorithm using an expanding ring search with a community strength factor $f = 1.0$. (a) The underlying network graph with seed node a selected. (b) Inclusion of all one-hop neighbours (i.e. nodes b, d, e, f, g) of seed node a as part of the community. (c) Identification of two-hops neighbours of seed node a and inclusion into the community if they satisfy Equation 1. Nodes c, h, i, m, n are added. (d) Continuation of the algorithm for three-hops neighbours where nodes j, l, o, q are added in the community. Node t is not added as it did not satisfy Equation 1. (e) Continuation of algorithm for four-hops neighbours where nodes p and r are included as part of the community while nodes k and u are excluded, according to Equation 1. (f) The algorithm terminates as no further five-hops neighbours can be found and the detected community centered at seed node a is returned (i.e. all yellow nodes).

Algorithm 1 Seed-centric Community Detection

Input: $G = (N, E)$: An undirected, unweighted social network graph, $s \in N$: the seed node

Output: detectedCommunity: A list of nodes in the community centered at the seed node s

```
begin
Add node $s$ to detectedCommunity
for all Neighbour $n_s$ of node $s$ do
    Add $n_s$ to detectedCommunity
end for
for all Neighbour $n_s$ of node $s$ do
    for all Neighbour $m_n$ of node $n_s$ do
        Add $m_n$ to listNeighbours
    end for
end for
while listNeighbours $\neq$ NULL do
    for all Node $n$ in listNeighbours do
        if $k_n^{in} > k_n^{out} \times f$ then
            Add $n$ to detectedCommunity
            Add $n$ to listNewMembers
        end if
    end for
    listNeighbours = NULL
    for all Node $n$ in listNewMembers do
        for all Neighbour $m_n$ of node $n$ do
            Add $m_n$ to listNeighbours
        end for
    end for
end while
return detectedCommunity
end
```

4.2.2 Iterative Expanding Ring Search

After identifying a seed node, Step 2 of Algorithm 1 is to include all neighbours of the seed node s as part of his/her community, which is reasonable as these neighbours are one-hop friends of seed node s who he/she is more likely to interact with frequently. Some authors have observed that topological links may not correspond to interaction, especially on online social networks [60, 49]. In this case, we can simply modify the definition of links to be based on interaction frequency rather than topological edges.

Following which, Steps 3 and 4 of Algorithm 1 are basically iterative steps that continuously include nodes (which satisfy Equation 1) in an expanding ring search. These steps continue iteratively until no further users can be added to the detected community, based on our definition of a community in Equation 1. This expanding ring search coupled with our definition of a community ensures that the search does not propagate too far, as nodes that do not satisfy this definition will not further propagate the search.

4.3 Summary

In this chapter, we proposed a seed-centric community detection algorithm for finding the community centered at an individual of interest, using an expanding ring search starting from this individual. At each progressive stage of the expanding ring search, we decide whether or not to add a user into this community based on our definition of a community. This definition is derived from the number of internal and external links of a user, coupled with an adjustable community strength factor. Our algorithm then continues iteratively until no further users can be added, thus resulting in the detected community comprising the list of added users.

Many studies have shown that social networks exhibit the characteristics of being small-world and scale-free [1, 48, 50]. A small-world network tend to consist of users who are closely clustered and separated by a small number of hops while scale-free networks comprise a small number of users with a large number of links. These small-world and scale-free characteristics of social networks allows for the efficient use of our algorithm on social networks. Firstly, the scale-free characteristics ensures that there would be sufficient influential users (i.e. users with many links) that can be selected as seed nodes. Secondly, the small-world characteristic ensures that our algorithm does not propagate too far to search for users linked to the seed node since most users are connected by a small number of hops.

One main advantage of this algorithm is that we are able to directly detect the community centered at an individual of interest, instead of having to first detect all communities (in the entire social network) followed by identifying those communities where this individual belongs to. Apart from its potential applications in viral marketing, counter-terrorism and epidemic modeling, this algorithm also overcomes the problem of detecting such communities on large-scale OSNs (e.g. Twitter and Facebook). For online social networks with a strict privacy policy (e.g. Facebook), researchers have to perform individual web-crawls of links in a breath-first search for the entire connected component, followed by detecting all communities and selecting those communities that the individual of interest belongs to. Our algorithm minimizes the number of individual web-crawls needed as we start directly from an individual of interest and progressively add users if they satisfy Equation 1. For users that are not added, we need not perform further web-crawls thus reducing the overall number web-crawls needed to detect the community centered at the individual of interest.

Chapter 5

Investigating the Effects of Interest on the Topological Structures of Communities[1]

In this chapter, we will evaluate the performance of the CICD method (proposed in Chapter 3) on the Twitter social network. In our evaluation, we use topological measures such as community size, clustering coefficient, average path length and average degrees to determine the effectiveness of our method to detect cohesive and well-connected communities. In addition to studying the topological characteristics of communities with common interests, we also investigate the effects of deepening or specialization of interest on these communities.

[1] This chapter is based on the following publications by Kwan Hui Lim and Amitava Datta: "Following the Follower: Detecting Communities with Common Interests on Twitter", *in Proceedings of the 23rd ACM Conference on Hypertext and Social Media (HT'12)* [39]; and "Finding Twitter Communities with Common Interests using Following Links of Celebrities", *in Proceedings of the 3rd International Workshop on Modeling Social Media (MSM'12), in-conjunction with HT'12* [38].

5.1 Data and Methods

The Twitter dataset collected by Kwak et al. [33] is used for our experimentations. This dataset was collected from 6th to 31st June 2009, comprising 41.7 million Twitter users and 1.47 billion links. In addition, the profiles of users with more than 10,000 followers are included and these profiles include details such as user ID, screen name, real name, location, etc. Kwak et al. have made the dataset publicly available [32].

We first study community detection and structure among individuals with a common interest in Section 5.2. We infer the interest of users based on the celebrities followed (as described in the CICD method) as users are unable to explicitly state their interests in Twitter. For this purpose, we identified six celebrities for each interest category, resulting in a total of 30 celebrities representing five categories. As a control group, we randomly chose 200,858 users to represent the group with no shared interest.[2] This control group allows us to compare the community structure of users with no common interest against users with a shared interest.

In Section 5.3, we further examine how the deepening and specialization of interest affects community structure. For this purpose, we compare communities with varying levels of interest in the specialized Country Music category against the general Music category. We selected seven winners of the Country Music Awards[3] from 2001 to 2008 as celebrities for the Country Music category based on their number of followers. Winners from 2009 onwards were not selected as the Twitter dataset [33] only comprises data until 31st June 2009. The control group in this case is the users interested in the Music category described in the previous paragraph.

Our evaluation uses topological measures such as clustering coefficient, average path length, average degree and diameter. The clustering coefficient of a node is the number

[2] This choice of 200,858 users ensures that the control group is larger in size compared to the users with a common interest (detected using our proposed method). This control group allows us to demonstrate that our proposed method is able to detect more communities with common interests that are also larger and more cohesive compared to those in the control group, despite the control group comprising a larger number of users.

[3] http://cmaawards.cmaworld.com/nominees/view-past-winners

of 3-node cliques (which includes this node) out of the total possible number of such 3-node cliques. In our experiments, we use the average clustering coefficient of all nodes in a community. Average path length is the average number of hops between all possible pair of nodes, while average degree refers to the average number of links each node has. Diameter is defined as the maximum value out of all shortest paths among every possible pair of nodes (i.e. the longest shortest path).

5.2 Communities with Common Interest

The Merriam-Webster dictionary defines a community as “a group of people with a common characteristic or interest living together within a larger society” [46]. Our CICD method builds on this definition and detects like-minded communities based on individuals sharing common interests. We evaluate our CICD method by comparing the detected communities (with common interest) to our control group comprising communities with no common interest. This comparison shows that our proposed CICD method effectively detects larger and more cohesive communities, which comprise users who share common interests.

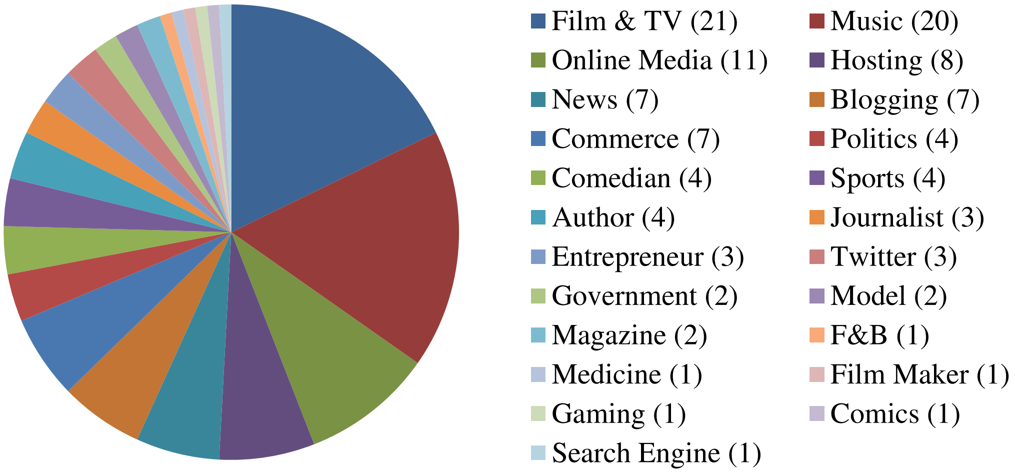


Figure 3: Popular Categories on Twitter

For our study, we selected Film & TV, Music, Hosting, News and Blogging as categories of interest due to their popularity. These categories are selected by first identifying the

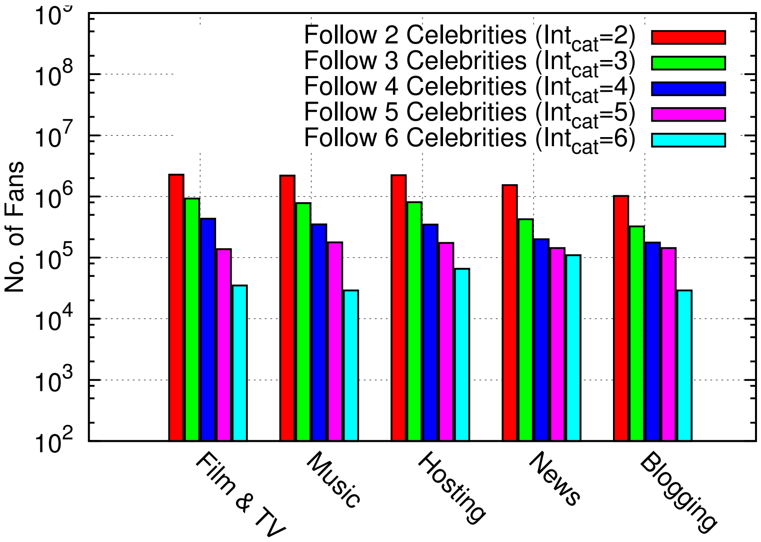


Figure 4: Fans Following Multiple Celebrities in a Category

top 100 celebrities based on their number of followers. Next, we used information on Google[4] and Wikipedia to determine the various categories these celebrities belong to. Following which, we build a list of categories based on the frequency of celebrities belonging to a category. Fig. 3 shows the popular categories on Twitter and we selected the five most popular categories among them.[5] For each category, we selected the six most popular celebrities based on their number of followers as listed in Table 3.[6] Also, a celebrity may belong to multiple categories (e.g. Miley Cyrus belongs to both the Music and Film & TV categories).

The next step of our proposed CICD method involves identifying individuals with common interests, where the interest of a user Int_{cat} is derived from the number of celebrities of category cat followed by the user. We now retrieve the list of users with $Int_{cat} > 1$, for $cat \in \{Film\&TV, Music, Hosting, News, Blogging\}$. A summary of users with $Int_{cat} > 1$ is shown in Fig. 4. In particular, we are interested in users with $Int_{cat} = 6$ as this indicates the most interest in a given category (and corresponds to users who are

[4] http://www.google.com/

[5] Some categories were not included due to the diversity of content within these categories (e.g. Online Commerce).

[6] Choosing six celebrities gives us an ideal number of followers (such that it is a sufficient number for us to detect meaningful communities from). While choosing a higher number of celebrities results in users with a higher level of interest, it also results in less number of followers.

Table 3: Twitter Celebrities

Screen Name	Real Name	Category
aplusk	Ashton Kutcher	Film & TV
mrskutcher	Demi Moore	Film & TV
jimmyfallon	Jimmy Fallon	Film & TV / Hosting
mileycyrus	Miley Cyrus	Film & TV / Music
PerezHilton	Mario A. Lavandeira, Jr	Blogging / Film & TV
50cent	Curtis James Jackson III	Music / Film & TV
britneyspears	Britney Spears	Music
johncmayer	John Mayer	Music
iamdiddy	Sean John Combs	Music
mileycyrus	Miley Cyrus	Film & TV / Music
coldplay	Coldplay	Music
souljaboytellem	DeAndre Cortez Way	Music
TheEllenShow	Ellen DeGeneres	Hosting
Oprah	Oprah Winfrey	Hosting
RyanSeacrest	Ryan Seacrest	Hosting
jimmyfallon	Jimmy Fallon	Film & TV / Hosting
chelsealately	Chelsea Handler	Hosting
Veronica	Veronica Belmont	Hosting
cnnbrk	CNN Breaking News	News
nytimes	The New York Times	News
TheOnion	The Onion	News
GMA	Good Morning America	News
Nightline	ABC News Nightline	News
BreakingNews	Breaking News	News
PerezHilton	Mario A. Lavandeira, Jr	Blogging / Film & TV
mashable	Mashable	Blogging
dooce	Dooce	Blogging
anamariecox	Ana Marie Cox	Blogging
BJMendelson	Brandon Mendelson	Author / Blogging
sockington	Sockington	Blogging

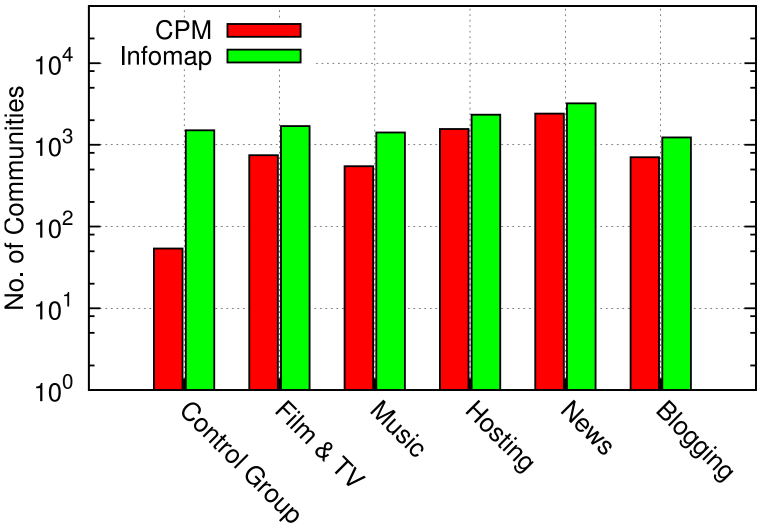


Figure 5: Total Communities Detected

most interested in the product/service).

Table 4: Reciprocity Among Interest Groups

Category	Film & TV	Music	Hosting	News	Blogging
Reciprocity	17.9%	18.2%	15.0%	17.3%	19.6%

We now examine reciprocity based on link information among users with $Int_{cat} = 6$, for $cat \in \{Film\&TV, Music, Hosting, News, Blogging\}$, as shown in Table 4. Reciprocity is obtained based on the number of friendship links out of all links. The reciprocity of 15.0% to 19.6% across all categories corresponds to observations by Cha et al. and Kwak et al. of 10% and 22% respectively for the entire Twitter population [13, 33]. This shows that reciprocity among users with common interests is similar to reciprocity among the general population.

Next, we use the CPM and Infomap algorithms to detect communities among users with $Int_{cat} = 6$, for $cat \in \{Film\&TV, Music, Hosting, News, Blogging\}$. Similarly, we detect communities among our control group comprising users with no common interest. We now compare the communities with common interests against the control group (i.e.

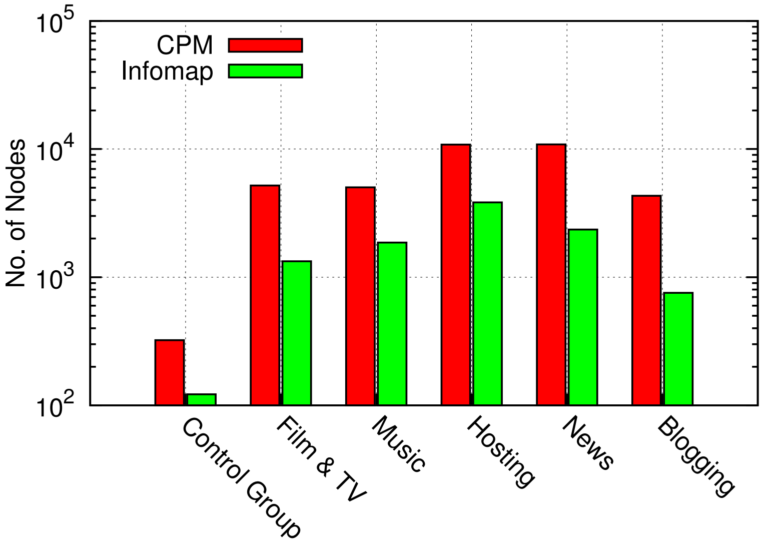


Figure 6: Size of Largest Community Detected

random users with no common interest) in terms of the total number of communities, size of largest community, and average community size as shown in Fig. 5, 6 and 7 respectively.

Fig. 5 and 6 show that users with common interests form more and larger communities than users without a common interest in the control group, regardless of whether CPM or Infomap was used. This is also despite the fact that the control group has a larger population of 200,858 users compared to users with a common interest, which ranges from 29,092 users ($Int_{Music} = 6$) to 109,779 users ($Int_{News} = 6$). Similarly, users with common interests form larger communities on an average as shown in Fig. 7. The exception is the News category detected using CPM as many of the detected communities were cliques of three nodes thus decreasing the average community size. However, our focus is on the largest community detected as this community provides the most benefit for any application of targeted advertising and viral marketing.

The k-value chosen for CPM affects the number and size of communities detected but in all cases, we detect larger and more communities for users with a common interest compared to users without a common interest (given the same k-values). We were able to detect communities with k-values of up to 25 for the News category and we could also detect communities with k-values of 9 or higher for the other categories. For the

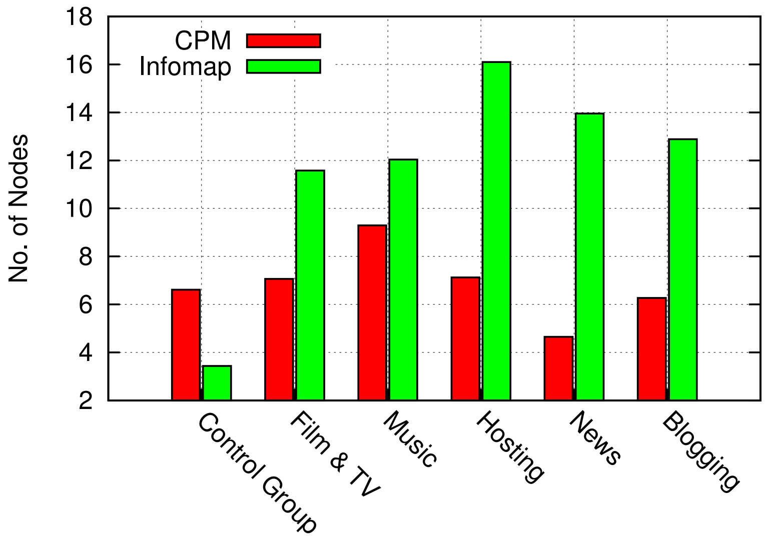


Figure 7: Average Size of Communities Detected

Table 5: Network Statistics of the Communities

Category	Control Group	Film & TV	Music	Hosting	News	Blogging
Average Path Length	2.83	3.03	2.82	3.09	3.35	3.09
Avg. Clustering Coefficient	0.60	0.62	0.63	0.59	0.58	0.62
Diameter	6	7	8	8	8	7
Average Degree	7.81	6.80	7.29	8.17	9.15	7.51

control group, we were unable to detect any communities with k-values higher than 6 which further proves that users with common interest form larger and more communities than users with no common interest. While the k-values affect community detection, this observation shows that our proposed CICD method performs better than the control group (experiment) given the same k-values.

Users with common interests also form communities that are more cohesive than those without common interest. Table 5 shows this trend where the communities with common interest have a higher clustering coefficient than our control group with no common interest, except the Hosting and News categories. However, users interested in Hosting and News have a higher average degree of links which shows that these users are better connected than users in the control group.

These results show that our proposed CICD method detects communities that are both larger and more cohesive. More importantly, our method efficiently detects communities with common interests without the need to perform large scale community detection on the entire social network. Thus, our method is less computationally intensive and compares favourably with existing methods that detect all communities then identify the interests of the communities [27, 37]. These results are also supported by observations of other authors that people with similar interests are more likely to be friends than those with dissimilar interests [19, 62].

5.3 Specialization and Deepening of Interest

Communities that share the same set of interests are likely to be more connected [37, 65] and interact on a more frequent basis [52]. As an extension of that argument, we show that users sharing a specialized interest form a more tightly-coupled community than users sharing a general interest. We show this by comparing users interested in the specialized category of Country Music against users interested in the general category of Music. The control group is the users interested in the general Music category as discussed in Section 5.2. The celebrities representing the Country Music category are seven Country Music singers who have won various awards at the Country Music Awards between 2001 to 2008 and have more than 10,000 followers. These celebrities (representing the Country Music category) are listed in Table 6.

Similar to Section 5.2, we used both CPM and Infomap to detect communities among users with $Int_{Country} > 1$.[7] Due to the smaller population of users following Country Music singers, the absolute number of communities detected by CPM are small (e.g. only 230 users with $Int_{Country} = 7$). We first focus on users with the most interest in

[7] We do not detect communities for users with $Int_{Country} = 1$ as this would mean all fans of any celebrity and this user group would not be meaningful for detecting communities with common interest.

Table 6: Country Music Celebrities

Screen Name	**Real Name**
cunderwood83	Carrie Underwood
KeithUrban	Keith Urban
KennyAChesney	Kenny Chesney
martinamcbride	Martina McBride
paisleyofficial	Brad Paisley
TimMcGrawArtist	Tim McGraw
tobykeithmusic	Toby Keith

Country Music, $Int_{Country} = 7$. For this user group, we detected five communities comprising 23 distinct users as shown in Fig. 8. The five communities are differentiated by nodes that are coloured green, orange, blue, yellow and purple. The grey nodes represent users that belong to multiple communities and serve as middlemen connecting the various communities. We also observed similar trends in the communities detected by Infomap.

5.3.1 Effects of Interest Specialization

In this section, we investigate the changes in the formation of communities and their topological structures as users specialize in their common interest (i.e. specializing in Country Music from the general Music category). To provide a relative comparison among users with $Int_{Music} = 6$ and $Int_{Country} = x,\ for\ 2 \leq x \leq 7$, we normalize the results by the number of users in each group. This normalization gives us an accurate representation of the community characteristics of each interest group without the biases of the base population size (e.g. 800 users with $Int_{Country} = 6$ compared to 29,092 users with $Int_{Music} = 6$).

The normalized average size of communities indicates the likelihood of large communities being formed among users with common interests. This measure allows us to compare if users with specialized interests form larger communities than users with a

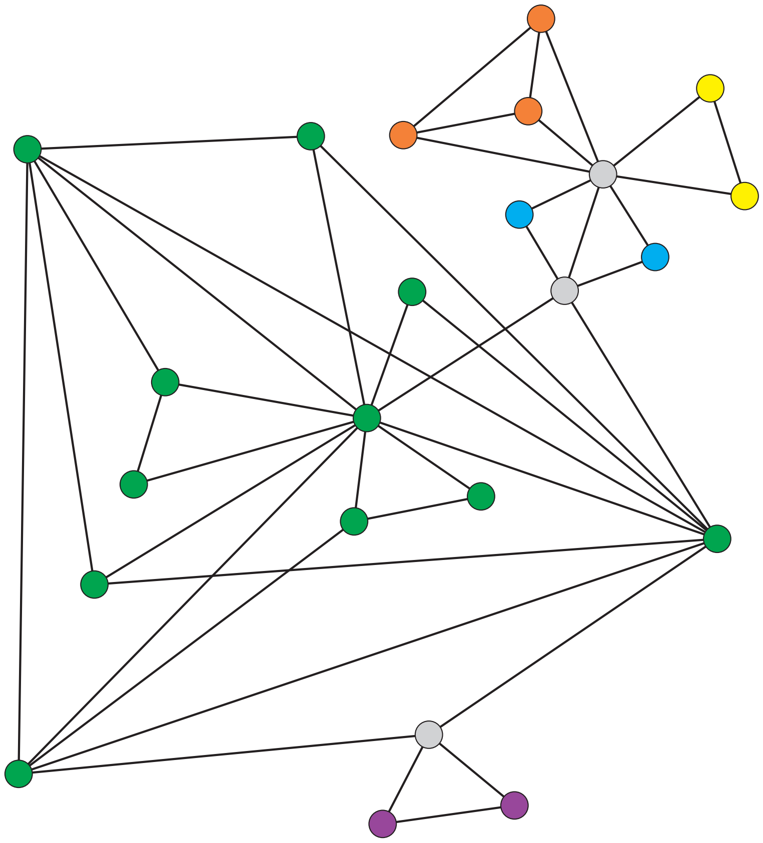


Figure 8: Community Graph of Fans who follow all Seven Country Singers

general interest. Comparing two user groups with the same level of interest in different categories (i.e. $Int_{Music} = 6$ and $Int_{Country} = 6$), we observe that the normalized average community size of the $Int_{Country} = 6$ group is 23 and 28 times larger than the $Int_{Music} = 6$ group using CPM and Infomap respectively, as shown in Fig. 9. This result shows that users sharing the same level of interest form larger communities if that interest is more specialized.

Even among users with a lower level of interest in a specialized category, they are more likely to form larger communities on an average compared to users with a higher level of interest in a general category. Fig. 9 shows that users with a lower interest in the specialized Country Music category ($Int_{Country} = 3$) have a normalized average community size that is up to two times larger than that of users with more interest in the general Music category ($Int_{Music} = 6$).

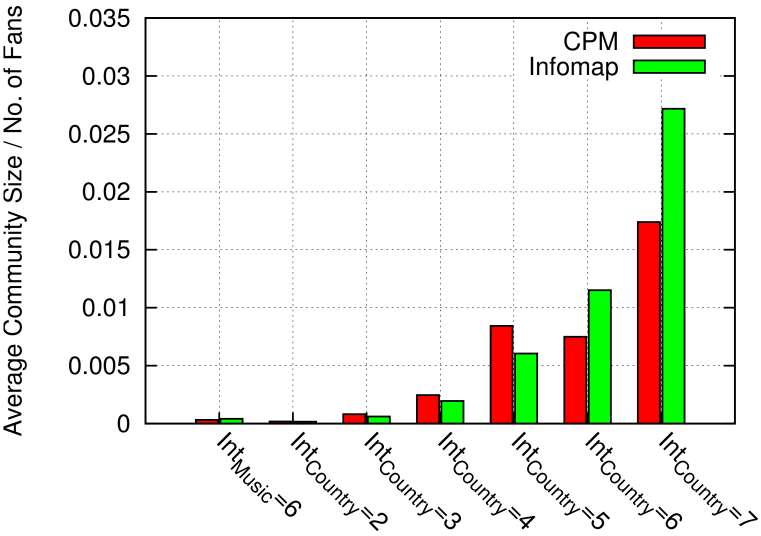


Figure 9: Normalized Average Community Size for Music and Country Music Categories

Table 7: Comparison of General and Specialized Interest

Category	General (Music)	Specialized (Country)
Average Path Length	2.82	2.10
Avg. Clustering Coefficient	0.63	0.76
Diameter	8	4
Average Degree	7.29	5.52
Reciprocity	18.2%	20.1%

Communities comprising users with a specialized interest are also more cohesive and well-connected than those with a more general interest. Table 7 best illustrates this where users with a specialized interest in Country Music form communities with a shorter average path length and diameter but higher clustering coefficient compared to those with a general interest in Music. In addition, users with $Int_{Country} = 6$ displayed a higher reciprocity of 20.1% compared to 18.2% for users with $Int_{Music} = 6$. This result shows that users with a specialized interest are more likely to be mutual followers of each other (i.e. be mutual friends) compared to users with a general interest.

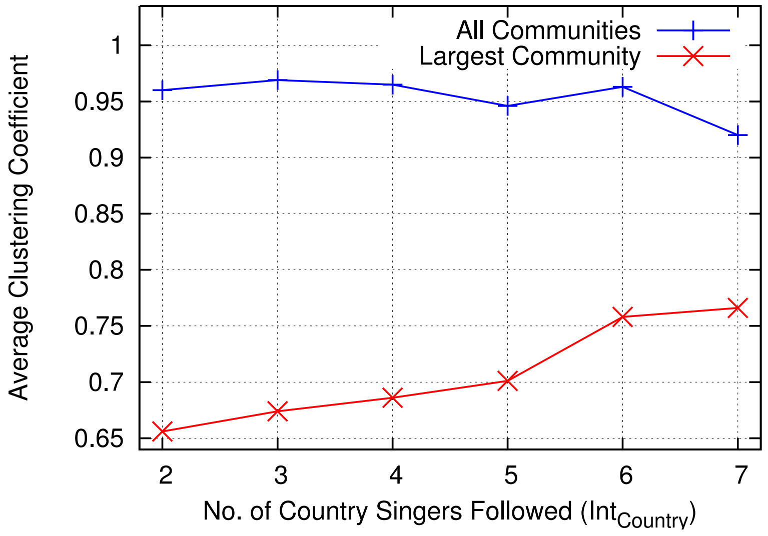


Figure 10: Average Clustering Coefficient of Country Music Category

5.3.2 Effects of Interest Deepening

Next, we investigate the changes in communities as their interest in a category grows deeper, which is indicated by an increasing Int_{cat} value. Specifically, we report on the changes in number of communities, normalized average community size, average clustering coefficient and average path length among users as their interest deepens. The size and number of communities shows how likely users with common interests form communities while clustering coefficient and path length gives an indication of connectedness within the communities.

An increase in interest level among users corresponds to an increase in their normalized average community size. Fig. 9 shows an increasing average community size with increasing $Int_{Country}$ values. This result supports our original observation that communities are more likely to be formed among like-minded individuals. In addition, the average size and number of communities formed increases as the interest level of the users increases.

Communities comprising users with common interests also get more tightly coupled as their level of interest increases. Fig. 10 shows a gradual increase in clustering coefficient among the largest communities with increasing $Int_{Country}$ values. While the average

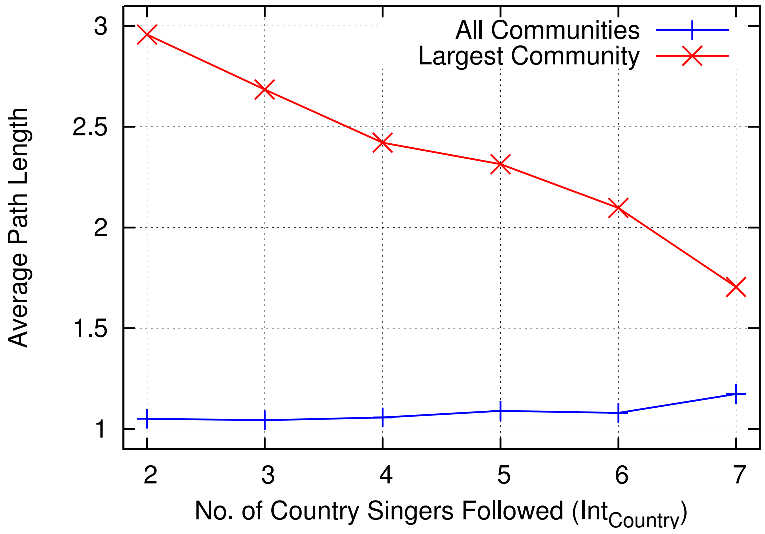


Figure 11: Average Path Length of Country Music Category

clustering coefficient of all communities remains relatively constant (from $Int_{Country}$ = 2 to $Int_{Country}$ = 6), this is due to the large number of small cliques detected at low $Int_{Country}$ values which increases the average clustering coefficient significantly. For example, out of 539 communities detected (with $Int_{Country} = 2$), 397 communities are cliques of three users thus having a clustering coefficient of one. At higher $Int_{Country}$ values, less of such cliques are detected thus they have less influence on the average clustering coefficient. We are most interested in the largest community (which shows an increasing clustering coefficient) as this community has the most potential for targeted advertising and viral marketing due to its size and cohesiveness.

Fig. 11 shows an average path length of 1.7 to 3.0 hops within the largest communities at varying values of $Int_{Country}$, illustrating that users sharing common interests form communities that are better connected. This compares well with Milgram's "six degrees of separation" which states that everyone is connected by six hops of acquaintances [47]. Similarly, studies on the Microsoft Messenger social network also show that their users are separated by an average of 6.6 hops [36]. Although we compare average path length of communities and not the entire population, the largest community for $Int_{Country} = 2$ comprising 3,725 users still shows a short average path length of three hops.

These experiments show that an increasing level of interest in a category correlates with detecting larger and more communities on average, who display characteristics of a higher clustering coefficient and shorter path length. This observation supports our initial claim that a community becomes more cohesive and tightly-coupled as its users share a deeper level of interest in a category.

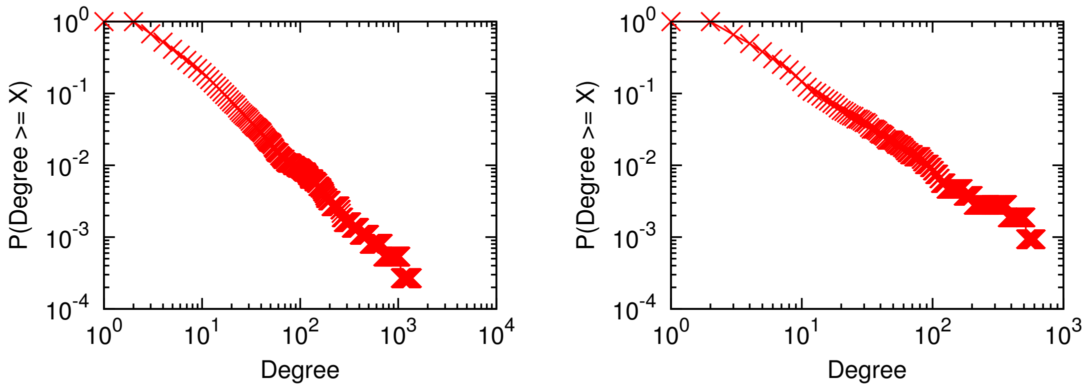


Figure 12: Degree Complementary Cumulative Distribution Function of Largest Community with $Int_{Country} = 2$ (left) and $Int_{Country} = 4$ (right)

The detected communities also display the characteristics of scale-free networks as shown in Fig. 12, which plots the Complementary Cumulative Distribution Function of the degree distribution of users with $Int_{Country} = 2$ and $Int_{Country} = 4$ respectively. The communities with other $Int_{Country}$ values also displayed similar trends. Upon closer examination, we observe that many individuals with large degree distribution are also country music artists but with less fans than the celebrities we have chosen (i.e. less than our threshold of 10,000 fans/followers). The fact that there are other minor country singers among these communities shows that our method effectively detects communities comprising users with a common interest. Using the Twitter API [25], we retrieved the profiles of 1,164 users (with $Int_{Country} = 2$), the remaining user profiles could not be retrieved due to locked or inactive accounts. Examining the retrieved user profiles, we observed that more than 7.7% of these users are from Nashville, Tennessee, a town that is closely associated with country music and hosts the annual Country Music Association Music Festival. This result shows a possible correlation between the interest of a user and his/her geographic location.

5.4 Summary

In this chapter, we evaluated our proposed CICD method that aims to efficiently detect like-minded communities comprising individuals with common interests, for applications in targeted advertising and viral marketing. This method detects communities that are larger, more cohesive and only comprise users that share a common interest, instead of detecting all communities. As Twitter has no explicit options for users to state their interests, we derived a measurement of interest based on the number of celebrities in an interest category that the user follows. Given the large-scale and growth rate of Twitter (and other OSNs), our method is very scalable for identifying communities sharing common interests as it only requires topological information.

In addition, the CICD method can also be applied to other online social networking sites by adapting to the unique characteristics of each site and their representations of celebrities and links. For example, in Facebook[8], celebrities could be defined as the respective Facebook pages of these celebrities and followership links as the individual user "likes" on these pages. Thereafter, our method could be applied as described in the paper using these Facebook pages (celebrities) and user "likes" (followership links).

From a sociology perspective, we also studied the characteristics among users with a common interest compared to users without a shared interest, particularly in the way they form communities and the topological structures of these communities. Also, we observed how their community structures become more connected and cohesive with deepening interest in a given category, as indicated by an increasing clustering coefficient and decreasing path length. Similarly, the communities become more connected and cohesive as users specialize in their interest (e.g. from the general Music category to the specialized Country Music category). These observations along with our proposed method of community detection provide a tool for the implementation of targeted advertising and viral marketing, especially for products with a niche or specialized audience.

[8] www.facebook.com

Chapter 6

Analyzing the Activity Patterns and Behaviour of Like-minded Communities based on Topological and Interaction Links[1]

After evaluating the CICD method in Chapter 5, we now evaluate the HICD method in this chapter. In addition to topological measures, we also compare interaction measures such as the frequency and content of messages sent among users in the communities detected using our proposed methods. This evaluation is performed on the Twitter social network and provides an insight into the communication behaviour and patterns of these communities with common interests. Also, we performed a preliminary study into the evolution of links in these communities over time.

[1] This chapter is based on the following publication by Kwan Hui Lim and Amitava Datta: "Tweets Beget Propinquity: Detecting Highly Interactive Communities on Twitter using Tweeting Links", *in Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence (WI'12)* [40].

6.1 Data and Methods

Using the Twitter API, we retrieved the user profiles, linkages, tweets and retweets of 17,941 Twitter users identified as four different Set $\mathcal{P}$ of the country music, tennis and basketball (Mavericks and Bulls teams) categories.[2] Each Twitter API call allows us to retrieve the last 200 tweets of any (unlocked) user. In total, we retrieved and analyzed 1.9 million tweets and retweets from 17 Nov 11 to 14 Jan 12.

Similar to Chapter 5, we evaluate the HICD method using topological measures such as community size, average clustering coefficient, average path length and average degree. Although we have previously evaluated the CICD method, we also perform this evaluation on the CICD method as a comprehensive comparison between the CICD and HICD methods. As previously mentioned in Chapter 3, the communities detected by the CICD and HICD methods are denoted Com_{CICD} and Com_{HICD} respectively.

In addition, we also evaluate the performance of our HICD method by analyzing the frequency and content of tweets among the detected communities, specifically on the usage of @mentions, #hashtags, URLs and keywords. @mentions, #hashtags and URLs are easily identified in tweets by respectively searching for the '@', '#' and "http://" prefixes to any word. On the other hand, keywords require some pre-processing to filter out commonly used words that have no significant meaning, such as pronouns, prepositions and conjunctions as listed in Table 8.

6.2 Comparison of Topological Structures of Communities

For our study, we demonstrate the effectiveness of our HICD method across different communities with common interests in country music, tennis and basketball respectively.

[2] While we selected these four categories, the CICD and HICD methods can be effectively applied to other categories by selecting celebrities that are representative of other interest categories. Instead of using the same dataset in Chapter 5, we collected our own dataset as the HICD method requires tweeting information which is not available in the former dataset (by Kwak et al. [33]).

Table 8: Words to filter

Type	Examples
Pronoun	I, you, she, he, it, we, you, they, me, her, him, it, us, you, them, mine, yours, hers, his, its, ours, theirs, this, that, etc
Preposition	about, above, across, after, against, among, around, at, before, behind, below, beneath, beside, between, beyond, but, etc
Conjunction	after, although, as, because, before, if, once, since, than, that, though, till, until, when, where, whether, while, etc

Table 9: Representative celebrities for interest categories

Country Music	Tennis	Dallas Mavericks	Chicago Bulls
Taylor Swift	Serena Williams	Lamar Odom	C. J. Watson
Brad Paisley	Rafael Nadal	Jason Terry	Carlos Boozer
Blake Shelton	Andy Murray	Dirk Nowitzki	Luol Deng
Miranda Lambert	Novak Djokovic	Shawn Marion	Kyle Korver
Kenny Chesney	Caroline Wozniacki	Vince Carter	Taj Gibson
Keith Urban	Venus Williams	Jason Kidd	Ronnie Brewer
Martina McBride	Andy Roddick	Brian Cardinal	Jimmy Butle
Tim McGraw	Sania Mirza-Malik		
Toby Keith	Kim Clijsters		

We selected nine country music celebrities based on winners of the Country Music Association Awards[3] from 2001 to 2011, with more than 90,000 followers. Similarly, we selected nine prominent tennis players for the tennis category based on their number of followers on Twitter. For the basketball category, we focused on two different National Basketball Association (NBA) teams: the Dallas Mavericks and Chicago Bulls. We selected seven players from each NBA team based on the team's current player roster. The list of celebrities representing each interest category is listed in Table 9.

Next, we retrieve the set of users following all celebrities in each category, Set $\mathcal{P}$ as

[3] http://cmaawards.cmaworld.com/nominees/view-past-winners

Table 10: Set $\mathcal{P}$ for the various interest categories

Category	**No. of Users**
Country Music	5,969
Tennis	2,708
Dallas Mavericks (Basketball)	4,457
Chicago Bulls (Basketball)	4,807

described in Chapter 3. The number of users in Set $\mathcal{P}$ of each category is shown in Table 10. Using the CICD method, we first modify Set $\mathcal{P}$ by removing all links that are not reciprocal. Following which, we run CPM and Infomap on the modified Set $\mathcal{P}$, resulting in communities with a common interest (Com_{CICD}) in the country music, tennis and basketball (Mavericks and Bulls) categories as shown in Fig. 13. From these detected communities, we selected the largest community (of each category) to analyze their tweeting and retweeting patterns within the community.

Using our HICD method, we determine the tweet-based community (Com_{HICD}) based on the Set $\mathcal{P}$ of users mentioned in the previous paragraph. For this purpose, we define the weight threshold T as 1, for constructing the Set $\mathcal{Q}$ of users. Similarly, we run CPM and Infomap on Set $\mathcal{Q}$ and concentrate on the largest community (of each category) for our study. The number of detected communities and size of the largest community are shown in Fig. 13 and 14 respectively.

The number of communities detected by our HICD method is dependent on the duration of the tweets collected. A longer period of tweet collection results in a larger number of communities detected, as there is a higher probability of users @mentioning each other. This observation is reflected by Fig. 13 where our HICD method (Com_{HICD}) detects more country music communities than the CICD method (Com_{CICD}). This result is due to Com_{HICD} of country music being detected using tweets from 17 Nov 11 to 14 Jan 12 whereas Com_{HICD} of tennis and basketball are only based on the past 200 tweets collected on 12 Jan 12.[4] Regardless of whether CPM or Infomap was used, Fig. 14 shows

[4] Even when the tweets are collected on a single day, the tweets dated more than six months back as

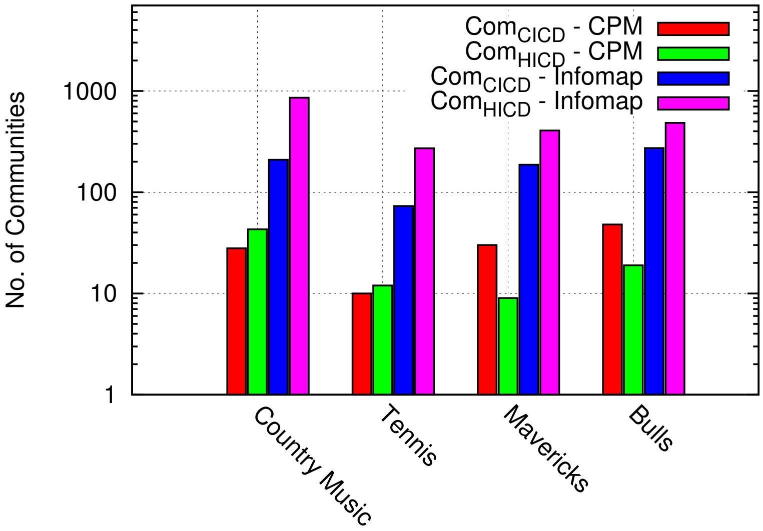


Figure 13: Total communities detected

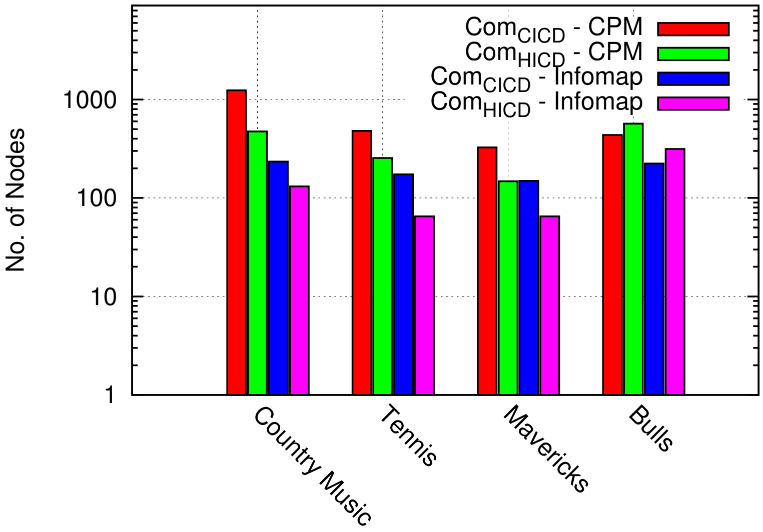


Figure 14: Size of largest community detected

a similar trend in the largest community detected (e.g. communities detected by CPM are larger than that selected by Infomap or vice versa, given the same interest category).[5]

As our HICD method uses implicit links derived from communication frequency, it is possible to detect communities that are not detectable using topological information of follower/following links. Fig. 14 best illustrates this phenomenon where the Com_{HICD} of Bulls is larger than its Com_{CICD} counterpart. This observation shows that our HICD method is able to detect communities based on communication links, even when there are no follower/following links present. Even if these users eventually form follower/following links because of their frequent communication, our HICD method is able to detect such users before they form these topological links. Furthermore, our HICD method filters out users that are topologically connected but otherwise do not communicate with each other. We now compare Set $\mathcal{P}$, Com_{CICD} and Com_{HICD} of the different categories, in terms of topological measures (clustering coefficient, average path length and average degree) to evaluate the effectiveness of our method.

Our HICD method detects communities (Com_{HICD}) that are more connected and cohesive than Set $\mathcal{P}$ and Com_{CICD} across all categories as shown in Fig. 15. Our HICD method outperforms the CICD method as indicated by a higher clustering coefficient of Com_{HICD} compared to Com_{CICD}. Despite the improvement, it is challenging to achieve a clustering coefficient close to one as only a fully-connected sub-graph (i.e. a clique) has a clustering coefficient of one. The Com_{CICD} and Com_{HICD} of all categories also have a clustering coefficient two times or more than Set $\mathcal{P}$ of their respective categories.

Similarly, Fig. 16 shows a shorter average path length for Com_{HICD} compared to Com_{CICD}, for the Mavericks and Bulls categories. As Set $\mathcal{P}$ contains disconnected segments of the network, the average path length could not be calculated. Com_{HICD} of country music has a longer path length than Com_{CICD} due to our choice of one for the threshold T of $I_{i,j}$ value. Once this threshold value is increased, Com_{HICD} progressively gets a

the most recent 200 tweets were collected. This meant that the country music group had two months more of tweets compared to the tennis and basketball groups.

[5] The largest community provides the most potential for targeted advertising and viral marketing and is the one we are interested in.

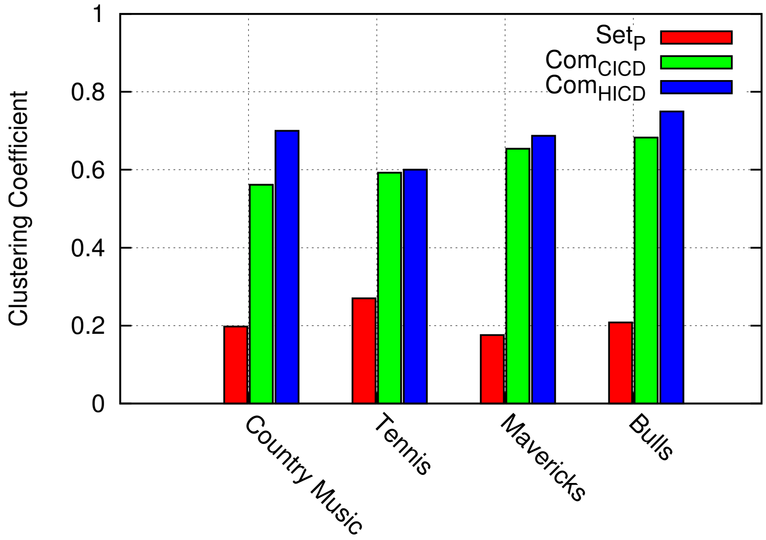


Figure 15: Clustering coefficient

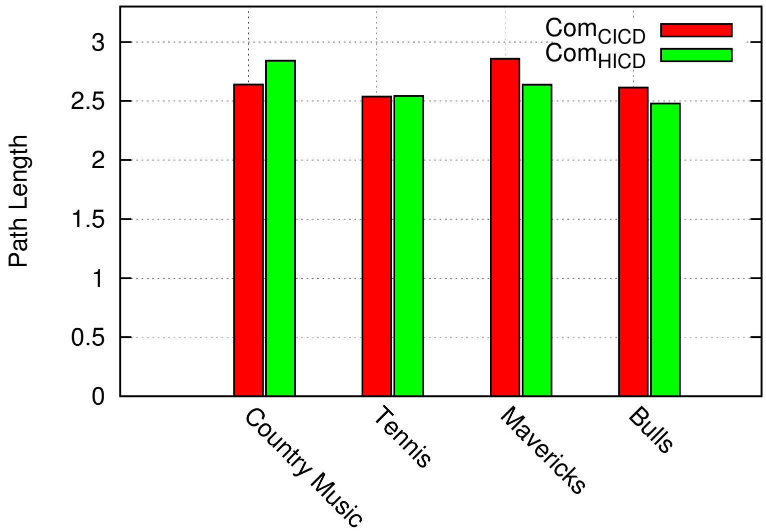


Figure 16: Average path length

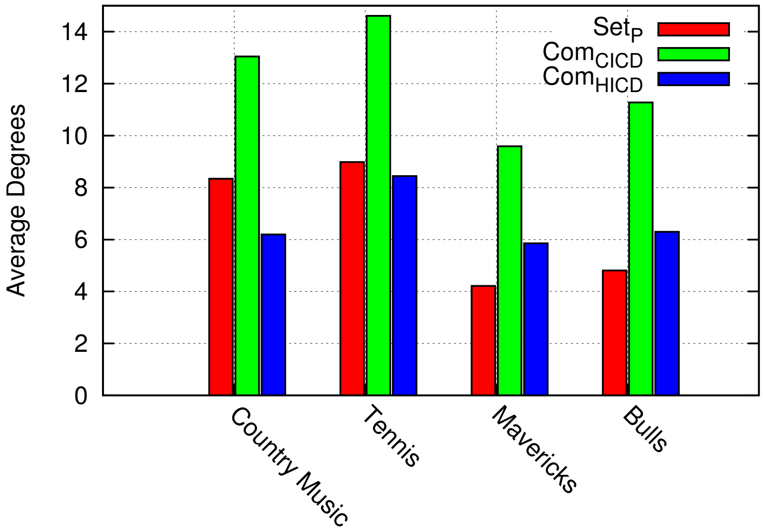


Figure 17: Average degree

shorter average path length compared to Com_{CICD} as shown in Table 12. The shorter average path length and higher clustering coefficient show that our HICD method detects communities that are more cohesive and connected.

Fig. 17 shows that Com_{HICD} has an average degree of links more similar to Set $\mathcal{P}$ (than Com_{CICD}) and significantly lower than Com_{CICD}. However, Com_{HICD} also has a higher clustering coefficient than Com_{CICD}, despite the lower average degree of Com_{HICD}. This observation shows that while Com_{HICD} has less average links, most of its links are connected to nodes within the same community. On the contrary, Com_{CICD} has more average links but many of the links are connected to nodes outside the community. These results show the effectiveness of our HICD method in detecting highly cohesive and connected communities.

Table 11 shows the top three locations stated in the profiles of users in Set $\mathcal{P}$, Com_{CICD} and Com_{HICD} of each category. The top location of each category is consistent throughout the user groups and is representative of the respective category. For country music, Nashville is home to many country music events such as the CMA Music Festival and CMA Awards. As for tennis, London is the venue of the popular Wimbledon Tennis Championships. Similarly for Mavericks and Bulls, their teams are based in Dallas and

Table 11: Top 3 user locations

Category	**Set** $\mathcal{P}$	Com_{CICD}	Com_{HICD}
Country Music	Nashville Quito Canada	Nashville Quito Canada	Nashville Quito Boston/Charlotte
Tennis	London Greenland Quito	London Paris Melbourne	London Paris Melbourne
Dallas Mavericks	Dallas Quito Philippines	Dallas Toronto Fort Worth	Dallas Fort Worth Various Texas Cities
Chicago Bulls	Chicago Quito Melbourne	Chicago New Jersey Melbourne	Chicago Aurora/Quito Melbourne

Chicago respectively. This result shows that members of such communities are geographically collocated and likely to know each other personally. Hence they may tweet to each other even when they are not connected through topological follower/following links. In addition, authors such as Java et al. [27] also noticed that the probability of two persons being connected is negatively correlated with their geographic distance. However, it should be noted that more than 20% of the examined users do not provide a specific location in their user profiles. Also, many users provide only general country locations (e.g. USA, Canada) or non-existent places (e.g. “Mother Ship castaway”, “Over here!”).

Next, we study the effects of increasing the threshold T of $I_{i,j}$ values, one of which is a corresponding increase in the cohesiveness and connectedness of the detected communities. This observation is supported by the trend of a decreasing path length and diameter, and increasing clustering coefficient with an increasing threshold T for the country music category, as shown in Table 12. This general trend is consistent with an increasing threshold T, apart for a minor deviation at a threshold T of 5. On the other hand, an increasing threshold T results in smaller communities being detected. This result shows

Table 12: Effects of increasing threshold T of $I_{i,j}$ for country music category

Threshold T of $I_{i,j}$	**1**	**2**	**3**	**4**	**5**	**6**
No. of Nodes	474	313	188	108	70	42
Average Path Length	2.84	2.63	2.64	2.52	2.68	2.49
Avg. Clustering Coefficient	0.70	0.72	0.74	0.77	0.75	0.77
Diameter	6	6	6	5	5	4
Average Degree	6.20	6.27	5.67	5.28	4.66	4.52

a trade-off between detecting more cohesive communities (at high threshold T) or larger communities (at low threshold T). For the rest of the chapter, we focus on the country music communities detected using a threshold T of 1 as we are most interested in the largest community.

6.3 Analysis of Communication Behaviour and Patterns

As a holistic approach to identifying highly interactive communities with common interest, it is necessary to consider their communication frequency and content. However, the CICD method considers only the topological information of the social network. Our HICD method improves upon the CICD method by considering the frequency of direct communication (via the use of @mentions in tweets) between individuals. We now examine the results from our HICD method based on a comparison of the top 10 #hashtags, @mentions, URLs and keywords among the three groups of users: Set $\mathcal{P}$, Com_{CICD} and Com_{HICD} of the country music category.

From a topical aspect, our HICD method detects communities that tweet more frequently about the common interest (i.e. country music). This statistic is determined based on the #hashtags that are most frequently used. Table 13 shows that among the top 10 #hashtags of Com_{HICD}, five #hashtags are related to country music (denoted by *). This result compares favourably with Com_{CICD} and Set $\mathcal{P}$ which have only two and one #hashtags related to country music, respectively. It is also important to note that the five country

Table 13: Top 10 #hashtags

Set $\mathcal{P}$	Com_{CICD}	Com_{HICD}
#FF	#FF	#FF
#fb	#fb	*#CMAawards**
#NowPlaying	#NowPlaying	#nowplaying
#nowplaying	*#CMAawards**	#fb
*#CMAawards**	#nowplaying	#PeoplesChoice
#iTunes	#jesustweeters	*#cmchat**
#PeoplesChoice	#iTunes	#ff
#ff	*#concert**	*#CMTAOTY**
#jesustweeters	#DT	*#countryartist**
#concert	#Nashville	*#ACAs**

music #hashtags of Com_{HICD} are related to country music in general and not to any specific country singer used in the initial seed of celebrities. This observation shows that our HICD method detects communities that are interested in the general category instead of just a specific celebrity representing that category.

Table 14: Top 10 @mentions

Set $\mathcal{P}$	Com_{CICD}	Com_{HICD}
youtube	youtube	*blakeshelton**
*blakeshelton**	*blakeshelton**	*davidnail**
YouTube	YouTube	*Miranda_Lambert**
GetGlue	*taylorswift13**	*ladyantebellum**
*taylorswift13**	*Miranda_Lambert**	GetGlue
justinbieber	*davidnail**	*ScottyMcCreery**
*Miranda_Lambert**	GetGlue	*ChrisYoungMusic**
*ScottyMcCreery**	*BradPaisley**	*Lauren_Alaina**
*BradPaisley**	*JimmyWayne**	*taylorswift13**
*jakeowen**	*jakeowen**	SUGARLAND4EVER

Likewise, our HICD method detects communities that make more @mentions of country music artists. Table 14 best illustrates this where eight of the top 10 @mentions of Com_{HICD} are country singers (denoted by *). Comparatively, Com_{CICD} and Set $\mathcal{P}$

has less @mentions of country music artists at a count of seven and six respectively. It is also worthwhile to note that five out of eight country singers (in the top 10 @mentions of Com_{HICD}) were not used as the initial seed of representative celebrities to construct Com_{HICD}. This observation shows that our HICD method is able to detect communities that frequently interact about country music in general, and not just about country singers in the initial seed of celebrities used. We also observed similar trends for the tennis and basketball categories.

Table 15: Top 10 *URLs*

Set $\mathcal{P}$	Com_{CICD}	Com_{HICD}
*Kickin Country Radio**	*Kickin Country Radio**	Branson Shows Ticket Booking
Trapier Blog	Trapier Blog	Branson Restaurant Discounts
GetGlue Invitation	B-93.7 FM Radio	People's Choice Voting
B-93.7 FM Radio	Youtube Video	GetGlue Invitation - User A (Anonymized)
Youtube Video	Escape Dates	TwittaScope - Taurus
Escape Dates	Branson Shows Ticket Booking	World Wrestling Entertainment
Lynzie Taylor Barton Blog	Branson Restaurant Discounts	GetGlue Invitation - User B (Anonymized)
Tax Reform	People's Choice Voting	People's Choice Voting
Lynzie Taylor Barton Blog	B-93.7 FM Radio	World Wrestling Entertainment
GetGlue Follow	TwittaScope - Virgo	UStream Video Streaming

We now examine the top 10 URLs used and present the broad title of the websites, instead of TinyURL addresses which do not have any textual meaning. TinyURLs are short versions of URLs and are often used in tweets to overcome the 140-character limit. Table 15 shows the top 10 websites that Set $\mathcal{P}$, Com_{CICD} and Com_{HICD} of the country music category use in their tweets. While Set $\mathcal{P}$ and Com_{CICD} have one URL related to country music, the exchange of URLs in Com_{HICD} is of a more personal nature. Examples are the two GetGlue invitations to join existing members, which indicate a friendship relationship that also exist outside of Twitter.

In addition, we also analyze the top 10 keywords for the three groups of users with the filtering criteria described in Section 6.1. Even after filtering out pronouns, prepositions, conjunctions and interjections, we did not notice any significant trends in keywords used. However, we observe that the ":)" and ".." character sequences were among the top 10 keywords used, even though these are not textual keywords.

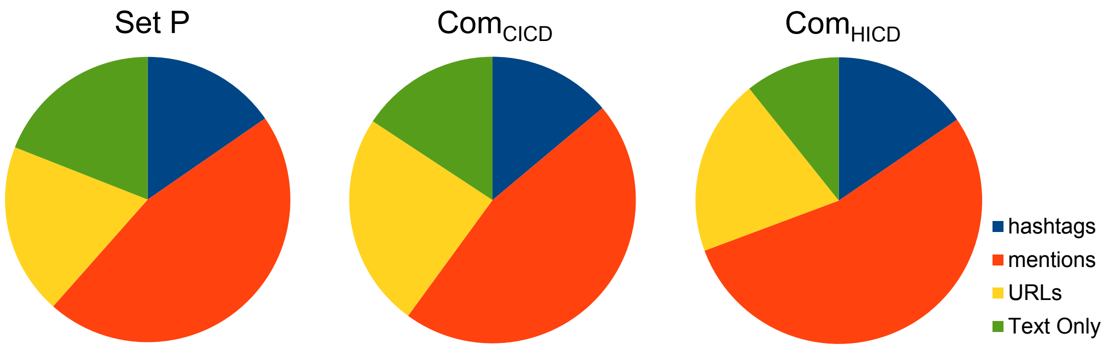


Figure 18: Type of tweets

6.3.1 Trends in Tweeting

We investigate tweeting trend by first examining the type of content covered in the tweets posted by Set $\mathcal{P}$, Com_{CICD} and Com_{HICD}, as illustrated in Fig. 18. The type of content in tweets can be any combination of textual information, #hashtags, @mentions and/or URLs. Fig. 18 shows the distribution of these content types for Set $\mathcal{P}$, Com_{CICD} and Com_{HICD} of the country music category. Set $\mathcal{P}$ and Com_{CICD} use similar allocation of the content types in their tweets with Set $\mathcal{P}$ using more text-based tweets and Com_{CICD} using more URLs. As our HICD method detects communities based on frequent direct communication, Com_{HICD} uses mostly @mentions in their tweets. We next investigate trends in the timings of tweets.

Across Set $\mathcal{P}$, Com_{CICD} and Com_{HICD}, Fig. 19 shows a slight increase in tweeting activities from 0900hrs to 1100hrs. On the contrary, tweeting activities decrease drastically from 1200hrs to 1700hrs before hitting a low between 1700hrs to 1800hrs. The minimum of tweeting activities is more pronounced for Com_{HICD} detected by our HICD method. For all three groups, tweeting activities gradually increase from 1800hrs to 2300hrs. As more than 65% of Twitter users are between the age of 15 - 24 years old [24], a possible explanation is that Twitter users are either at school or at work between 1200hrs and 1700hrs. Hence, they do not tweet as much during that period but tweeting activities gradually increase once they return home after school or work.

Another important area to examine is the relation between number of tweets posted by

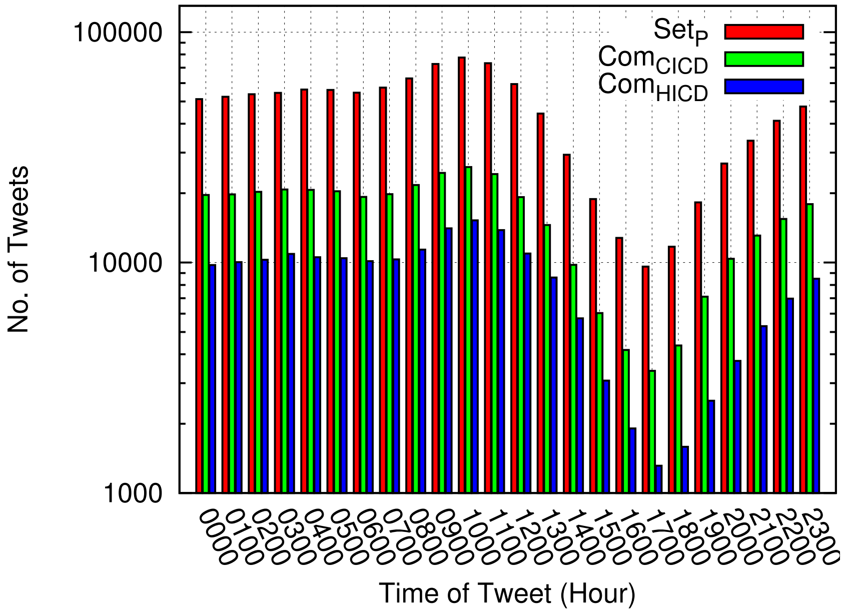


Figure 19: Time distribution of tweets

a user to his/her number of followers and followings. Fig. 20 and 21 show a scatterplot of the number of tweets to followers and followings, respectively. Both the CICD and HICD methods tend to select users (Com_{CICD} and Com_{HICD}) who have a high number of followers and followings, as shown in Fig. 20 and 21.

In addition, Fig. 20 and 21 also show that our HICD method tend to select users (Com_{HICD}) that tweet more often than users in Set $\mathcal{P}$ and Com_{CICD}. These results further support how our HICD method detects communities that are highly interactive and well-connected, based on their frequent tweets and high number of followers and followings.

6.4 Temporal Analysis of Link Creation and Deletion

Now, we study the formation and deletion of links over time for the three groups of users: Set $\mathcal{P}$, Com_{CICD} and Com_{HICD}. We retrieved the follower list of users in these groups on four-day intervals between 28 Nov 11 and 07 Jan 12. Thereafter, we study the number of links created and deleted at time intervals of four days. The results of the average number of links created and deleted at each time interval are shown in Fig. 22 and 23

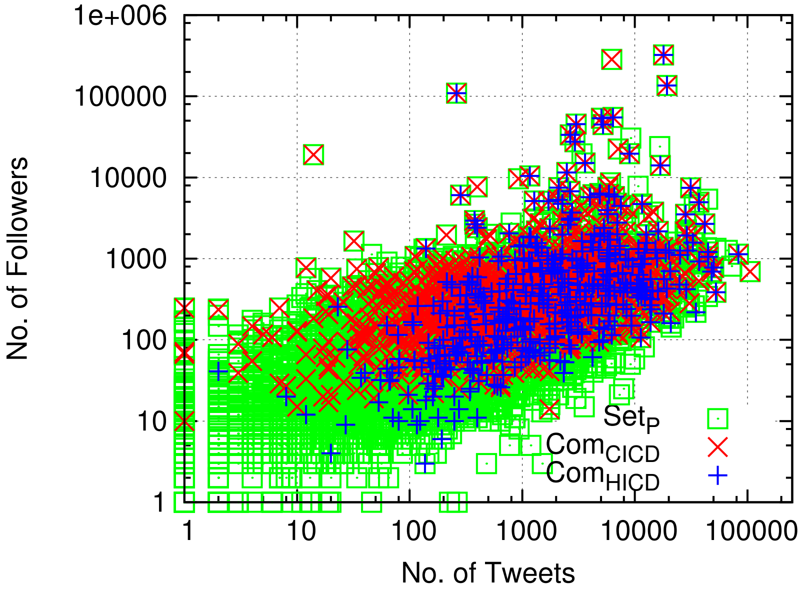


Figure 20: Comparison of tweets to followers

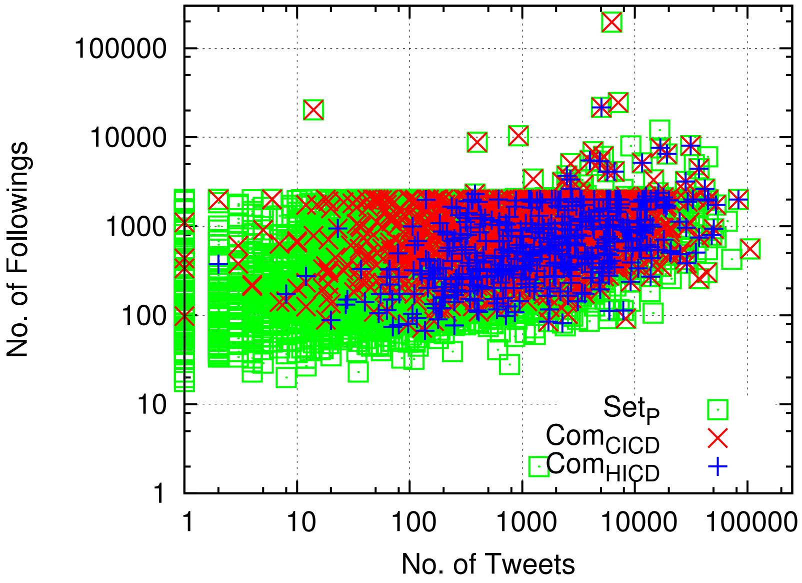


Figure 21: Comparison of tweets to followings

respectively.

Fig. 22 and 23 show that users selected by our HICD method are more active in following new users or unfollowing existing ones, compared to the CICD method. Following or unfollowing a user corresponds to creating or deleting a link to that user, respectively. Users in Com_{HICD} both create and delete more links on average than users in Set $\mathcal{P}$ and Com_{CICD}. It is interesting to note that Com_{HICD} creates almost three times the links that it deletes whereas Set $\mathcal{P}$ creates less than two times the links that it deletes. This observation points to a trend where links in Com_{HICD} are more persistent than those in Set $\mathcal{P}$ and Com_{CICD}, as users in Com_{HICD} are less likely to unfollow another user once the following link is created. This result serves as a preliminary analysis and we plan to further investigate on the motivating factors behind a user's choice in following/unfollowing other users (e.g. similar interests, common friends, etc).

6.5 Summary

In this chapter, we evaluated the HICD method for detecting highly interactive communities that are both topologically more cohesive and connected, and also frequently communicate about a specific interest. Our approach uses the frequency of direct tweets between users to construct a network of weighted links. Using these weighted links, we then detect the highly interactive communities based on a pre-determined threshold. In addition, we studied the topology and communications patterns among these users and showed that our approach detects communities that are more cohesive and connected, and communicate frequently about the specific interests based on the content of #hashtags and @mentions. Thus, given the availability of tweeting data, our HICD method would be more beneficial for targeted advertising and viral marketing compared to the CICD method.

As discussed in Chapter 3, the CICD method would be preferred if there is a restriction in data collection as this method only requires a single snapshot of topological information (i.e. follower/following links). On the other hand, the HICD method requires

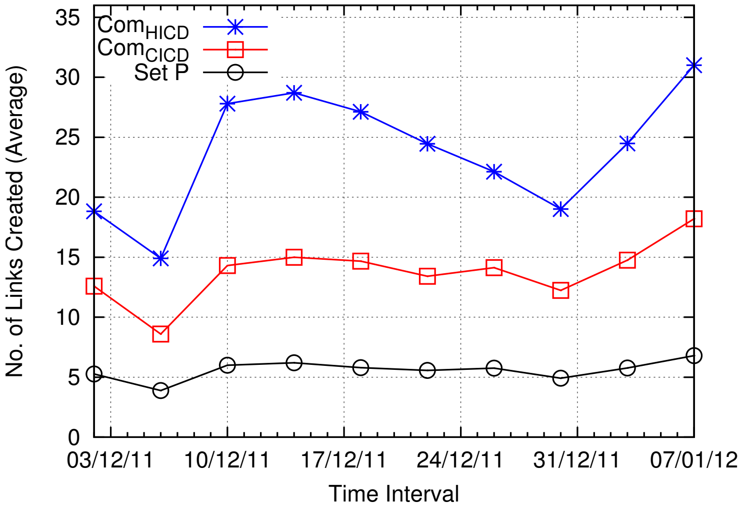


Figure 22: Time analysis of created links

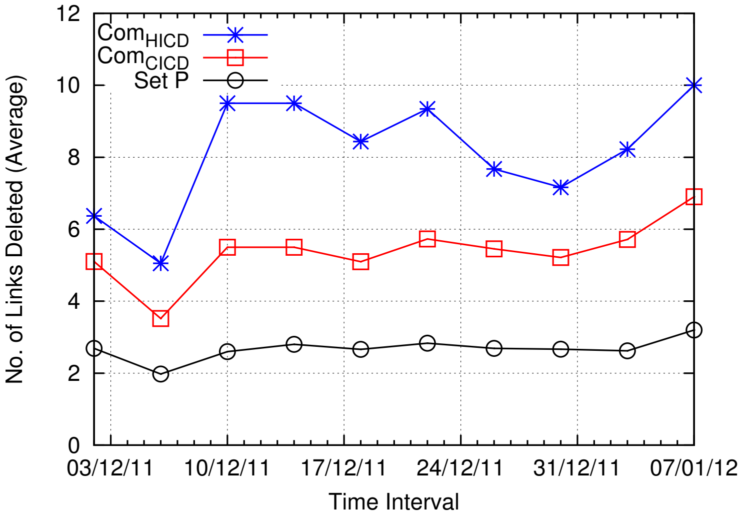


Figure 23: Time analysis of deleted links

multiple snapshots of communication data (i.e. tweeting links at specific time intervals) but results in highly interactive communities that frequently communicate about their common interest. While the communities detected by the CICD method may not be as interactive (about their common interest) as those detected by the HICD method, they are still relatively interactive compared to users in Set $\mathcal{P}$. The content of tweets (specifically the #hashtags and @mentions) posted by users in Com_{CICD} further supports this observation.

Our HICD method also presents an interesting perspective to community detection on Twitter where we build communities that may not reflect existing follower/following links. Instead, we detect communities using direct tweeting links between users. We also found that many tweeting links do not correspond to follower/following links and this may be indicative of real-life relationships where the users are geographically collocated and know each other personally. This observation is further supported by our study on user location which shows that many users reside in a geographic location that is closely affiliated with their common interest (e.g. Nashville for Country Music fans).

We also studied the trends and patterns in how people behave on Twitter, particularly in the way they tweet, follow and unfollow other users. We found trends in tweeting which reflect real-life working/schooling hours, where there is a reduction in tweeting activities from 1200hrs to 1700hrs. Our preliminary link analysis of Twitter users over time shows that users follow other users at a rate of two to three times as they unfollow other users. This finding presents an interesting area for future work on investigating the trends in how users follow/unfollow one another.

Chapter 7

Evaluating the Performance of Seed-centric Community Detection using an Expanding Ring Search[1]

In this chapter, we will evaluate our proposed seed-centric community detection algorithm and demonstrate its ability to find communities that correspond to their real-life counterparts. First, we validate our algorithm on three real-life social networks where we are able to compare the communities detected by our algorithm with that of their real-life communities. Next, we also evaluate our algorithm on a large-scale YouTube social network where we approximate real-life communities using the YouTube groups that users are part of. In addition, we further validate our results using topological measures of average clustering coefficient, average path length, average degree and diameter.

[1] This chapter is based on the following publication by Kwan Hui Lim and Amitava Datta: "A Seed-Centric Community Detection Algorithm based on an Expanding Ring Search", *in Proceedings of the 1st Australasian Web Conference (AWC'13)* [41].

7.1 Data and Methods

In order to validate the correctness of the communities detected by our algorithm, it is important to evaluate our community detection algorithm on social networks where we know the ground truth (i.e. the real-life communities). For this purpose, we selected the Zachary Karate Club, Doubtful Sound Dolphins and Santa Fe Institute Collaboration datasets which have also been used by many authors to establish the correctness of their community detection algorithms [22, 4, 10, 66]. These datasets are chosen as we know the ground truth of the actual real-life communities and can compare them to the results produced by our algorithm.

Next, we also evaluate our algorithm on a large-scale OSN based on the YouTube social network. The main challenge with evaluating community detection algorithms on OSNs is the verification of actual real-life communities (i.e. establishing the ground truth). In this case, we adopt the best approximation of ground truth by using the YouTube groups that the users belong to. Users who are members of the same YouTube group are inferred to be members of the same real-life community. In addition, we further validate our algorithm using topological measures such as average clustering coefficient, average path length, average degree and diameter as an evaluation of the topological structure of the detected communities.

7.1.1 Description of Datasets

The Zachary Karate Club dataset comprises 34 nodes which are further divided into two communities of 18 and 16 nodes respectively [64]. Over a period of three years, Zachary observed a university's karate club and noted the social relationships and interactions among its members. At one point, there was a disagreement between the club's president and instructor over the pricing of lessons. This disagreement eventually resulted in the instructor resigning and setting up his own private karate club thus effectively dividing the original club into two communities: those who stayed at the university's karate club; and those who joined his private karate club.

The Doubtful Sound Dolphin dataset comprises 62 nodes which are further divided into two communities of 42 and 20 nodes respectively [44, 43]. Lusseau et al. spent seven years at Doubtful Sound, New Zealand observing a school of bottlenose dolphins and their association with each other. A pair of dolphins are deemed to be associated with each other if they have been seen together frequently and in a non-coincidental manner. Subsequently, the disappearance of a few dolphins resulted in the school of dolphins being divided into two distinct communities.

The Santa Fe Institute Collaboration dataset comprises 118 nodes which are further divided into four communities of 26, 51, 24 and 17 nodes respectively [22]. These 118 nodes represent the scientists working at Santa Fe Institute and constitute the largest connected component of a complete set of 271 nodes (for the entire institute). A pair of scientists are assigned a link if they have co-authored a publication together. This largest component of 118 scientists is divided into their four fields of research, namely Structure of RNA, Statistical Physics, Mathematical Ecology and Agent-based Models.

The YouTube social network dataset comprises 1.1 million nodes that are joined by 2.9 million edges [56, 57]. A pair of YouTube users are connected by an edge if either of them sent a friend request that was approved by the other user. In addition, there exists 47 different YouTube groups that a YouTube user may be part of. A YouTube user may join a particular group if he/she produces or uploads videos which are related to that group. These groups are used as our approximation of the ground truth. This dataset is publicly available at http://socialcomputing.asu.edu/datasets/YouTube2.

7.2 Evaluation on Real-life Social Networks

We begin our evaluation on the three real-life social networks by first selecting the seed nodes for each social network. For the Zachary Karate Club, we chose the club president and instructor as the two respective seed nodes for our algorithm. These two seed nodes are also the nodes with the highest number of links. Similarly, for the Doubtful Sound Dolphins, we chose two nodes with the highest number of links in their respective communities as the seed nodes for our algorithm. Likewise for the Santa Fe Institute Collaboration Network, we selected one seed node from each field of research who also has one of the highest number of links to other scientists. It is important to re-iterate that our algorithm was not designed to detect all communities in a network but rather meaningful communities centered at an individual of interest. In this case, the choices of the seed nodes correspond to individuals of interest given their large degree of links and the social roles they play in their respective communities.

7.2.1 Overview of Results

We first evaluate the correctness of our algorithm by examining the precision and recall results on these three real-life social network datasets. Precision refers to the number of correct nodes classified out of all nodes classified while recall indicates the number of correct nodes classified out of all actual nodes in the community. The precision and recall results of our algorithm are as listed in Table 16.

Table 16: Precision and Recall

Dataset	**Precision**	**Recall**
Zachary Karate Club	84.9%	100%
Doubtful Sound Dolphins	97.6%	100%
Santa Fe Institute	87.2%	98.5%

On the whole, our algorithm performs well for the measurements of recall and precision. In terms of recall, our algorithm is able to detect almost all nodes ($\geq$98.5%) that belong to their respective communities. We were able to achieve 100% recall for both the Doubtful Sound Dolphins and Zachary Karate Club datasets. The recall rate for the Santa Fe Institute dataset was also high at 98.5%. Similarly, the results for precision are also relatively high with our algorithm correctly classifying 97.6%, 87.2% and 84.9% of nodes into their actual communities for the Doubtful Sound Dolphins, Santa Fe Institute and Zachary Karate Club datasets, respectively. While the results for precision are high, it is

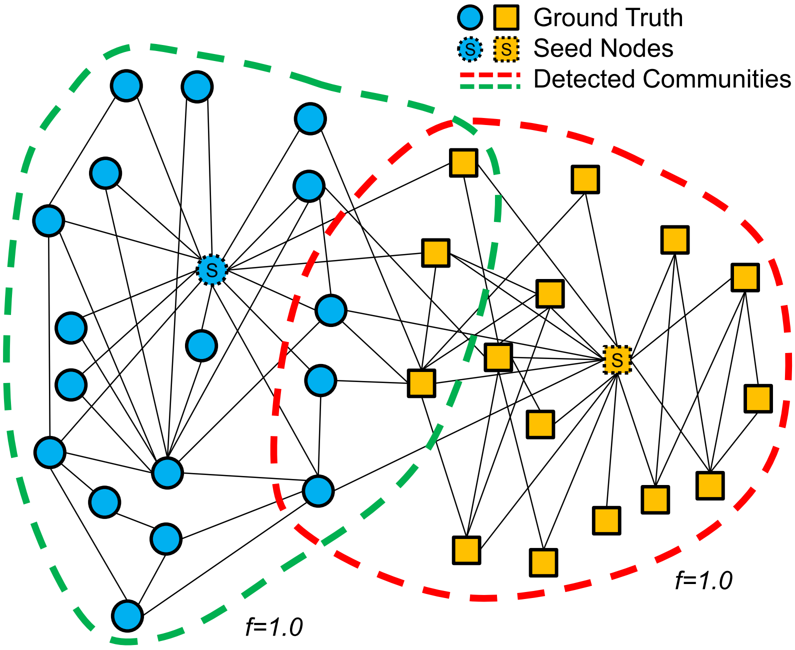


Figure 24: Zachary Karate Club

worthwhile to better understand the reasons behind the incorrect classification of the few nodes. We begin by examining these nodes and their position in the entire network.

7.2.2 Further Analysis of Results

We now analyze the Zachary Karate Club dataset where Fig. 24 shows the communities detected (circled with a dashed line) by our algorithm compared to the ground truth of actual communities (indicated by different shapes and colour of the nodes). For the six nodes that were mis-classified into the wrong community, four of them (a two-third majority) had direct links to both seed nodes in the two respective communities (i.e. the club president and instructor). The remaining two nodes were directly linked to the seed node in one community while being one-hop away from the seed node of the other community. This close proximity of the mis-classified nodes to the seed nodes of the two communities show that the mis-classified nodes actually act as effective bridges or middle-men between the two communities. As such, they would be better classified as members of both communities rather than just belonging to a single community.

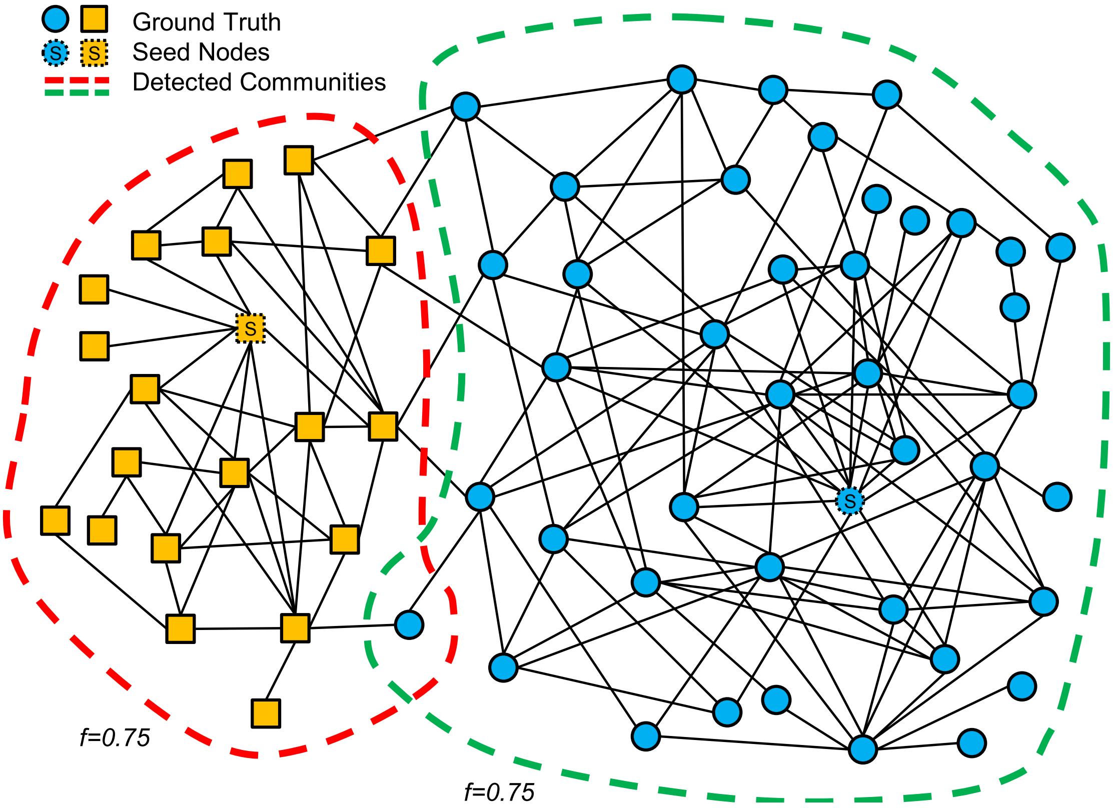


Figure 25: Doubtful Sound Dolphins

While the results are different from the ground truth, this is consistent with the observations of many authors that there are overlapping communities in social networks and individuals may belong to multiple communities [51, 35, 20]. Furthermore, in Zachary's study of the karate club, he also noted that "not all individuals in the network were solidly members of one faction or the other", thus further supporting the results of our algorithm [64]. Although the GN algorithm [22] performs better and only mis-classifies one node, it should be noted that the GN algorithm is designed to detect all communities in the network and is unable to detect overlapping communities. On the other hand, our proposed algorithm is able to detect such overlapping communities and specifically for communities that are centered at a particular seed node.

Similarly, the one mis-classified node for the Doubtful Sound Dolphins acts as such a bridge between the two communities. As shown in Fig. 25, this node is on the edge of both communities and have one link into each community. Hence, this node can

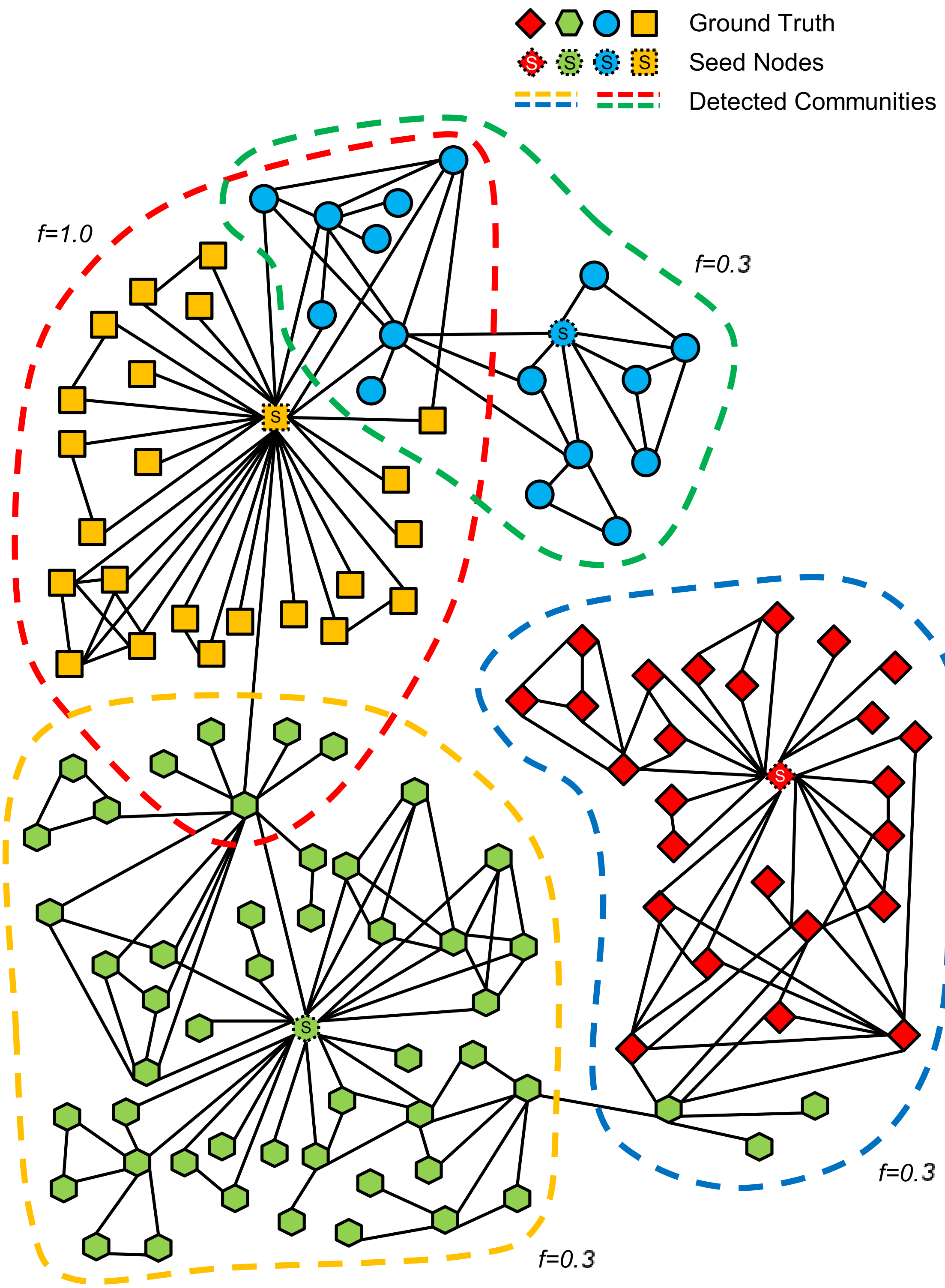


Figure 26: Santa Fe Institute Collaboration Network

easily belong to either community and would be better classified as belonging to both communities, considering its topological links and position in the network. Other authors also shared similar views that if a node has only a single link to a community, it should be classified as part of that community [22]. These results show that while the precision of our algorithm does not fully match that of the ground truth, the detected communities are reasonable and meaningful groupings of the nodes.

While we achieved a high recall rate of 98.5% for the Santa Fe Institute dataset, the unsuccessful 1.5% is attributed to three (green, hexagon) nodes being mis-classified, as shown in Fig. 26. These three nodes were classified as part of the diamond community while the ground truth dictates that they belong to the hexagon community. However, an analysis of the actual topological links implies that these three nodes are better suited as members of the diamond community. Specifically, one of these nodes has four links to the diamond community but only one link to the hexagon community (while the other two misclassified nodes have only one link to this node). Based on this topological analysis, these nodes would be better classified as part of the diamond community.

Similarly, while we achieved a relatively high precision rate of 87.2% for the Santa Fe Institute dataset, the unsuccessful 12.8% was largely due to the mis-classification of members of the hexagon and circle communities as part of the square community. Like the Zachary Karate Club dataset, almost half of these mis-classified (intermediate) nodes are directly linked to the seed node and thus should also be classified as part of the square community. For the remaining nodes, they are directly linked to these intermediate nodes and all of them have more links to these intermediate nodes than to other nodes. Therefore, they should also belong to the same community as these intermediate nodes (i.e. the square community).

7.3 Evaluation on YouTube Social Network

After evaluating our algorithm on three real-life social networks, we now evaluate it on the large-scale YouTube social network dataset. The main challenge in this evaluation is

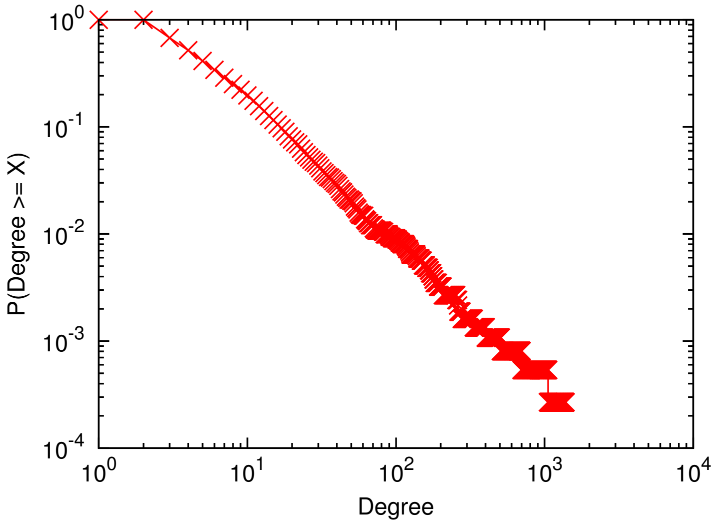


Figure 27: Degree Distribution of YouTube Users

the lack of an established ground truth of real-life communities, unlike the three real-life social networks previously evaluated. As such, we best approximate this ground truth using YouTube groups where users belonging to the same group are deemed to be in the same real-life community.

As YouTube groups are an approximation of the ground truth of real-life communities, we further evaluate the communities detected by our algorithm using topological measures of average clustering coefficient, average path length, average degree and diameter. These are suitable metrics for evaluation as communities display typical characteristics of a high clustering coefficient and average degree with low average path length and diameter, especially when compared to the overall network.

7.3.1 Experimental Setup

In the YouTube social network dataset, there exist users who do not join any YouTube groups. Since YouTube groups serve as ground truth for our evaluation, we consider only users who have joined at least one YouTube group. Based on this criterion, there are 22,693 users who have joined at least one YouTube group. This set of users will be used to evaluate our algorithm as we are able to compare the detected communities with the

Table 17: Network Statistics of YouTube Dataset

Network Property	**Detected Community**			**Control**
	Minimum	***Maximum***	***Average***	**Group**
No. of Nodes	701	7241	1676	22180
YouTube Group Overlap	67.8%	93.7%	78.0%	N.A.
Average Degree of Links	3.54	13.97	8.25	8.66
Average Clustering Coefficient	0.14	0.36	0.28	0.13
Average Path Length	2.14	3.53	2.82	4.08
Diameter	4	7	5.4	11

actual YouTube groups they belong to.

As a control group for comparing topological measures, we selected the largest connected component from this set of 22,693 users (who have joined at least one YouTube group). This largest connected component comprises 22,180 users and would be used as the control group to compare against the detected communities (of our algorithm) in terms of average clustering coefficient, average path length, average degree and diameter. An ideal community detection algorithm would detect communities that exhibit a higher clustering coefficient, and shorter average path length and diameter compared to the overall network (i.e. the control group).

Similar to the selection of seed nodes for the three real-life social networks, we selected seed nodes for the YouTube dataset based on users with a high number of links. This selection criterion corresponds to the aim of our algorithm which is to detect communities centered at individuals of interest, such as influential or well-connected individuals. Fig. 27 shows the link degree distribution among all the users who have joined at least one YouTube group.

We first identify a set of users that is in the top 1% of the dataset, in terms of the number of links in this set. From this set of users, we selected 10 users as the seed nodes for our algorithm. Using our algorithm, we then attempt to detect communities centered at each of these 10 users followed by calculating the topological measures of the resulting

10 communities. In particular, we compare the average topological measures of these communities against that of the control group. Using the average result (from these 10 communities) avoids the effect of random or outlier results that may be unique to any particular community.

7.3.2 Comparison of Network Statistics

Table 17 shows the (minimum, maximum and average) topological measures of our detected communities compared to that of the control group. The YouTube group overlap measures how many other users in the detected community belong to the same YouTube group as the seed user. The high average result of 78% show that our algorithm is able to accurately detect communities where most of its users belong to the same YouTube group (as the seed user), an approximation of their real-life communities.

The YouTube group overlap result is not 100% due to the unique nature of YouTube groups where users who join such groups are producers/uploaders of videos related to that group. On the contrary, there are users who are only interested in viewing such videos but do not produce/upload videos. These users simply become friends with members of such groups and are able to be alerted about their new videos without having to join their YouTube groups. Even with such users, our algorithm is able to detect communities that are up to 93.7% accurate compared to the real-life communities

Despite the small average size of the detected communities, the average degree of links of these communities is very similar to that of the control group (differing only by 4.7%). This result shows that the detected communities comprise users who are well-connected among themselves, as indicated by a high average degree of links, despite having an average community size that is less than 8% of the control group. This well-connectedness is another network property of a strong community, where its users have more links within the community than outside of the community [53].

In addition to being well-connected, the detected communities are also highly cohesive based on an average clustering coefficient that is two times higher than that of the control group. Another observation is the lower average path length and diameter of the detected communities compared to that of the control group. A lower average path length and diameter means that nodes within these communities are able to reach each other in a smaller number of steps, which is also an indication of a cohesive and well-connected community.

Based on our approximation of ground truth using YouTube groups, our proposed algorithm is able to detect communities that closely resemble real-life communities (up to 93.7%). The topological measures of these detected communities further illustrate the effectiveness of our algorithm. Specifically, the high clustering coefficient and average degree of links, and low average path length and diameter (of the detected community) indicate that our algorithm detects communities which are highly cohesive and well-connected, especially when compared to the control group.

7.4 Summary

In this chapter, we evaluated our seed-centric community detection algorithm for finding a community centered at an individual of interest, using an expanding ring search starting from this individual. We first used three real-life social networks (Zachary Karate Club, Doubtful Sound Dolphins and Santa Fe Institute) to compare the detected communities to the ground truth of actual real-life communities. The results show that our algorithm is able to detect communities (that strongly correspond to their real-life counterparts) at a high level of precision and recall rate of up to 97.6% and 100% respectively. Further analysis also shows that while some nodes are “incorrectly” classified (according to the ground truth), the detected communities are still reasonable and meaningful grouping of these nodes (based on their topological links).

We also conducted experiments on a large-scale YouTube social network to evaluate our algorithm. The results show that our algorithm is able to detect communities that closely resemble real-life communities (based on YouTube groups), up to an accuracy of 93.7%.

Our evaluation using topological measures of average clustering coefficient, average degree of links, average path length, average degree and diameter also indicates that the detected communities are highly cohesive and well-connected. The high clustering coefficient and average degree of links, and low average path length and diameter (of the detected community) further support that our algorithm is able to detect communities that reflect the topological structure and characteristics of real-life communities.

Chapter 8

Conclusion

In conclusion, we will now summarize our key contributions in this thesis which include the proposal of four algorithms and their corresponding evaluation. In addition, we also discuss some of the limitations of our proposed algorithms and suggest future research directions in the areas of community detection and social network analysis.

8.1 Summary of Contributions

In this thesis, we have proposed four algorithms for various purposes such as identifying users with common interest, detecting like-minded communities and detecting communities centered at individuals of interest. Also, we have performed a comprehensive set of experiments to evaluate the effectiveness of these algorithms. In summary, our key contributions are as follows:

- We proposed a selection algorithm for identifying users with common interest (Chapter 3). As the users do not explicitly state their interest, we infer this interest based on their following of celebrities that are representative of the various interest categories. In addition, we can also adjust the sensitivity of this selection algorithm based on the interest level inferred by the number of celebrities (of each interest category) followed by the users.

- We also proposed the CICD method that uses topological information for detecting like-minded communities comprising users with common interests (Chapter 3). This method is able to detect such communities using only a single snapshot of the topological links among users in the network and is suitable when there is a constraint on data collection (e.g. API calls limit).

- Similarly, we proposed the HICD method that uses communication information for detecting highly interactive communities where its members frequently communicate about their common interests (Chapter 3). While this method is able to detect such highly interactive communities, its effectiveness is dependent on multiple snapshots of communication data (i.e. messages sent at specific time intervals).

- Also, we proposed a seed-centric community detection algorithm for finding the community centered at an individual of interest (Chapter 4). This algorithm utilizes an expanding ring search where we iteratively include users into the community at increasing number of hops. The criteria for including such users are based on our proposed definition of a community, derived from the number of internal and external links of a user coupled with an adjustable strength factor.

- Next, we evaluated our CICD method on the Twitter social network and showed that our method detects larger and more communities, which are also topologically cohesive and well-connected (Chapter 5). We also studied Twitter communities that share common interests and found that the deepening and specialization of interests result in these communities becoming more cohesive and well-connected.

- We also performed an evaluation of the HICD method on the Twitter social network using both topological and interaction measures (Chapter 6). We observed that the detected communities are not only cohesive and well-connected, the members of such communities also frequently communicate about their common interests, based on the frequency and content (#hashtags and @mentions) of their tweets. Our preliminary study on the temporal evolution of links also show that links among the detected communities are more persistent, and these users create more new links than delete existing one.

- Finally, we evaluated our seed-centric community detection algorithm on three real-life social networks and the large-scale YouTube social network (Chapter 7). The results show that the detected communities strongly correspond to their real-life counterparts on the respective social networks. In addition, we further evaluated our algorithm on the YouTube social network using topological measures and found that the detected communities also display topological characteristics which strongly resemble real-life communities, such as a high clustering coefficient and average degree of links, and low average path length and diameter.

8.2 Limitations

The key limitation of our selection algorithm (for identifying users with common interest) is with the correct classification of celebrities into their interest categories. This classification task could be very subjective and its accuracy affects the set of users (with common interest) being identified. As such, we try to limit this subjective judgment using information on Wikipedia, specifically the "occupation" field and description of these celebrities. However, this still requires a certain amount of judgment (albeit significantly less) to associate each "occupation" with its corresponding interest category.

The CICD method uses topological links to detect like-minded communities which may not necessarily correspond to communities that are highly interactive. However, the CICD method requires just a single snapshot of the topological structure of the OSN and is ideal when there are data collection constraints (e.g. API calls limit). Furthermore, we observe that the detected communities are still fairly interactive based on the frequency that community members communicate about their common interest. Our HICD method further improves upon the CICD method by using communication links to detect highly interactive communities that frequently communicate about their common interest.

However, the key limitation of our HICD method is in the use of communication links to detect such highly interactive communities. The collection of communication (links) data is an intensive process as we have to retrieve this communication data at multiple time intervals (i.e. multiple snapshots instead of only a single snapshot of topological links like the CICD method). While this data collection may be intensive, the resulting communities are highly interactive, cohesive and well-connected. Given that our CICD and HICD methods overcome the limitation of each other, the choice of a method should be based on data collection constraints or the requirement for highly interactive communities. For example, the CICD method should be used when data collection is an issue and the HICD method is preferred when the interactivity of the detected communities is more important than constraints in data collection.

8.3 Future Directions

One area for future work is to further improve our selection algorithm (for identifying users with common interest). This future work would involve creating an automated system where the process of selecting and classifying celebrities into their respective categories is automated using information from Wikipedia. This system would be pre-loaded with a library of keywords and their associated interest categories. Thereafter, the system can automatically grab keywords used in the celebrities' Wikipedia articles and automatically classify them into their respective interest categories based on the pre-loaded library of keywords. This automated process would overcome our method's main limitation of needing to manually select and classify celebrities into their respective categories.

Another possible area for future work is to perform further studies into the temporal evolution of links among users on OSNs. This study would involve examining the correlation between communication frequency with the formation of links. The key idea is to propose a model for predicting the formation of links based on the communication patterns between two individuals and subsequently, allowing us to study how and why links are formed within communities. In addition, we would study if there exists different patterns in the formation/deletion of links among users with different interests.

Bibliography

[1] Lada Adamic, Orkut Buyukkokten and Eytan Adar. A social network caught in the web. *First Monday*, Volume 8, Number 6, Jun 2003.

[2] Reid Andersen, Fan Chung and Kevin Lang. Local graph partitioning using pagerank vectors. In *FOCS '06: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 475–486, Oct 2006.

[3] Reid Andersen and Kevin J. Lang. Communities from seed sets. In *WWW '06: Proceedings of the 15th International Conference on World Wide Web*, pages 223–232, May 2006.

[4] Alexandre Arenas, Alberto Fernández and Sergio Gómez. Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics*, Volume 10, Number 5, pages 053039, May 2008.

[5] Martin Atzmueller and Folke Mitzlaff. Efficient descriptive community mining. In *FLAIRS '11: Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference*, pages 459–464, May 2011.

[6] Hemant Balakrishnan and Narsingh Deo. Discovering communities in complex networks. In *ACMSE '06: Proceedings of the 44th Annual Southeast Regional Conference*, pages 280–285, Mar 2006.

[7] Hila Becker, Mor Naaman and Luis Gravano. Beyond trending topics: Real-world event identification on twitter. In *ICWSM '11: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, pages 438–441, May 2011.

[8] Engineering Blog. The engineering behind twitters new search experience. Internet, Jul 2012. Available from: http://engineering.twitter.com/2011/05/engineering-behind-twitters-new-search.html.

[9] YouTube Blog. Holy nyans! 60 hours per minute and 4 billion views a day on youtube. Internet, Sep 2012. Available from: http://youtube-global.blogspot.com.au/2012/01/holy-nyans-60-hours-per-minute-and-4.html.

[10] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Grke, Martin Hoefer, Zoran Nikoloski and Dorothea Wagner. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, Volume 20, Number 2, pages 172–188, Feb 2008.

[11] Fred Brauer. An introduction to networks in epidemic modeling. In *Mathematical Epidemiology*, Volume 1945 of *Lecture Notes in Mathematics*, pages 133–146. Springer Berlin / Heidelberg, 2008.

[12] Carlos Castillo, Marcelo Mendoza and Barbara Poblete. Information credibility on twitter. In *WWW '11: Proceedings of the 20th International Conference on World Wide Web*, pages 675–684, Mar 2011.

[13] Meeyoung Cha, Hamed Haddadi, Fabrcio Benevenuto and Krishna P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM '10: Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 10–17, May 2010.

[14] Meeyoung Cha, Alan Mislove, Ben Adams and Krishna P. Gummadi. Characterizing social cascades in Flickr. In *WOSN '08: Proceedings of the First Workshop on Online Social Networks*, pages 13–18, Aug 2008.

[15] Hyunwoo Chun, Haewoon Kwak, Young-Ho Eom, Yong-Yeol Ahn, Sue Moon and Hawoong Jeong. Comparison of online social relations in volume vs interaction: a case study of cyworld. In *IMC'08: Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement*, pages 57–70, Oct 2008.

[16] Aaron Clauset. Finding local community structure in networks. *Physical Review E*, Volume 72, Number 2, pages 026132, Aug 2005.

[17] ComScore. It's a social world: Top 10 need-to-knows about social networking and where its headed. Internet, Dec 2011. Available from: http://www.comscore.com/Insights/Presentations_and_Whitepapers/2011/it_is_a _social_world_top_10_need-to-knows_about_social_networking.

[18] Nan Du, Bin Wu, Xin Pei, Bai Wang and Liutong Xu. Community detection in large-scale social networks. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, pages 16–25, Aug 2007.

[19] Timothy La Fond and Jennifer Neville. Randomization tests for distinguishing social influence and homophily effects. In *WWW '10: Proceedings of the 19th International Conference on World Wide Web*, pages 601–610, Apr 2010.

[20] Santo Fortunato. Community detection in graphs. *Physics Reports*, Volume 486, Number 3-5, pages 75–174, 2010.

[21] Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic and Wolfgang Kellerer. Outtweeting the Twitterers - Predicting information cascades in microblogs. In *WOSN '10: Proceedings of the 3rd International Workshop on Online Social Networks*.

[22] Michelle Girvan and Mark E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, Volume 99, Number 12, pages 7821–7826, Jun 2002.

[23] Jeff Huang, Katherine M. Thornton and Efthimis N. Efthimiadis. Conversational tagging in twitter. In *HT '10: Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, pages 1079–1088, Jun 2010.

[24] Sysomos Inc. Inside twitter: An in-depth look inside the twitter world. Internet, Jun 2009. Available from: http://www.sysomos.com/docs/Inside-Twitter-BySysomos.pdf.

[25] Twitter Inc. Twitter API. Internet, Sep 2011. Available from: https://dev.twitter.com.

[26] Ganesh Iyer, David Soberman and J. Miguel Villas-Boas. The targeting of advertising. *Marketing Science*, Volume 24, Number 3, pages 461–476, 2005.

[27] Akshay Java, Xiaodan Song, Tim Finin and Belle Tseng. Why we twitter: Understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, pages 56–65, Aug 2007.

[28] Xin Jin, Chi Wang, Jiebo Luo, Xiao Yu and Jiawei Han. Likeminer: A system for mining the power of 'like' in social media networks. In *KDD '11: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 753–756, Aug 2011.

[29] Andreas M. Kaplan and Michael Haenlein. Two hearts in three-quarter time: How to waltz the social media/viral marketing dance. *Business Horizons*, Volume 54, pages 253–263, 2011.

[30] Mohit Naresh Kewalramani. Community Detection in Twitter. Master's thesis, University of Maryland, 2011.

[31] Efthymios Kouloumpis, Theresa Wilson and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg! In *ICWSM '11: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, pages 538–541, May 2011.

[32] Haewoon Kwak, Changhyun Lee, Hosung Park and Sue Moon. Twitter dataset. Internet, Jun 2009. Available from: http://an.kaist.ac.kr/traces/WWW2010.html.

[33] Haewoon Kwak, Changhyun Lee, Hosung Park and Sue Moon. What is twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th International Conference on World Wide Web*, pages 591–600, Apr 2010.

[34] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: A comparative analysis. *Physical Review E*, Volume 80, Number 5, pages 056117, Nov 2009.

[35] Andrea Lancichinetti, Santo Fortunato and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, Volume 11, Number 3, pages 033015, Mar 2009.

[36] Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *WWW '08: Proceedings of the 17th International Conference on World Wide Web*, pages 915–924, Apr 2008.

[37] Daifeng Li, Bing He, Ying Ding, Jie Tang, Cassidy Sugimoto, Zheng Qin, Erjia Yan, Juanzi Li and Tianxi Dong. Community-based topic modeling for social tagging. In *CIKM '10: Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1565–1568, Oct 2010.

[38] Kwan Hui Lim and Amitava Datta. Finding Twitter communities with common interests using following links of celebrities. In *MSM '12: Proceedings of the 3rd International Workshop on Modeling Social Media*, pages 25–32, Jun 2012.

[39] Kwan Hui Lim and Amitava Datta. Following the follower: Detecting communities with common interests on Twitter. In *HT '12: Proceedings of the 23th ACM Conference on Hypertext and Social Media*, pages 317–318, Jun 2012.

[40] Kwan Hui Lim and Amitava Datta. Tweets beget propinquity: Detecting highly interactive communities on twitter using tweeting links. In *WI '12: Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 214–221, Dec 2012.

[41] Kwan Hui Lim and Amitava Datta. A seed-centric community detection algorithm based on an expanding ring search. In *AWC '13: Proceedings of the 1st Australasian Web Conference*, pages 21–26, Jan 2013.

[42] Feng Luo, James Z. Wang and Eric Promislow. Exploring local community structures in large networks. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 233–239, Dec 2006.

[43] David Lusseau and Mark E. J. Newman. Identifying the role that animals play in their social networks. *Proceedings of the Royal Society London B (Supplementary)*, Volume 271, pages S477–S481, 2004.

[44] David Lusseau, Karsten Schneider, Oliver J. Boisseau, Patti Haase, Elisabeth Slooten and Steve M. Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting association. can geographic isolation explain this unique trait? *Behavioral Ecology and Sociobiology*, Volume 54, Number 4, pages 396–405, 2003.

[45] Sofus A. Macskassy and Matthew Michelson. Why do people retweet? anti-homophily wins the day! In *ICWSM '11: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, pages 209–216, May 2011.

[46] Merriam-Webster. Merriam-webster dictionary and thesaurus. Internet, Oct 2011. Available from: http://www.merriam-webster.com/dictionary/community.

[47] Stanley Milgram. The small world problem. *Psychology Today*, Volume 2, pages 60–67, 1967.

[48] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *IMC '07: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pages 29–42, Oct 2007.

[49] Bimal Viswanathand Alan Mislove, Meeyoung Cha and Krishna P. Gummadi. On the evolution of user interaction in Facebook. In *WOSN '09: Proceedings of the 2nd ACM Workshop on Online Social Networks*, pages 37–42, Aug 2009.

[50] Katarzyna Musial, Marcin Budka and Krzysztof Juszczyszyn. Creation and growth of online social network. *World Wide Web*, Volume 16, Number 4, pages 421–447, 2013.

[51] Gergely Palla, Imre Derényi, Illés Farkas and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, Volume 435, pages 814–818, Jun 2005.

[52] Barbara Poblete, Ruth Garcia, Marcelo Mendoza and Alejandro Jaimes. Do all birds tweet the same? characterizing twitter around the world. In *CIKM '11: Proceedings of the 20th ACM Conference on Information and Knowledge Management*, pages 1025–1030, Oct 2011.

[53] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto and Domenico Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*, Volume 101, Number 9, pages 2658–2663, Mar 2004.

[54] Daniel M. Romero, Brendan Meeder and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *WWW '11: Proceedings of the 20th International Conference on World Wide Web*, pages 695–704, Mar 2011.

[55] Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, Volume 105, Number 4, pages 1118–1123, 2008.

[56] Lei Tang and Huan Liu. Relational learning via latent social dimensions. In *KDD '09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 817–826, Jun 2009.

[57] Lei Tang and Huan Liu. Scalable learning of collective behavior based on sparse social dimensions. In *CIKM '09: Proceedings of the 18th ACM International Conference on Information and Knowledge Management*, pages 1107–1116, Nov 2009.

[58] All Twitter. Twitter to surpass 500 million registered users on wednesday. Internet, Jul 2012. Available from: http://www.mediabistro.com/alltwitter/500-million-registered-users_b18842.

[59] Clifford Weinstein, William Campbell, Brian Delaney and Gerald OLeary. Modeling and detection techniques for counter-terror social network analysis and intent recognition. In *2009 IEEE Aerospace Conference*, pages 1–16, Mar 2009.

[60] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P. N. Puttaswamy and Ben Y. Zhao. User interactions in social networks and their implications. In *EuroSys'09: Proceedings of the 4th ACM European Conference on Computer Systems*, pages 205–218, Apr 2009.

[61] Jiang Yang and Scott Counts. Predicting the speed, scale, and range of information diffusion in twitter. In *ICWSM '10: Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 355–358, May 2010.

[62] Shuang Hong Yang, Bo Long, Alex Smola, Narayanan Sadagopan, Zhaohui Zheng and Hongyuan Zha. Like like alike - joint friendship and interest propagation in social networks. In *WWW '11: Proceedings of the 20th International Conference on World Wide Web*, pages 537–546, Mar 2011.

[63] Zi Yang, Jingyi Guo, Keke Cai, Jie Tang, Juanzi Li, Li Zhang and Zhong Su. Understanding retweeting behaviors in social networks. In *CIKM '10: Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1633–1636, Oct 2010.

[64] Wayne W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, Volume 33, Number 4, pages 452–473, 1977.

[65] Dejin Zhao and Mary Beth Rosson. How and why people twitter: the role that micro-blogging plays in informal communication at work. In *GROUP ’09: Proceedings of the ACM 2009 International Conference on Supporting Group Work*, pages 243–252, May 2009.

[66] Kun Zhao, Shao-Wu Zhang and Quan Pan. Fuzzy analysis for overlapping community structure of complex network. In *2010 Chinese Control and Decision Conference*, pages 3976–3981, May 2010.