A Topological Approach for Detecting Twitter Communities with Common Interests

Kwan Hui Lim and Amitava Datta

School of Computer Science and Software Engineering
The University of Western Australia
Crawley, WA 6009, Australia
kwanhui@graduate.uwa.edu.au, datta@csse.uwa.edu.au

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 methods involve large scale community detection on the entire social network before determining the interests of individuals within these communities. This approach is both computationally intensive and may result in communities without a common interest. We propose an efficient topological-based approach for detecting communities that share common interests on Twitter. Our approach involves first identifying celebrities that are representative of an interest category before detecting communities based on linkages among followers of these celebrities. We also study the network characteristics and tweeting behaviour of these communities, and the effects of deepening or specialization of interest on their community structures. In particular, our evaluation on Twitter shows that these detected communities comprise members who are well-connected, cohesive and tweet about their common interest.

Keywords: Twitter, Social Network Analysis, Community Detection

1 Introduction

Twitter is a popular micro-blogging platform that allows short messages of up to 140 characters (called tweets) to be posted and received by registered users. The popularity of Twitter is seen from its social network comprising 500 million users who produce 2,200 tweets per second [1, 2]. The popularity of Twitter and availability of data have created plenty of interest in its academic study in recent years [3–5]. In particular, this large user base and high activity level provide tremendous opportunities for companies to effectively reach out to a large group of potential customers.

One key consideration for such companies applying targeted advertising and viral marketing to online social networks is the efficient identification of communities with common interests in large social networks [6, 7]. These communities would serve as potential target audience, given their common interest (in the specific product/service). However, most of the current approaches involve first

detecting all communities, followed by determining the interests of these communities [8, 9]. These approaches involve a lengthy and intensive process of detecting communities for the entire social network, which is growing daily. Furthermore, many of the detected communities may not share the interest we are looking for.

Our study offers a method to identify communities comprising like-minded individuals with common interests on Twitter. This method differs from existing ones that first detect all communities, followed by identifying the topics they are interested in [8, 9]. Also, our method does not unnecessarily detect communities that do not share any specific interest. Instead, our method allows for the efficient detection of only communities sharing a common interest and can be applied to targeted advertising and viral marketing (for identifying a target audience). In addition, our method is able to detect communities at different levels of interest. While there have been recent studies on detecting communities with common interest [10–12], these are interaction-based methods which use tweeting behaviour between users. On the other hand, we propose a topological-based method that uses topology links between users which are easier to collect (than the large volume of tweeting data), and also allow us to detect communities with common interest even if the users are not active in tweeting [13, 14]. Our main contributions in this chapter include the following:

- An efficient topological-based approach for detecting Twitter communities comprising users that share common interests.
- A study of the network characteristics and tweeting behaviour of Twitter communities that share common interests.
- An investigation into the effects of deepening or specialization of interest on these communities.

This chapter is structured as follows: Section 2 covers background information on Twitter; Section 3 discusses related work in the field; Section 4 describes our data and methods; Section 5 highlights our findings on community detection based on common interests; Section 6 investigates the effects of deepening or specialization of interest on these communities; and Section 7 summarizes and concludes the chapter.

2 Description of Twitter

Twitter allows registered users to post and receive short messages of up to 140 characters. 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).

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 [15]. This API allows developers to create applications for Twitter and researchers to study the characteristics of an online social network from the individual to community level. Currently, there is a rate limit on the number of API calls that can be executed within a specific time interval.

3 Related Work

Social networks have been intensively studied in recent years due to the availability and scale of online social networking sites. As our proposed approach aims to detect entire communities comprising users with common interests, we first discuss some related work on modeling and detecting user interests on online social networks. Next, we proceed to describe some proposed methods for detecting communities with common interests, which can be further divided into topological-based and interaction-based methods. The topological-based methods also include tag-based approaches that utilizes the tagging behaviour of users on various items to build a network graph for community detection.

3.1 User Interest Detection

One such study on user interest detection resulted in the LikeMiner system which identifies popular topics on online social networks based on the explicit "likes" indicated by users [16]. 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. Our approach differs from this system as we infer interest based on a user's followings instead of requiring the explicit "like" by a user. More importantly, the LikeMiner system identifies individuals whereas our approach identifies communities with common interests.

Similarly, the Friendship and Interest Propagation (FIP) model identifies interests of an individual and potential friendship links with other users [17]. 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. The main difference with our method is that we identify an entire community sharing a common interest whereas the FIP model identifies an individual user's interest and recommends friendships. Also, this study was conducted on Yahoo! Pulse (pulse.yahoo.com) whereas ours is based on Twitter. Furthermore, interests are explicitly stated for the FIP model whereas our model implicitly infers interests based on a user's followings.

3.2 Topological-based Community Detection

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 Clique Percolation Method (CPM) to detect overlapping communities on the Twitter social network [8]. 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 interests, 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. The difference with our approach is that we do not detect all communities then determine their interest but rather, focus directly only on communities sharing specific interests that we are interested in.

Li et. al. proposed the TTR-LDA community detection algorithm using the Latent Dirichlet Allocation model and Girvan-Newman algorithm with an inference mechanism for topic distribution [9]. They used the TTR-LDA algorithm to first detect all communities among the top 50,000 taggers in Delicious (delicious.com), 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. Similar to Java et. al., they detect all communities first before determining their interest. Also, their data is based on only the top users of Delicious whereas ours is based on the full dataset of Twitter.

Using BibSonomy (www.bibsonomy.org), Atzmueller and Mitzlaff demonstrated an approach for mining communities with common descriptive features [18]. 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 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. While this approach 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 infers 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.

3.3 Interaction-based Community Detection

The Highly Interactive Community Detection (HICD) method is an interaction-based approach for detecting communities where its members share a common interest and frequently interact with each other regarding this interest [10]. The

HICD method uses interaction (tweeting) links to build such communities based on a threshold for their communication frequency. In the same spirit, Correa et. al. developed the iTop algorithm which uses interactions between Twitter users (@mentions and retweets) to detect topic-centric communities [11]. The iTop algorithm models the social network as a weighted graph and tries to detect the topic-centric community from this weighted graph based on the greedy maximization of local modularity. While the HICD method and iTop algorithm are able to detect interactive communities, the collection of such interaction data is a potentially time-consuming and tedious process (requiring the consistent monitoring of messages sent among such users). Our proposed approach differ mainly in the use of topological links (instead of interaction links), which is preferable when there are data collection constraints such as API call limits.

Similarly, Palsetia et. al. used interactions among Facebook and Twitter users for detecting communities comprising users with the same social interest [12]. The form of interactions used are wall posts on Facebook and tweets that mention specific Twitter users. The authors then model an undirected graph with these interactions as links and assign weights to the links based on a similarity coefficient between two users sharing a common link. Next, a modified version of the Clauset, Newman, and Moore (CNM) algorithm [19] is used in a recursive fashion to detect communities from the earlier constructed graph. Like the HICD method and iTop algorithm, this algorithm uses interaction links whereas our proposed approach uses topological links, which are smaller in volume (than interaction data) and easier to collect, especially with the stringent API call limits imposed on many online social networking sites.

4 Data and Methods

The Twitter dataset collected by Kwak et. al. [4] 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 [20]. We also used the Twitter API to collect the profiles and tweets of users whom belong to either the control group or communities with common interest (as detected by our proposed approach).

We model the Twitter social network as a directed graph, $G = (U, L)$ where U refers to the set of users and L refers to the set of links. 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. Also, we can adjust this required number of followers to select celebrities at varying levels of popularity.

4.1 Proposed Method

Our proposed approach for detecting communities with common interest involves the following steps:

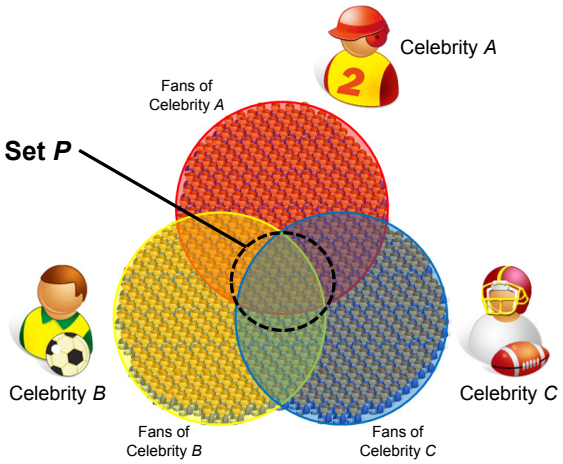

Fig. 1. Illustration of Set P

1. For a specific interest category, select a set of celebrities that represents this particular category.
2. Based on the set of identified celebrities, select the users who follow all of these celebrities.
3. Retrieve the topology links among these users and detect communities among them.

Step 1: Representing Interest using Celebrities. Our first step is to identify a set of celebrities that represents an interest category cat, Int_{cat} and we infer the interest of a user in this category based on 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 (en.wikipedia.org) to classify these celebrities into their respective categories.

As described in [21], this process can be automated by utilizing a keyword-to-interest mapping on keywords used in either the "occupation" field or main (textual) description of the celebrity's Wikipedia article page. In particular, this keyword-to-interest mapping uses a library of 179 keywords and Table 1 gives an example of which keywords are mapped to which interest categories. This automated interest classification of a celebrity has been evaluated on a group of 1,000 celebrities with an accuracy of 83.9%. This automated process allows us to overcome the need to manually classify celebrities into their respective categories. Coupled with a secondary (manual) verification of the automated

Table 1. Example of Keyword to Interest Category Mapping

Interest	**Keywords**
Business	entrepreneur, founder, chairman, owner, etc
Fashion	fashion designer, model, clothing designer, etc
Film & TV	actor, actress, film producer, movie director, etc
Music	singer, songwriter, dancer, band, composer, etc
Publishing	writer, author, columnist, novelist, etc

classification, we can further minimize the chances of classifying celebrities into the wrong category.

Step 2: Identifying Users with Common Interests. Our next step is to retrieve the set of Twitter users who follow all celebrities in a given category. Suppose we identify a set of n celebrities $c_1, c_2, ..., c_n$. We next identify all the followership links for the individual celebrities 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 illustration of Set $\mathcal{P}$, which (in this case) are fans who follow all three sports celebrities.

Step 3: Detecting Communities using Topology Links. For the next step of community detection, 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 $\mathcal{P}$ next using the CPM algorithm developed by Palla et. al. [22]. 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. We 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 series of communities formed based on the k-cliques and adjacency criteria. 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 [23]. 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 walks 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. 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, particularly in the selection of nodes that constitute the detected communities.

4.2 Experiments and Evaluation

We first study community detection and structure among individuals with a common interest in Section 5. We infer the interest of users based on the celebrities followed 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.[1] This control group allows us to compare the community structure of users with no common interest against users with a shared interest.

Next, we further examine how the deepening and specialization of interest affects community structure in Section 6. 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 [24] 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 [4] 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.

In our experiments, we measure the effectiveness of our proposed approach using the metrics of reciprocity, clustering coefficient, average path length, average degree and diameter. Reciprocity is defined as the number of friendship (bi-directional) links out of the total number of links. The clustering coefficient of a node is based on the number of triangular sub-graph that includes this node, out of all possible triangular sub-graphs. As we are interested in measuring the (entire) detected community, we take the average clustering coefficient of all nodes. In addition, we also measure the average number of links (of all nodes) and average path length (between all possible pair of nodes). Lastly, the diameter of a community (sub-graph) is based on the maximum length among all possible shortest paths. In terms of user behaviour, we also analyze the tweets posted

[1] 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.

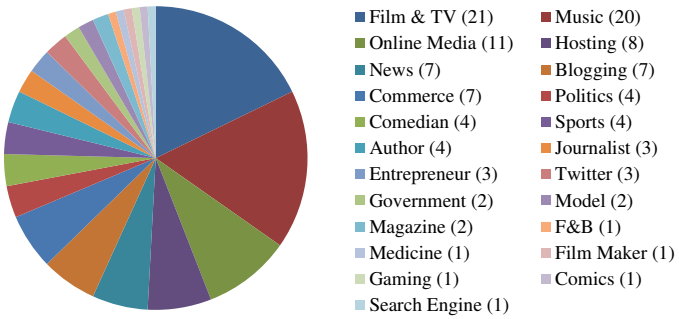

Fig. 2. Popular Categories on Twitter

by users in the detected communities, with a focus on their usage of #hashtags (representing their interest topics).

5 Investigating Communities with Common Interests

The Merriam-Webster dictionary defines a community as "a group of people with a common characteristic or interest living together within a larger society" [25]. Building on this definition, we propose a community detection approach based on individuals sharing common interests. We evaluate our approach by comparing the detected communities (with common interest) to our control group comprising communities with no common interest. This comparison shows that our approach of community detection based on common interests results in larger and more cohesive communities, comprising users who share common interests. Furthermore, we also show that the detected communities exhibit evidence of these common interests in the tweets they post.

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 top 100 celebrities based on their number of followers. Next, we used information on Wikipedia and Google[2] 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. 2 shows the popular categories in Twitter and we selected the five most popular categories among them.[3] For each category, we selected the six most popular celebrities based on their number of followers as listed in Table 2.[4] Also, a celebrity may belong to

[2] If the celebrity's Wikipedia article is unavailable or not comprehensive enough, Google is used as a secondary source (e.g. news articles, fan club pages, etc).

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

[4] 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 2. 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

multiple categories (e.g. Miley Cyrus belongs to both the Music and Film & TV categories).

The next step of our community detection approach 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. 3. 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 most interested in the product/service).

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 3. 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

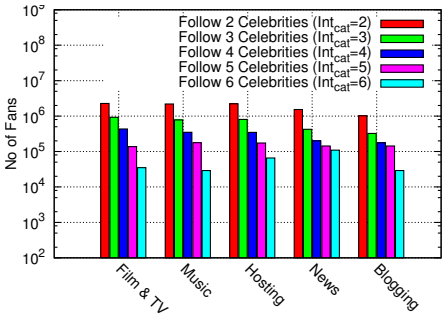

Fig. 3. Fans Following Multiple Celebrities in a Category

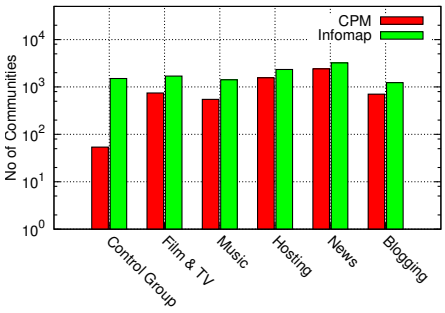

Fig. 4. Total Communities Detected

Table 3. Reciprocity Among Interest Groups

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

for the entire Twitter population [3, 4]. This shows that reciprocity among users with common interests is similar to reciprocity among the general population.

5.1 Analysis of Community Structure

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. community with no common interest) in terms of the total number of communities, size of largest community, and average community size as shown in Fig. 4, 5 and 6 respectively.

Fig. 4 and 5 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. 6. The exception is the News category detected using CPM as many cliques of three nodes were detected as communities 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 size and number of communities detected but in all cases, we detect larger and more communities for users with

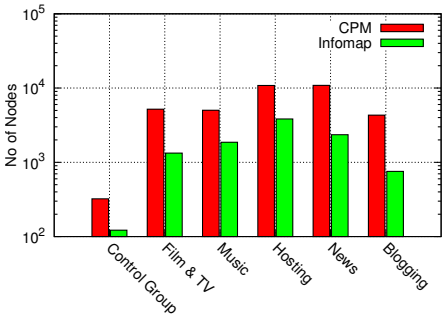

Fig. 5. Size of Largest Community Detected

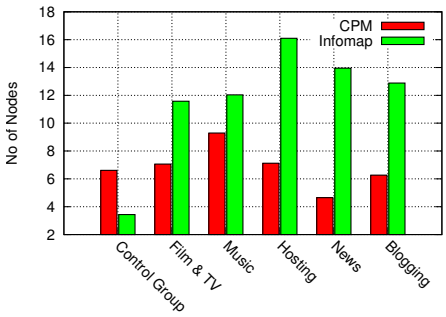

Fig. 6. Average Size of Communities Detected

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 control group, we were unable to detect any communities at 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 affects community detection, this observation shows that our approach performs better than the control group (experiment) given the same k-values. In addition, the detection of communities at a high k-value of up to 25 also shows that our proposed approach effectively selects users who are tightly-coupled in the first place (as a k-clique is a sub-graph comprising k nodes that are fully inter-connected).

Table 4. Network Statistics of the Communities

Category	**Control Group**	**Film & TV**	**Music**	**Hosting**	**News**	**Blogging**
Avg. 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
Avg. Degree	7.81	6.80	7.29	8.17	9.15	7.51

Users with common interests also form communities that are more cohesive than those without any common interest. Table 4 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.

5.2 Analysis of Tweeting Behaviour

Apart from studying the topology structure of the detected communities (with common interests), we further evaluate their interest by analyzing their tweeting behaviour. Specifically, we study their use of #hashtags which serves as a topical label of their tweets. Using the Twitter API, we collected the most recent (last 200) tweets posted by each individual user in both the control group and detected communities with common interests.[5] Next, we compare the usage of #hashtags in the communities with common interest against that of the control group.

Fig. 7. Hashtag Cloud - Control

Fig. 8. Hashtag Cloud - Music

Fig. 7 and 8 illustrate the #hashtag clouds for the control group and Music community respectively. From these #hashtag clouds, we observe that the control group does not tweet about a common topic, as indicated by their #hashtags which are generally unrelated and do not show a common theme (except #hotels and #travel which are related). On the other hand, the Music community tweet frequently about the Music topic, which is evident in the use of #soundcloud, #music, #nowplaying, #itunes, #grammys, etc in their tweets. While the other communities also use #hashtags about a common topic, it is to a lesser extent compared to the Music community (the main reason being that music-related topics are more popular). Another trend we observe is that all communities also display a similar interest in gaming as shown in their use of gaming-related #hashtags such as #140mafia, #mobsterworld, #gameinsight, etc.

These results show that our community detection approach detects communities that are larger, more cohesive and actively tweet about their common interests. More importantly, our approach efficiently detects communities with common interests without the need to perform large scale community detection on the entire social network. Thus, our approach is less computationally intensive (since it directly detects like-minded communities) and compares favourably to existing approaches that detect all communities then identify the interests of

[5] While these tweets may be collected at a different time compared to Kwak et. al.'s dataset (topology links), it provides us with insight to the tweeting behaviour of these users. These tweets also show that detected communities are persistent in their common interest, despite the tweeting data being collected a few years after Kwak et. al.'s dataset.

the communities [8, 9]. 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 [26, 17].

Table 5. 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

6 Specialization and Deepening of Interests

Communities that share the same set of interests are likely to be more connected [9, 27] and interact on a more frequent basis [5]. 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. 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 5.

Similar to Section 5, we used both CPM and Infomap to detect communities among users with $Int_{Country} > 1$.[6] 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 Country Music, $Int_{Country} = 7$. For this user group, we detected five communities comprising 23 distinct users as shown in Fig. 9. 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.

[6] 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.

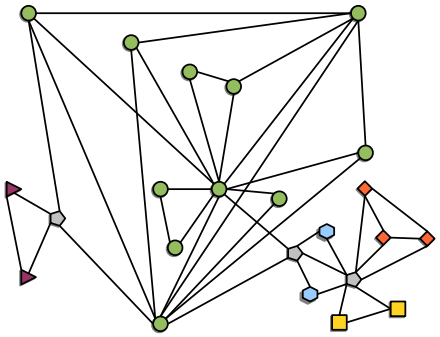

Fig. 9. Community Graph of Fans who follow all Seven Country Singers

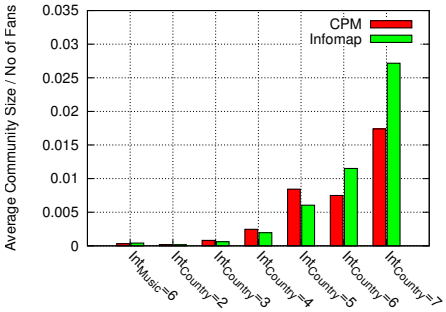

Fig. 10. Normalized Average Community Size for Music and Country Music Categories

6.1 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 respective 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 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. 10. 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 average compared to users with a higher level of interest in a general category. Fig. 10 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 ($Int_{Music} = 6$).

Communities comprising users with a specialized interest are also more cohesive and well-connected than those with a more general interest. Table 6 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

Table 6. Comparison of General and Specialized Interest

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

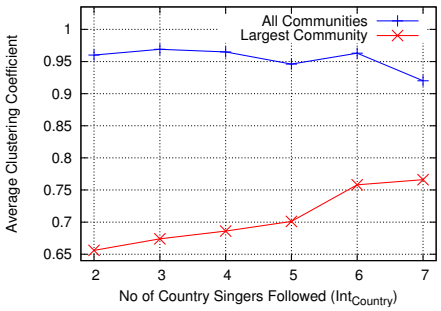

Fig. 11. Average Clustering Coefficient of Country Music Category

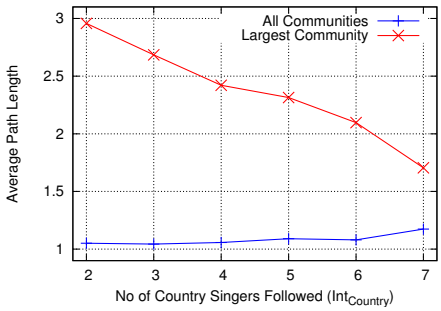

Fig. 12. Average Path Length of Country Music Category

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.

6.2 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. 10 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. 11 shows a gradual increase in clustering coefficient among the largest communities with increasing $Int_{Country}$

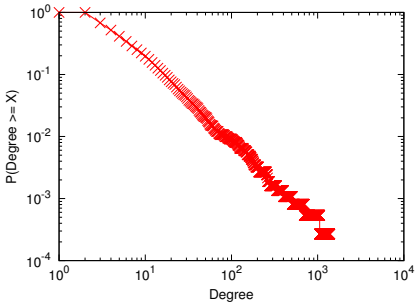

Fig. 13. Degree Complementary Cumulative Distribution Function of Largest Community with $Int_{Country} = 2$

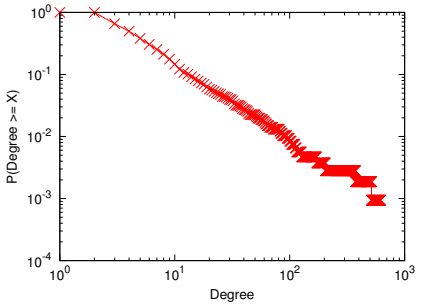

Fig. 14. Degree Complementary Cumulative Distribution Function of Largest Community with $Int_{Country} = 4$

values. While the average 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. 12 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 [28]. Similarly, studies on the Microsoft Messenger social network also show that their users are separated by an average of 6.6 hops [29]. 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. These detected communities also 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.

The detected communities also display the characteristics of scale-free networks as shown in Fig. 13 and 14, 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, 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. Thus, a possible future direction is to further enhance the detection of like-minded communities by considering geolocation information.

7 Conclusion

In this chapter, we proposed a topological-based method to efficiently detect like-minded communities comprising individuals with common interests (Section 4), for applications in targeted advertising and viral marketing. Our method was not developed to detect all communities on Twitter. Instead, it detects larger and more cohesive communities that only comprise users who share a common interest and actively tweet about this interest. 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 online social networking sites), our method is very scalable for identifying communities sharing common interests as it only requires topological information (and Wikipedia/Google for interest classification). The main advantage of our method is that it directly detects communities with common interests instead of having to perform a large scale community detection on the entire social network (then select communities with the common interests).

In addition, this 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 (www.facebook.com), 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 chapter 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, the topological structure of these communities and their tweeting behaviour (Section 5). 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 (Section 6). 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.

Some future areas that we are working on include the geographical analysis of communities comprising like-minded individuals and enhancing our approach to also consider geolocation data for detecting such communities. Also, we intend to perform a temporal analysis of link formation and deletion within these communities, and better understand the contributing factors of individuals joining and/or leaving communities. In addition to studying the deepening of interest based on the number of celebrities followed, we would also like to explore other definitions of deepening interest such as the celebrities' popularity (no. of followers) and the duration of this celebrity following relationship (i.e. is the user a new follower or a long-time fan).

8 Acknowledgments

Kwan Hui Lim was supported by the Australian Government, University of Western Australia (UWA) and School of Computer Science and Software Engineering (CSSE) under the International Postgraduate Research Scholarship, Australian Postgraduate Award, UWA CSSE Ad-hoc Top-up Scholarship and UWA Safety Net Top-Up Scholarship.

References

1. 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.
2. 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.
3. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring user influence in Twitter: The million follower fallacy. In: ICWSM '10: Proceedings of the 4th International AAAI Conference on Weblogs and Social Media. (May 2010) 10–17
4. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: WWW '10: Proceedings of the 19th International Conference on World Wide Web. (Apr 2010) 591–600
5. Poblete, B., Garcia, R., Mendoza, M., Jaimes, A.: Do all birds Tweet the same? Characterizing Twitter around the world. In: CIKM '11: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. (Oct 2011) 1025–1030
6. Iyer, G., Soberman, D., Villas-Boas, J.M.: The targeting of advertising. Marketing Science **24**(3) (2005) 461–476
7. Kaplan, A.M., Haenlein, M.: Two hearts in three-quarter time: How to waltz the social media/viral marketing dance. Business Horizons **54** (2011) 253–263

8. Java, A., Song, X., Finin, T., Tseng, B.: Why we Twitter: Understanding microblogging usage and communities. In: WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis. (Aug 2007) 56–65
9. Li, D., He, B., Ding, Y., Tang, J., Sugimoto, C., Qin, Z., Yan, E., Li, J., Dong, T.: Community-based topic modeling for social tagging. In: CIKM '10: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. (Oct 2010) 1565–1568
10. Lim, K.H., Datta, A.: 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. (Dec 2012) 214–221
11. Correa, D., Sureka, A., Pundir, M.: iTop - Interaction based topic centric community discovery on twitter. In: PIKM '12: Proceedings of the 5th Ph.D. Workshop on Information and Knowledge. (Nov 2012) 51–58
12. Palsetiay, D., Patwary, M.M.A., Zhang, K., Lee, K., Moran, C., Xie, Y., Honbo, D., Agrawal, A., keng Liao, W., Choudhary, A.: User-interest based community extraction in social networks. In: SNA-KDD '12: Proceedings of the 6th SNA-KDD Workshop on Social Network Mining and Analysis. (Aug 2012)
13. Lim, K.H., Datta, A.: Following the follower: Detecting communities with common interests on Twitter. In: HT '12: Proceedings of the 23th ACM Conference on Hypertext and Social Media. (Jun 2012) 317–318
14. Lim, K.H., Datta, A.: Finding Twitter communities with common interests using following links of celebrities. In: MSM '12: Proceedings of the 3rd International Workshop on Modeling Social Media. (Jun 2012) 25–32
15. Twitter: Twitter API. Internet (Sep 2011) Available from: https://dev.twitter.com.
16. Jin, X., Wang, C., Luo, J., Yu, X., Han, J.: 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. (Aug 2011) 753–756
17. Yang, S.H., Long, B., Smola, A., Sadagopan, N., Zheng, Z., Zha, H.: Like like alike - Joint friendship and interest propagation in social networks. In: WWW '11: Proceedings of the 20th International Conference on World Wide Web. (Mar 2011) 537–546
18. Atzmueller, M., Mitzlaff, F.: Efficient descriptive community mining. In: FLAIRS '11: Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference. (May 2011) 459–464
19. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Physical Review E **70**(6) (Dec 2004) 066111
20. Kwak, H., Lee, C., Park, H., Moon, S.: Twitter dataset. Internet (Jun 2009) Available from: http://an.kaist.ac.kr/traces/WWW2010.html.
21. Lim, K.H., Datta, A.: Interest classification of Twitter users using Wikipedia. In: WikiSym+OpenSym '13: Proceedings of the 9th International Symposium on Wikis and Open Collaboration. (Aug 2013)
22. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435** (Jun 2005) 814–818
23. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences **105**(4) (2008) 1118–1123

24. CMA: CMA Award Winners 1967-2011 (Jul 2013) Available from: http://www.cmaworld.com/cma-awards/winners/past-winners.
25. Merriam-Webster: Merriam-webster dictionary and thesaurus. Internet (Oct 2011) Available from: http://www.merriam-webster.com/dictionary/community.
26. Fond, T.L., Neville, J.: Randomization tests for distinguishing social influence and homophily effects. In: WWW '10: Proceedings of the 19th International Conference on World Wide Web. (Apr 2010) 601–610
27. Zhao, D., Rosson, M.B.: 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. (May 2009) 243–252
28. Milgram, S.: The small world problem. Psychology Today **2** (1967) 60–67
29. Leskovec, J., Horvitz, E.: Planetary-scale views on a large instant-messaging network. In: WWW '08: Proceedings of the 17th International Conference on World Wide Web. (Apr 2008) 915–924