An Interaction-based Approach to Detecting Highly Interactive Twitter Communities using Tweeting Links

Kwan Hui Lim * and Amitava Datta
School of Computer Science and Software Engineering
The University of Western Australia
Crawley, WA 6009, Australia
E-mail: kwanhui@graduate.uwa.edu.au, datta@csse.uwa.edu.au

Abstract.
The immense popularity and rapid growth of Online Social Networks (OSN) have attracted the interest of researchers and companies, particularly in how users group together to form communities online. While many community detection algorithms have been developed to detect communities on such OSNs, most of these algorithms are based only on topological links and researchers have observed that many topological links do not translate to actual user interaction. As such, many members of the detected communities do not communicate frequently to each other. This inactivity creates a problem in targeted advertising and viral marketing, which require the community to be highly active so as to facilitate the diffusion of product/service information. We propose an approach to detect highly interactive Twitter communities that share common interests, based on the frequency and patterns of direct tweeting among users, rather than the topological information implicit in follower/following links. Our experimental results show that communities detected by our proposed approach are more cohesive and connected within different interest groups, based on topological measures. We also show that the detected communities actively interact about the specific interests, based on the high frequency of #hashtags and @mentions related to this interest. In addition, we study the trends in their tweeting patterns such as how they follow and unfollow other users, and observe that our approach detects communities comprising users whose links are more persistent compared to those in other groups of users.

Keywords: Twitter, Tweets, Social Network Analysis, Community Detection, Like-minded Communities, Interaction Links, Common Interests

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" [9]. With the rapid proliferation of OSNs, many companies have embraced social media as a new outlet for their targeted advertising and viral marketing efforts. Twitter is one such OSN given its large user base and high user activity. However, one main problem in targeted advertising and viral marketing is identifying the right target audience, comprising 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 [18] and the connectedness of this group facilitates word-of-mouth advertising [20].

[*] Corresponding author. K. H. Lim is now with the Department of Computing and Information Systems, University of Melbourne, Australia. Email: limk2@student.unimelb.edu.au.

Most community detection algorithms consider only topological information (such as follower/following links) but not user activity (such as tweeting patterns) [19]. 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 [6,40]. Thus, it is necessary to consider user activity in addition to topological information for community detection, especially for advertising and marketing purposes. We propose a method for identifying communities where its members not only share common interests but actively and frequently communicate about the common interests. This approach involves identifying community members based on their frequency of direct communication with other users in the community.

1.1. Contributions

Our main contributions in this paper include the following:

- *Proposing an approach for detecting highly interactive communities that frequently communicate about their common interests*. Our proposed approach models the frequency of direct tweets between users as a network of weighted links. Using these weighted links, we then detect the highly interactive communities based on a pre-determined threshold.
- *Evaluating the topological structures of these communities*. We use topological measures such as clustering coefficient, average path length, average degree and diameter, to measure the effectiveness of our proposed approach. Our results indicate that our approach detects communities that are more cohesive and connected, as supported by a shorter average path length and higher clustering coefficient.
- *Studying the communication behaviour and patterns of these communities*. Specifically, we track the users in these detected communities and observe that our approach detects interactive communities where its users communicate frequently about the specific interests based on the content of #hashtags and @mentions.
- *Performing a preliminary study into the temporal evolution of links among these communities*. 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. In particular, we observe that links among users in our detected communities are more persistent compared to those in other groups of users.

This article is an extended version of an earlier conference version in Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence [25].

1.2. Structure and Organization

We first give a description of Twitter in Section 2 and discuss some related work in Section 3. Following which, we further elaborate on our proposed method and dataset used in Section 4. Next, we evaluate our proposed method in terms of network topological measures and communication/interaction patterns in Sections 5 and 6 respectively. Finally, we discuss our findings in Section 7 before concluding the paper in Section 8.

2. Background Information

Twitter is an OSN that allows users to post short messages (called tweets) of up to 140 characters. A user can follow another user to receive the tweets that he/she posts. Also, tweets posted by a user can be forwarded to all users following him/her, a process known as retweeting. Users can retweet by either manually adding the “RT @username” prefix in front of the original tweet or use the built-in “retweet” button. In addition, tweets can also contain @mentions and #hashtags for mentioning other users and tagging interesting topics respectively. Replies are similar to mentions as the tweets contains @username but the tweet starts off with @username and is replied by using the ”reply” button.

All of these Twitter-related data and statistics can be retrieved using the publicly accessible Twitter Application Programming Interface (API) [37]. The availability of the Twitter API has stirred immense interest in the academic study of the Twitter social network. For example, 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 [14,32,43]. Romero et al. [34] and Huang et al. [17] studied the diffusion of #hashtags on Twitter and investigated the factors behind the mass adoption of #hashtags and their subsequent dying off.

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. [4] proposed a method to determine the credibility of tweets. Similarly, Becker et al. [2] presented a real-time system to detect tweets that describe real-life events, while Zhu et al. [44] used a Latent Dirichlet Allocation based topic model that utilizes word pairs to understand people's activity patterns. Also, Kouloumpis et al. [21] studied the sentiment of tweets based on the usage of #hashtags, emoticons, caps and punctuation. While these studies analyze tweeting patterns and contents, they do not use tweeting links to detect communities with common interests.

3. Related Work

As our proposed approach is an interaction-based community detection method, we first discuss some related studies on user interactions within OSNs, followed by a broad overview of community detection research. Thereafter, we focus on some related work that also aims to detect communities using interaction-based metrics and highlight the differences between our proposed approach and these earlier work.

3.1. Studying Interaction Among Communities

Many researchers have used the interaction frequency among users of OSNs to study information diffusion and the topological characteristics of entire OSNs. For example, various researchers constructed interaction graphs to study the general structure and behaviour of users on OSNs such as Cyworld and Facebook [6,40,38]. Similarly, the frequency of co-purchased media and book items were used to form networks for identifying local communities of co-purchased items on Amazon [29,30]. The interaction activity between users has also been used to construct networks for the purpose of studying information diffusion on Twitter and Flickr [34,5,42].

The main difference of our proposed approach (from these related work) is that we use interaction frequency to detect highly-interactive communities with common interests, while most of these researchers use it 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 (frequency of @mentions) with each other.

3.2. General Community Detection

The effective detection of community structures from an underlying network graph has been an important and frequently studied problem, due to its potential application in complex networks ranging from biological networks to OSNs [13]. Apart from applications on OSNs, community detection is also a common research problem on real-life social networks, such as scientific collaboration networks [1,12,27]. However, these methods consider only topological links to detect community structures, which may not translate to interactive communities [6,40].

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, instead of using only certain aspects of communication (e.g., only #hashtags). Our study also differs from the related work which examines communication behaviour of the general Twitter population instead of a specific group (with a common interest). Also, the related work considers only certain aspects of communication (e.g., only #hashtags), instead of the entire spectrum of information available.

3.3. Interaction-based Community Detection

Most related to our research would be other community detection algorithms that use interaction-based metrics. For example, Correa et al. developed the iTop algorithm to find topic-centric communities on Twitter, based on the @mentions and retweets exchanged among its users [10]. Specifically, the iTop algorithm constructs a weighted graph of user interactions and greedily maximizes the local modularity in order to detect these topic-centric communities based on a set of seed users. On the same note, Palsetia et al. proposed an algorithm to detect communities that share the same social interest [33]. Using wall posts (Facebook) and tweets (Twitter) as interactions links, this algorithm first constructs an undirected graph, weighted based on a similarity coefficient between users of the interaction link, before trying to detect communities using a modified version of the Clauset, Newman, and Moore (CNM) algorithm [7] that calls itself recursively.

While the algorithms proposed by Correa et al. and Palsetia et al. are based on an underlying network graph modeled from interaction links, these algorithms are generally modularity-based methods that require an additional user pre-processing step. Both methods are recursively executed in an attempt to detect communities, until a particular modularity metric is achieved. More importantly, [10] requires an additional step of a "warm start" process where a set of seed users are identified based on a textual search of their Twitter user profile biography, whereas our proposed approach does not require this additional process. Similarly, [7] requires users' interests to be explicitly determined based on their Twitter profiles or Facebook walls. On the other hand, our proposed approach does not utilize such a modularity metric in a recursive nature nor require an explicit search of user interest on Twitter profiles or Facebook walls, but instead we pre-select users based on their followings of celebrities as a proxy for their implicit interest. Thereafter, we attempt to detect communities from this pre-selected group of users based on their interaction links with a thresholding mechanism. Due to the recursive calling of the CNM algorithm, [7] also results in small communities (with the largest community comprising only 21 users in their experiments), whereas our proposed approach detects larger communities which are better suited for targeted advertising and viral marketing purposes.

4. Methodology

We model topological links in the Twitter social network as followership links where a link (i, j) represents that user i is a follower of user j. These followership links can also be reciprocal and such a bi-directional link between users i and j are represented as a friendship link $Fr_{i,j}$. The interest of a user is represented by the number of celebrities (of the same interest category) that he/she follows. Here, we define celebrities as users with 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.

Table 1
Notations and Definitions

Notations	Definitions
(i, j)	A followership (uni-directional) link from user i to 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

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 this paper.

4.1. Common Interest Community Detection

We extend upon the Common Interest Community Detection (CICD) method [28] which is used for detecting communities comprising only individuals with common interests, using only topological links. The main strategy employed by the CICD method to detect communities with common interests is to select users with common interests (based on their following of celebrities), then detect communities using the common links among these users (with common interests). Thus, the CICD method comprises of the following three steps:

1. Identify an interest category (and its set of representative celebrities).
2. Select a group of users with common interest (based on their following links of celebrities).
3. Detect communities based on the common topology links among this group of users.

The first step is to select a set of n celebrities $c_1, c_2, ..., c_n$, that belongs to a common interest category.[1] We ascertain the interest category of a celebrity

[1] 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.

based on the Wikipedia article (particularly the "occupation" field) describing this celebrity. This classification using Wikipedia minimizes the mis-classification of celebrities into the wrong interest categories, especially for celebrities who belong to multiple categories (e.g., a singer-actress with her own fashion line would belong to three interest categories, namely Music, Film & TV and Business). In [26], we have also demonstrated how this classification of celebrities can be automatically performed using a bag-of-words concept along with content on Wikipedia. This bag-of-words concept is essentially a mapping of keywords (found on the celebrity's Wikipedia article page) to the respective interest categories. Similarly, other researchers [39] have demonstrated how celebrities can be mapped to their respective interest categories, using the descriptions of celebrities and WordNet domain hierarchy [3].

In step 2, we retrieve the list of users following each celebrity $c_j, 1 \leqslant j \leqslant n$, and select the group of users following all n celebrities. In short, we retrieve the set:

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

Basically, we construct Set $\mathcal{P}$ out of users who follow all n celebrities in an interest category. Set $\mathcal{P}$ is also the intersection set among the fans of each celebrity.

Following which (in step 3), we retrieve all bi-directional links among Set $\mathcal{P}$ then use the Clique Percolation Method (CPM) [11] and Infomap algorithm [35] to detect communities among Set $\mathcal{P}$.[2] CPM detects communities based on a series of adjacent cliques (fully-connected sub-graphs), while Infomap uses the frequent paths of a random walker to detect communities. These detected communities shall be referred to as the link-based communities, Com_{CICD}. The criteria for the CICD method can also be relaxed such that we select users who follow x out of n celebrities, where the value of x determines the interest level of the resulting Set $\mathcal{P}$. For the purpose of this paper, we select users who follow all celebrities to construct a Set $\mathcal{P}$ with the most interest in the given category.

4.2. Highly Interactive Community Detection

Our proposed approach, the Highly Interactive Community Detection (HICD) method detects a highly interactive community using the communication pattern and frequency among the users [25]. Basically, we adapt steps 1 and 2 of the CICD method and modify step 3 to consider interaction links (with a filtering threshold) instead of using topology links. As described earlier, we define $M_{i,j}$ as a tweet posted by user i that contains a @mention of user j. Next, we model the communication intensity of user i to j as the number of @mentions user i makes of user j, denoted:

$$I_{i,j} = M_{i,j}, for\ i, j \in \mathcal{P} \qquad (2)$$

Essentially, $I_{i,j}$ is the number of times user i @mentions user j in his/her tweets. Next, we 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 pre-determined 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. Finally, we 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.

The main difference between the CICD and HICD methods is 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, our proposed 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. As pointed out in [31], @mention links are a stronger measure of interaction activity, compared to follower/following links. Due to this different usage of links, the communities detected by the CICD and HICD methods may overlap but are unlikely to be

[2] Using both CPM and Infomap demonstrate that the results obtained by our proposed methods are independent of the community detection algorithm chosen. 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 [13]. Refer to [11] and [35] for more information on CPM and Infomap respectively.

Table 2

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		

Table 3

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, etc
Preposition	about, above, across, after, against, among, around, at, before, behind, below, beneath, beside, between, etc
Conjunction	after, although, as, because, before, if, once, since, than, that, though, whether, until, when, where, etc

a subset of one another (as users may @mention each other even when they are not topologically connected).

4.3. Evaluation

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 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).

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 3.

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.[3] 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.

5. Interactive Communities with Common Interests: A Topological Evaluation

In this section, we are interested in the topological evaluation of our proposed HICD method compared to the CICD method. We first describe the overall process of using our HICD method to detect highly interactive communities before describing some key characteristics of these detected communities. Following which, we evaluate these detected communities based on topological measures and discuss the effects of a threshold parameter used in the HICD method.

5.1. Detection of Communities

For our study, we demonstrate the effectiveness of our approach across different communities with common interests in country music, tennis and basketball respectively. We selected nine country music celebrities based on winners of the Country Music Association Awards [8] from 2001 to 2011, with more than

[3] 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.

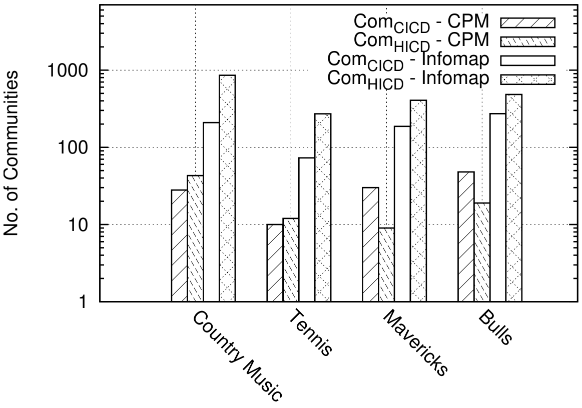

Fig. 1. Total communities detected

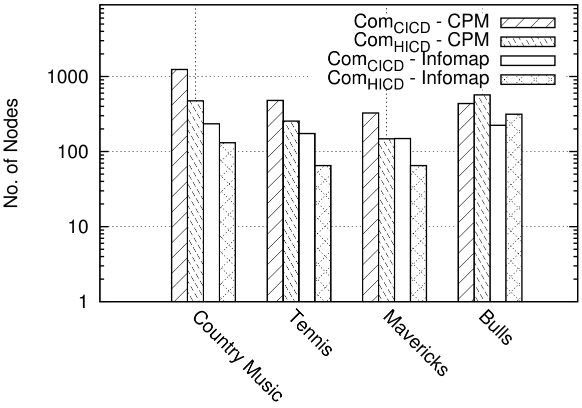

Fig. 2. Size of largest community detected

Table 4

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

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

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 (as of 2012). The list of celebrities representing each interest category is listed in Table 2.

Next, we retrieve the set of users following all celebrities in each category, Set $\mathcal{P}$ as described in Equation (1). The number of users in Set $\mathcal{P}$ of each category is shown in Table 4. 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 in the country music, tennis and basketball (Mavericks and Bulls) categories as shown in Fig. 1. From these detected communities, we selected the largest community (of each category) to analyze their tweeting and retweeting patterns within the community. These link-based communities shall be referred to as Com_{CICD} for each of the categories, in the rest of the paper.

Using our HICD method, we determine the tweet-based community (denoted 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. 1 and 2 respectively.

5.2. General Community Characteristics

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. 1 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. 2 shows a similar trend in the largest community detected (e.g., communities detected by CPM are larger than that selected by Infomap or vice versa,

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

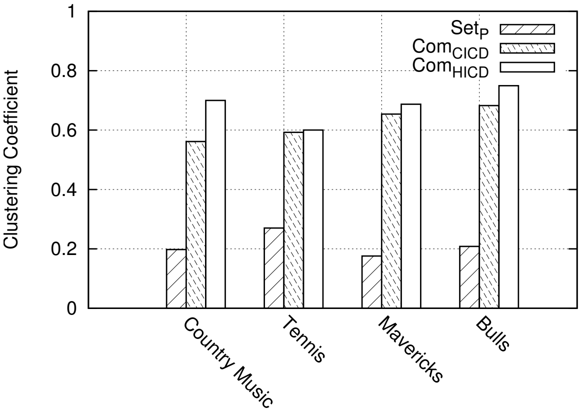

Fig. 3. Clustering coefficient

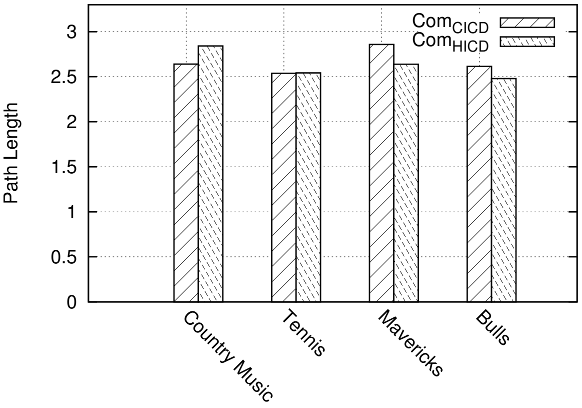

Fig. 4. Average path length

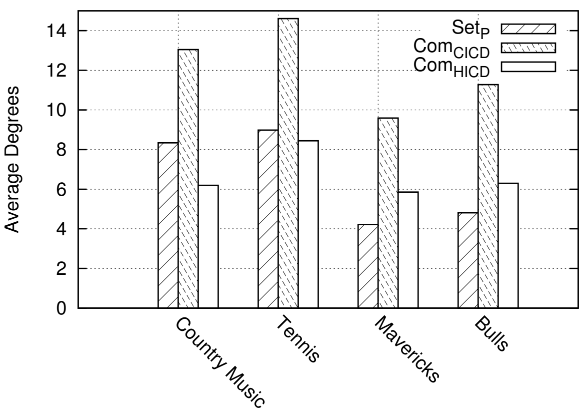

Fig. 5. Average degree

given the same interest category). In this work, we are most interested in the largest community as it provides the most potential for targeted advertising and viral marketing, compared to smaller communities.

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. 2 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 next 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.

5.3. Topological Characteristics

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. 3. 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 subgraph (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. 4 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 shorter average path length compared to Com_{CICD} as shown in Table 5. The shorter average path length and higher clustering coefficient show that our HICD method detects communities that are more cohesive and connected.

Table 5
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
Average 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

Table 6
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
Mavericks	Dallas Quito Philippines	Dallas Toronto Fort Worth	Dallas Fort Worth Various Texas Cities
Bulls	Chicago Quito Melbourne	Chicago New Jersey Melbourne	Chicago Aurora/Quito Melbourne

Fig. 5 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.

5.4. Geographical Location of Users

Table 6 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 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, researchers such as Java et al. [19] also noticed that the probability of two persons being connected is negatively correlated with their geographic distance. Similarly, Kwak et al. [23] noticed that users with less than 1,000 friends tend to be geographically nearer to each other. 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!”).

5.5. Effects of Increasing Threshold T of $I_{i,j}$

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 5. 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 a trade-off between detecting more cohesive communities (at high threshold T) or larger communities (at low threshold T). Thus, any user of our HICD method has the flexibility to select this threshold parameter according to the usage scenario (i.e., is he/she interested in smaller but more co-

Table 7
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*_**

Table 8
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

hesive communities, or larger but less cohesive communities). For the rest of the paper, we focus on the country music communities detected using a threshold T of 1 as we are most interested in the largest community.

6. Interactive Communities with Common Interests: An Interaction-based Evaluation

After the topological evaluation in Section 5, we now study and evaluate our proposed HICD method using various interaction-based measures. Some of these measures include: (i) the content of tweets used, i.e., #hashtags, @mentions, URLs and keywords; (ii) the temporal trend in tweeting behaviour; and (iii) the way users follow and unfollow one another over time.

6.1. Content of Tweets

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.

6.1.1. Usage of #hashtags

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 7 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 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 interest category. Thus, this result also shows that using the celebrity followings of an user is a good proxy for his/her interests.

6.1.2. Usage of @mentions

Likewise, our HICD method detects communities that make more @mentions of country music artists. Table 8 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}.

Similar to the analysis on #hashtags usage, this observation shows that our HICD method is able to detect communities that frequently interact about coun-

Table 9

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

try 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. These results show the effectiveness of our HICD method in detecting communities comprising interactive users who communicate frequently about the specific interests based on their usage of #hashtags and @mentions.

6.1.3. *Usage of URLs and Keywords*

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 9 shows the top 10 websites that Set $\mathcal{P}$, Com_{CICD} and Com_{HICD} of the country music category use in their tweets. Some websites are repeated as different TinyURLs may also point to the same website (or different sub-pages of the same parent website). 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 4. 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.

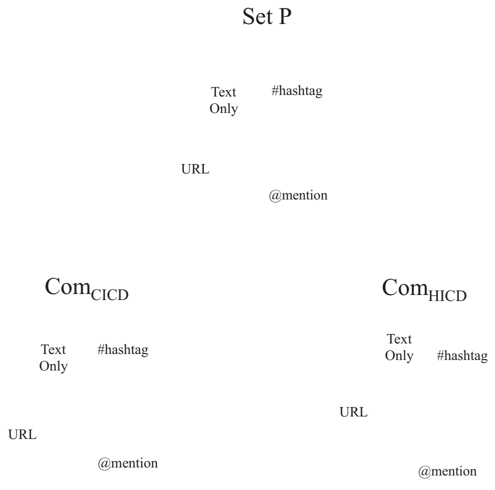

Fig. 6. Type of tweets

6.2. *Trends in Tweeting Behaviour*

We investigate tweeting trends 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. 6. The type of content in tweets can be any combination of textual information, #hashtags, @mentions and/or URLs. Fig. 6 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

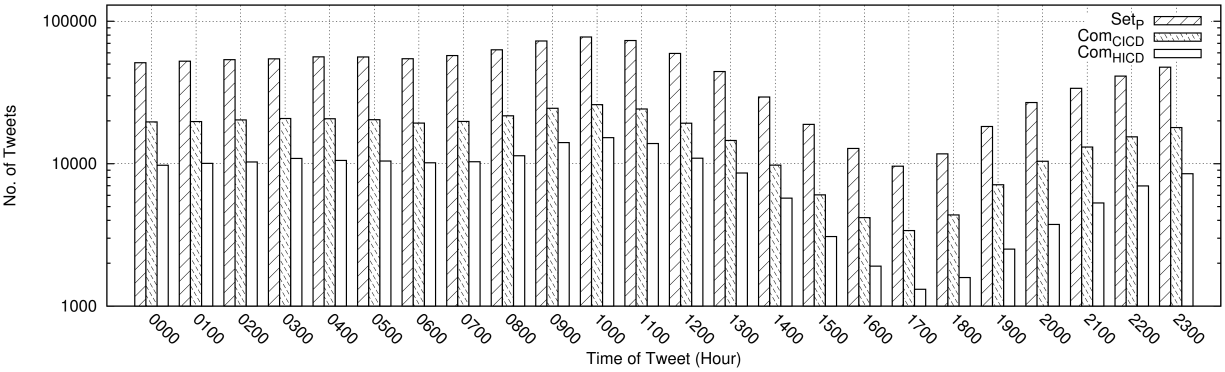

Fig. 7. Time distribution of tweets

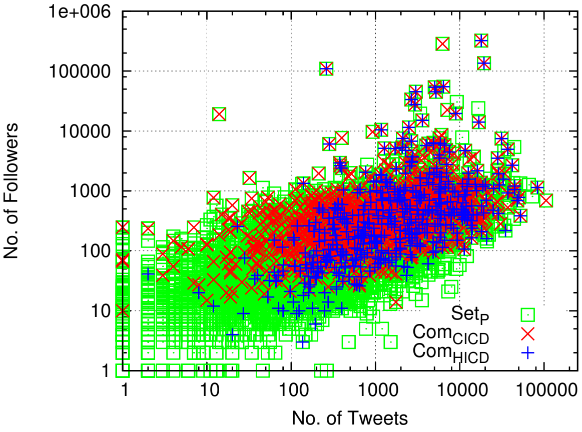

Fig. 8. Comparison of tweets to followers (Best viewed in colour)

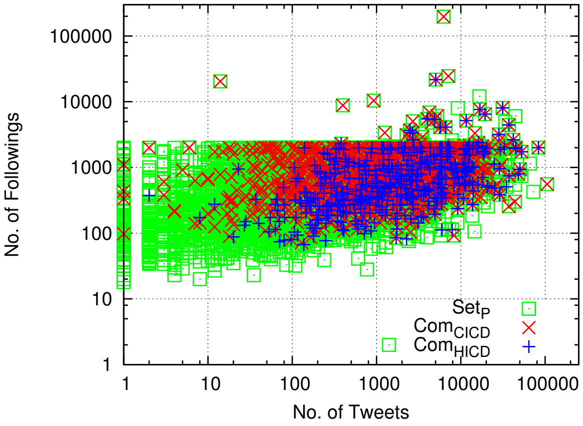

Fig. 9. Comparison of tweets to followings (Best viewed in colour)

tweets. We next investigate trends in the timings of tweet postings.

Across Set $\mathcal{P}$, Com_{CICD} and Com_{HICD}, Fig. 7 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 [36], 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 a user to his/her number of followers and followings. Fig. 8 and 9 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. 8 and 9.

In addition, Fig. 8 and 9 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.

6.3. Temporal Analysis of Links

As the focus of our paper is on detecting highly interactive communities with common interest, we only perform a preliminary analysis on link formation and deletion over time. We leave the more detailed analysis for future work where we plan to investigate on

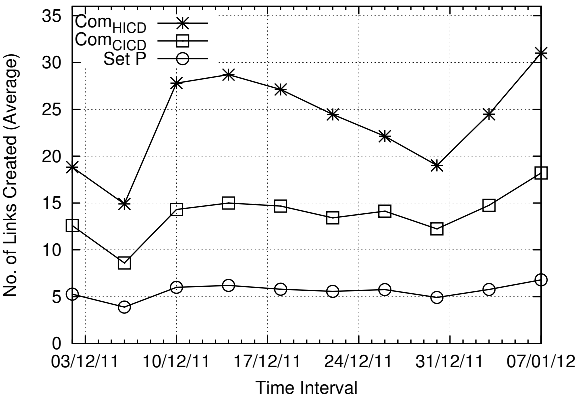

Fig. 10. Time analysis of created links

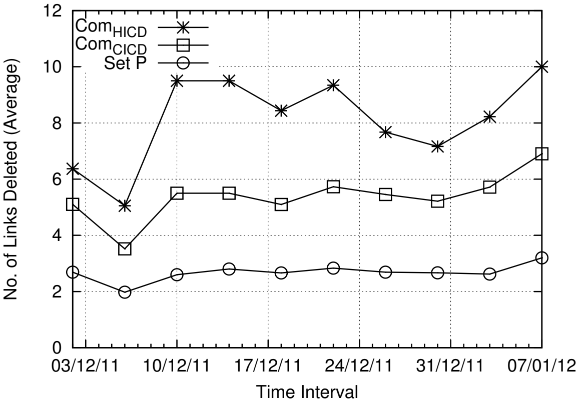

Fig. 11. Time analysis of deleted links

the motivating factors behind a user's choice in following/unfollowing other users (e.g., similar interests, common friends, etc). For now, this preliminary analysis serves to provide readers with a general overview of following/unfollowing behavior among users in Set $\mathcal{P}$, Com_{CICD} and Com_{HICD}.

Our preliminary analysis involves studying 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. 10 and 11 respectively.

Fig. 10 and 11 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. Thus, this result shows our approach effectively detects communities where users share more persistent links, compared to those in other groups.

Even though our HICD method detects communities with more persistent links, it may seem that Twitter users create and delete links at a high frequency. While alarming at first, this trend of frequent unfollowing has also been studied and observed by other researchers [24,41,22]. One contributing factor is the "low cost" of unfollowing as Twitter users are able to easily unfollow with just a single click and the users being unfollowed are not notified of this action [22]. Of particular note, Kwak et al. observed that users are less likely to unfollow each other if they engage in frequent interactions [24]. This observation further supports why our HICD method is able to detect communities with more persistent links, as these communities comprise users who share a common interest and frequently interact about this interest.

7. Discussion

While we describe our proposed HICD method on the basis of Twitter, this method can also be applied to other OSNs by adapting the notations and definitions to the unique nature of 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. For example, our proposed HICD method could be used in Facebook by defining $I_{i,j}$ as the number of posts a user i writes on the wall/timeline of user j.

There are also distinct advantages and disadvantages to both the CICD and HICD 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 levels of detected communities.

8. Conclusion

In this paper, we proposed 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.

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.

Another possible area for future work involves examining the correlation between communication frequency and message content with the formation of links. This would provide a useful model for predicting the formation of links based on the communication patterns between two individuals and subsequently, allow us to study how and why links are formed within communities. As a user's personal interest may differ from the interest of his/her community [16,15], another interesting area is to investigate how a user's personal interest may influence him/her in joining or leaving a community in future.

9. 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. The authors thank the editors and anonymous reviewers for their useful comments.

References

[1] H. Balakrishnan and N. Deo. Discovering communities in complex networks. In *Proceedings of the 44th annual Southeast Regional Conference (ACMSE '06)*, pages 280–285, Mar 2006.

[2] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on Twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM '11)*, pages 438–441, May 2011.

[3] L. Bentivogli, P. Forner, B. Magnini, and E. Pianta. Revising the WordNet domains hierarchy: Semantics, coverage and balancing. In *Proceedings of the 2004 Workshop on Multilingual Linguistic Resources (MLR '04)*, pages 101–108, Aug 2004.

[4] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*, pages 675–684, Mar 2011.

[5] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi. Characterizing social cascades in Flickr. In *Proceedings of the 1st Workshop on Online Social Networks (WOSN '08)*, pages 13–18, Aug 2008.

[6] H. Chun, H. Kwak, Y.-H. Eom, Y.-Y. Ahn, S. Moon, and H. Jeong. Comparison of online social relations in volume vs interaction: A case study of Cyworld. In *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement (IMC '08)*, pages 57–70, Oct 2008.

[7] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, Dec 2004.

[8] CMA. CMA Award Winners 1967-2011, Jul 2013. Available from: http://www.cmaworld.com/cma-awards/winners/past-winners.

[9] ComScore. It's a social world: Top 10 need-to-knows about social networking and where it's 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.

[10] D. Correa, A. Sureka, and M. Pundir. iTop - Interaction based topic centric community discovery on Twitter. In *Proceedings of the 5th Ph.D. Workshop on Information and Knowledge (PIKM '12)*, pages 51–58, Nov 2012.

[11] I. Derényi, G. Palla, and T. Vicsek. Clique percolation in random networks. *Physical Review Letters*, 94(16):240–253, Apr 2005.

[12] N. Du, B. Wu, X. Pei, B. Wang, and L. Xu. Community detection in large-scale social networks. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (WebKDD/SNA-KDD '07)*, pages 16–25, Aug 2007.

[13] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, Feb 2010.

[14] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the Twitterers - Predicting information cascades in microblogs. In *Proceedings of the 3rd International Workshop on Online Social Networks (WOSN '10)*, Jun 2010.

[15] T.-A. Hoang. Modeling user interest and community interest in microbloggings: An integrated approach. In *Proceedings of the 19th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '15)*, pages 708–721, May 2015.

[16] T.-A. Hoang and E.-P. Lim. On joint modeling of topical communities and personal interest in microblogs. In *Proceedings of the 6th International Conference on Social Informatics (SocInfo '14)*, pages 1–16, Nov 2014.

[17] J. Huang, K. M. Thornton, and E. N. Efthimiadis. Conversational tagging in Twitter. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia (HT '10)*, pages 1079–1088, Jun 2010.

[18] G. Iyer, D. Soberman, and J. M. Villas-Boas. The targeting of advertising. *Marketing Science*, 24(3):461–476, Aug 2005.

[19] A. Java, X. Song, T. Finin, and B. Tseng. Why we Twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (WebKDD/SNA-KDD '07)*, pages 56–65, Aug 2007.

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

[21] E. Kouloumpis, T. Wilson, and J. Moore. Twitter sentiment analysis: The Good the Bad and the OMG! In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM '11)*, pages 538–541, May 2011.

[22] H. Kwak, H. Chun, and S. Moon. Fragile online relationship: A first look at unfollow dynamics in Twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*, pages 1091–1100, May 2011.

[23] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*, pages 591–600, Apr 2010.

[24] H. Kwak, S. Moon, and W. Lee. More of a receiver than a giver: Why do people unfollow in Twitter? In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM '12)*, pages 499–502, Jun 2012.

[25] K. H. Lim and A. 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 and Intelligent Agent Technology (WI-IAT '12)*, pages 214–221, Dec 2012.

[26] K. H. Lim and A. Datta. Interest classification of Twitter users using Wikipedia. In *Proceedings of the 9th International Symposium on Wikis and Open Collaboration (WikiSym+OpenSym '13)*, Aug 2013.

[27] K. H. Lim and A. Datta. A seed-centric community detection algorithm based on an expanding ring search. In *Proceedings of the 1st Australasian Web Conference (AWC '13)*, pages 21–26, Jan 2013.

[28] K. H. Lim and A. Datta. A topological approach for detecting Twitter communities with common interests. In *Ubiquitous Social Media Analysis*, volume 8329 of *Lecture Notes in Computer Science*, pages 23–43, Dec 2013.

[29] F. Luo, J. Z. Wang, and E. Promislow. Exploring local community structures in large networks. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT '06)*, pages 233–239, Dec 2006.

[30] F. Luo, J. Z. Wang, and E. Promislow. Exploring local community structures in large networks. *Web Intelligence and Agent Systems*, 6(4):387–400, Oct 2008.

[31] M.-D. Luu and A. C. Thomas. Beyond mere following: Mention network, a better alternative for researching user interaction and behavior. In *Proceedings of the 8th International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (SBP '15)*, pages 362–368, Mar 2015.

[32] S. A. Macskassy and M. Michelson. Why do people retweet? Anti-homophily wins the day! In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM '11)*, pages 209–216, May 2011.

[33] D. Palsetiay, M. M. A. Patwary, K. Zhang, K. Lee, C. Moran, Y. Xie, D. Honbo, A. Agrawal, W. keng Liao, and A. Choudhary. User-interest based community extraction in social networks. In *Proceedings of the 6th SNA-KDD Workshop on Social Network Mining and Analysis (SNA-KDD '12)*, Aug 2012.

[34] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in

the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*, pages 695–704, Mar 2011.

[35] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, Jan 2008.

[36] 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.

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

[38] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the evolution of user interaction in Facebook. In *Proceedings of the 2nd ACM Workshop on Online Social Networks (WOSN '09)*, pages 37–42, Aug 2009.

[39] D. Wang, K. Kwon, and I.-J. Chung. Domain classification for celebrities using spreading activation and reasoning on semantic network. In *Proceedings of the 5th International Conference on Ubiquitous and Future Networks (ICUFN '13)*, pages 744–749, Jul 2013.

[40] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European Conference on Computer Systems (EuroSys '09)*, pages 205–218, Apr 2009.

[41] B. Xu, Y. Huang, H. Kwak, and N. S. Contractor. Structures of broken ties: Exploring unfollow behavior on Twitter. In *Proceedings of the 16th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '13)*, pages 871–876, Feb 2013.

[42] J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in Twitter. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM '10)*, pages 355–358, May 2010.

[43] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su. Understanding retweeting behaviors in social networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10)*, pages 1633–1636, Oct 2010.

[44] D. Zhu, Y. Fukazawa, E. Karapetsas, and J. Ota. Activity-based topic discovery. *Web Intelligence and Agent Systems*, 12(2):193–209, Oct 2014.