Tweets Beget Propinquity: Detecting Highly Interactive Communities on Twitter 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
Email: kwanhui@graduate.uwa.edu.au, datta@csse.uwa.edu.au

***Abstract*—Many community detection algorithms have been developed to detect communities on Online Social Networks (OSN). However, 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 requires the community to be highly active so as to allow 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. From a topological aspect, we show that our method detects communities that are more cohesive and connected within different interest groups. We also show that the detected communities interact actively 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.**

I. Introduction

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 [1] and the connectedness of this group facilitates word-of-mouth advertising [2].

Most community detection algorithms consider only topological information (such as follower/following links) but not user activity (such as tweeting patterns) [3]. 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 [4], [5]. 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.

Our contributions to this paper include the following:

- An approach for detecting highly interactive communities that frequently communicate about their common interests.
- A study into the communication behaviour and patterns of these communities.
- A preliminary study into the evolution of links among these communities over time.

We first give a description of Twitter and discuss some related work in Section II. Following which, we further elaborate on our proposed methods and data used in Section III. Next, we evaluate our proposed methods in terms of network topology and communication patterns in Section IV. Finally, we discuss our findings and conclude the paper in Section V.

II. Background and Related Work

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 other users, 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. All of these Twitter-related data and statistics can be retrieved using the publicly accessible Twitter Application Programming Interface (API)[1].

The availability of the Twitter API has stirred immense interest in the academic study of the Twitter social network. 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

[1] https://dev.twitter.com

timings [6], [7], [8]. Romero et al. [9] and Huang et al. [10] studied the diffusion of #hashtags on Twitter and investigated the factors behind the mass adoption of #hashtags and their subsequent dying off.

In addition, tweets have been analyzed to determine their credibility, sentiments and relation to real-life events. Using the tweeting patterns of a user, tweet content and external references, Castillo et al. [11] proposed a method to determine the credibility of tweets. Similarly, Becker et al. [12] presented a real-time system to detect tweets that are describing real-life events. Also, Kouloumpis et al. [13] 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.

Many authors have also used the interaction frequency among users of OSNs to study information diffusion and the topological characteristics of entire OSNs. Various authors constructed interaction graphs to study the general structure and behaviour of users on OSNs such as Cyworld and Facebook [4], [5], [14]. Similarly, the interaction activity between users has also been used to construct networks for the purpose of studying information diffusion on Twitter and Flickr [9], [15], [16]. The main difference of our work (from these studies) is that we use interaction frequency to detect highly-interactive communities with common interests while these authors use it only for studying information diffusion on the overall structure of OSNs.

Community detection is also a common research problem on other real-life social networks, such as scientific collaboration networks [17], [18]. However, these methods consider only topological links to detect community structures, which does not translate to interactive communities [4], [5]. 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 only certain aspects of communication (e.g. only #hashtags).

III. 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. 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.

We extend upon the Common Interest Community Detection (CICD) method [19], [20] 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), determine the common links among these users, then detect communities using these links. The first step is to select a set of k celebrities $c_1, c_2, ..., c_k$, that belongs to a common interest category. Next, we retrieve the list of users following each celebrity $c_j, 1 \leqslant j \leqslant k$, and select the group of users following all k celebrities. In short, we retrieve the set:

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

Basically, we construct Set $\mathcal{P}$ out of users who follow all k celebrities in an interest category. Following which, we retrieve all bi-directional links among Set $\mathcal{P}$ then use the Clique Percolation Method (CPM) [21] and Infomap algorithm [22] 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 k celebrities, where the value of x would determine 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.

Our proposed model, the Highly Interactive Community Detection (HICD) method detects a highly interactive community using the communication pattern and frequency among the users. We first 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 by:

$$I_{i,j} = M_{i,j}, for\ i, j \in \mathcal{P}$$

Essentially, $I_{i,j}$ is the number of times user i @mentions user j in his/her tweets.[3] 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.

[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 [23]. Refer to [21] and [22] for more information on CPM and Infomap respectively.

[3]Our proposed HICD method can also be applied to other OSNs by adapting the definition of $I_{i,j}$ (e.g. this method could be used in Facebook by defining $I_{i,j}$ as the number of posts a user i writes on the wall of user j).

The two methods differ 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, 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. Due to this different usage of links, the communities detected by the CICD and HICD methods may overlap but are unlikely to be a subset of one another.

In addition, we evaluate the performance of our method by analyzing the 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.

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.[4] 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.

IV. Community with Common Interests

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

Next, we retrieve the set of users following all celebrities in each category, Set $\mathcal{P}$ as described in Equation (1). 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.

[4]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.

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

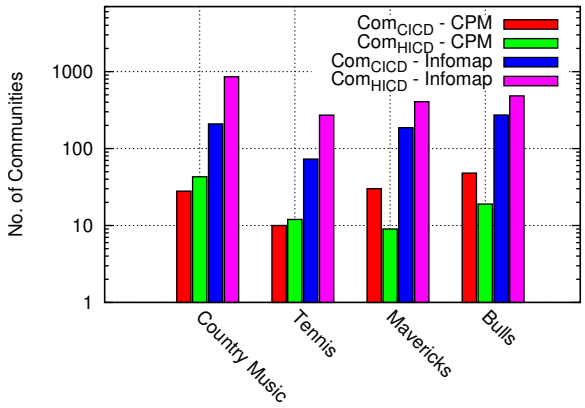

Fig. 1. Total communities detected

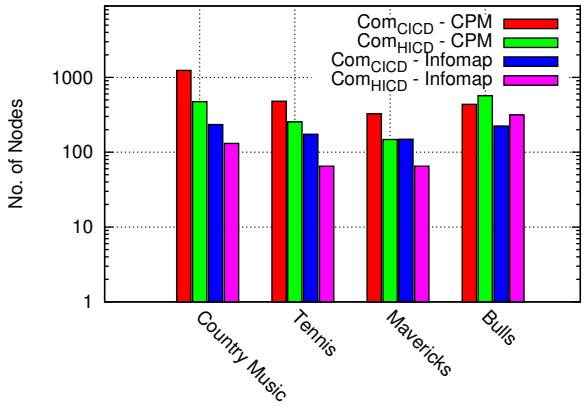

Fig. 2. Size of largest community detected

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.

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

TABLE I
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		

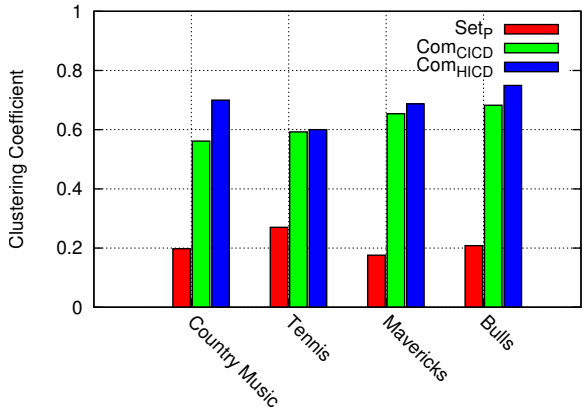

Fig. 3. Clustering coefficient

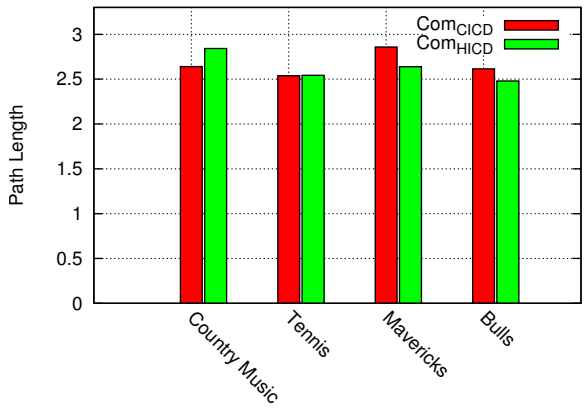

Fig. 4. Average path length

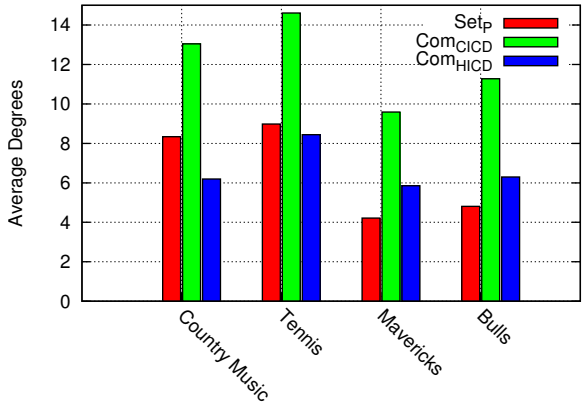

Fig. 5. Average degree

on the past 200 tweets collected on 12 Jan 12.[6] 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 by Infomap or vice versa, given the same interest category).[7]

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 now compare Set $\mathcal{P}$, Com_{CICD} and Com_{HICD} of the different categories, in terms of network characteristics to evaluate the effectiveness of our method.

Our HICD method detects communities (Com_{HICD}) that are more connected and cohesive than Set $\mathcal{P}$ and Com_{CICD} across all categories as shown in Fig. 3. Our HICD method outperforms the CICD method as indicated by a higher clustering coefficient[8] of Com_{HICD} compared to Com_{CICD}. Despite the improvement, it is challenging to achieve a clustering coefficient close to one as only a fully-connected sub-graph (i.e. a clique) has a clustering coefficient of one. The Com_{CICD} and Com_{HICD} of all categories also have a clustering coefficient two times or more than Set $\mathcal{P}$ of their respective categories.

[6]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.

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

[8]The clustering coefficient of a node is the number of 3-node cliques (which includes this node) out of the total possible number. In our experiments, we use the average clustering coefficient of all nodes in a community.

TABLE II
TOP 3 USER LOCATIONS

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

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. While Com_{HICD} of country music has a longer path length than Com_{CICD}, this is due to an $I_{i,j}$ value of 1 being chosen. Once the $I_{i,j}$ value is increased, Com_{HICD} progressively gets a shorter average path length compared to Com_{CICD} as shown in Table III. The shorter average path length and higher clustering coefficient show that our approach detects communities that are more cohesive and connected.

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.

Table II 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 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. 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!").

TABLE III
EFFECTS OF INCREASING THRESHOLD T OF $I_{i,j}$ FOR COUNTRY MUSIC CATEGORY

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

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 III. 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). 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.

A. Content of Tweet

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 this method by considering the frequency of direct communication (via the use of @mentions in tweets) between individuals. We now examine the results from our approach 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.

TABLE IV
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**

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 IV shows that among the top 10 #hashtags of Com_{HICD}, five #hashtags are related to country music (denoted by *). This result compares

TABLE VI
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

TABLE V
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

favourably to 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 category.

Likewise, our HICD method detects communities that make more @mentions of country music artists. Table V best illustrates this where eight of the top 10 @mentions of Com_{HICD} are country singers (denoted by *). Comparatively, Com_{CICD} and Set $\mathcal{P}$ has less @mentions of country music artists at a count of seven and six respectively. It is also worthwhile to note that five out of eight country singers (in the top 10 @mentions of Com_{HICD}) were not used as the initial seed of representative celebrities to construct Com_{HICD}. This observation shows that our HICD method is able to detect communities that frequently interact about country music in general, and not just about country singers in the initial seed of celebrities used. We also observed similar trends for the tennis and basketball categories.

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 VI shows the top 10 websites that Set $\mathcal{P}$, Com_{CICD} and Com_{HICD} of the country music category use in their tweets. While Set $\mathcal{P}$ and Com_{CICD} have one URL related to country music, the exchange of URLs in

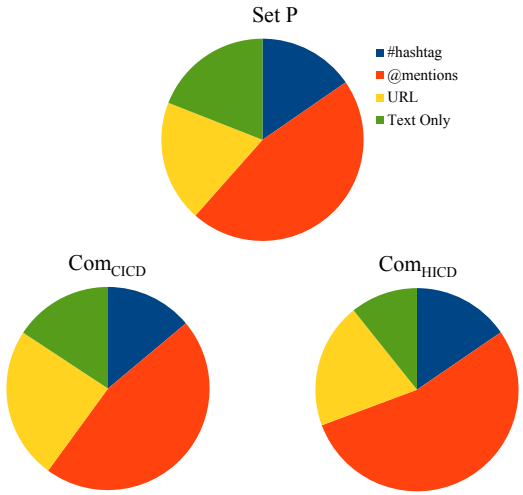

Fig. 6. Type of tweets

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 III. 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.

B. Trends in Tweeting

We investigate tweeting trend by first examining the type of content covered in the tweets posted by Set $\mathcal{P}$, Com_{CICD} and Com_{HICD}. 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 tweets. We next investigate trends in the timings of tweets.

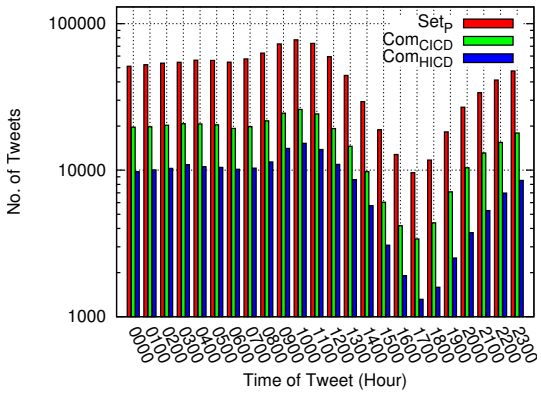

Fig. 7. Time distribution of tweets

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 increases from 1800hrs to 2300hrs. As more than 65% of Twitter users are between the age of 15 - 24 years old [24], a possible explanation is that Twitter users are either at school or work from 1200hrs to 1700hrs. Hence, they do not tweet as much during that period but tweeting activities gradually increases 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.

C. Temporal Analysis of Links

Now, we study the formation and deletion of links over time for the three groups of users: Set $\mathcal{P}$, Com_{CICD} and Com_{HICD}. We retrieved the follower list of users in these groups on four-day intervals, from 28 Nov 11 to 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

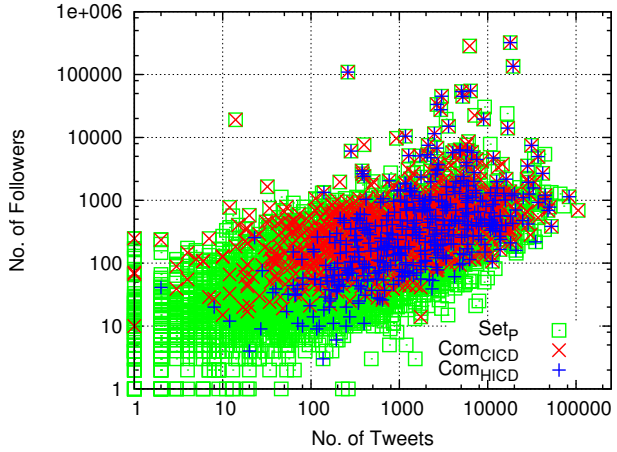

Fig. 8. Comparison of tweets to followers (Best viewed in colour)

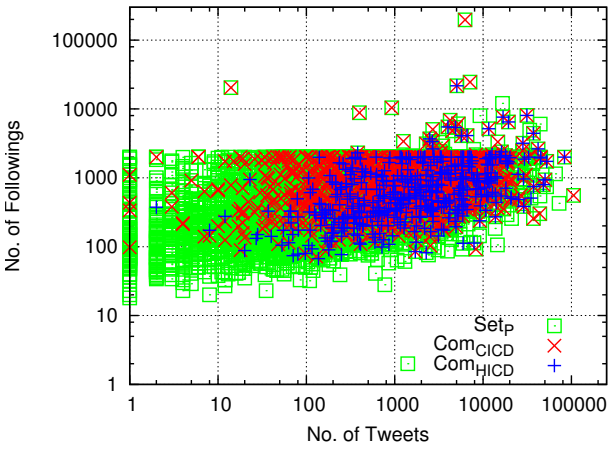

Fig. 9. Comparison of tweets to followings (Best viewed in colour)

and delete more links on average than users in Set $\mathcal{P}$ and Com_{CICD}. It is interesting to note that Com_{HICD} creates almost three times the links that it deletes whereas Set $\mathcal{P}$ creates less than two times the links that it deletes. This observation points to a trend where links in Com_{HICD} are more persistent than those in Set $\mathcal{P}$ and Com_{CICD}, as users in Com_{HICD} are less likely to unfollow another user once the following link is created. This result serves as a preliminary analysis and we plan to further investigate on the motivating factors behind a user's choice in following/unfollowing other users (e.g. similar interests, common friends, etc).

V. 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

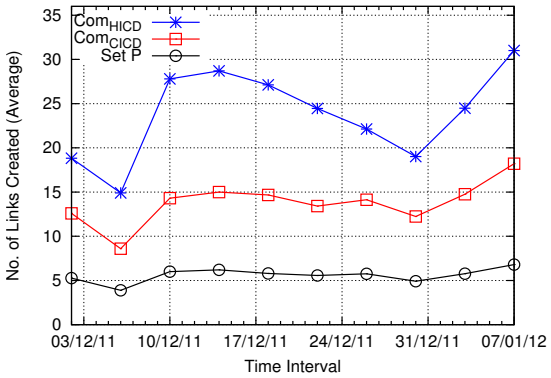

Fig. 10. Time analysis of created links

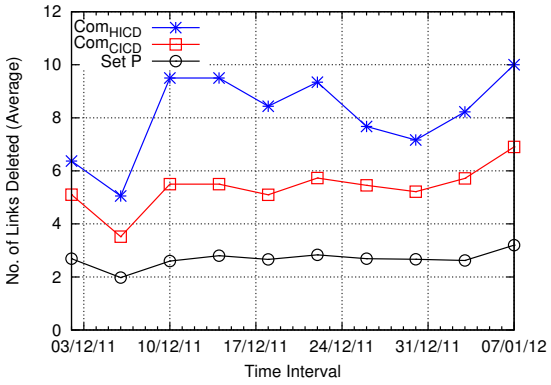

Fig. 11. Time analysis of deleted links

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.

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 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.

VI. 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] G. Iyer, D. Soberman, and J. M. Villas-Boas, "The targeting of advertising," *Marketing Science*, vol. 24, no. 3, pp. 461–476, 2005.

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

[3] A. Java, X. Song, T. Finin, and B. Tseng, "Why we Twitter: Understanding microblogging usage and communities," in *Proc. of WebKDD/SNA-KDD '07*, Aug 2007, pp. 56–65.

[4] 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 *Proc. of IMC'08*, Oct 2008, pp. 57–70.

[5] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao, "User interactions in social networks and their implications," in *Proc. of EuroSys'09*, Apr 2009, pp. 205–218.

[6] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer, "Outtweeting the Twitterers - Predicting information cascades in microblogs," in *Proc. of WOSN '10*.

[7] S. A. Macskassy and M. Michelson, "Why do people retweet? Anti-homophily wins the day!" in *Proc. of ICWSM '11*, May 2011, pp. 209–216.

[8] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su, "Understanding retweeting behaviors in social networks," in *Proc. of CIKM '10*, Oct 2010, pp. 1633–1636.

[9] 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 *Proc. of WWW '11*, Mar 2011, pp. 695–704.

[10] J. Huang, K. M. Thornton, and E. N. Efthimiadis, "Conversational tagging in Twitter," in *Proc. of HT '10*, Jun 2010, pp. 1079–1088.

[11] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proc. of WWW '11*, Mar 2011, pp. 675–684.

[12] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on Twitter," in *Proc. of ICWSM '11*, May 2011, pp. 438–441.

[13] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The Good the Bad and the OMG!" in *Proc. of ICWSM '11*, May 2011, pp. 538–541.

[14] B. V. A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in Facebook," in *Proc. of WOSN '09*, Aug 2009, pp. 37–42.

[15] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi, "Characterizing social cascades in Flickr," in *Proc. of WOSN '08*, Aug 2008, pp. 13–18.

[16] J. Yang and S. Counts, "Predicting the speed, scale, and range of information diffusion in Twitter," in *Proc. of ICWSM '10*, May 2010, pp. 355–358.

[17] H. Balakrishnan and N. Deo, "Discovering communities in complex networks," in *Proc. of ACMSE '06*, Mar 2006, pp. 280–285.

[18] N. Du, B. Wu, X. Pei, B. Wang, and L. Xu, "Community detection in large-scale social networks," in *Proc. of WebKDD/SNA-KDD '07*, pp. 16–25.

[19] K. H. Lim and A. Datta, "Following the follower: Detecting communities with common interests on Twitter," in *Proc. of HT '12*, Jun 2012, pp. 317–318.

[20] K. H. Lim and A. Datta, "Finding Twitter communities with common interests using following links of celebrities," in *Proc. of MSM '12*, Jun 2012, pp. 25–32.

[21] I. Derényi, G. Palla, and T. Vicsek, "Clique percolation in random networks," *Physical Review Letters*, vol. 94, no. 16, pp. 240–253, 2005.

[22] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proc. of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.

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

[24] S. 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.