# Finding Twitter Communities with Common Interests using Following Links of Celebrities*

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

## ABSTRACT

One important problem in target advertising and viral marketing on online social networking sites is the efficient identification of communities with common interests in large social networks. Existing methods involve large scale community detection on the entire social network before determining the interests of individuals within these communities. This approach is both computationally intensive and may result in communities without a common interest. We propose an efficient approach for detecting communities that share common interests on Twitter. Our approach involves first identifying celebrities that are representative of an interest category before detecting communities based on linkages among followers of these celebrities. We also study the characteristics of these communities and the effects of deepening or specialization of interest.

## Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and behavioral sciences

## General Terms

Theory

## Keywords

Twitter, Social Networks, Community Detection, Graph Mining

## 1. INTRODUCTION

Twitter is a popular micro-blogging service that allows messages of up to 140 characters (called tweets) to be posted and received by registered users. Tweets form the basis of social interactions in Twitter where a user is kept updated of the tweets of someone he/she is following. The popularity of Twitter is seen from its daily usage of 200 million tweets, as of 1st Aug 2011 [18]. The popularity of Twitter and availability of data have created plenty of interest in its academic study in recent years [2, 9, 15].

*This paper is an extended version of the poster paper listed in [11].


*MSM'12,* June 25, 2012, Milwaukee, Wisconsin, USA.


One important problem in the application of target advertising and viral marketing to online social networks is the efficient identification of communities with common interests in large social networks [4, 7]. Most of the current approaches involve first detecting all communities, followed by determining the interests of these communities [5, 10]. These approaches involve a lengthy and intensive process of detecting communities for the entire social network, which is growing daily. Furthermore, many of the detected communities may not share the interest we are looking for.

Our study offers a method to identify communities comprising like-minded individuals with common interests on Twitter. This method differs from existing ones that first detect all communities, followed by identifying the topics they are interested in [5, 10]. Also, our method does not unnecessarily detect communities that do not share any specific interest. Instead, our method allows for the efficient detection of only communities sharing a common interest and can be applied to target advertising and viral marketing. In addition, our method is able to detect communities at different levels of interest. As far as we are aware, there has been no prior study on the detection of communities with common interest on Twitter.

Our main contributions in this paper include the following:

- An efficient approach for detecting Twitter communities that share common interest.
- A study of the characteristics of Twitter communities that share common interests.
- An investigation into the effects of deepening or specialization of interest on these communities.

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

## 2. DESCRIPTION OF TWITTER

Twitter allows registered users to post and receive messages of up to 140 characters. These messages are called tweets and they can be posted via the Twitter website, short messaging services or third party applications. Tweets form the basis of social interactions in Twitter where a user is kept updated of the tweets of someone he/she is following. A user can also forward the tweets of others to another user, which is called retweeting. In addition, users can

@mention each other in their tweets (via @username) or #hashtag keywords or topics for easy search by others (via #topic).

Twitter also provides an Application Programming Interface (API) with the functionality to collect data such as user profiles, linkages among users, tweets, retweets and @mentions [17]. This API allows developers to create applications for Twitter and researchers to study the characteristics of an online social network from the individual to community level. Currently, there is an hourly rate limit on the number of API calls that can be executed.

# 3. RELATED WORK

Social networks have been intensively studied in recent years due to the availability and scale of online social networks. One such study resulted in the LikeMiner system which identifies popular topics on online social networks based on the explicit "likes" indicated by users [6]. LikeMiner is then able to predict the interests of a user based on the interests of his/her friends. Our approach differs from this system as we infer interest based on a user's followings instead of requiring the explicit "like" by a user. More importantly, the LikeMiner system identifies individuals whereas our approach identifies communities with common interests.

Similarly, the Friendship and Interest Propagation (FIP) model identifies interests of an individual and potential friendship links with other users [19]. FIP determines the interests of an individual user based on the interests of his/her friends and recommends friends based on those sharing similar interests. The main difference with our method is that we identify an entire community sharing a common interest whereas the FIP model identifies an individual user's interest and recommends friendships. Also, this study was conducted on Yahoo! Pulse[1] whereas ours is based on Twitter. Furthermore, interests are explicitly stated for the FIP model whereas our model implicitly infer interests based on a user's followings.

In their study of Twitter, Java et al. used the Hyperlink-Induced Topic Search algorithm to detect communities based on a set of hubs and authority, and the Clique Percolation Method to detect overlapping communities on Twitter [5]. Through tweet analysis, they found that such communities share common interest, which are further divided into formal and informal ones. The difference with our approach is that we do not detect all communities then determine their interest but rather, focus directly only on communities sharing specific interests that we are interested in.

Li et al. proposed the TTR-LDA community detection algorithm using the Latent Dirichlet Allocation model and Girvan-Newman algorithm with an inference mechanism for topic distribution [10]. They used the TTR-LDA algorithm to detect communities among the top 50,000 taggers in Delicious[2], determine interest topics of the communities and model the temporal evolution of these topics. They observed that communities share common interests which divide into defined sub-categories over time. Similar to Java et al., they detect all communities first before determining their interest. Also, their data is based on only the top users of Delicious whereas ours is based on the full dataset of Twitter.

Using BibSonomy[3], Atzmueller and Mitzlaff demonstrated an approach for mining communities with common descriptive features [1]. This approach integrates a database (of user attributes) and topological graph (of user links) into a dataset comprising only links connecting two users with the same attribute. Communities are then detected based on the desired attribute using this new collection of links. While this approach can be applied to detect communities with common interest, our method is able to detect communities with varying levels of interest. Furthermore, our method implicitly infer a user's interests based on his/her followings while Atzmueller and Mitzlaff build user attributes using explicit tags on BibSonomy.

[1]http://pulse.yahoo.com
[2]http://delicious.com/
[3]http://www.bibsonomy.org/

# 4. DATASET AND METHODS

The Twitter dataset collected by Kwak et al. [9] is used for our experimentations. This dataset was collected from 6th to 31st June 2009, comprising 41.7 million Twitter users, 1.47 billion links, and the profiles of users with more than 10,000 followers. Kwak et al. have made the dataset publicly available at [8].

We model the Twitter social network as a directed graph, $G = (U, L)$ where $U$ refers to the set of users and $L$ refers to the set of links. A followership link $(i, j) \in L$ indicates that user $i \in U$ is a follower of user $j \in U$, while a friendship link $Fr_{i,j}$ indicates $(i, j) = (j, i)$. We classify a Twitter user as a celebrity if he/she has more than 10,000 followers.

The interest of a user in a category $cat$, $Int_{cat}$ is inferred by the number of celebrities (of category $cat$) that the user follows. Although $Int_{cat}$ represents the interest level of a user in a category, this metric is subjective due to the celebrities selected. The accuracy of $Int_{cat}$ is dependent on the correct classification of celebrities into their respective categories, which is subjective as some celebrities loosely belong to multiple categories (e.g. a singer that has starred in some movies). We minimize this subjective judgment by using information on Wikipedia[4] to classify these celebrities into their respective categories. On the Wikipedia page of a celebrity, there is an "occupation" field which we use to determine the categories this celebrity belong to. Thus, this process minimizes the chances of classifying celebrities into the wrong category.

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

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

$\mathcal{P}$ is the set of fans who follow all the $k$ celebrities in the set $\bigcup c_j, for\ 1 \leqslant j \leqslant k$. We consider only friendship links (among Set $\mathcal{P}$) for community detection as friendship links are stronger and more reflective of real-life interactions. Using this set of friendship links (which corresponds to an undirected graph), we try to detect communities among the members of $\mathcal{P}$ next using the Clique Percolation Method (CPM) developed by Palla et al. [14]. The CPM defines a community as one with a series of adjacent $k$-cliques, where a $k$-clique comprises $k$ nodes that are interconnected. We first identify all $k$-cliques in the network and connect them if they are adjacent. Two $k$-cliques are adjacent if they share $(k-1)$ common nodes. This procedure of connecting $k$-cliques continues iteratively until no adjacent $k$-cliques can be found. The result is a series of communities formed based on the $k$-cliques and adjacency

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

criteria. For our experiments, we use CPM with a $k$-value of 3 as this produces the best results in detecting communities compared to other $k$-values.

Similarly, we also detect communities among the members of $\mathcal{P}$ next using the Infomap algorithm by Rosvall and Bergstrom [16]. Infomap approaches community detection as a coding or compression problem where the network graph can be compressed to retain its key structures. These key structures represent communities or clusters that are found within the network graph. Infomap uses random walks on the network graph to analyze information flow where the random walker is more likely to traverse within a cluster of nodes belonging to the same community. Using both CPM and Infomap show that our proposed method produces results that are independent of the chosen community detection algorithm and their unique characteristics.

We first study community detection and structure among individuals with a common interest in Section 5. We infer the interest of users based on the celebrities followed as users are unable to explicitly state their interests in Twitter. For this purpose, we identified six celebrities for each interest category, resulting in a total of 30 celebrities covering five categories. As a control group, we randomly chose 200,858 users to represent the group with no shared interest. This control group allows us to compare the community structure of users with no common interest against users with a shared interest.

Next, we further examine how the deepening and specialization of interest affects community structure in Section 6. For this purpose, we compare communities with varying levels of interest in the specialized Country Music category against the general Music category. We selected seven winners of the Country Music Awards[5] from 2001 to 2008 as celebrities for the Country Music category based on their number of followers. Winners from 2009 onwards were not selected as the Twitter dataset only comprises data until 31st June 2009. The control group chosen is the users interested in the Music category described in the previous paragraph.

# 5. INVESTIGATING COMMON INTERESTS

The Merriam-Webster dictionary defines a community as "a group of people with a common characteristic or interest living together within a larger society" [12]. Building on this definition, we propose a community detection approach based on individuals sharing common interests. We evaluate our approach by comparing the detected communities (with common interest) to our control group comprising communities with no common interest. This comparison shows that our approach of community detection based on common interests results in larger and more cohesive communities.

For our study, we selected Film & TV, Music, Hosting, News and Blogging as categories of interest due to their popularity. These categories are selected by first identifying the top 100 celebrities based on their number of followers. Next, we used information on Google[6] and Wikipedia to determine the various categories these celebrities belong to. Following which, we build a list of categories based on the frequency of celebrities belonging to a category. Fig. 1 shows the popular categories in Twitter and we selected the five most popular categories among them.[7] For each category, we selected the six most popular celebrities based on their number of followers as listed in Table 1.[8] Also, a celebrity may belong to multiple categories (e.g. Miley Cyrus belongs to both the Music and Film & TV categories).

[5]http://cmaawards.cmaworld.com/nominees/view-past-winners
[6]http://www.google.com/
[7]Some categories were not included due to the diversity of content within these categories (e.g. Online Commerce)

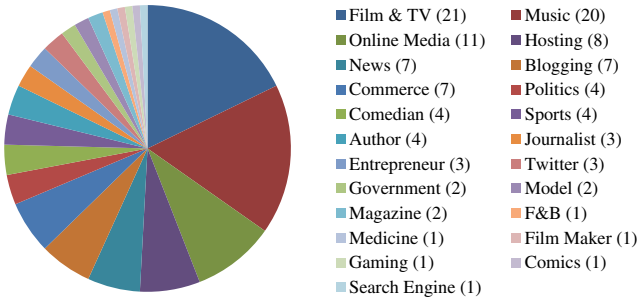


**Figure 1: Popular Categories on Twitter**

**Table 1: Twitter Celebrities**

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

## 5.1 Communities with Common Interests

The next step of our community detection approach involves identifying individuals with common interests, where the interest of a user $Int_{cat}$ is derived from the number of celebrities of category $cat$ followed by the user. We now retrieve the list of users with $Int_{cat} > 1$, for $cat \in \{Film\&TV, Music, Hosting, News, Blogging\}$. A summary of users with $Int_{cat} > 1$ is shown in

[8]Choosing six celebrities gives us an optimal number of followers. While choosing a higher number of celebrities results in users with a higher level of interest, it also results in less number of followers.

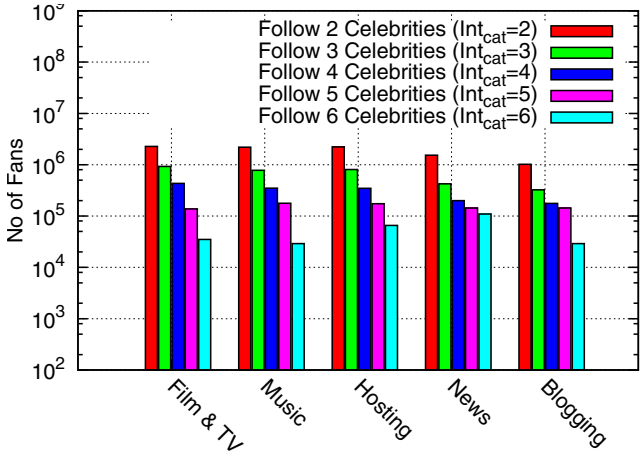


**Figure 2: Fans Following Multiple Celebrities in a Category**

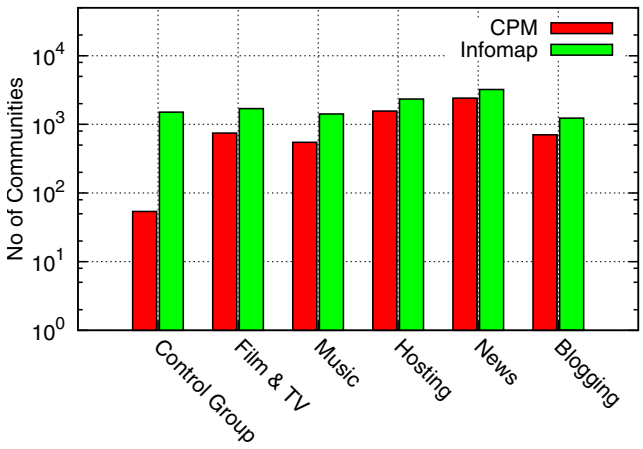


**Figure 3: Total Communities Detected**

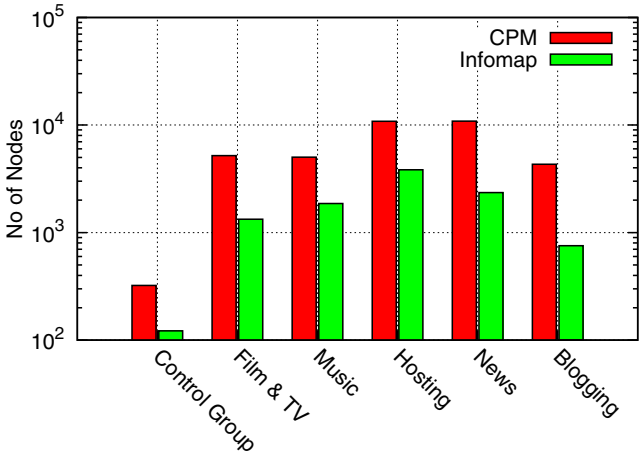


**Figure 4: Size of Largest Community Detected**

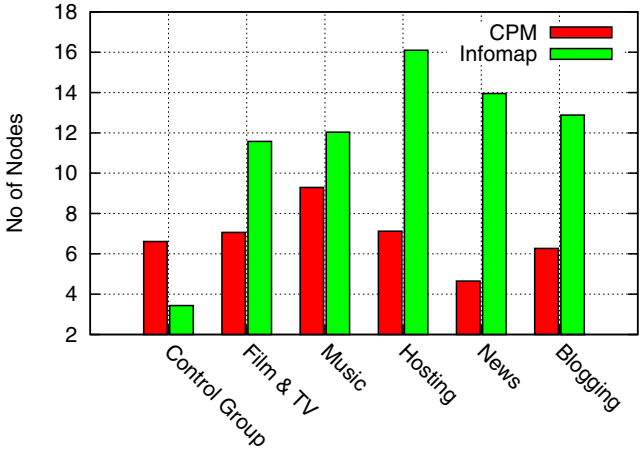


**Figure 5: Average Size of Communities Detected**

Fig. 2. In particular, we are interested in users with $Int_{cat} = 6$ as this indicates the most interest in a given category.

**Table 2: Reciprocality Among Interest Groups**

| Category | Reciprocality |
|---|---|
| Film & TV | 17.9% |
| Music | 18.2% |
| Hosting | 15.0% |
| News | 17.3% |
| Blogging | 19.6% |

We now examine reciprocality based on link information among users with $Int_{cat} = 6$, for $cat \in \{Film\&TV, Music, Hosting, News, Blogging\}$, as shown in Table 2. Reciprocality is obtained based on the number of friendship links out of all links. The reciprocality of 15.0% to 19.6% across all categories corresponds to observations by Cha et al. and Kwak et al. of 10% and 22% respectively for the entire Twitter population [2, 9]. This shows that reciprocality among users with common interests is similar to reciprocality among the general population.

Next, we use the CPM and Infomap to detect communities among users with $Int_{cat} = 6$, for $cat \in \{Film\&TV, Music, Hosting, News, Blogging\}$. Similarly, we detect communities among our control group comprising users with no common interest. We now compare the communities with common interests against the control group (i.e. community with no common interest) in terms of the total number of communities, size of largest community, and average community size as shown in Fig. 3, 4 and 5 respectively.

Fig. 3 and 4 show that users with common interests form more and larger communities than users without a common interest in the control group, regardless of whether CPM or Infomap was used. This is also despite the fact that the control group has a larger population of 200,858 users compared to users with a common interest, which ranges from 29,092 users ($Int_{Music} = 6$) to 109,779 users ($Int_{News} = 6$). Similarly, users with common interests form larger communities on average as shown in Fig. 5. The exception is the News category detected using CPM as many cliques of three nodes were detected as communities thus decreasing the average community size. However, our focus is on the largest community detected as this provides the most benefit for any application of target advertising and viral marketing.

The $k$-value chosen for CPM affects the number and size of communities detected but in all cases, we detect larger and more communities for users with a common interest compared to users without a common interest (given the same $k$-values). We were able to

detect communities at $k$-values of up to 25 for the News category and we could also detect communities at $k$-values of 9 or higher for the other categories. For the control group, we were unable to detect any communities at $k$-values higher than 6 which further proves that users with common interest form larger and more communities than users with no common interest. While the $k$-values affects community detection, this observation shows that our approach performs better than the control experiment given the same $k$-values.

Users with common interests also form communities that are more cohesive than those without common interest. Table 3 shows this trend where the communities with common interest have a higher clustering coefficient than our control group with no common interest, except the Hosting and News categories. However, users interested in Hosting and News have a higher average degree of links which shows that these users are better connected than users in the control group.

These results show that our community detection approach finds communities that are both larger and more cohesive. More importantly, our approach efficiently detects communities with common interests without the need to perform large scale community detection on the entire social network. Thus, our approach is less computationally intensive and compares favourably to existing approaches that detect all communities then identify the interests of the communities [5, 10]. These results are also supported by observations of other authors that people with similar interests are more likely to be friends than those with dissimilar interests [3, 19].

# 6. SPECIALIZATION AND DEEPENING OF INTERESTS

Communities that share the same set of interests are likely to be more connected [10, 20] and interact on a more frequent basis [15]. As an extension of that argument, we show that users sharing a specialized interest form a more tightly-coupled community than users sharing a general interest. We show this by comparing users interested in the specialized category of Country Music against users interested in the general category of Music. The control group is the users interested in general Music category as discussed in Section 5. The celebrities representing the Country Music category are seven Country Music singers who have won various awards at the Country Music Awards between 2001 to 2008 and have more than 10,000 followers. These celebrities (representing the Country Music category) are listed in Table 4.

**Table 4: Country Music Celebrities**

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

Similar to Section 5, we used both CPM and Infomap to detect communities among users with $Int_{Country} > 1$.[9] Due to the smaller population of users following Country Music singers, the absolute number of communities detected by CPM are small (e.g. only 230 users with $Int_{Country} = 7$). We first focus on users with the most interest in Country Music, $Int_{Country} = 7$. For this user group, we detected five communities comprising 23 distinct users as shown in Fig. 6. The five communities are differentiated by nodes that are coloured green, orange, blue, yellow and purple. The grey nodes represent users that belong to multiple communities and serve as middlemen connecting the various communities. We also observed similar trends in the communities detected by Infomap.

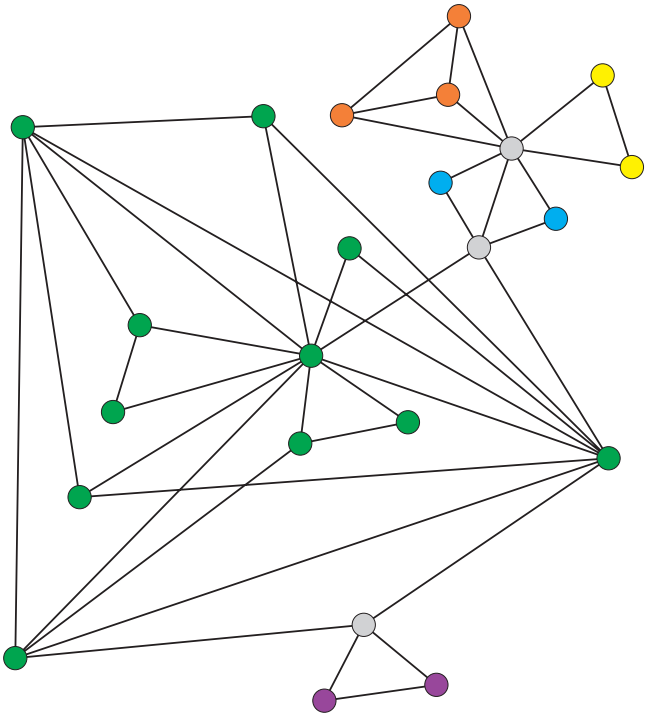

**Figure 6: Community Graph of Fans who follow all Seven Country Singers**

## 6.1 Effects of Interest Specialization

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

The average size of communities indicates the likeliness of large communities being formed among users with common interests. This allows us to compare if users with specialized interests form larger communities than users with a general interest. Comparing two user groups with the same level of interest in different categories (i.e. $Int_{Music} = 6$ and $Int_{Country} = 6$), we observe that the normalized average community size of the $Int_{Country} = 6$

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

**Table 3: Network Statistics of the Communities**

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

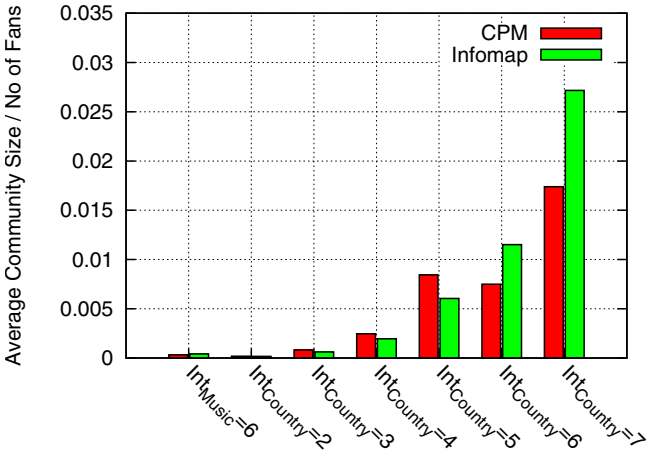


**Figure 7: Normalized Average Community Size for Music and Country Music Categories**

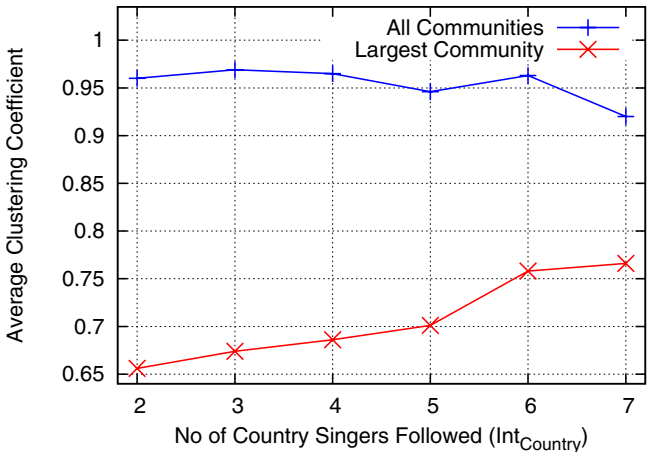


**Figure 8: Average Clustering Coefficient of Country Music Category**

group is 23 and 28 times larger than the $Int_{Music} = 6$ group using CPM and Infomap respectively, as shown in Fig. 7. This result shows that users sharing the same level of interest form larger communities if that interest is more specialized.

Even among users with a lower level of interest in a specialized category, they are more likely to form larger communities on average compared to users with a higher level of interest in a general category. Fig. 7 shows that users with a lower interest in the specialized Country Music category ($Int_{Country} = 3$) have a normalized average community size that is up to two times larger than that of users with more interest in the general Music category ($Int_{Music} = 6$).

**Table 5: Comparison of General and Specialized Interest**

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

Communities comprising users with a specialized interest are also more cohesive and well-connected than those with a more general interest. Table 5 best illustrates this where users with a specialized interest in Country Music form communities with a shorter average path length and diameter but higher clustering coefficient compared to those with a general interest in Music. In addition, users with $Int_{Country} = 6$ displayed a higher reciprocality of 20.1% compared to 18.2% for users with $Int_{Music} = 6$.

## 6.2 Effects of Interest Deepening

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

An increase in interest level among users corresponds to an increase in their average community size. Fig. 7 shows an increasing average community size with increasing $Int_{Country}$ values. This result supports our original observation that communities are more likely to be formed among like-minded individuals. In addition, the average size and number of communities formed increases as the interest level of the users increases.

Communities comprising users with a common interest get more tightly coupled as their level of interest increases. Fig. 8 shows a gradual increase in clustering coefficient among the largest communities with increasing $Int_{Country}$ values. While the average clustering coefficient of all communities remains relatively constant (from $Int_{Country}$ = 2 to $Int_{Country}$ = 6), this is due to the large number of small cliques detected at low $Int_{Country}$ values which increases the average clustering coefficient significantly. For example, out of 539 communities detected (with $Int_{Country} = 2$), 397 communities are cliques of three users thus having a clustering coefficient of one. At higher $Int_{Country}$ values, less of such cliques are detected thus they have less influence on the average clustering coefficient. We are most interested in the largest community (which shows an increasing clustering coefficient) as this

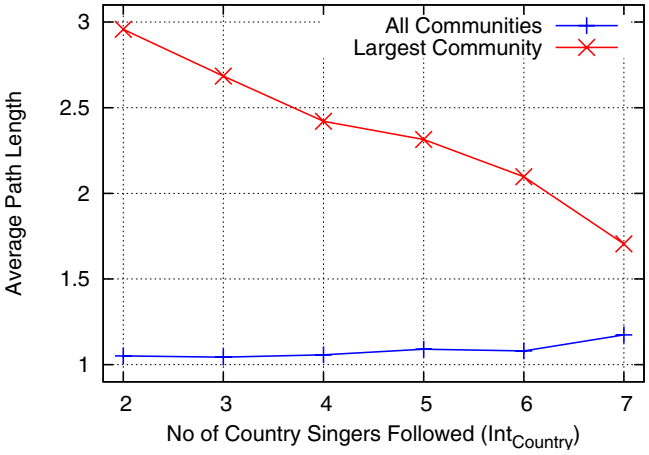


**Figure 9: Average Path Length of Country Music Category**

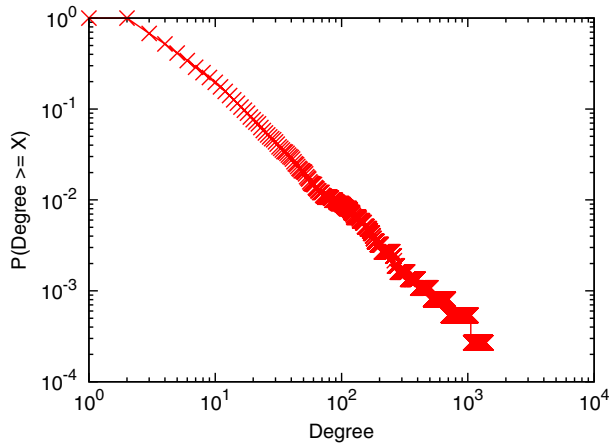


**Figure 10: Degree Complementary Cumulative Distribution Function of Largest Community with $Int_{Country} = 2$**

community has the most potential for target advertising and viral marketing due to its size and cohesiveness.

Fig. 9 shows an average path length of 1.7 to 3.0 hops within the largest communities at varying values of $Int_{Country}$, illustrating that users sharing common interests form communities that are better connected. This compares well with Milgram's "six degrees of separation" which states that everyone is connected by six hops of acquaintances [13]. Although we compare average path length of communities and not the entire population, the largest community for $Int_{Country} = 2$ comprising 3,725 users still shows a short average path length of three hops.

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

The detected communities also display the characteristics of scale-free networks as shown in Fig. 10 and 11, which plots the Complementary Cumulative Distribution Function of the degree distribu-

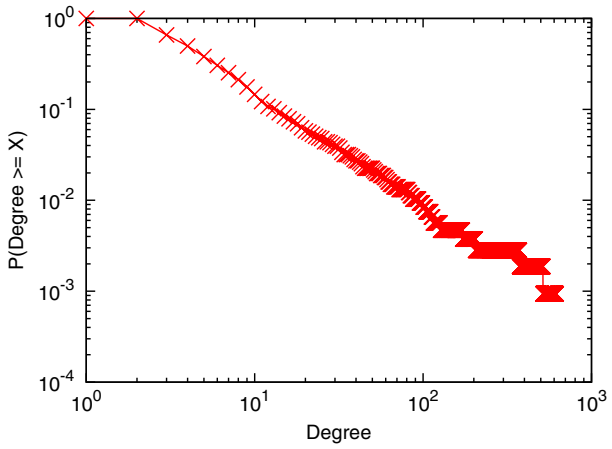


**Figure 11: Degree Complementary Cumulative Distribution Function of Largest Community with $Int_{Country} = 4$**

tion of users with $Int_{Country} = 2$ and $Int_{Country} = 4$ respectively. The communities with other $Int_{Country}$ values also displayed similar trends. Upon closer examination, we observe that many individuals with large degree distribution are also country music artists but with less fans than the celebrities we have chosen. The fact that there are other minor country singers among these communities shows that our method effectively detects communities comprising users with a common interest. Using the Twitter API [17], we retrieved the profiles of 1,164 users (with $Int_{Country} = 2$), the remaining user profiles could not be retrieved due to locked or inactive accounts. Examining the retrieved user profiles, we observed that more than 7.7% of these users are from Nashville, Tennessee, a town that is closely associated with country music and hosts the annual Country Music Association Music Festival.

## 7. CONCLUSION

In this paper, we proposed a method to efficiently detect communities comprising individuals with common interests for applications in target advertising and viral marketing. Our method was not developed to detect all communities on Twitter. Instead, it detects communities that are larger, more cohesive and only comprise users that share a common interest. As Twitter has no explicit options for users to state their interest, we derived a measurement of interest based on the number of celebrities in an interest category that the user follows. Given the large scale and growth rate of Twitter, our method is very scalable for identifying communities sharing common interests as it only requires topological information.

In addition, this method can also be applied to other online social networking sites by adapting to the unique characteristics of each site and their representations of celebrities and links. For example, in Facebook[10], celebrities could be defined as the respective Facebook pages of these celebrities and followership links as the individual user "likes" on these pages. Thereafter, our method could be applied as described in the paper using these Facebook pages (celebrities) and user "likes" (followership links).

From a sociology perspective, we also studied the characteristics among users with a common interest compared to users without a shared interest, particularly in the way they form communities and the structure of these communities. Also, we observed how

[10]www.facebook.com

their community structures become more connected and cohesive with deepening interest in a given category, as indicated by an increasing clustering coefficient and decreasing path length. These observations along with our proposed method of community detection provide a tool for the implementation of target advertising or viral marketing, especially for products with a niche or specialized audience.

Some future areas that we are working on include the geographical analysis of communities comprising like-minded individuals and temporal analysis to determine the evolution of community structure, specifically the trends of individuals joining and leaving communities. Also, we are working on an automated system where the process of selecting and classifying celebrities into their respective categories is automated using information from Wikipedia. This automated process would overcome our method's main limitation, which is the need to manually select and classify celebrities into their respective categories.

## 8. ACKNOWLEDGMENTS

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

## 9. REFERENCES

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

[2] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in Twitter: The million follower fallacy. In *ICWSM '10: Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 10–17, May 2010.

[3] T. L. Fond and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In *WWW '10: Proceedings of the 19th International Conference on World Wide Web*, pages 601–610, Apr 2010.

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

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

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

[7] 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, 2011.

[8] H. Kwak, C. Lee, H. Park, and S. Moon. Twitter dataset. Internet, Jun 2009. Available from: http://an.kaist.ac.kr/traces/WWW2010.html.

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

[10] D. Li, B. He, Y. Ding, J. Tang, C. Sugimoto, Z. Qin, E. Yan, J. Li, and T. Dong. Community-based topic modeling for social tagging. In *CIKM '10: Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1565–1568, Oct 2010.

[11] K. H. Lim and A. Datta. Following the follower: Detecting communities with common interests on Twitter. In *HT '12: Proceedings of the 23th ACM Conference on Hypertext and Social Media*, page to appear, Jun 2012.

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

[13] S. Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

[14] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818, Jun 2005.

[15] B. Poblete, R. Garcia, M. Mendoza, and A. Jaimes. Do all birds Tweet the same? Characterizing Twitter around the world. In *CIKM '11: Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 1025–1030, Oct 2011.

[16] 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, 2008.

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

[18] Twitter. Twitter Blog - Your world, more connected. Internet, Aug 2011. Available from: http://blog.twitter.com/2011/08/your-world-more-connected.html.

[19] S. H. Yang, B. Long, A. Smola, N. Sadagopan, Z. Zheng, and H. Zha. Like like alike - Joint friendship and interest propagation in social networks. In *WWW '11: Proceedings of the 20th International Conference on World Wide Web*, pages 537–546, Mar 2011.

[20] D. Zhao and M. B. Rosson. How and why people Twitter: The role that micro-blogging plays in informal communication at work. In *GROUP '09: Proceedings of the ACM 2009 International Conference on Supporting Group Work*, pages 243–252, May 2009.