# Why We Twitter: Understanding Microblogging Usage and Communities

Akshay Java
University of Maryland Baltimore County
1000 Hilltop Circle
Baltimore, MD 21250, USA
aks1@cs.umbc.edu

Xiaodan Song
NEC Laboratories America
10080 N. Wolfe Road, SW3-350
Cupertino, CA 95014, USA
xiaodan@sv.nec-labs.com

Tim Finin
University of Maryland Baltimore County
1000 Hilltop Circle
Baltimore, MD 21250, USA
finin@cs.umbc.edu

Belle Tseng
NEC Laboratories America
10080 N. Wolfe Road, SW3-350
Cupertino, CA 95014, USA
belle@sv.nec-labs.com

## ABSTRACT

Microblogging is a new form of communication in which users can describe their current status in short posts distributed by instant messages, mobile phones, email or the Web. Twitter, a popular microblogging tool has seen a lot of growth since it launched in October, 2006. In this paper, we present our observations of the microblogging phenomena by studying the topological and geographical properties of Twitter's social network. We find that people use microblogging to talk about their daily activities and to seek or share information. Finally, we analyze the user intentions associated at a community level and show how users with similar intentions connect with each other.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval - Information Filtering; J.4 [**Computer Applications**]: Social and Behavioral Sciences - Economics

## General Terms

Social Network Analysis, User Intent, Microblogging, Social Media

## 1. INTRODUCTION

Microblogging is a relatively new phenomenon defined as *"a form of blogging that lets you write brief text updates (usually less than 200 characters) about your life on the go and send them to friends and interested observers via text messaging, instant messaging (IM), email or the web."* [1]. It is provided by several services including Twitter[2], Jaiku[3] and more recently Pownce[4]. These tools provide a light-weight, easy form of communication that enables users to broadcast and share information about their activities, opinions and status. One of the popular microblogging platforms is Twitter [29]. According to ComScore, within eight months of its launch, Twitter had about 94,000 users as of April, 2007 [9]. Figure 1 shows a snapshot of the first author's Twitter homepage. Updates or posts are made by succinctly describing one's current status within a limit of 140 characters. Topics range from daily life to current events, news stories, and other interests. IM tools including Gtalk, Yahoo and MSN have features that allow users to share their current status with friends on their buddy lists. Microblogging tools facilitate easily sharing status messages either publicly or within a social network.


**Figure 1: An example Twitter homepage with updates talking about daily experiences and personal interests.**

[1] http://en.wikipedia.org/wiki/Micro-blogging

[2] http://www.twitter.com
[3] http://www.jaiku.com
[4] http://www.pownce.com


Joint 9th WEBKDD and 1st SNA-KDD Workshop '07 , August 12, 2007 , San Jose, California , USA . Copyright 2007 ACM 1-59593-444-8...$5.00.

Compared to regular blogging, microblogging fulfills a need for an even faster mode of communication. By encouraging shorter posts, it lowers users' requirement of time and thought investment for content generation. This is also one of its main differentiating factors from blogging in general. The second important difference is the frequency of update. On average, a prolific bloger may update her blog once every few days; on the other hand a microblogger may post several updates in a single day.

With the recent popularity of Twitter and similar microblogging systems, it is important to understand *why* and *how* people use these tools. Understanding this will help us evolve the microblogging idea and improve both microblogging client and infrastructure software. We tackle this problem by studying the microblogging phenomena and analyzing different types of user intentions in such systems.

Much of research in user intention detection has focused on understanding the intent of a search queries. According to Broder [5], the three main categories of search queries are navigational, informational and transactional. Understanding the intention for a search query is very different from user intention for content creation. In a survey of bloggers, Nardi et al. [26] describe different motivations for "why we blog". Their findings indicate that blogs are used as a tool to share daily experiences, opinions and commentary. Based on their interviews, they also describe how bloggers form communities online that may support different social groups in real world. Lento et al. [21] examined the importance of social relationship in determining if users would remain active in a blogging tool called Wallop. A user's retention and interest in blogging could be predicted by the comments received and continued relationship with other active members of the community. Users who are invited by people with whom they share pre-exiting social relationships tend to stay longer and active in the network. Moreover, certain communities were found to have a greater retention rate due to existence of such relationships. Mutual awareness in a social network has been found effective in discovering communities [23].

In computational linguists, researchers have studied the problem of recognizing the communicative intentions that underlie utterances in dialog systems and spoken language interfaces. The foundations of this work go back to Austin [2], Stawson [32] and Grice [14]. Grosz [15] and Allen [1] carried out classic studies in analyzing the dialogues between people and between people and computers in cooperative task oriented environments. More recently, Matsubara [24] has applied intention recognition to improve the performance of automobile-based spoken dialog system. While their work focusses on the analysis of ongoing dialogs between two agents in a fairly well defined domain, studying user intention in Web-based systems requires looking at both the content and link structure.

In this paper, we describe how users have adopted a specific microblogging platform, Twitter. Microblogging is relatively nascent, and to the best of our knowledge, no large scale studies have been done on this form of communication and information sharing. We study the topological and geographical structure of Twitter's social network and attempt to understand the user intentions and community structure in microblogging. From our analysis, we find that the main types of user intentions are: daily chatter, conversations, sharing information and reporting news. Furthermore, users play different roles of information source, friends or information seeker in different communities.

The paper is organized as follows: in Section 2, we describe the dataset and some of the properties of the underlying social network of Twitter users. Section 3 provides an analysis of Twitter's social network and its spread across geographies. Next, in Section 4 we describe aggregate user behavior and community level user intentions. Section 5 provides a taxonomy of user intentions. Finally, we summarize our findings and conclude with Section 6.

## 2. DATASET DESCRIPTION

Twitter is currently one of the most popular microblogging platforms. Users interact with this system by either using a Web interface, IM agent or sending SMS updates. Members may choose to make their updates public or available only to friends. If user's profile is made public, her updates appear in a "public timeline" of recent updates. The dataset used in this study was created by monitoring this public timeline for a period of two months starting from April 01, 2007 to May 30, 2007. A set of recent updates were fetched once every 30 seconds. There are a total of 1,348,543 posts from 76,177 distinct users in this collection.

Twitter allows a user, A, to "follow" updates from other members who are added as "friends". An individual who is not a friend of user A but "follows" her updates is known as a "follower". Thus friendships can either be reciprocated or one-way. By using the Twitter developer API[5], we fetched the social network of all users. We construct a directed graph $G(V, E)$, where $V$ represents a set of users and $E$ represents the set of "friend" relations. A directed edge $e$ exists between two users $u$ and $v$ if user $u$ declares $v$ as a friend. There are a total of 87,897 distinct nodes with 829,053 friend relation between them. There are more nodes in this graph due to the fact that some users discovered though the link structure do not have any posts during the duration in which the data was collected. For each user, we also obtained their profile information and mapped their location to a geographic coordinate, details of which are provided in the following section.

## 3. MICROBLOGGING IN TWITTER

This section describes some of the characteristic properties of Twitter's Social Network including it's network topology and geographical distribution.

### 3.1 Growth of Twitter

Since Twitter provides a sequential user and post identifier, we can estimate the growth rate of Twitter. Figure 2 shows the growth rate for users and Figure 3 shows the growth rate for posts in this collection. Since, we do not have access to historical data, we can only observe its growth for a two month time period. For each day we identify the maximum value for the user identifier and post identifier as provided

[5]http://twitter.com/help/api

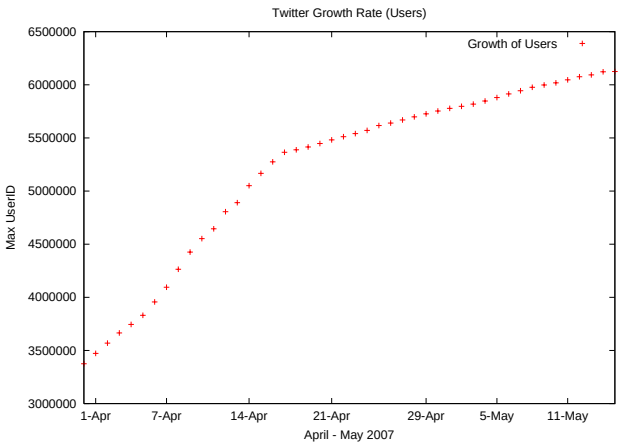



**Figure 2: Twitter User Growth Rate. Figure shows the maximum userid observed for each day in the dataset. After an initial period of interest around March 2007, the rate at which new users are joining Twitter has slowed.**

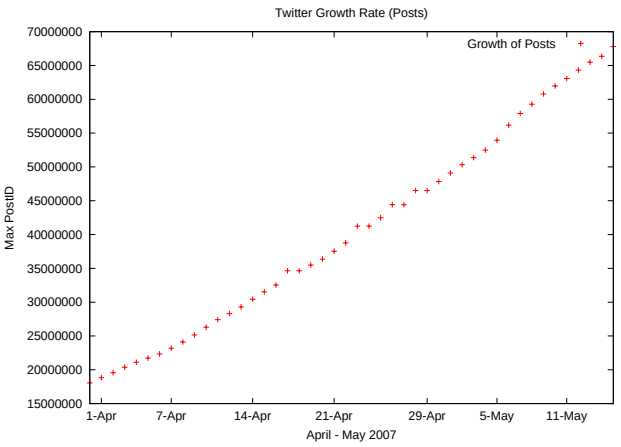



**Figure 3: Twitter Posts Growth Rate. Figure shows the maximum post ID observed for each day in the dataset. Although the rate at which new users are joining the network has slowed, the number of posts are increasing at a steady rate.**

by the Twitter API. By observing the change in these values, we can roughly estimate the growth of Twitter. It is interesting to note that even though Twitter launched in 2006, it really became popular soon after it won the South by SouthWest (SXSW) conference Web Awards[6] in March, 2007. Figure 2 shows the initial growth in users as a result of interest and publicity that Twitter generated at this conference. After this period, the rate at which new users are joining the network has slowed. Despite the slow down, the number of new posts is constantly growing, approximately doubling every month indicating a steady base of users generating content.

Following Kolari et al. [18], we use the following definition of user activity and retention:

**Definition** *A user is considered active during a week if he or she has posted at least one post during that week.*

**Definition** *An active user is considered retained for the given week, if he or she reposts at least once in the following X weeks.*

Due to the short time period for which the data is available and the nature of Microblogging we decided to use X as a period of one week. Figure 4 shows the user activity and retention for the duration of the data. About half of the users are active and of these half of them repost in the following week. There is a lower activity recorded during the last week of the data due to the fact that updates from the public timeline are not available for two days during this period.

## 3.2 Network Properties

The Web, blogosphere, online social networks and human contact networks all belong to a class of "scale-free networks" [3] and exhibit a "small world phenomenon" [33]. It has been shown that many properties including the degree distributions on the Web follow a power law distribution [19, 6]. Recent studies have confirmed that some of these properties also hold true for the blogosphere [31].

| Property | Twitter | WWE |
|---|---|---|
| Total Nodes | 87897 | 143,736 |
| Total Links | 829247 | 707,761 |
| Average Degree | 18.86 | 4.924 |
| Indegree Slope | -2.4 | -2.38 |
| Outdegree Slope | -2.4 | NA |
| Degree correlation | 0.59 | NA |
| Diameter | 6 | 12 |
| Largest WCC size | 81769 | 107,916 |
| Largest SCC size | 42900 | 13,393 |
| Clustering Coefficient | 0.106 | 0.0632 |
| Reciprocity | 0.58 | 0.0329 |

**Table 1: Twitter Social Network Statistics**

Table 1 describes some of the properties for Twitter's social network. We also compare these properties with the corresponding values for the Weblogging Ecosystems Workshop (WWE) collection [4] as reported by Shi et al. [31]. Their study shows a network with high degree correlation (also shown in Figure 6) and high reciprocity. This implies that there are a large number of mutual acquaintances in the graph. New Twitter users often initially join the network on invitation from friends. Further, new friends are added to the network by browsing through user profiles and adding other known acquaintances. High reciprocal links has also been observed in other online social networks like Livejournal [22]. Personal communication and contact network such as cell phone call graphs [25] also have high degree correlation. Figure 5 shows the cumulative degree distributions [27, 8] of Twitter's network. It is interesting to note that the slopes $\gamma_{in}$ and $\gamma_{out}$ are both approximately -2.4. This value for the power law exponent is similar to that found for

[6]http://2007.sxsw.com/

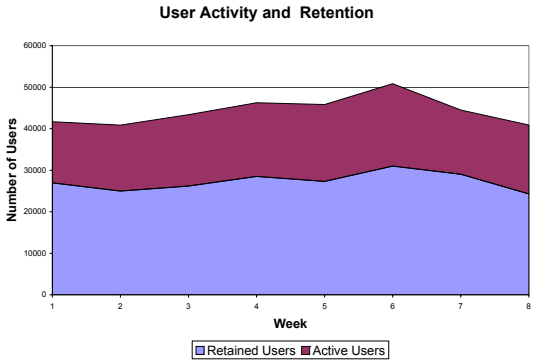



**Figure 4: Twitter User Activity and Retention**

| Continent | Number of Users |
|---|---|
| North America | 21064 |
| Europe | 7442 |
| Asia | 6753 |
| Oceania | 910 |
| South America | 816 |
| Africa | 120 |
| Others | 78 |
| Unknown | 38994 |

**Table 2: Table shows the geographical distribution of Twitter users. North America, Europe and Asia have the highest adoption of Twitter.**

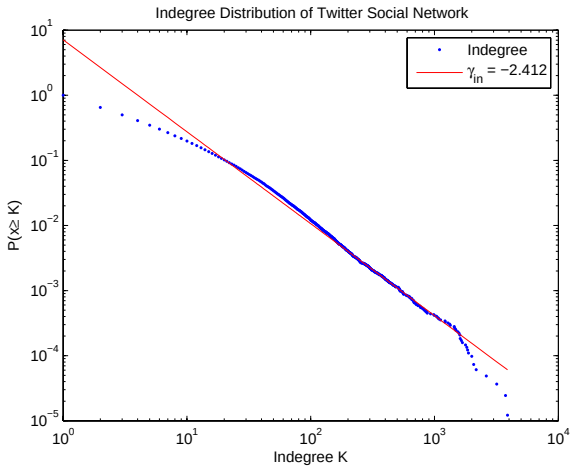
the Web (typically -2.1 for indegree [11]) and blogosphere (-2.38 for the WWE collection).

## 3.3 Geographical Distribution

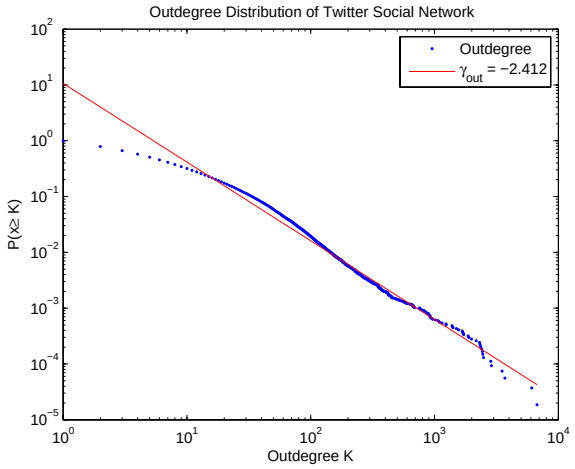
Twitter provides limited profile information such as name, bio, timezone and location. For the 76K users in our collection about 39K had specified locations that could be parsed correctly and resolved to their respective latitude and longitudinal coordinates (using Yahoo! Geocoding API[7] ). Figure 7 and Table 2 shows the geographical distribution of Twitter users and the number of users in each continent. Twitter is most popular in US, Europe and Asia (mainly Japan). Tokyo, New York and San Francisco are the major cities where user adoption of Twitter is high [16].

Twitter's popularity is global and the social network of its users crosses continental boundaries. By mapping each user's latitude and longitude to a continent location we can extract the origin and destination location for every edge. Table 3 shows the distribution of friendship relations across major continents represented in the dataset. Oceana is used to represent Australia, New Zealand and other island nations. A significant portion (about 45%) of the Social Network still lies within North America. Moreover, there are more intra-

[7]http://developer.yahoo.com/maps/




(a) Indegree Distribution




(b) Outdegree Distribution

**Figure 5: Twitter social network has a power law exponent of about -2.4 which is similar to the Web and blogosphere.**

continent links than across continents. This is consistent with observations that the probability of friendship between two users is inversely proportionate to their geographic proximity [22].

In Table 4, we compare some of the network properties across these three continents with most users: North America, Europe and Asia. For each continent the social network is extracted by considering only the subgraph where both the source and destination of the friendship relation belong to the same continent. Asian and European communities have a higher degree correlation and reciprocity than their North American counterparts. Language plays an important role is such social networks. Many users from Japan and Spanish speaking world connect with others who speak the same language. In general, users in Europe and Asia tend to have higher reciprocity and clustering coefficient values in their corresponding subgraphs.

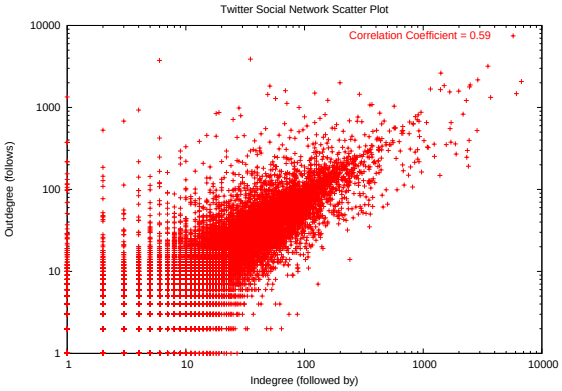



**Figure 6: Scatter plot showing the degree correlation of Twitter social network. A high degree correlation signifies that users who are followed by many people also have large number of friends.**

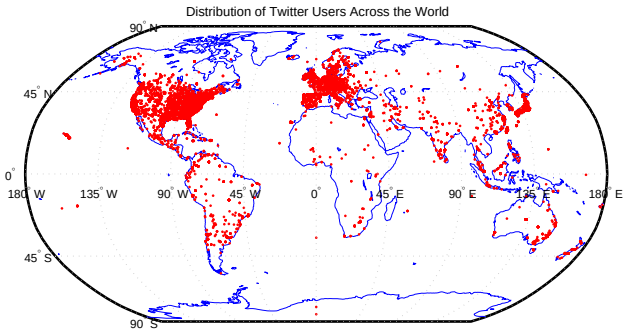



**Figure 7: Figure shows the global distribution of Twitter users. Though initially launched in US Twitter is popular across the world.**

## 4. USER INTENTION

In this paper, we propose a two-level framework for user intention detection. First, we used the HITS algorithm [17] to find the hubs and authorities in the network. Hubs and authorities have a mutually reinforcing property and are computed as follows: $H(p)$ represents the hub value of the page p and $A(p)$ represents the authority value of a page p.

$$Authority(p) = \sum_{v \in S, v \to p} Hub(v)$$

And

$$Hub(p) = \sum_{u \in S, p \to u} Authority(u)$$

Table 5 shows a listing of top ten hubs and authorities. From this list, we can see that some users have high authority score, and also high hub score. For example, Scobleizer, JasonCalacanis, bloggersblog, and Webtickle who have many followers and friends in Twitter are located in this category. Some users with very high authority scores have relatively low hub score, such as Twitterrific, ev, and springnet. They have many followers while less friends in Twitter, and thus are located in this category. Some other users with very high hub scores have relatively low authority scores, such as dan7, startupmeme, and aidg. They follow many other users while have less friends instead. Based on this rough categorization, we can see that user intention can be roughly categorized into these 3 types: information sharing, information seeking, and friendship-wise relationship.

After the hub/authority detection, we identify communities within friendship-wise relationships by only considering the bidirectional links where two users regard each other as friends. A community in a network can be vaguely defined as a group of nodes more densely connected to each other than to nodes outside the group. Often communities are topical or based on shared interests. To construct web communities, Flake et. al. [12] proposed a method using HITS

| from-to | Asia | Europe | Oceana | N.A | S.A | Africa |
|---|---|---|---|---|---|---|
| Asia | **13.45** | 0.64 | 0.10 | **5.97** | 0.005 | 0.01 |
| Europe | 0.53 | **9.48** | 0.25 | 6.16 | 0.17 | 0.02 |
| Oceana | 0.13 | 0.40 | 0.60 | 1.92 | 0.02 | 0.01 |
| N.A | **5.19** | **5.46** | 1.23 | **45.60** | 0.60 | 0.10 |
| S.A | 0.06 | 0.26 | 0.02 | 0.75 | 0.62 | 0.00 |
| Africa | 0.01 | 0.03 | 0.00 | 0.11 | 0.00 | 0.03 |

**Table 3: Table shows the distribution of Twitter social network links across continents. Most of the social network lies within North America. (N.A = North America, S.A = South America)**

| Property | N.A | Europe | Asia |
|---|---|---|---|
| Total Nodes | 16,998 | 5201 | 4886 |
| Total Edges | 205,197 | 42,664 | 60519 |
| Average Degree | 24.15 | 16.42 | 24.77 |
| Degree Correlation | 0.62 | 0.78 | 0.92 |
| Clustering Coefficient | 0.147 | 0.54 | 0.18 |
| Percent Reciprocity | 62.64 | 71.62 | 81.40 |

**Table 4: Network properties of social networks within a continent. Europe and Asia have a higher reciprocity indicating closer ties in these social networks. (N.A = North America)**

| User | Authority | User | Hub |
|---|---|---|---|
| Scobleizer | 0.002354 | Webtickle | 0.003655 |
| Twitterrific | 0.001765 | Scobleizer | 0.002338 |
| ev | 0.001652 | dan7 | 0.002079 |
| JasonCalacanis | 0.001557 | startupmeme | 0.001906 |
| springnet | 0.001525 | aidg | 0.001734 |
| bloggersblog | 0.001506 | lisaw | 0.001701 |
| chrispirillo | 0.001503 | bhartzer | 0.001599 |
| darthvader | 0.001367 | bloggersblog | 0.001559 |
| ambermacarthur | 0.001348 | JasonCalacanis | 0.001534 |

**Table 5: Top 10 Hubs and Authorities in Twitter. Some of the top authorities are also popular bloggers. Top hubs include users like startupmeme and aidg which are microblogging versions of a blogs and other web sites.**

and maximize flow/minimize cut to detect communities. In social network area, Newman and Girvan [13, 7] proposed a metric called modularity to measure the strength of the community structure. The intuition is that a good division of a network into communities is not merely to make the number of edges running between communities small; rather, the number of edges between groups is smaller than expected. Only if the number of between group edges is significantly lower than what would be expected purely by chance can we justifiably claim to have found significant community structure. Based on the modularity measure of the network, optimization algorithms are proposed to find good divisions of a network into communities by optimizing the modularity over possible divisions. Also, this optimization process can be related to the eigenvectors of matrices. However, in the above algorithms, each node has to belong to one community, while in real networks, communities often overlap. One person can serve a totally different functionality in different communities. In an extreme case, one user can serve as the information source in one community and the information seeker in another community.

People in friendship communities often know each other. Prompted by this intuition, we applied the Clique Percolation Method (CPM) [28, 10] to find overlapping communities in networks. The CPM is based on the observation that a typical member in a community is linked to many other members, but not necessarily to all other nodes in the same community. In CPM, the k-clique-communities are identified by looking for the unions of all k-cliques that can be reached from each other through a series of adjacent k-cliques, where two k-cliques are said to be adjacent if they share k-1 nodes. This algorithm is suitable for detecting the dense communities in the network.

Here we list a few specific examples of how communities form in Twitter and why users consist of these communities - what user intentions are in each community. Figure 8 illustrates a representative community with 58 users closely communicating with each other through Twitter service. The key terms they talk about include work, Xbox, game, and play. It looks like some users with gaming interests getting together to discuss the information about certain new products on this topic or sharing gaming experience. When we go to specific users website, we also find this type of conversation: *"BDazzler@Steve519 I don't know about the Jap PS3's. I think they have region encoding, so you'd only be able to play Jap games. Euro has no ps2 chip"* or *"BobbyBlackwolf Playing with the PS3 firmware update, can't get WMP11 to share MP4's and the PS3 won't play WMV's or AVI's...Fail."* We also noticed that users in this community also share with each other their personal feeling and daily life experiences in addition to comments on "gaming". Based on our study of the communities in Twitter dataset, we observed that this is a representative community in Twitter network: people in one community have certain common interests and they also share with each other about their personal feeling and daily experience.

Using CPM, we are able to find how communities connected to each other by overlapped components. Figure 9 illustrates two communities with podcasting interests where GSPN and pcamarata are the ones who connected these two communities. In GSPN's bio, he mentioned he is the Producer of the Generally Speaking Podcast Network[8]; while in pcamarata's bio, he mentioned he is a family man, a neurosurgeon, and a a podcaster. By looking at the top key terms of these two communities, we can see that the focus of the green community is a little more diversified: people occasionally talk about podcasting, while the topic of the red community is a little more focused. In a sense, the red community is like a professional community of podcasting while the green one is a informal community about podcasting.

Figure 10 illustrates five communities connected by Scobleizer, who is a Tech geek blogger. People follow his posts to get technology news. People in different communities share different interests with Scobleizer. Specifically, AndruEdwards, Scobleizer, daryn, and davidgeller get together to share video related news. CaptSolo et al. have some interests on Semantic Web. AdoMatic et al. are engineers and have interests with coding related issues.

Studying intentions at a community level, we observe users participate in communities which share similar interests. Individuals may have different intentions for joining these communities. While some act as information providers, others are merely looking for new and interesting information. Next, we analyze aggregate trends across users spread over many communities, we can identify certain distinct themes. Often there are recurring patterns in word usages. Such patterns may be observed over a day or a week. For example Figure 11 shows the trends for the terms "friends" and "school" in the entire corpus. While school is of interest during weekdays, friends take over on the weekends. The

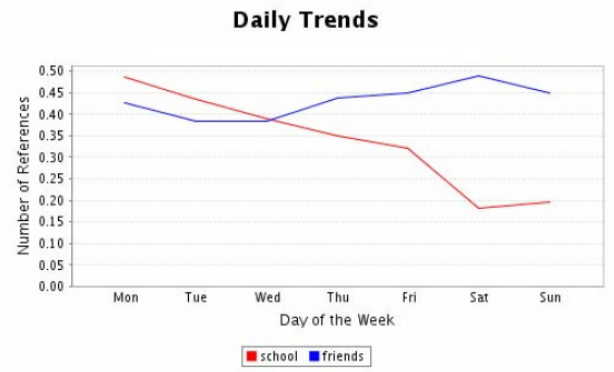



**Figure 11: Daily Trends for terms "school" and "friends". The term school is more frequent during the early week, friends take over during the weekend.**

log-likelihood ratio is used to determine terms that are of significant importance for a given day of the week. Using a technique described by Rayson and Garside [30], we create a contingency table of term frequencies for a given day and the rest of the week.

[8]http://ravenscraft.org/gspn/home

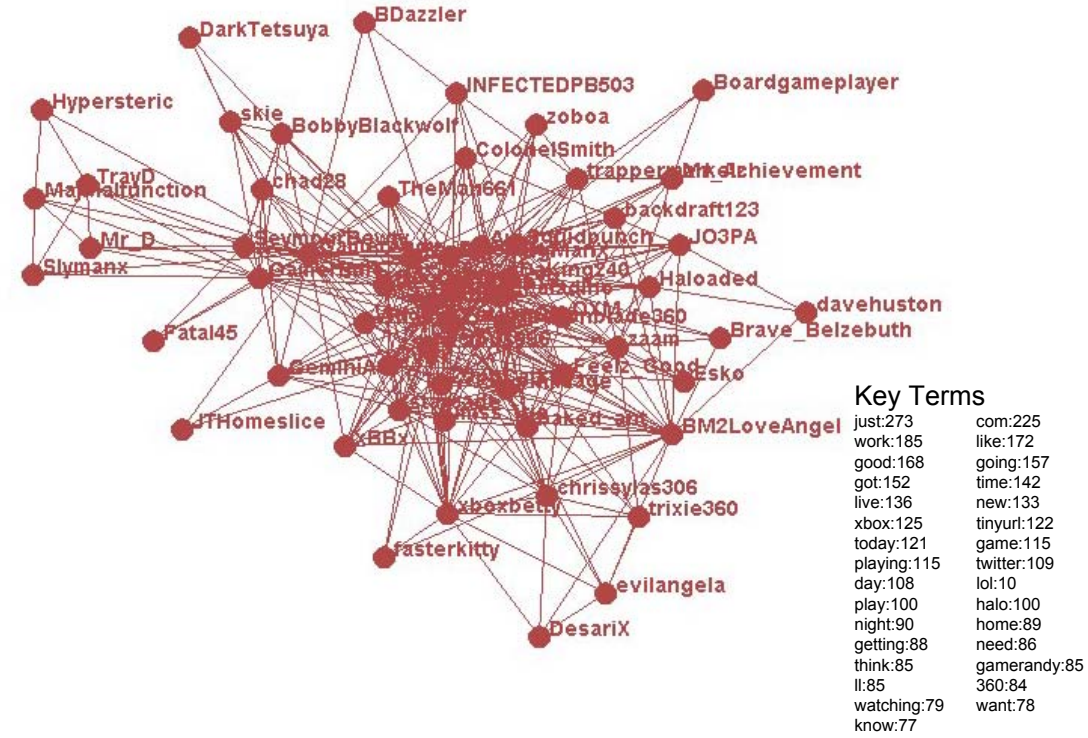



**Figure 8: An example of a "gaming" community who also share daily experiences.**

| | Day | Rest of the Week | Total |
|---|---|---|---|
| Freq of word | a | b | a+b |
| Freq of other words | c-a | d-b | c+d-a-b |
| Total | c | d | c+d |

Comparing the terms that occur on a given day with the histogram of terms for the rest of the week, we find the most descriptive terms. The log-likelihood score is calculated as follows:

$$LL = 2 * (a * log(\frac{a}{E1}) + b * log(\frac{b}{E2}))$$

where $E1 = c * \frac{a+b}{c+d}$ and $E2 = d * \frac{a+b}{c+d}$

Figure 12 shows the most descriptive terms for each day of the week. Some of the extracted terms correspond to recurring events and activities significant for a particular day of the week for example "school" or "party". Other terms are related to current events like "easter" and "EMI".

## 5. DISCUSSION

Following section presents a brief taxonomy of user intentions on Twitter. The apparent intention of a Twitter post was determined manually by the first author. Each post was read and categorized. Posts that were highly ambiguous or for which the author could not make a judgement were placed in the category UNKNOWN. Based on this analysis we have found following are some of the main user intentions on Twitter:

- *Daily Chatter* Most posts on Twitter talk about daily routine or what people are currently doing. This is the largest and most common user of Twitter
- *Conversations* In Twitter, since there is no direct way for people to comment or reply to their friend's posts, early adopters started using the @ symbol followed by a username for replies. About one eighth of all posts in the collection contain a conversation and this form of communication was used by almost 21% of users in the collection.

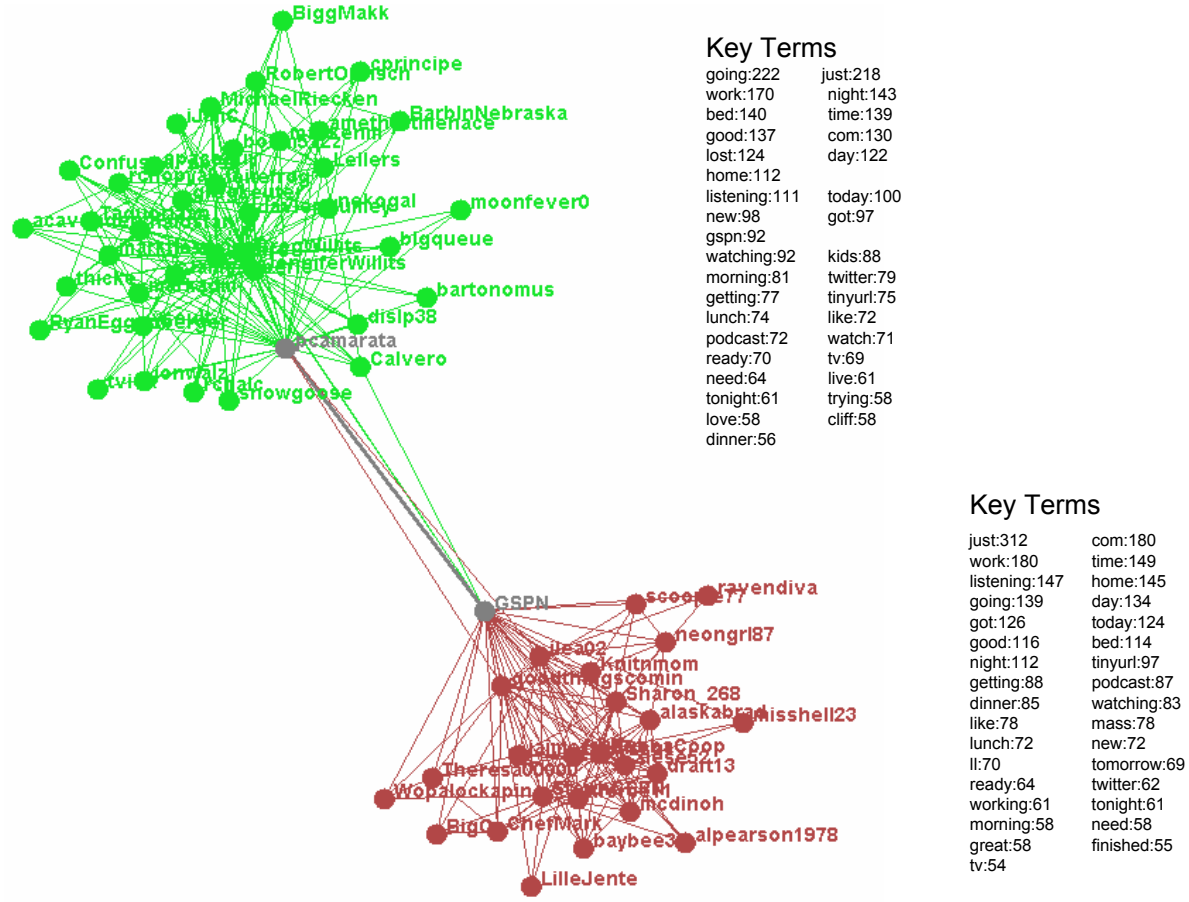



**Figure 9: Example of how two communities connect to each other**

- *Sharing information/URLs* About 13% of all the posts in the collection contain some URL in them. Due to the small character limit a URL shortening service like TinyURL[9] is frequently used to make this feature feasible.

- *Reporting news* Many users report latest news or comment about current events on Twitter. Some automated users or agents post updates like weather reports and new stories from RSS feeds. This is an interesting application of Twitter that has evolved due to easy access to the developer API.

Using the link structure, following are the main categories of users on Twitter:

- *Information Source* An information source is also a hub and has a large number of followers. This user may post updates on regular intervals or infrequently. Despite infrequent updates, certain users have a large number of followers due to the valuable nature of their updates. Some of the information sources were also found to be automated tools posting news and other useful information on Twitter.

- *Friends* Most relationships fall into this broad category. There are many sub-categories of friendships on Twitter. For example a user may have friends, family and co-workers on their friend or follower lists. Sometimes unfamiliar users may also add someone as a friend.

- *Information Seeker* An information seeker is a person who might post rarely, but follows other users regularly.

Our study has revealed different motivations and utilities of microblogging platforms. A single user may have multiple intentions or may even serve different roles in different communities. For example, there may be posts meant to

[9]http://www.tinyurl.com

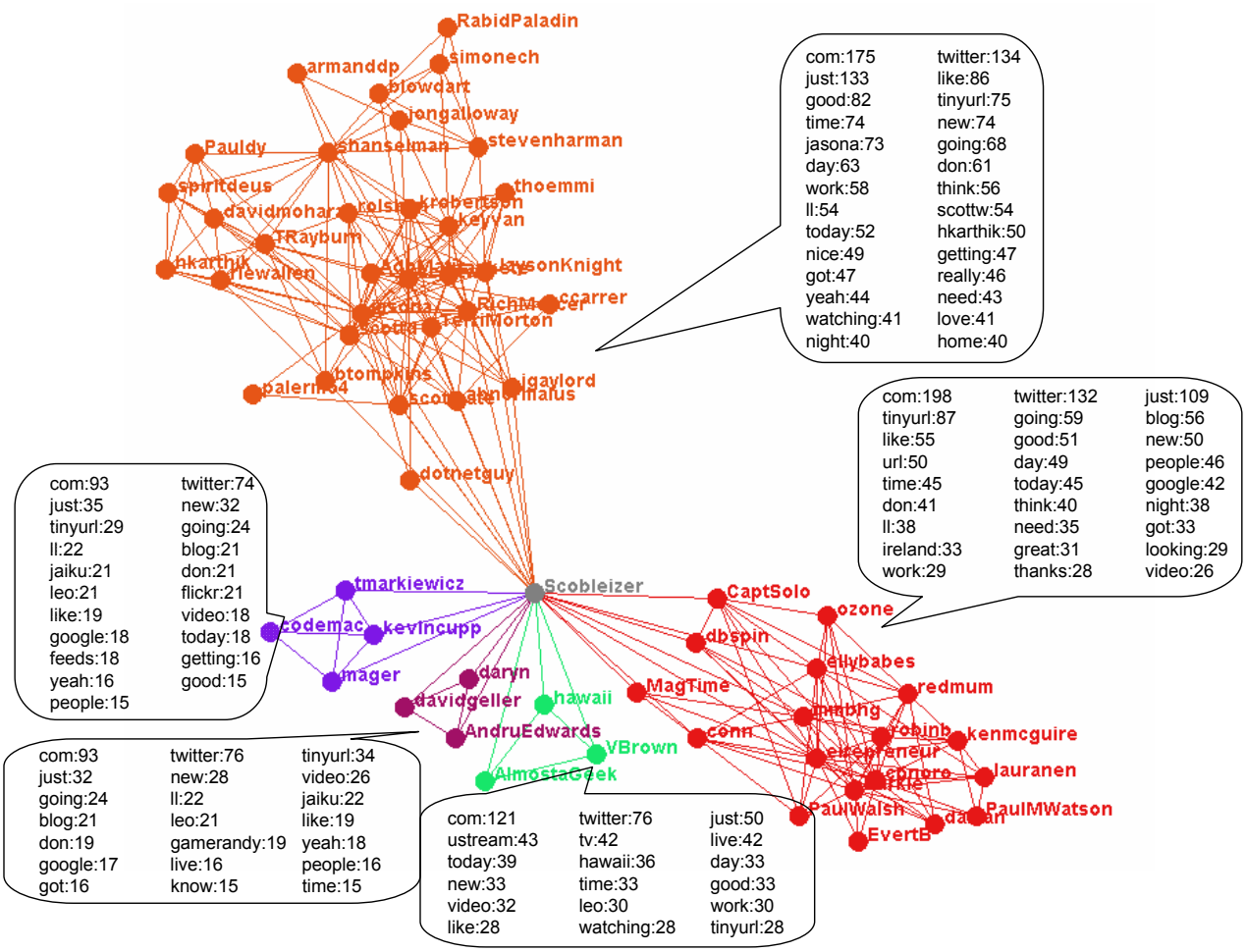



**Figure 10: Example Communities in Twitter Social Network. Key terms indicate that these communities are talking mostly about technology. The user Scobliezer connects multiple communities in the network.**

update your personal network on a holiday plan or a post to share an interesting link with co-workers. Multiple user intentions have led to some users feeling overwhelmed by microblogging services [20]. Based on our analysis of user intentions, we believe that the ability to categorize friends into groups (e.g. family, co-workers) would greatly benefit the adoption of microblogging platforms. In addition features that could help facilitate conversations and sharing news would be beneficial.

## 6. CONCLUSION

In this study we have analyzed a large social network in a new form of social media known as microblogging. Such networks were found to have a high degree correlation and reciprocity, indicating close mutual acquaintances among users. While determining an individual user's intention in using such applications is challenging, by analyzing the aggregate behavior across communities of users, we can describe the community intention. Understanding these intentions and learning *how* and *why* people use such tools can be helpful in improving them and adding new features that would retain more users. In this work, we have identified different types of user intentions and studied the community structures. Currently, we are working on automated approaches of detecting user intentions with related community structures.

## 7. ACKNOWLEDGEMENTS

We would like to thank Twitter Inc. for providing an API to their service and Pranam Kolari, Xiaolin Shi and Amit Karandikar for their suggestions.

## 8. REFERENCES

[1] J. Allen. Recognizing intentions from natural language utterances. *Computational Models of Discourse*, pages 107–166, 1983.

[2] J. Austin. *How to Do Things with Words*. Oxford University Press Oxford, 1976.

[3] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999.

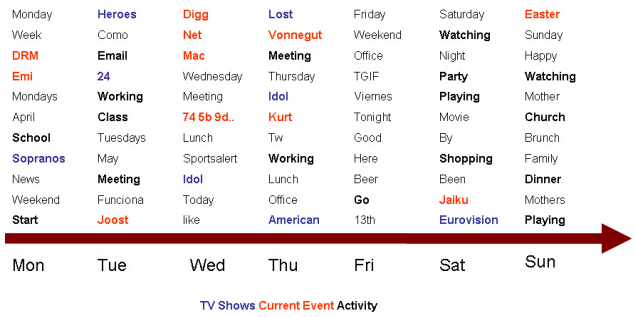



**Figure 12: Distinctive terms for each day of the week ranked using Log-likelihood ratio.**


[4] Blogpulse. The 3rd annual workshop on weblogging ecosystem: Aggregation, analysis and dynamics, 15th world wide web conference, May 2006.

[5] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[6] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications netowrking*, pages 309–320, Amsterdam, The Netherlands, The Netherlands, 2000. North-Holland Publishing Co.

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

[8] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data, Jun 2007.

[9] Comscore. `http://www.usatoday.com/tech/webguide/2007-05-28-social-sites_N.htm`.

[10] I. Derenyi, G. Palla, and T. Vicsek. Clique percolation in random networks. *Physical Review Letters*, 94:160202, 2005.

[11] D. Donato, L. Laura, S. Leonardi, and S. Millozzi. Large scale properties of the webgraph. *European Physical Journal B*, 38:239–243, March 2004.

[12] G. W. Flake, S. Lawrence, C. L. Giles, and F. Coetzee. Self-organization of the web and identification of communities. *IEEE Computer*, 35(3):66–71, 2002.

[13] M. Girvan and M. E. J. Newman. Community structure in social and biological networks, Dec 2001.

[14] H. Grice. Utterers meaning and intentions. *Philosophical Review*, 78(2):147–177, 1969.

[15] B. J. Grosz. *Focusing and Description in Natural Language Dialogues*. Cambridge University Press, New York, New York, 1981.

[16] A. Java. `http://ebiquity.umbc.edu/blogger/2007/04/15/global-distribution-of-twitter-users/`.

[17] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[18] P. Kolari, T. Finin, Y. Yesha, Y. Yesha, K. Lyons, S. Perelgut, and J. Hawkins. On the Structure, Properties and Utility of Internal Corporate Blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*, March 2007.

[19] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for emerging cyber-communities. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1481–1493, 1999.

[20] A. Lavallee. Friends swap twitters, and frustration - new real-time messaging services overwhelm some users with mundane updates from friends, March 16, 2007.

[21] T. Lento, H. T. Welser, L. Gu, and M. Smith. The ties that blog: Examining the relationship between social ties and continued participation in the wallop weblogging system, 2006.

[22] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences,*, 102(33):11623–1162, 2005.

[23] Y.-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. Tseng. Discovery of Blog Communities based on Mutual Awareness. In *Proceedings of the 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 15th World Wid Web Conference*, May 2006.

[24] S. Matsubara, S. Kimura, N. Kawaguchi, Y. Yamaguchi, and Y. Inagaki. Example-based Speech Intention Understanding and Its Application to In-Car Spoken Dialogue System. *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7, 2002.

[25] A. A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi. On the structural properties of massive telecom call graphs: findings and implications. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 435–444, New York, NY, USA, 2006. ACM Press.

[26] B. A. Nardi, D. J. Schiano, M. Gumbrecht, and L. Swartz. Why we blog. *Commun. ACM*, 47(12):41–46, 2004.

[27] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323, 2005.

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

[29] J. Pontin. From many tweets, one loud voice on the internet. *The New York Times*, April 22, 2007.

[30] P. Rayson and R. Garside. Comparing corpora using frequency profiling, 2000.

[31] X. Shi, B. Tseng, and L. A. Adamic. Looking at the blogosphere topology through different lenses. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*, March 2007.

[32] P. Strawson. Intention and Convention in Speech Acts. *The Philosophical Review*, 73(4):439–460, 1964.

[33] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.