Is it Really About Me? Message Content in Social Awareness Streams

Mor Naaman, Jeffrey Boase, Chih-Hui Lai
Rutgers University, School of Communication and Information
4 Huntington St., New Brunswick, NJ 08901, USA
{mor, jboase, chihhui}@rutgers.edu

ABSTRACT
In this work we examine the characteristics of social activity and patterns of communication on Twitter, a prominent example of the emerging class of communication systems we call "social awareness streams." We use system data and message content from over 350 Twitter users, applying human coding and quantitative analysis to provide a deeper understanding of the activity of individuals on the Twitter network. In particular, we develop a content-based categorization of the type of messages posted by Twitter users, based on which we examine users' activity. Our analysis shows two common types of user behavior in terms of the content of the posted messages, and exposes differences between users in respect to these activities.

Author Keywords
Social media, Twitter, communication systems.

ACM Classification Keywords
H.4.3. Information Systems Applications: Communications Applications.

General Terms
Human Factors

INTRODUCTION
The rise of social media services has contributed to the altering of many people's communication patterns and social interaction. In particular, semi-public communication platforms such as the Facebook "Newsfeed," Twitter, and FriendFeed represent a new class of communication technologies. In these systems, participants post short status messages or pointers to resources like links to articles, photos and videos. The posted messages are often available publicly, or semi-publicly (e.g., restricted to the user's designated contacts). The postings are consumed by readers in "streams" of messages published by the various users that they follow.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CSCW 2010, February 6–10, 2010, Savannah, Georgia, USA.
Copyright 2010 ACM 978-1-60558-795-0/10/02...$10.00.

These *social awareness streams* (SAS), as we call them, are typified by three factors distinguishing them from other communication: a) the public (or personal-public) nature of the communication and conversation; b) the brevity of posted content; and, c) a highly connected social space, where most of the information consumption is enabled and driven by articulated online contact networks.

To date, the CSCW community has not developed a strong understanding of these emerging communication systems (often referred to as Micro-blogging communities). While a number of different research efforts examined activities on Facebook and other services where SAS are available [1,2,7,8], most of these studies did not address the unique aspects of SAS. Recently, a number of studies from other computing disciplines examined data patterns and trends on Twitter. Huberman et al. showed that the rate of user activity on Twitter is influenced by social connectivity and the user's network (i.e., number of contacts) [4]. Java et al. provide many descriptive statistics about Twitter use [5], and hypothesize that the differences between users' network connection structures suggest three types of *distinct* user activities: information seeking, information sharing, and social activity. Krishnamurthy et al. also performed a descriptive analysis of the Twitter network, suggesting that frequent updates might be correlated with high overlap between friends and followers [6]. Honeycutt and Herring [3] examined the functions and uses of the @ ("reply/mention") sign on Twitter and the coherence of exchanges on Twitter. Using content analysis, they developed a categorization of the functional use of @, and analyzed the content of the reply messages.

To address the gap, in this exploratory study, we aim to acquire an initial understanding of the type of content activities commenced by individuals in SAS. To this end, we use system data and both qualitative and quantitative analysis to examine the activity of participants in the Twitter network. The contributions of this work are:

- A characterization of the content of messages posted on social awareness streams.
- An examination of how message content varies by user characteristics, personal networks, and usage patterns.

We begin with a general introduction of Twitter, a popular SAS that is the focus of this work.

TWITTER AND SOCIAL AWARENESS STREAMS

Twitter is a popular social media service, with millions of registered users as of September 2009. Twitter's core functions represent a very simple social awareness stream model. Twitter users can post short messages, or *tweets*, which are up to 140 characters long. The messages are displays as a "stream" on the user's Twitter page. In terms of social connectivity, Twitter allows a user to *follow* any number of other users, called "friends." The Twitter contact network is asymmetric, meaning that if Rohan follows Yumi, Yumi need not follow Rohan. The users *following* a Twitter user are called "followers." A user can set their privacy preferences such that their updates are available only to the user's followers. By default, the posted messages are available to anyone; in this work, we only consider messages posted publicly on Twitter. Users consume messages mostly by viewing a core page showing a stream of the latest messages from all their friends, listed in reverse chronological order.

Twitter supports posting of messages via SMS, Web and mobile Web services, but also allows users to use different "third party" applications to post (and consume) Twitter messages. As a result, an array of applications and avenues for posting to Twitter are available, ranging from mobile, web-based, desktop and other applications, including posts on behalf of the user from automated agents.

Finally, Twitter users can reference other users in posted messages by using the @ symbol, effectively creating a link from their message to the referenced user's account. Such reference messages (known as "reply" or "mention", depending on the use) appear in the referenced user's account so that users can keep track of messages mentioning them.

METHOD

Our analysis is based on Twitter data, downloaded using the Twitter API; a qualitative categorization of Twitter messages; and a quantitative examination of the system data and message coding. The downloaded data included profile details, messages, and data about the sampled users' friends and followers. We first report on how we analyzed Twitter content to generate a coding scheme, and then describe the details of the dataset and the coding process.

Generating Content Categories

To characterize the type of messages posted on Twitter we used a grounded approach to thematize and code a sample of 200 public posts ("messages") downloaded from Twitter. First, the three authors independently assigned categories to the downloaded messages. We then proceeded to analyze the affinity of the emerging themes to create an initial set of coding categories. Next, we downloaded a second set of 200 posts, categorized them, then reflected on and adapted the initial categories based on the additional input. Finally, during the subsequent coding process we added two other message categories based on feedback from our coders. It is important to note that the stated goal of our coding was to provide a descriptive evaluation of the message content, not to hypothesize on the intent of the user posting the message (e.g., making conversation, maintaining ties and so forth). The resulting categories and sample messages for each are presented in Table 1, and are similar to categories derived by Honeycutt and Herring [3].

Given the short format and message content, selecting a single coding category was not always possible or desirable. For instance, a message stating, "I love driving to Richmond" could be a random statement, but might also implies that the user is currently at the wheel. Other messages, for example, both share information and express opinion about the information being shared. For this reason, when suitable, more than one category was assigned to a single message.

Dataset

To obtain a random sample of Twitter users we recorded the IDs of all users who had messages posted on Twitter's public timeline, which displays a subset of Twitter users' public messages. This process resulted in a sampling frame of 125,593 unique user IDs whose updates were marked as public at the time of posting.

From the sampling frame, we set out to select a sample of active 'personal' Twitter users. In other words, we selected users who are active participants in the Twitter network, and who are not organizations, marketers or those who 'have something to sell'. To operationalize the notion of active users, we selected users who had at least 10 friends, 10 followers, and had posted at least 10 messages. We randomly selected users that fit those criteria from our sampling frame. We manually examined each user's profile details, and coded them for 'personal use', ruling out in the process commercial entities, as well as people solely promoting their services or businesses. The process left us with 911 users, out of which we randomly selected 350 users for analysis.

For each of the selected users we downloaded their lists of friend and followers, as well as all the user's messages available from the Twitter API at the time of the crawl (April 2009). The Twitter API limits downloads to (roughly) three weeks of activity and a maximum of 1500 messages. As seen in other work [4,5], participation and social connectivity parameters in our dataset (e.g., the follower count) resemble power law distributions.

Coding

To analyze the content of messages, we randomly selected 10 messages for each user in our database (we only selected messages that were not "replies"; 13 users had fewer than ten such messages in the downloaded dataset), for a total of 3379 messages. Selecting 10 messages per user allows us to make predictions with a 95% confidence level and a 20% confidence interval for typical users, accurate enough to detect trends. Eight coders independently categorized the content of each of the 3379 messages as discussed above

Code	Example(s)
Information Sharing (IS)	"15 Uses of WordPress <URL REMOVED>"
Self Promotion (SP)	"Check out my blog I updated 2day 2 learn abt tuna! <URL REMOVED>"
Opinions/Complaints (OC)	"Go Aussie $ go!" "Illmatic = greatest rap album ever"
Statements and Random Thoughts (RT)	"The sky is blue in the winter here" "I miss New York but I love LA..."
Me now (ME)	"tired and upset" "just enjoyed speeding around my lawn on my John Deere. Hehe :)"
Question to followers (QF)	"what should my video be about?"
Presence Maintenance (PM)	"i'm backkkk!" "gudmorning twits"
Anecdote (me) (AM)	"oh yes, I won an electric steamboat machine and a steam iron at the block party lucky draw this morning!"
Anecdote (others) (AO)	"Most surprised <user> dragging himself up pre 7am to ride his bike!"

Table 1. Message Categories.

(Table 1). As mentioned, the coders were allowed to assign multiple categories to each message. Each message was assigned to two coders; to resolve discrepancies between coders we simply assigned to each message a union of categories assigned by the coders. The short length of Twitter messages meant a lack of context that did not permit a simple resolution to coder differences. Instead, we opted to consider all interpretations of the messages by coders. Over-coding was not a problem as messages had 1.3 categories assigned on average.

Beyond message content, we manually coded the gender of users and the type of application used to post each message. As gender information is not available from the Twitter user profile, we coded it by examining the picture and details of the users' profiles (48% female, 52% male, with four cases undetermined). We also manually categorized the 196 different applications used to post messages into types (mobile, web, desktop, etc.), and classified each message by its application type. For example, we found that 25% of messages were posted from mobile applications.

ANALYSIS

Our main objective in this work is to identify different types of user activity, specifically focusing on message content and its relationship to patterns of use. We address the following research questions:

RQ1 What types of messages are commonly posted and how does message type relate to other variables?

RQ2 What are the differences between users in terms of the types and diversity of messages that they usually post?

RQ3 How are these differences between users' content practices related to other user characteristics?

Let us start with RQ1; Figure 1 displays the breakdown of content categories in our coded dataset. As the figure shows, the four dominant categories were information sharing (IS; 22% of messages were coded in that category), opinions/complaints (OC), statements (RT) and "me now" (ME), with the latter dominating the dataset (showing that, indeed, "it's all about me" for much of the time).

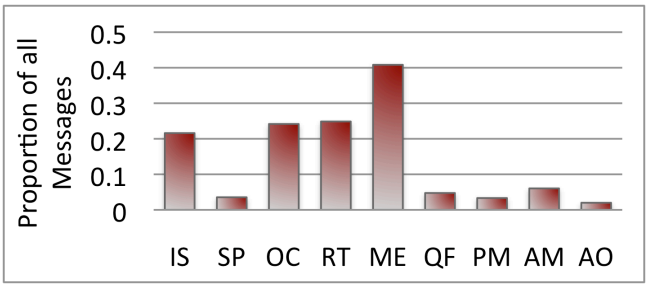

Figure 1. Message Category Frequency.

Figure 2 considers the *proportion* of users' activity dedicated to each type of content out of 10 messages coded for each user. The figure focuses on the four most popular categories shown above, and the blue area in each section represents all users. For example, the ME histogram shows that 14% of all users had 0-10% (left-most column) of their messages in the "Me Now" category; on average, users had 41% of their messages in "Me Now". The figure contrasts the span of activities of the network: most people engage in some scale of ME activity, while relatively few undertake information sharing as a major activity.

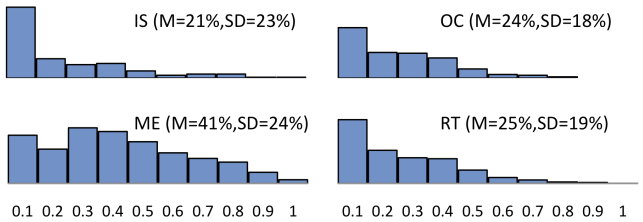

Figure 2. Message category as proportion of users' content for categories IS, OC, ME and RT.

To further address RQ1, we examine the difference between males and females in terms of the types of message they post as percentage of the user's messages. Our results show that females are more likely to post "me now" messages (M=45% of a user's messages) than males (M=37%), and that this difference is statistically significant (two-tailed t-test; $t(344)=3.12$, $p<0.005$). We also examine the relationship between message type and the use of mobile devices to post messages. We find that overall, 51% of mobile-posted messages are "me now" messages, compared to the 37% of "me now" messages posted from non-mobile applications. A Pearson Chi-square analysis shows that this difference is statistically significant ($\chi^2=49.7$, $p<.0001$).

To address RQ2, we use Ward's linkage cluster analysis to categorize users based on the types of messages that they typically post. We then use Kalensky's analysis to detect the optimum number of clusters that minimizes the differences within groups and maximizes differences between groups. The analysis resulted in two clusters, which we labeled "Informers" (20% of users) and – to

suggest a new term – "Meformers" (80%). Figure 3 shows the mean of the average proportion of messages in the top four categories for each user. For instance, on average Informers had 53% of their messages in the IS category, while a significant portion (M=48%) of the messages posted by Meformers were "Me Now" messages. Indeed, the figure suggests that while Meformers typically post messages relating to themselves or their thoughts, Informers post messages that are informational in nature.

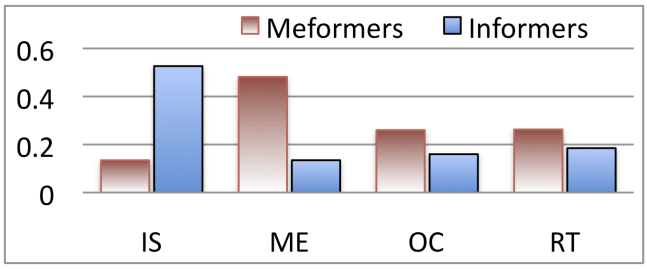

Figure 3. Mean user message proportions for the four main categories, breakdown by cluster.

For RQ3, we examined how Meformers and Informers are different in respect to several independent variables. We found that Informers have more friends ($Median_1$=131) and followers ($Median_2$=112) than Meformers ($Median_1$=61, $Median_2$= 42), in a statistically significant manner (we report medians and use the non-parametric Mann-Whitney test due to the skewed distribution of network ties; z_1=-5.1, z_2=-3.97; p<.0001 for both; outliers removed). Informers also have a higher proportion of mentions of other users in their messages (M=54% vs. M=41%, t(349) =4.12, p<.001).

Finally, we looked at diversity of user content in context of RQ2 and RQ3. We represented the diversity of messages typically posted by each user by calculating a standard entropy scale from the user's content categories proportions. Larger values on this scale indicate a greater diversity of messages posted by individual users. The resulting distribution (range 0-25, M=14, slightly positive normal distribution) suggests that differences between users in this dimension exist, but are not pronounced. We then (RQ3) correlated entropy with variables such as number of friends and frequency of posting. We find a negative association between entropy and the average number of messages posted per hour (r=-0.19, p<.01) and a positive association between entropy and proportion of mentions/replies posted by the user (r=0.15, p<.01). These findings indicate that users who post more restricted span of messages tend to post more frequently; and that the more balanced posters are more likely to interact with other users via their messages.

DISCUSSION AND CONCLUSIONS

We have performed an analysis of the content of messages posted by individuals on Twitter, a popular social awareness stream service, representing a new and understudied communication technology. Our analysis extends the network-based observations of Java et al. [5], showing that Twitter users represent two different types of "content camps": a majority of users focus on the "self", while a smaller set of users are driven more by sharing information. Note that although the Meformers' self focus might be characterized by some as self-indulgent, these messages may play an important role in helping users maintain relationships with strong and weak ties. Our findings suggest that the users in the "information sharing" group tend to be more conversational, posting mentions and replies to other users, and are more embedded in social interaction on Twitter, having more social contacts. We note that the direction of the causal relationship between information sharing behavior and extended social activity is not clear. One hypothesis is that informers prove more "interesting" and therefore attract followers; an alternative explanation is that informers seek readers and attention for their content and therefore make more use of Twitter's social functions; or that an increased amount of followers encourages user to post additional (informative) content [4]. A longitudinal study may help us address these alternatives.

Finally, we did not address in this work the relationship between social network structure and social influence to the type of content posted by users. It is certainly possible that users are subject to social learning, and are influenced by the activity of others they observe on the service [1]. We assume that theories such as social presence and social capital can help inform a theoretical understanding of the type and characteristics of content published in the service. We intend to explore these associations in future work.

REFERENCES

1. Burke, M., Marlow, C., and Lento, T. Feed me: motivating newcomer contribution in social network sites. In Proc. CHI '09. ACM Press (2009), 945-954.
2. Ellison, N.B., Steinfield, C. and Lampe, C. The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. Comp.-Mediated Comm. 12, 4 (2007), 1143-1168
3. Honeycutt, C., & Herring, S. Beyond microblogging: Conversation and collaboration via Twitter. *In Proc. HICSS '09*. IEEE Press (2009).
4. Huberman, B., Romero, D., and Wu, F. Social networks that matter: Twitter under the microscope. *First Monday [Online] 14*, 1 (2008).
5. Java, A., Song, X., Finin, T., and Tseng, B. Why we twitter: understanding microblogging usage and communities. In *Proc. WebKDD/SNA-KDD '07*, ACM Press (2007).
6. Krishnamurthy, B., Gill, P., and Arlitt, M. A few chirps about twitter. *In Proc. WOSP '08*. ACM Press (2008).
7. Lampe, C., Ellison, N.B., and Steinfield, C. Changes in use and perception of facebook. In *Proc. CSCW '08*, ACM Press (2008), 721-730.
8. Sun, E., Rosenn, I., Marlow, C., Lento, T. Gesundheit! Modeling Contagion through Facebook News Feed. In *Proc. ICWSM '09*, AAAI (2009)