# Towards Understanding Twitter Use in Software Engineering: Preliminary Findings, Ongoing Challenges and Future Questions

Gargi Bougie, Jamie Starke, Margaret-Anne Storey, Daniel M. German
University of Victoria
{gbougie, jstarke, mstorey, dmg}@uvic.ca

## ABSTRACT

There has been some research conducted around the motivation for the use of Twitter and the value brought by micro-blogging tools to individuals and business environments. This paper explores the phenomenon as it affects the population which birthed the technology: Software Engineers. We find that the Software Engineering community extensively leverages Twitter's capabilities for conversation and information sharing and that use of the tool is notably different between distinct Software Engineering groups. Our work exposes topics for future research and outlines some of the challenges in exploring this type of data.

## 1. INTRODUCTION

The micro-blogging phenomenon began at the beginning of 2006 when Facebook launched status messages. The communication medium took off in the latter half of the year with the introduction of the first major micro-blogging service: *Twitter* [1]. To date, there are several platforms for exchanging short messages with friends, colleagues or strangers, including Free and Open Source Software (FOSS) versions, such as *Identi.ca* [3].

What we seek to discover in this research is the way in which the Software Engineering community has itself embraced these forms of media, with a specific focus on Twitter. Social networks are, in many arenas, replacing more traditional forms of communication, such as email [2]. This trend is reinforced by a comment from a member of the Eclipse community that we study in this work:

> Many teams will "tweet" when they publish new technologies, tutorials, blog posts, etc. and Twitter provides me with an easy way to scan this information. This information is often published in other mediums, but it's usually repeated on Twitter – so instead of subscribing to a variety of mailing lists, I can simply "watch twitter".

This paper takes an exploratory approach to examining the use of micro-blogging tools by Software Engineers and their effect on communication within respective Software Engineering groups. Through archival and grounded theory analysis, we examine the conversation and community of Software Engineers on Twitter by building up an understanding of their usage characteristics and discussion topics.

### 1.1 Related Work

Zhao et al. examine the reasons behind a person's choice of tools like Twitter over other forms of social media, such as blogs [20]. They introduce a framework for studying the benefits of informal communication and examine the features of Twitter according to these guidelines. The authors conclude that the inherent brevity of micro-blogs "reduces the cost of sharing". In addition, many people now turn to Twitter for news and information updates because of its real-time nature and because the sources are people they've come to know and trust. In other words, micro-blogs act as "people-based RSS feeds".

Java et al. explore the intentions of users when they post to Twitter and identify four main categories: daily chatter, conversations (indicated by use of the direct messages), sharing information (indicated by the inclusion of a link), and reporting news [15]. The authors also outline the user roles of information seeker and information source, where a source tends to have more followers than followees (people they follow). In addition, Java et al. touch on the idea of communities being detectable based on the key words they use in their posts.

Research by Honeycutt et al. examine the duration and coherency of interactive Twitter exchanges [13]. The study begins by using a grounded theory approach to analyze a corpus of public Twitter messages. From this, the authors categorize the messages and the use of the '@' sign, discovering that though there are various uses for the symbol, its major function is to direct messages to specific individuals. In terms of conversation attributes, the findings indicate that most conversations consist of three to five messages exchanged between two people over less than a half hour. Honeycutt et al. also produce evidence that conversation and collaboration involving several people does occur via Twitter. However, these conversations are complex, as several sub-threads are seen to develop initially before the communication becomes more centrally focused.

Huberman et al. examine the relationship of the number of followers a person has to how active he or she is on Twitter [14]. The results show that it is not so much followers that dictates activity, but rather interactions. Huberman et al. define a person's "friend" as someone they have sent at least two direct posts to. By this definition, even though it is unidirectional, a much clearer relationship is seen between the number of friends a person has and their level of activity on Twitter. Ninety-eight percent of users involved in the study had fewer friends than followees. By these findings, the authors conclude that though an individual may have a seemingly large network of followers and followees, the friend network is the more influential social network and is much smaller.

Ehrlich et al. compare the use of an internal corporate micro-blogging tool with the public tool, Twitter, in order to understand the difference in information sharing and communication that arises [10]. The most notable difference between the two mediums is the use of the internal tool for soliciting information from and providing information to colleagues, despite their own use of and their colleagues' presence on Twitter. Additionally, the findings of Zhao et al. [20] and Skeels et al. [18] are supported in this study by the fact that participants used status updates to promote and maintain "ambient awareness", especially to signal a change in availability, such as returning from vacation. Another interesting finding of Ehrlich et al. is the frequency of brief "conversations" that arose in the internal tool as compared to Twitter. Short conversations comparable to an email thread with brief messages passed back and forth occurred significantly more frequently over the internal tool than through Twitter, possibly due to an implied common ground and smaller user group.

## 1.2 Motivation and Exploratory Questions

Moving from what we know about the Twitter community and the value it provides for the general population, we look to examine the use of this micro-blogging technology by Software Engineers. As we can see from previous research, there has been little focus to date on understanding Twitter use by a distinct, interest-sharing group. The broader goal of this paper is to begin to explore the potential of social media to influence tool support for Software Engineers. There are already some research projects that explore the use of micro-blogging in an IDE [17, 12]. We begin our preliminary work with the following broad questions relating to Twitter:

- How do Software Engineers make use of Twitter to support communication in their community?
- What do Software Engineers talk about over Twitter?

# 2. METHOD AND DATA SOURCES

In this research, we use archival analysis to quantify some basic parameters of Twitter use by Software Engineers, such as the amount of direct messages sent from one user to another. We compare this to prior findings by Java et al [15] concerning the general population of Twitter users. We also use grounded theory to manually code 600 "tweets" from our sample set in order to learn what topics are being discussed by Software Engineers over Twitter.

For our exploratory purposes, we elected to use *wefollow* [7], a website that lists the most prominent Twitter users under specific tags. From this site, we selected the top 30 individuals for the topics, *Linux* and *Eclipse*. We chose these two topics based on their potential to expose "tweeters" from a large operating system community as well as an IDE development community. We also decided to investigate a project for which all committers use Twitter. Through a colleague, we were informed that the MXUnit project lists the Twitter user names for all eight of its committers. The MXUnit project [4] is a small, open source ColdFusion test framework that is written as an Eclipse plugin.

Our unit of analysis in this paper is a single "tweet", a message posted by a Twitter user that can consist of up to 140 characters. We collected a total of 11,679 tweets that were made by or referenced Twitter users in our sample set. Table 1 shows the two time periods during which tweets were collected, along with the number of tweets collected for each of the three Software Engineering communities during these time periods. We collected data at two different time periods to minimize the possibility of any date-sensitive phenomena dominating our findings.

At the time of this work, the basic Twitter API states that it only allows for the retrieval of the most recent 1,500 tweets for any given search. In addition, data older than one to two weeks is often not available programmatically. Therefore, we collected data in a continuous fashion. Custom Perl scripts, as well as a free program called *The Archivist* [5] were used to collect tweets approximately once every two days over the indicated time periods.

For the grounded theory analysis [11, 19, 9], we selected the first 100 tweets from our sample for each community and each time period listed in Table 1, for a total of 600 tweets. We did this, as opposed to sampling tweets randomly, in order to preserve the coherency and context of discussion over time. We first used an open coding process on a subset of 300 tweets, 100 from each June/July group, to develop a set of codes for the major topics being discussed. These were: *Software Engineering*, *gadgets and technology*, *current events*, and *chatter*. Following this, we used these codes to perform closed, or fixed coding on the full sample set of 600 tweets described above. This allowed us to obtain the proportions of tweets relating to each major topic, which we discuss later.

| Community | Number of users | Number of tweets collected in June/July 2010 | Number of tweets collected in January 2011 |
|---|---|---|---|
| Eclipse | 30 | 541 tweets over 6 days (90/day) | 812 tweets over 7 days (116/day) |
| Linux | 30 | 7244 tweets over 15 days (482/day) | 2545 tweets over 7 days (363/day) |
| MXUnit | 8 | 400 tweets over 12 days (33/day) | 137 tweets over 7 days (19/day) |

**Table 1: The communities selected for our research along with the number of users in our sample and the number of tweets collected. (The numbers indicated in brackets represent the number of tweets normalized over days.)**

# 3. PRELIMINARY FINDINGS

Our archival and grounded theory analysis produced mutually supporting results in several areas. In this section and in our discussion, we are able to paint a first picture of the Software Engineering Twitter culture. Names of Twitter users are anonymized by appending an assigned number to the community name in which a user belongs.

## 3.1 Use of Twitter

**Conversation and Information Sharing.** We found that the amount of conversation (as determined by direct messages indicated with the '@' symbol) is drastically higher (by approximately 50% - 67%, as seen in Table 2) for all three Software Engineering communities studied than was reported for the 2007 Twitter corpus collected by Java et al. [15]. Information sharing (as determined by URLs) is also more frequent by Software Engineers (by approximately 6% - 24%, as seen in Table 2) than was seen in the findings of Java et al.

**Retweets.** In their research on blog communities, Lin et al. [16] showed that discovering communities is highly dependent on mutual awareness throughout a social network. As such, we examined "retweets" (tweets reposted by someone other than the original poster, with credit to the original poster) to determine whether they are a common means of promoting awareness among Software Engineers on Twitter. However, the number of tweets in our sample set containing retweets was quite low (ranging from approximately 6% to 22% of tweets, as seen in Table 3). In most cases, retweets were used to spread interesting or important announcements relating to current events or technology. For example:

> RT @[Linux#1]: Even as SCO dies, the company lies http://bit.ly/dvdS6m Ack! Yet more unfounded SCO #IBM #Unix copyright claims against #Linux

**Hashtags.** The occurrence of hashtags, a convention developed informally among Twitter users for categorizing the content of posts [6], was again low. Hashtags are made up of a single word that begins with the '#' symbol and are placed anywhere in a 140-character post. Tweets containing at least one hashtag ranged from approximately 13% to 23%, as seen in Table 3 . In the instances where hashtags were used by Software Engineers, it seemed clear that they were attempting to reach a broader community. For example, *#fifa* was a common hashtag during the 2010 FIFA World Cup.

As illustrated in Figure 1, the percentages of tweets containing conversation, information sharing, hashtags and retweets are not significantly different between the June/July 2010 and January 2011 data sets.

| Group | % @ | % URLs |
|---|---|---|
| Twitter corpus collected by Java et al. | 12.5 | 13 |
| Linux June/July 2010 | 79.7 | 37.3 |
| Linux Jan. 2011 | 68.8 | 34.3 |
| Eclipse June/July 2010 | 76.3 | 27.5 |
| Eclipse Jan. 2011 | 62.1 | 31.9 |
| MXUnit June/July 2010 | 76.5 | 23.8 |
| MXUnit Jan. 2011 | 72.3 | 19.7 |

**Table 2: The percentage of conversation (@) and information sharing (URLs) seen in the three communities at the two different time periods, compared to the 2007 Twitter corpus of Java et al.**

| Group | % Retweets | % Hashtags |
|---|---|---|
| Linux June/July 2010 | 21.8 | 23.6 |
| Linux Jan. 2011 | 21.5 | 23.1 |
| Eclipse June/July 2010 | 22.9 | 18.9 |
| Eclipse Jan. 2011 | 21.4 | 16.1 |
| MXUnit June/July 2010 | 6.5 | 13.8 |
| MXUnit Jan. 2011 | 13.9 | 23.4 |

**Table 3: The percentage of retweets and hashtags seen in the three communities at the two different time periods.**

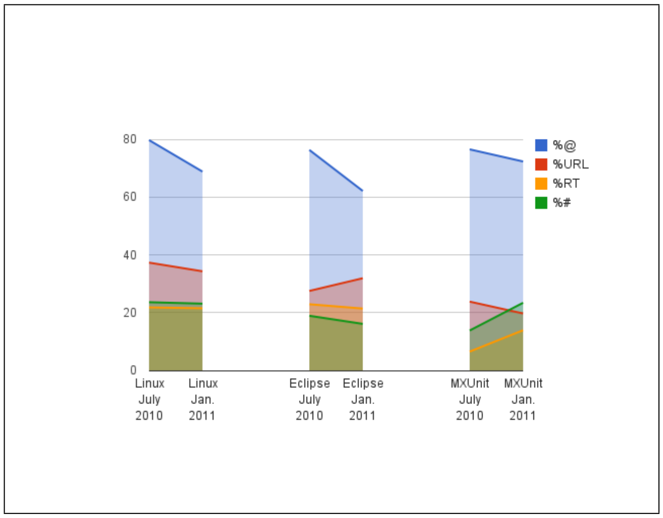

**Figure 1:** The change in amount of conversation (@), information sharing (URL's), retweets and hashtags seen from June/July 2010 to January 2011.

## 3.2 Topics of Discussion on Twitter

From our manual qualitative analysis of the messages posted by Software Engineers in our sample set, the following four categories of messages emerged:

1. Software engineering-related (often also work-related) topics (e.g. projects being worked on, seeking or providing technical help)
2. Gadgets and technological topics (e.g. iPhone 4, product news from Microsoft)
3. Current events outside of technical topics (e.g. FIFA World Cup, Ottawa earthquake in summer of 2010)
4. Daily chatter (e.g. family, weekend activities)

The Software Engineering categories we found overlap with those of Java *et al.*[15] in the first category, *daily chatter* and the fourth category, *reporting news* (in our case, *current events*). However, the second and third categories identified by Java *et al.* were conversation and information sharing, respectively. As outlined in the previous section, a significant portion of tweets made by Software Engineers fell into these two categories. Since Java *et al.* did not develop topic-related categories that were specific enough for comparison to our findings, we cannot draw distinctions between Twitter discussion among the general population of Twitter users and Twitter discussion among the Software Engineers in our sample set. We simply conclude that the communities of Software Engineers in our study tweet often about topics relating to Software Engineering and technology, as described below.

**Software Engineering-related topics.** In support of Honeycutt *et al.* [13], we found evidence of coherent conversations around Software Engineering topics taking place over Twitter. Software Engineering-related tweets accounted for about 23% of the tweets that were qualitatively analyzed. These cases included discussion of the current tasks developers are working on, and in several cases, attempts to find solutions to pertinent issues they encounter. For example, we found that *MXUnit#1* tweeted about a piece of code that identifies the current time on a system. After looking at the code, *MXUnit#2* tweets "@[*MXUnit#1*] almost positive you're right," and a few minutes later adds "@[*MXUnit#1*] my quickie test comparing it with the java way of doing it shows that getTickCount() works as you expect: now().getTime()". *MXUnit#3* then joins the conversation, identifying that the time returned by this piece of code might be offset by the local time zone. *MXUnit#4* adds his voice, with the comment that the getTickCount() function they are using is documented as having "... no meaning".

**Gadgets and technology topics.** Technology-related tweets accounted for about 29% of the tweets analyzed. These tweets consisted of tweets related to high-tech devices, topics and events (e.g., the release of a new product). For example, in July 2010, Apple made an announcement about the poor reception on the iPhone4 and how the company would give customers a protective case called a *Bumper* to reduce the reception issues. Discussion on this topic crossed all three communities in our sample set, with one Software Engineer asking, "what the [****] is a Bumper" (*MXUnit#2*). Much criticism of this as an acceptable solution followed.

**Current events.** Current events accounted for only 8% of the tweets we analyzed. These tweets were composed primarily of discussions covering politics, or major world events (e.g., B.P. Oil Spill, World Cup, etc.).

**Chatter.** The chatter topic accounted for about 36% of the tweets we analyzed. Java *et al.*[15] described daily chatter as "talk about daily routine or what people are currently doing"[15]. We use our chatter topic in a slightly different manner, as it refers to all tweets that don't fit into our other three categories. In many cases, this topic was made up of mentions of what developers do or think about besides work, such as "Heading out for an evening sail with family" (*Eclipse#1*), or suggesting that today should be an "early beer day" (*MXUnit#2*). In other cases, developers would talk about their upcoming schedules. For example, one developer mentioned that he would be leaving town for "3 days of Agile training starting tomorrow" (*MXUnit#2*). This "chatter" brings the personality of individuals into the community. It also provides an "ambient awareness" [20, 18] of a given developer's availability. Future work will involve exploring the tweets that fall into this category in more depth.

# 4. DISCUSSION AND FUTURE WORK

Although this research has provided significant insight into how Software Engineers use Twitter, it is far from finished, as many new questions and challenges have been discovered.

The main challenges and threats to validity within our study have to do with how we decided which data sets to examine. Through the use of *wefollow*, we were able to quickly locate communities of users discussing specific topics, such as Eclipse and Linux. However, the size of these communities differs greatly. For instance, the Eclipse *wefollow* list contains 141 users, whereas the Linux *wefollow* list contains 1,756 users. Additionally, the members of the Linux community tended to greatly "out tweet" the Eclipse users. When the number of tweets are normalized over days, the Linux tweets outnumber the Eclipse tweets by a factor of 4 on average (see Table 1). By reading the public profiles, we also discovered that although many of the top 30 Eclipse users appear to be Software Engineers or members of the Eclipse Foundation, the top 30 Linux users appear to include several Linux enthusiasts rather than developers. This raises an important question: How do we identify a Software Engineer? In other words, can Software Engineers be distinguished on Twitter from other users based on their topics of discussion or how they present themselves? Additionally, can these users be identified in an automated fashion? These are still open and challenging questions.

While performing qualitative analysis of Twitter conversations, it is important to be able to rebuild the context of conversations by way of exploring several previous messages that form the conversation threads. The tools currently available for exploring and analyzing Twitter data have proved frustrating. In many cases, accessing tweets that are several weeks or months old is difficult, both programmatically and manually, whether through the Twitter site or via current Twitter tools. Without a good set of tools, it will be difficult to use grounded methods to increase our knowledge of how different groups communicate through mediums like Twitter.

This paper explores micro-blogging usage by Software Engineers through Twitter alone. However, during our exploration, we became aware of a similar tool called *Identi.ca*, which is a micro-blogging service based on the Free Software tool *StatusNet*[3]. Our search on *wefollow* found 1,756 Twitter users who tweet about Linux. In contrast, the Linux group on Identi.ca contains 11,921 users. This raises additional questions, such as why the Linux community has chosen to adopt Identi.ca in greater numbers than was seen on Twitter. Does the fact that Identi.ca is Free and Open Source cause users from the FOSS community to gravitate towards it as opposed to a closed source alternative?

Previous studies have identified the existence of social hierarchies among OSS participants on developer mailing lists [8]. There are many open questions about the structures of communication within communities on micro-blogs and how they may also be affected by some form of "status'". For example, we wonder what types of structures might define Software Engineering communication through micro-blogs, and whether these structures would vary from community to community. Additionally, are the communication and social structures seen in the micro-blogging communities similar to those found in other forms of communication used by the same groups? These structures are of interest because, for example, if the Linux community has a similar communication structure on Twitter as it does on its mailing lists, this might suggest that a communication structure is defined by the community. However, if the structures of the two mediums differ greatly, it may imply that communication structures are influenced by the tools being used.

Finally, there are yet further questions relating to how Twitter might fill gaps that exist in other communication tools. The Eclipse developer whom we quoted in the introduction to this paper stated that he feels Twitter adds "an almost 'watercooler-like atmosphere' to the Eclipse community":

> Most of the existing mediums we use are strictly 'work related' (newsgroups, mailing lists, etc..), however, twitter brings a personal element to the whole thing. Reading about what hockey team someone cheers for, or what beer someone likes to drink provides an almost 'watercooler like atmosphere' to the Eclipse community – and in a distributed team, this is very important.

## 5. CONCLUSION

To the best of our knowledge, this is the first exploration into the use of micro-blogging services by Software Engineers. We have used both archival and grounded theory analysis to understand the conversation and community of three Twitter groups related to Software Engineering. We first compared some basic parameters of Twitter usage by Software Engineers to those of "average" Twitter users. We then manually analyzed a selection of 600 tweets to learn about the types of conversation and discussion taking place among Software Engineers over Twitter. We have also presented the challenges we faced in investigating the use of Twitter by a specific, interest-sharing population. Furthermore, this paper identifies a number of interesting questions and areas of future research that remain open.

Our preliminary findings indicate that Software Engineers make up a highly interactive micro-blogging population, with distinct sub-communities based around specific topic areas. Further data collection, as well as an online survey and interviews with Software Engineers who use micro-blogging services may help us develop a more in-depth understanding of their motivations for micro-blogging, as well as their personal approaches to and opinions of the technology. This knowledge has the potential to influence social media-enabled tool support to meet the needs of Software Engineers in a variety of settings.

## 6. REFERENCES

[1] *A Brief History of Microblogging. Available online at* `http://www.technologyreview.com/computing/21227`.

[2] *E-mail's Big Demographic Split. Available online at* `http://bits.blogs.nytimes.com/2010/12/21/e-mails-big-demographic-split/?hpw`.

[3] *Identi.ca. Available online at* `http://identi.ca/`.

[4] *MXUnit. Available online at* `http://wiki.mxunit.org`.

[5] *The Archivist - Save and Analyze Tweets. Available online at* `http://visitmix.com/labs/archivist-desktop`.

[6] *Twitter Hashtags. Available online at* `http://twitter.pbworks.com/Hashtags`.

[7] *wefollow. Available online at* `http://wefollow.com`.

[8] C. Bird, A. Gourley, P. Devanbu, M. Gertz, and A. Swaminathan. Mining email social networks. In *Proceedings of the 2006 international workshop on Mining software repositories*, pages 137–143. ACM, 2006.

[9] J. Corbin and A. Strauss. *Basics of qualitative research: Techniques and procedures for developing grounded theory.* Sage Publications, Inc, 2008.

[10] K. Ehrlich and N. Shami. Microblogging Inside and Outside the Workplace. 2010.

[11] B. Glaser and A. Strauss. *The discovery of grounded theory: Strategies for qualitative research.* Aldine Transaction, 2007.

[12] A. Guzzi, M. Pinzger, and A. van Deursen. Combining micro-blogging and IDE interactions to support developers in their quests. In *Software Maintenance (ICSM), 2010 IEEE International Conference on*, pages 1–5. IEEE.

[13] C. Honeycutt and S. Herring. Beyond Microblogging: Conversation and Collaboration in Twitter. *Proc 42nd HICSS.*

[14] B. Huberman, D. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1):8, 2009.

[15] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[16] Y. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. Tseng. Discovery of blog communities based on mutual awareness. In *Proceedings of the WWW06 Workshop on Web Intelligence.* Citeseer, 2006.

[17] W. Reinhardt. Communication is the key–Support Durable Knowledge Sharing in Software Engineering by Microblogging. In *Proc. of the SENSE Workshop, Software Engineering within Social Software Environments*, 2009.

[18] M. Skeels and J. Grudin. When social networks cross boundaries: a case study of workplace use of facebook and linkedin. In *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 95–104. ACM, 2009.

[19] A. Strauss and J. Corbin. *Basics of qualitative research.* Sage Newbury Park, CA, 1990.

[20] D. Zhao and M. Rosson. How and why people Twitter: the role that micro-blogging plays in informal communication at work. In *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 243–252. ACM, 2009.