Proceedings of the Ninth International AAAI Conference on Web and Social Media

Characterizing Silent Users in Social Media Communities

Wei Gong, Ee-Peng Lim, Feida Zhu
School of Information Systems
Singapore Management University
{wei.gong.2011, eplim, fdzhu}@smu.edu.sg

Abstract

Silent users often constitute a significant proportion of an online user-generated content system. In the context of social media such as Twitter, users can opt to be silent all or most of the time. They are often called the invisible participants or *lurkers*. As lurkers contribute little to the online content, existing analysis often overlooks their presence and voices. However, we argue that understanding lurkers is important in many applications such as recommender systems, targeted advertising, and social sensing. This research therefore seeks to characterize lurkers in social media and propose methods to profile them. We examine 18 weeks of tweets generated by two Twitter communities consisting of more than 110K and 114K users respectively. We find that there are many lurkers in the two communities, and the proportion of lurkers in each community changes with time. We also show that by leveraging lurkers' neighbor content, we are able to profile them with accuracy comparable to that of profiling active users. It suggests that user generated content can be utilized for profiling lurkers and lurkers in Twitter are after all not that "invisible".

Introduction

There has been a growing interest in an important group of users on social media sites such as Twitter and Facebook who choose to be silent most of the time and are therefore known as the *lurkers*. Lurkers contribute very little or no content, and prefer to consume content or perform other non-content-generating activities quietly. We call this the *lurking* behavior. In the paper, we call the other non-lurking users *active users*. As active users contribute most of the social media content, most of the existing social media research have focused on them but not the lurkers (Hannon, Bennett, and Smyth 2010; Uysal and Croft 2011; Nguyen et al. 2013). For example, when mining the topical interests and sentiments of users, one often does not consider lurkers as they do not generate sufficient content. This obviously leads to biased representation of topical interests and sentiments.

In many applications, it is very important to identify the lurkers, and their demographic attributes, interests and opinions. Despite their online silence, lurkers (like active users) are individuals with interests and preferences. They pay attention to topics of interest to them and will seek for relevant content. They have preferences that can potentially be expressed as ratings and reviews on consumer products. They are also potential customers for targeted marketing. It is possible for lurkers to have different demographic and opinion distribution from active users. Failing to account for lurkers could therefore lead to misjudgement of overall population-level interests. For example, Gayo-Avello (2012) pointed out that one of the main reasons that has caused the low election prediction accuracy using social media (i.e., Twitter) data is that "*The silent majority is a huge problem. Very little has been studied in this regard and this should be another central part of future research*".

Research Objectives. In this work, we study lurkers with two research goals. Our first goal is to define lurkers and characterize them in Twitter, which is chosen because it is where lurkers could most easily occur as a result of the ease of following others and, accordingly, the convenience of silent information consumption. We focus on 110,907 Twitter users from a Singapore-based community and 114,576 Twitter users from an Indonesia-based community. We examine the proportion of lurkers in these two Twitter communities, lurkers' social links with others and the motivations that may cause them to break silence. We identify the characteristics of lurkers by comparing them with active users. This gives us new insights into the lurking behavior and lurker's motivation of using Twitter. Note that this analysis is only possible with the availability of user tweets over a significant period of time as well as the follow relationships among the users. Therefore, we crawled all tweets posted by the users from the above two communities over 18 weeks and the follow links involving these users.

We define a lurker on Twitter as a user who is silent most of the time, i.e., he/she posts very few tweets during a given time interval. Using our Twitter datasets, we find that there are many lurkers in both communities. Compared with active users, lurkers have much fewer followers and followees. Both active users and lurkers are more likely to connect with active users. By sampling tweets and manually annotating them, we also found that a lurker breaks silence mainly to share information such as breaking news and updates of personal life.

Our second goal is to profile lurkers, i.e., to predict who the lurkers are and what they think. Unlike many existing user profiling works that exclude lurkers from their empirical studies due to their inadequate content data (Rao et al. 2010; Nguyen et al. 2013), we propose to utilize their neighbors' (their one-hop connected users) content to infer latent attributes including marital status, religion, and political orientation. We invest significant efforts in human annotation to obtain the ground truth labels. In our experiments, we compare the user profiling accuracy of lurkers with that of active users. The results show that using neighbors' content, we can predict lurkers' profile labels as accurate as active users' profile labels. It suggests that even lurkers do not generate much content, their profile attributes can still be uncovered from their neighbors. Therefore, it is indeed possible to personalize services for lurkers.

Related Work

Lurking in Online Communities

In traditional printed news media, lurking is almost the only activity it allows as all news articles are written by professional journalists leaving very few selected reader comments to appear in special news columns. Social media, in contrast, depends largely on users to contribute and share content. At the first glance, lurking might not be a desired user behavior on social media. Without enough users actively contributing content, the social media user community may shrink. In practice, however, lurking is a very common behavior found in many content providing sites (Nonnecke and Preece 2000; Muller et al. 2010; Antin and Cheshire 2010; Benevenuto et al. 2009; Bernstein et al. 2013). Benevenuto el al. (2009) in their work on user behavior in online social networks (such as myspace.com and LinkedIn), concluded that browsing actions (i.e., lurking) constitute 92% of all user actions. Only very few users contributed content. Nonnecke and Preece (2000) examined online discussion lists and showed 46% and 82% of users in health-support and software-support discussion lists respectively are lurkers. Muller et al. (2010) showed that 72.2% of users are lurkers in an enterprise file-sharing service. All these studies conclude that lurking is a common behavior among online users.

As online communities are interesting when there are many users contributing content, lurkers are described as free-riders (Kollock and Smith 1996). Free-riding carries a negative connotation especially in online community sites that require users to collaboratively generate or select content. There are therefore a few research works that focus on ways to motivate lurkers to contribute (Preece and Shneiderman 2009; Zhu et al. 2013). On the other hand, researchers also consider lurking as a passive form of participation which allows the online content to reach out to a wide audience (Soroka and Rafaeli 2006; Antin and Cheshire 2010).

Whether lurkers are negative free-riders or positive participants, there is no doubt that they form a significant share of the online communities which makes the research on them worthwhile (Soroka and Rafaeli 2006). In this paper, we shall not delve into the advantages nor disadvantages of lurking behavior to online systems. Instead, we focus on characterizing lurkers.

Reasons behind Lurking

As lurkers make up a significant proportion of users in online communities, several studies (Nonnecke and Preece 2001; Preece, Nonnecke, and Andrews 2004; Lampe et al. 2010) have focused on the reasons for lurking. Preece et al. (2004) conducted interviews and reported reasons such as (i) no need to post, (ii) personal privacy and safety concerns, (iii) shyness over public posting, and (iv) poor system usability. *No need to post* appears to be the top reason. In a survey conducted on a user-generated encyclopedia called Everything2.com, Lampe et al. (2010) reported that many users choose to lurk because they are satisfied with "getting information", as opposed to "sharing information". Antin and Cheshire (2010) found that Wikipedia users choose to lurk so as to learn enough about the site before they could actively contribute content. Similar findings of de-lurking behavior were also reported in other works (Preece and Shneiderman 2009; Rafaeli, Ravid, and Soroka 2004).

The reasons for lurking are closely related to the reasons for lurkers to break their silence. An interesting research question here is whether these reasons are the same as those of active users contributing content. In the context of Twitter, Java et al. (2007) identified four reasons for general users to post tweets, namely (i) daily chatter (i.e., personal updates), (ii) conversations (i.e., interacting with people), (iii) information/URLs sharing, and (iv) news reporting. Naaman, Boase, and Lai (2010) manually coded 400 tweets with nine category labels which include information sharing, self promotion, me now (i.e., personal activities), opinions/complaints, statements and random thoughts, and others. By analyzing 350 users and their posts (for each user, they randomly selected 10 posts without replies for analysis), they concluded that most users focus on personal updates. Alhadi, Staab, and Gottron (2011) conducted a survey of tweeting reasons on 1000 randomly selected tweets using Amazons Mechanical Turk. They found that social interaction is the top reason, followed by emotion (which covers personal updates and me now in (Naaman, Boase, and Lai 2010)).

Although the above studies identify the possible reasons for lurking and tweeting, they did not study the reasons for lurkers breaking silence and whether these reasons are any different from those of active users tweeting. Our paper therefore fills this gap by examining the motivations for lurkers posting tweets which may suggest new ways to encourage lurkers to generate more content.

From User Profiling to Lurker Profiling

User profiling (Rao et al. 2010; Li et al. 2012) aims to infer a user's attributes such as age, marital status, religion, political orientation, home location and interests using the observed data generated by the user and others. These inferred attributes are useful in categorizing users and providing them personalized services.

Previous user profiling research has shown that users' latent attributes can be inferred with reasonable accuracy based on the user-generated data including users' posted

content, and their social networks (Rao et al. 2010; Nguyen et al. 2013). However, many of these research works often leave out the lurkers as they do not provide rich content features. Hence, we are not able to ascertain the accuracy of attribute profiling for lurkers, and whether the accuracy for lurkers and active users are very different. In this work, we attempt to answer these questions.

There are some user profiling methods that explore the use of social links and neighbors' attributes (Mislove et al. 2010; Yang et al. 2011). For example, in (Yang et al. 2011), a model is proposed to propagate attribute labels among users via their social links. Nevertheless, not all attributes can be propagated (Li et al. 2012), e.g., home location, marital status, and gender. Such propagation-based methods are also less effective when the attribute labels are sparse among neighbors.

Our proposed approach to profile lurkers is more closely related to those methods making use of social links and neighbors' content (Li et al. 2012; Zamal, Liu, and Ruths 2012). As shown in our later data analysis, lurkers are likely to follow active users whose content are abundant. User profiling using neighbors' content has been shown to perform well for active users only in previous studies (Zamal, Liu, and Ruths 2012). In this work, we show that a good level of accuracy can also be achieved for the lurkers. To the best of our knowledge, such application and evaluation of user profiling on lurkers have not been conducted earlier.

Characteristics of Lurkers in Twitter

In this section, we first describe the Twitter dataset used in this part of research. Secondly, we define lurkers and examine the extent of lurking behavior in our dataset. Thirdly, we study the difference between lurkers and active users in terms of their social links. Finally, we examine the motivations behind lurkers breaking silence.

Data

We focus on two communities in Twitter: a Singapore-based community and an Indonesia-based community. We crawled these two communities using the following strategy. We started the crawling process with 69 and 123 popular seed users from Singapore and Indonesia respectively. We then added users who are one hop and two hops away from the seed users. They are the seed users' followers and followees and the followers and followees of the seed users' followers and followees. Finally, we chose the users who declare Singapore (or Indonesia) as their locations in the biography fields and share their tweets and social links in the public domain. We then obtained 140,851 Singapore-based users and 126,047 Indonesia-based users. This research requires a full set of tweets generated by users during a target study period which is very different from many other research projects that were performed on sampled tweet data. We then crawled **all** their tweets generated during an 18-week period which our analysis will focus on. For Singapore users, we crawled from April 28th to August 31st, 2014. For Indonesia users, we crawled from June 16th to Oct 19th, 2014.

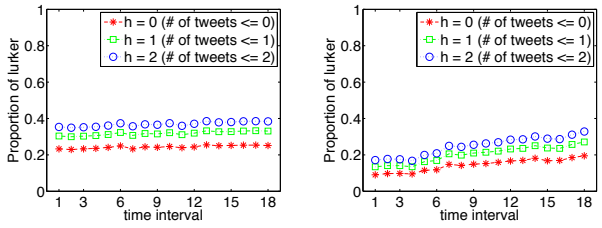
Removing churners A limitation of the above dataset is that it may include lots of users who already left Twitter. These users are known as *churners*. Churners do not post, so we may confound them with lurkers. Since we only have limited access to Twitter users' data (e.g., their tweets and connections), it is impossible for us to know *exactly* who are the churners. This problem has also been discussed in previous studies that aim to predict churners in social media (Oentaryo et al. 2012). Churners are then typically defined as the users who do not post for a significant long period time. Based on this definition, in order to filter away the churners from the above dataset, we use the following strategy. We crawled all the tweets that are posted by the 140,851 Singapore-based users and 126,047 Indonesia-based users for another 3 months after the 18-week period. If a user never posted during that 3 months, then we consider him/her as a churner and remove him/her from our dataset. In this way, we make sure that the users we analyze are confirmed "alive" during the 18-week period time. After removing the churners, we finally left with **110,907 Singapore-based users** and **114,576 Indonesia-based users**.

Lurkers in Twitter

Definition of lurker We say a user is *lurking* or a user is a *lurker* during a time interval with duration d, if the number of tweets he/she posts in the time interval is not more than a *lurking threshold* h. This definition caters to the time duration covered by the observed data. With the flexibility of varying time interval duration d and lurking threshold h, we are able to examine different degrees of posting behavior (i.e., never post or post only a few tweets) over time.

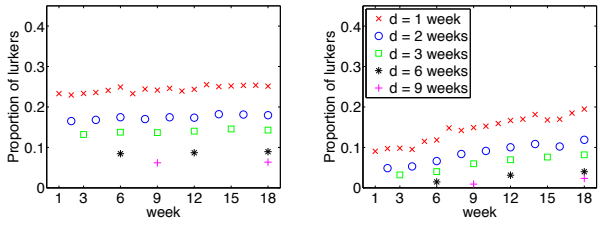
Proportion of lurkers in Twitter communities We empirically set d to be one week (d = 1 week) and vary h from 0 to 2, and derive the proportion of lurkers in the two communities across different disjoint time intervals over the 18 weeks. As shown in Figure 1(a), the proportion of lurkers in the Singapore community is nearly stable with a very small increasing trend. Every week, there are on average 24.4% of the users posting zero tweet (lurking threshold $h = 0$), 31.8% of the users posting no more than 1 tweet ($h = 1$), and 36.9% of the users posting no more than 2 tweets ($h = 2$). On the other hand, Figure 1(b) shows that the Indonesia community has smaller proportion of lurkers (e.g., on average 14.4% when $h = 0$), but the proportion increases steadily. Similar increasing trends are also observed when we use larger time interval duration d as shown in Figure 2. This figure shows that larger d has a smaller proportion of lurkers. Moreover, fewer users remain silent for longer time interval in both Twitter communities. It also implies that users may change their behavior from lurking to active between weeks.

Twitter users lurking behavior To explain the above findings, we model Twitter user behavior changes overtime in the following way. We use L_t (or A_t) to denote a user is lurking (or active) at time interval t. We use $x_t = P(L_{t+1}|L_t)$ to represent the probability that a user maintains lurking behavior from time t to $t + 1$. As a user who is lurking at t will either be lurking or active at $t + 1$,



(a) Singapore (b) Indonesia

Figure 1: Proportion of lurkers with $d = 1$ week.



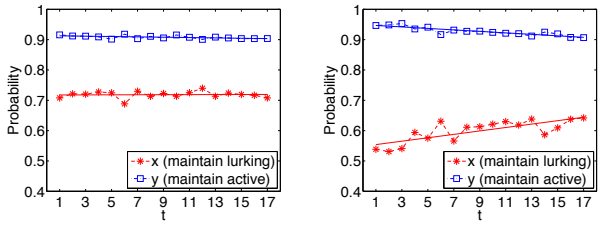
(a) Singapore (b) Indonesia

Figure 2: Proportion of lurkers with $h = 0$.

we therefore have $1 - x_t = P(A_{t+1}|L_t)$. Similarly, we use $y_t = P(A_{t+1}|A_t)$ to represent the probability that a user maintains active behavior from time t to $t + 1$, and $1 - y_t = P(L_{t+1}|A_t)$ to represent the probability of a user active at t but lurking at $t + 1$.

Given a set of users U and their lurking and active behavior from time 1 to T, i.e., $D = \langle \{U_A^1, U_L^1\}, \ldots, \{U_A^T, U_L^T\} \rangle$ where U_A^t and U_L^t represents the set of active users and the set of lurkers at t respectively, we can estimate x_t and y_t by $x_t = P(L_{t+1}|L_t) = \frac{|U_L^t \cap U_L^{t+1}|}{|U_L^t|}$ and $y_t = P(A_{t+1}|A_t) = \frac{|U_A^t \cap U_A^{t+1}|}{|U_A^t|}$. With the time interval duration d as one week and $h = 0$, Figure 3 shows the probability of maintaining lurking ($x_t = P(L_{t+1}|L_t)$) and the probability of maintaining active ($y_t = P(A_{t+1}|A_t)$) from $t = 1$ to $t = 17$ in the two communities. We also plot the trend line of x_t and y_t. Note that we have consistent findings with different duration d and lurking threshold h settings.

The result suggests that generally lurkers are more likely to stay lurking and active users are also more likely to stay active (i.e., $x_t > 1 - x_t$ and $y_t > 1 - y_t$). There are users changing their behavior between weeks but with a small probability (e.g., $1 - x_t < 0.5$ and $1 - y_t < 0.1$ in Indonesia community). It is more likely for users go from lurking to active than from active to lurking as $1 - x_t > 1 - y_t$. This trend however may not continue forever for our two Twitter communities. In both communities, the probability of user maintaining lurking (x) has an increasing trend, and the probability of user maintaining active (y) has a decreasing trend. Comparing the two communities, we see the probability of user maintaining lurking (x) from Indonesia community has a higher gradient (0.0001 for Singapore and 0.0056 for Indonesia), and the probability of user remaining active (y)



(a) Singapore (b) Indonesia

Figure 3: Probability of maintaining lurking and maintaining active.

from Indonesia also has a higher negative gradient (-0.0006 for Singapore and -0.0024 for Indonesia). This explains that we see the proportion of lurkers grows in both communities and the lurking behavior in Indonesia Twitter community shows a clear increasing trend in Figures 1 and 2.

In summary, we first observe there is a significant proportion of lurkers in our two Twitter user communities. Secondly, more users prefer to maintain their lurking or active behavior than to change their behavior. Finally, the proportions of lurkers in both communities are growing with different trends.

Lurker's Social Connections

A major difference between Twitter and other online community platforms such as Wikipedia is the presence of social connections among Twitter users. Lurkers in Twitter are therefore not entirely "invisible", as they may follow others or being followed by others. In Twitter, if user u follows another user v, we say u is v's follower, and v is u's followee. If v also follows u, we say they are friends with each other.

We first define the following social connection measures. User u's *in-reciprocity ratio* measures the proportion of u's followers who are followed back by u (i.e., $\frac{u\text{'s friend count}}{u\text{'s follower count}}$). User u's *out-reciprocity ratio* measures the proportion of u's followees who follow back u (i.e., $\frac{u\text{'s friend count}}{u\text{'s followee count}}$). User u's *lurker-follower ratio*, *lurker-followee ratio*, *lurker-friend ratio* measures the proportion of lurkers among u's followers, followees and friends respectively.

We then divide the two communities' users into lurkers and active users setting the time interval duration d as 18 weeks and $h = 5$. Thus we have 10,170 lurkers and 100,737 active users in the Singapore community, and 2,060 lurkers and 112,516 active users in the Indonesia community. We have tried other time interval durations and lurking thresholds and they do not affect the following findings.

Table 1 summarizes the lurkers and active users' social connections using different measures. The difference between lurkers and active users by every measure is statistically significant ($p < 0.05$ using Two-Sample t-test). We observe that lurkers are likely to have fewer followees and followers than active users. This suggests that lurkers are less interested in following others, and also are less attractive for others to follow. Despite this finding, lurkers have reasonable number of followees (median:85 in Singapore and median:145 in Indonesia) as they need to follow others to get information. We also notice that lurkers have followers (me-

	U	Singapore Med.	Singapore Mean	Indonesia Med.	Indonesia Mean
Followee count	L	85	189.6	145	266.3
	A	193	345.8	275	447.7
Follower count	L	42	166.4	60	233.5
	A	167	875.1	299	2400.8
Out-reciprocity ratio	L	0.25	0.31	0.21	0.29
	A	0.51	0.50	0.53	0.52
In-reciprocity ratio	L	0.56	0.53	0.59	0.55
	A	0.63	0.57	0.52	0.50
Lurker-followee ratio	L	0.04	0.1	0	0.01
	A	0.04	0.07	0	0.01
Lurker-follower ratio	L	0	0.16	0	0.03
	A	0.06	0.11	0	0.02
Lurker-friend ratio	L	0	0.12	0	0.02
	A	0.03	0.08	0	0.01

Table 1: Summary of social connections. L represents lurkers and A represents active users.

dian:42 in Singapore and median:60 in Indonesia) although they do not post many tweets. One possible reason is that a lurker is followed by the users who know him/her offline. Another possible reason is that a lurker could be active before, and gained the followers during that time.

Reciprocity of social links is a very prevalent pattern in social networks. Kwak et al. (2010) showed that link reciprocity ratio in Twitter is expected to be around 0.22. In other social networks (e.g., Flickr, Yahoo, etc.), the reciprocity ratio is much higher. In Table 1, our results show that the out- and in-reciprocity ratios are around 0.5 for the active users. These numbers are reasonable as we consider only follow links among users from the same community (Singapore or Indonesia). The findings on reciprocity ratio of lurkers are on the other hand quite different.

The out-reciprocity ratio result shows it is much less likely for users to follow back to a lurker than an active user because lurkers offer little information and social interactions. For in-reciprocity, the Singapore lurkers are slightly less likely to follow back to their followers than active users. Lurkers from Indonesia community are slightly more likely to follow back. This result may be caused some culture difference between the two communities in following back behavior.

Finally, we also observe that the proportion of lurkers among both lurkers and active users' followers, followees and friends are very small. It reveals both lurkers and active users prefer to connect with active users.

Lurker's Motivations for Speaking Out

Lurkers choose to remain silent and prefer to be an observer. However, although very infrequently, lurkers may be triggered to break silence. *What drives a user who prefers silent to speak out? Are the motivations to speak out different among the users with different activity levels (e.g., from lurkers, normal active users to very active users)?* We aim to answer these questions so as to gain insights about lurkers' behavior. This analysis will shed some lights on how to encourage lurkers to be more active in content generation. This part of study focused on Singapore users only as many Indonesia users do not write in English.

Motivations Based on the theory of Use and Gratification (U&G) (Papacharissi and Rubin 2000), we know that people like to contribute to a media product because using it gratifies their needs. From this theory, Rafaeli et al. (2009) derived three motivations for using and contributing to Wikipedia, and they are information seeking, information sharing and entertainment.

In the case of Twitter, users can see it as a social platform and/or a media. They therefore use Twitter to get information such as current news and friends' updates, to interact with other Twitter accounts such as friends, celebrities and organizations, to share messages relating to personal activities or thoughts, to share information such as breaking news and interesting videos, books and games, etc., and to do advertisement. Among them, getting information is likely the main motivation (or need) for lurkers using Twitter (Lampe et al. 2010). In order to interact with people, share (personal or public) information, and conduct advertisement, one needs to de-lurk.

Manual motivation labeling To carefully determine the motivations for lurkers breaking silence, we first assign motivation labels to their tweets. For example, consider a tweet about a conversation between the tweet author and his/her friends *"@<User Name> Yea, see you tomorrow! Good night!"*. This is motivated by the need for social interaction. The questions now are therefore "what are the different motivation labels out there?" and "how these labels can be assigned to the lurker and non-lurker's tweets?"

Even with the tweet content at hand, assigning motivations to tweets is not an easy task. Nagarajan et al. (2010) applied some simple heuristic rules to study user engagement in communities. For example, they defined a rule to assign conversational label to tweets that *"made references to other Twitter users utilizing the @user handle"*. The heuristic rule is however not always correct. For example, a tweet such as *"I love @<Celebrity Name>. She is doing great in the show!"* suggests that the user shares self opinions or thoughts rather than interacts with the celebrity. Therefore, most previous works (Java et al. 2007; Naaman, Boase, and Lai 2010; Alhadi, Staab, and Gottron 2011; Toubia and Stephen 2013) that attempt to understand user intentions in writing tweets have resorted to manual effort to label tweets.

We manually assign motivations to tweets using a multiple-round approach mentioned in (Naaman, Boase, and Lai 2010). We first randomly selected 100 tweets. Then two coders (who are the author and another experienced social media researcher) independently labeled them with a set of motivation labels while writing down the reasons for choosing a certain motivation. Note that coders can assign multiple labels to one tweet. The two coders discussed and modified the set of motivation labels and the reasons of choosing them. We performed the above tasks three rounds (each round with a new set of 100 tweets) before finalizing the motivation label set and a common interpretation of the labels.

We measure the agreement of two coders using Jaccard

Motivations	Resaons	Description
Information Sharing	News General Information	Share latest news, or trending events Share alerts, knowledge, videos, jokes and games, etc.
Personal Update	Activity Emotion Opinion Thought	Update activities and status Express emotions and feelings towards self Express opinions and feelings towards other things Express random thoughts and statements
Friend Interaction	Chat Mention	Chat with friends Mention friends to get their attention
Public Interaction	Request Voice	Queries or ask for feedback and advice Chat with celebrities, organizations or customers.
Advertisement	Commercial related. Non-commercial related	Post commercial related advertisements and promotions Promote charitable institutions and political organizations, etc.

Table 2: Motivations of sending tweets.

Coefficient which is commonly used to measure similarity between two sets. Given a tweet i, if one coder assigns a set of motivation labels A, and the other coder assigns another set B, then the Jaccard Coefficient of this tweet is $J_i = \frac{|A \cap B|}{|A \cup B|}$. The agreement of two coders for a set of tweets I is then the average Jaccard Coefficient among all tweets, i.e., $J = \frac{\sum_{i \in I} J_i}{|I|}$. In the third round, the coders achieved 0.82 average Jaccard Coefficient. We believe this is a reasonable agreement and therefore finalized the set of motivation labels as shown in Table 2. The labels are *information sharing*, *personal update*, *friend interaction*, *public interaction* and *advertisement* and are described in the table. In this table, we use 'information sharing' label for sharing information that are not personal while 'personal update' label for sharing personal information.

We then recruited three coders (two of them are not authors of this paper) to label a new and much larger set of tweets from users of different activity levels, namely, *lurkers, normal-low active users, normal-high active users and very active users*. They post $[1, 5]$, $[6, 200]$, $[201, 1000]$, and $[1001, +\infty]$ tweets respectively within our observed 18 weeks.

We sampled tweets to be labeled from the same time period (July 14 to July 27, 2014) in the following way. For users of each activity level, we first randomly selected 400 of them who published at least one tweet during the time period. Then for each user, we sampled one of his/her tweets. Among the 1600 tweets we sampled, 307 tweets are not written in English and were thus discarded. The agreements of every two out of three coders are 0.81, 0.83 and 0.81 respectively measured by average Jaccard Coefficient. We then use majority vote to determine the final motivation label(s) for each tweet, i.e., a label is assigned to a tweet if this label is agreed by at least two coders. We also discarded tweets (22 of them) that are assigned completely different labels. We were left with 326, 339, 303, and 303 tweets for lurkers, normal-low active users, normal-high active users and very active users respectively for motivation analysis.

Results For a user type U, the proportion of tweets triggered by motivation label m is defined as $\frac{\text{No. of tweets from } U \text{ with label } m}{\text{Total No. of tweets from } U}$. As one tweet can have multiple labels, the sum of the proportion of tweets triggered by different motivations is greater than or equal to 1. Figure 4 shows the proportion of tweets assigned with different labels for each user activity level.

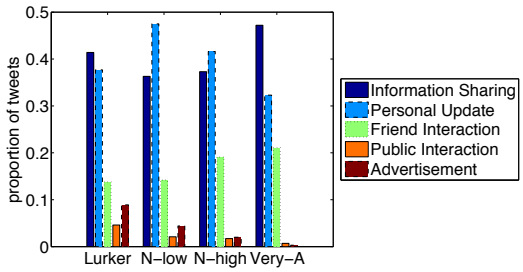


Figure 4: Proportion of tweets assigned with different motivation labels among lurkers (Lurker), normal-low active users (N-low), normal-high active users (N-high), and very active users (Very-A).

The result shows that information sharing and personal update are the top two motivations of speaking out across all user types. For lurkers and very active users, information sharing label is assigned to more tweets than personal update, whereas for normal-low and normal-high active users, personal update is assigned to slightly more tweets. Intuitively, a person is expected to have limited personal matter to update and limited number of friends to interact with. Tweets posted by very active users are therefore more likely motivated by information sharing than other motivations. For lurkers, the result suggests that lurkers are more likely to break silence when they encounter interesting matters and breaking news compared with other motivations.

Other than information sharing and personal update, the friend interaction label has been assigned to a significant proportion of tweets across all user types. Compared with active users, lurkers have the lowest proportion of tweets that are assigned with the friend interaction label. When users decide to post tweet, the active users are more likely to interact with friends.

Across all user types, public interaction and advertisement motivate the least proportion of tweets. Compare with

Date	Top popular words from all lurkers' tweets	Top popular hashtags from all users' tweets
July 14th	#singapore, team, deco, unzipped, hoodie, messi	#ger, #worldcup, #sosingaporean, #cfcinbrazil, #cfc, #singapore
July 15th	#glendaph, game, typhoon, concert, ko, @abscbnnews	#cfc, #welcomediego, #notosofitel, #boycottsofitelph, #singapore, #freepalestine
July 16th	sa, po, @youtube, pl, @lovehindishows, #geraldd987xijtpxhunterhayes	#cfc, #cfclive, #notosofitelday3, #boycottsofitelphday3, #singapore, #notosofitelday4
July 17th	**#mh17**, **mh17**, **malaysia**, **ukraine**, **airlines**, **plane**	**#mh17**, **#prayformh17**, **#ukraine**, **#malaysiaairlines**, #singapore, #prayforgaza
July 18th	**#mh17**, @youtube, **mh17**, sa, recruiting, @9vska	#cfc, **#mh17**, **#prayformh17**, #singapore, **#malaysiaairlines**, #gaza
July 19th	**#mh17**, inadh, **mh17**, chance, installed, battery	#mtvhottest, #cfclive, **#mh17**, #cfc, #zaynappreciationday, #singapore
July 20th	#prayforgaza, god, stats, @iam, #vaalutrailer, business	#mtvhottest, #liamappreciationday, #twitterpurge, **#mh17**, #sgxclusive, #singapore

Table 3: Top words from lurkers and Top hashtags from all users. The words and hastags are ordered according to the number of users adopting them.

active users, lurkers have the highest proportion of tweets that are labeled as public interaction. These are tweets for interacting with public figures, celebrities, or organizations which typically do not lead to further conversations. When users post tweets, compared with active users, lurkers are more likely to have conversations that do not enhance their social connectivity. Similarity, we also find that lurkers are more likely to post advertising tweets which again, typically do not enhance their social connectivity.

Popular topics among lurkers The above findings show a major reason that lurkers break silence is to share something interesting and breaking. We now look into the topics in tweets generated by lurkers in large scale. The purpose is to have a deeper understanding of what events or topics that are likely to motivate many lurkers to break silence. We compare the top popular words (excluding the stop words) posted by all lurkers and top popular hashtags used by all users each day from July 14th to July 20th, 2014 (see Table 3). Hashtags (i.e., #some-keyword) on Twitter are used to mark topics in tweets for categorization purposes. Very popular hashtags are often trending topics. Therefore, the top popular hashtags used by all users are the topics that draw the most interest.

During the period July 14th to July 20th, 2014, we observed that there are three common topics (or events) popular among both lurkers and all users. We marked them differently. The words in magenta also underlined are related to World Cup 2014 which ended on July 14th Singapore time (July 13rd in Brazil time). The words in blue and also **boldfaced** are related to a Malaysia airline crash tragedy on July 17th. And the words in red (also marked with boxes) are related to Gaza-Israel conflict 2014 which begins from July 8th. Hashtag #singapore is popular among Singapore Twitter users. We do not discuss it because it is often used for specifying the location of the events rather than describing topics.

The results show that when a global event such as the World Cup closing or Malaysia airline tragedy occurs, it becomes the top topic that triggers lurkers to break silence. In a normal day such as July 15th and 16th, lurkers do not follow general trends of hashtag adoption such as #cfc (the Chelsea football club). Gaza-Israel conflict 2014 as an event started about one week earlier was also popular among lurkers and other users, but in different dates. It suggests that this event was still globally aware but no longer "breaking".

Lurker Profiling

Another goal of this work is to profile lurkers and answer two questions: *How accurate are we able to profile lurkers? And are the performance of profiling lurkers and active users very different?* We choose to profile three attributes including marital status, religion, and political orientation. In this Section, we first describe the dataset used in each attribute profiling task. We then describe the methods of profiling lurkers. Finally we show the profiling performance.

Data

We use Singapore-based users with ground truth attribute labels in this part of research. To derive the ground truth of users' *marital status* and *religion*, we define several keywords and phrases related to the respective attribute label and use them to select the subsets of users for manual labeling (Nguyen and Lim 2014). For example, married users are likely to mention "wife", "husband", "my son", and "my daughter", while single users may mention "dating", "girlfriend", "my gf", "boyfriend", and "in a relationship". Christians are likely to mention "jesus", "christ", "protestant", "catholic", and "church", while Muslim users may often mention "allah", "muslim", "islam", and "mosque". We selected users whose biography field includes these keywords or phrases relevant to the marital status and religion and then assigned the attribute labels after manually reading the biographies. For religion attribute, we focus on profiling Christians and Muslim users, as much fewer Singapore Twitter users of other religions (e.g., Buddhists and Hindu, etc.) mention their religion beliefs in their biography fields.

	Marital status		Religion		Political orientation	
User group	Married	Single	Christian	Muslim	Opposition	Ruling party
0-MAX (All Users)	1329	1556	403	258	5002	2481
$[0, 5]$ (Lurkers)	331	268	70	29	1110	427
$[6, 50]$	361	310	94	31	1136	362
$[51, 200]$	302	284	108	38	1171	431
$[201, MAX]$	335	694	131	160	1585	1261

Table 4: Label distribution in our datasets

The above approach unfortunately does not work well when determining the ground truth labels of users' *political orientation*. This is because very few Singapore Twitter users publicly declare their political orientation. We therefore adopt a similar method that was first introduced in (Hoang et al. 2013) in which a few seed Twitter accounts owned by different political parties are used to propagate political affiliation labels. These seed accounts either belong to the *Ruling party* or the *Opposition*. If a user follows two or more seed political accounts and they all belong to only one party, we label the user with the respective political affiliation. Manual checks on a few labeled users verified that these assigned labels are accurate. In this way, we obtained the ground truth label of ruling party and opposition affiliated users.

Table 4 shows a summary of our datasets corresponding to the three attributes to be profiled. We obtained 2885, 661, and 7483 users with marital status, religion, and political labels respectively. To evaluate the accuracy of lurker profiling, we divide each set of labeled users into four groups according to their activity levels, i.e., the number of tweets they post during the 18 weeks from April 28, 2014 to August 31, 2014. For example, the lurker group is represented by the users who post no more than 5 tweets during the 18 weeks. We then have 331 married lurkers and 268 single lurkers.

As shown in Table 4, the distribution of users in different attribute classes is different for users with different activity levels. In the marital status dataset, among the most active users ($[201, MAX]$), there are much more single users than married users, whereas among the less active users ($[0, 5]$, $[6, 50]$, and $[51, 200]$), there are more married users. A similar situation also applies to the religion dataset. In the political orientation dataset, although opposition users are always the majority, they significantly outnumber the ruling party users in the less active user groups. This implies that lurkers may have very different profile composition compared with the active users.

Profiling Methods

We define four types of tweet content features to develop our profiling methods. These include the content of tweets posted by (a) the user, (b) the user's followees, (c) the user's followers and (d) the user's friends respectively. For lurkers, using their posted tweets is likely to give low accuracy. Our purpose is to evaluate methods using the tweets from the lurker's followees, followers or friends can help improve the profiling performance for lurkers. We also compare the accuracy of profiling lurkers against that of active users.

For each type of features, we apply Naive Bayes (NB) (Manning, Raghavan, and Schütze 2008) and Support Vector Machine (SVM) to learn classifiers. For SVM, we use LIBSVM package (Chang and Lin 2011) and TF-IDF of words in tweets as features. All the methods remove stop words from the tweets before training. We use F-score of the minority class to evaluate the profiling results since the datasets are skewed. In our experiment, we apply 5-fold cross validation to derive the average F-score. At each round, we train a classifier, then apply this classifier to different activity levels of the users in the testing set. In this way, we obtain the profiling results for the users in $[0, 5], \ldots, [201, MAX]$ and $[0, MAX]$ group. We use a random predictor as baseline. The F-score for a random predictor is computed as $\frac{\text{number of samples in the minority class}}{\text{total number of samples}}$. The minority class is determined from the training datasets. They are the married, Muslim and ruling party classes for marital status, religion, and political orientation attributes respectively. For our datasets, we find that NB can achieve comparable and often better performance than SVM. Therefore, to ease of the comparison, we only show the results using NB as the classification algorithm.

Profiling Results

Figures 5, 6 and 7 show the prediction results for marital status, religion and political orientation respectively. We summarize the main findings as follows. First of all, as we expected, using users' tweets to predict attributes does not work well for the lurkers who do not post enough tweets (see performance on lurker group $[0, 5]$). Especially in the prediction of marital status, using users' tweets (F-score $= 0.46$) performs much worse than the random predictor (F-score $= 0.55$). However, we find using one-hop neighbors' (i.e., followees, followers or friends) tweets can achieve significantly better performance than using the random predictor and users' tweets in predicting lurkers' attributes. On the other hand, for active users who posted tweets in the range of $[6, 50]$, $[51, 200]$ and $[201, MAX]$, using users' tweets and user neighbors' tweets outperforms the random predictor significantly.

Secondly, we find the methods using followees' tweets usually outperform the methods using followers and friends' tweets when predicting marital status and political orientation. However, the methods that predict religion using followers and friends' tweets perform better than followees' tweets. Previous works often believe followees can better reveal a user's attribute as users can control who they follow but cannot control who follow them (Zamal, Liu, and Ruths 2012). Our results show it is not always the case and suggest that we should also make use of the tweets generated

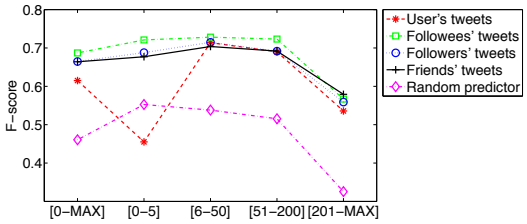


Figure 5: Marriage status prediction performance.

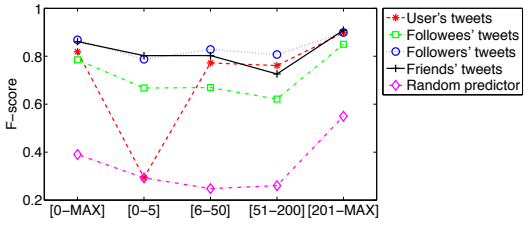


Figure 6: Religion prediction performance.

by followers and mutual friends in the future.

Last but not least, we find that inferring lurkers' attributes is not always harder than active users. In our results, we see that using followees' tweets for marital status, followers and friends' tweets for religion, and followees' tweets for political status can achieve similar performance as profiling active users.

Discussion

In this Section, we discuss our findings, the limitations of our work and the future directions.

Lurkers should not be neglected. The problem of *silent users* has been pointed out in a few previous works (Mustafaraj et al. 2011; Gayo-Avello 2012; Lin et al. 2013). As a major function of social media is making social connections, we believe that it is meaningful to study the silent users in a community setting. From our analysis on Singapore and Indonesia user communities, we find that lurkers make up a significant proportion of users (see Figures 1 and 2). It suggests a large number of lurkers can be easily overlooked when inferring opinion, interest, attitude or preference at the population level by aggregating tweets. Furthermore, as the size of lurker group is growing (see Figure 3), it is crucial for a social media to keep lurkers interested in returning. In other words, a healthy social media should be able to attract audience. Thus it is important to create a pleasant and interesting space to draw lurkers' attention continuously. For example, in Google+ (plus.google.com), there are *What's Hot and Recommended* messages and people *You may know* in a user's timeline. While existing research often focuses on recommending content for active users (Hannon, Bennett, and Smyth 2010; Uysal and Croft 2011), it is also important to do the same for lurkers.

Profiling lurkers is possible. To prevent misrepresentation of lurkers, we are compelled to use features beyond user

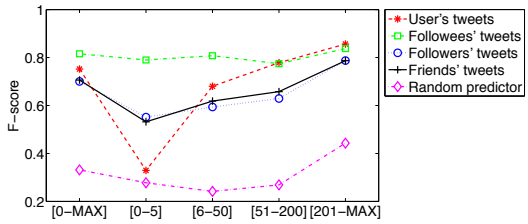


Figure 7: Politic orientation prediction performance.

generated tweets to profile them. Our study shows that they are still connected with many active users (see Table 1). We demonstrate that it is possible to profile lurkers by leveraging the content generated by active users and the links between active users and lurkers. For attributes 'marital status', 'religion', and 'political orientation', we are able to profile lurkers with accuracy comparable to that of profiling active users. This result suggests it is possible to infer other lurkers' latent attributes and the techniques introduced in this paper can be adopted. This will also enable lurkers to enjoy personalized services such as search, recommendation systems and advertising.

Limitations and Future Research

Our study has limitations which we hope can be addressed in the future research. Firstly, we do not differentiate the lurkers in different "lurking" levels. For example, some lurkers like to login Twitter and spend a lot of time reading, but some don't. Distinguishing them would be useful to the studies that aim to attract lurkers (i.e., audience) for a social media. For example, we could examine the factors that contribute to lurkers visiting Twitter often. Users' invisible activities such as login data and click history are needed to know users' "lurking" levels. However, collecting such data could lead to privacy concerns.

Secondly, our methods of profiling users' latent attributes are rudimentary. We infer lurkers attributes from the tweets generated by their one-hop neighbors. Future work could consider the network features of users to improve the lurker profiling accuracy.

Lastly, lurkers' behavior can be further explored in other aspects. For example, what makes a user become a lurker. Are lurkers born to be lurkers? If no, what causes active users to become lurkers? What are the differences between lurker and active user behavior outside of social media (West, Weber, and Castillo 2012)?

Conclusion

To the best of our knowledge, this is the first work that conducts a systematic study on lurkers and their behavior in Twitter. We also show that profiling lurkers can be as accurate as profiling active users. Considering 1) the size of lurker population is significant and growing, and 2) lurkers are the potential customers and audience, we suggest that future research could place more emphasis on understanding them so as to make social media a more desired place to keep lurkers engaged and possibly to make them active.

Acknowledgments

The authors gratefully thank two coders for their help in the tweets annotation task. This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA). This work is partially supported by Pinnacle Lab at Singapore Management University.

References

Alhadi, A. C.; Staab, S.; and Gottron, T. 2011. Exploring User Purpose Writing Single Tweets. In *WebSci 2011*.

Antin, J., and Cheshire, C. 2010. Readers Are Not Free-riders: Reading As a Form of Participation on Wikipedia. In *CSCW 2010*.

Benevenuto, F.; Rodrigues, T.; Cha, M.; and Almeida, V. 2009. Characterizing User Behavior in Online Social Networks. In *IMC 2009*.

Bernstein, M. S.; Bakshy, E.; Burke, M.; and Karrer, B. 2013. Quantifying the Invisible Audience in Social Networks. In *CHI 2013*.

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* 2(3):27:1–27:27.

Gayo-Avello, D. 2012. "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper"-A Balanced Survey on Election Prediction using Twitter Data. *arXiv preprint arXiv:1204.6441*.

Hannon, J.; Bennett, M.; and Smyth, B. 2010. Recommending Twitter Users to Follow Using Content and Collaborative Filtering Approaches. In *RecSys 2010*.

Hoang, T.-A.; Cohen, W. W.; Lim, E.-P.; Pierce, D.; and Redlawsk, D. P. 2013. Politics, Sharing and Emotion in Microblogs. In *ASONAM 2013*.

Java, A.; Song, X.; Finin, T.; and Tseng, B. 2007. Why We Twitter: Understanding Microblogging Usage and Communities. In *WebKDD/SNA-KDD 2007*.

Kollock, P., and Smith, M. 1996. Managing the Virtual Commons: Cooperation and Conflict in Computer Communities. In Herring, S., ed., *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives*.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is Twitter, a Social Network or a News Media? In *WWW 2010*.

Lampe, C.; Wash, R.; Velasquez, A.; and Ozkaya, E. 2010. Motivations to Participate in Online Communities. In *CHI 2010*.

Li, R.; Wang, S.; Deng, H.; Wang, R.; and Chang, K. C.-C. 2012. Towards Social User Profiling: Unified and Discriminative Influence Model for Inferring Home Locations. In *KDD 2012*.

Lin, Y.-R.; Margolin, D.; Keegan, B.; and Lazer, D. 2013. Voices of Victory: A Computational Focus Group Framework for Tracking Opinion Shift in Real Time. In *WWW 2013*.

Manning, C. D.; Raghavan, P.; and Schütze, H. 2008. *Introduction to Information Retrieval, Naive Bayes Text Classification*. New York, NY, USA: Cambridge University Press.

Mislove, A.; Viswanath, B.; Gummadi, K. P.; and Druschel, P. 2010. You Are Who You Know: Inferring User Profiles in Online Social Networks. In *WSDM 2010*.

Muller, M.; Shami, N. S.; Millen, D. R.; and Feinberg, J. 2010. We Are All Lurkers: Consuming Behaviors Among Authors and Readers in an Enterprise File-sharing Service. In *GROUP 2010*.

Mustafaraj, E.; Finn, S.; Whitlock, C.; and Metaxas, P. T. 2011. Vocal Minority Versus Silent Majority: Discovering the Opionions of the Long Tail. In *Privacy, security, risk and trust (passat), SocialCom 2011*.

Naaman, M.; Boase, J.; and Lai, C.-H. 2010. Is It Really About Me?: Message Content in Social Awareness Streams. In *CSCW 2010*.

Nagarajan, M.; Purohit, H.; and Sheth, A. P. 2010. A Qualitative Examination of Topical Tweet and Retweet Practices. In Cohen, W. W., and Gosling, S., eds., *ICWSM 2010*.

Nguyen, M.-T., and Lim, E.-P. 2014. On Predicting Religion Labels in Microblogging Networks. In *SIGIR 2014*.

Nguyen, D.; Gravel, R.; Trieschnigg, D.; and Meder, T. 2013. "How Old Do You Think I Am?" A Study of Language and Age in Twitter. In *ICWSM 2013*.

Nonnecke, B., and Preece, J. 2000. Lurker Demographics: Counting the Silent. In *CHI 2000*.

Nonnecke, B., and Preece, J. 2001. Why Lurkers Lurk. In *AMCIS 2001*.

Oentaryo, R. J.; Lim, E.-P.; Lo, D.; Zhu, F.; and Prasetyo, P. K. 2012. Collective Churn Prediction in Social Network. In *ASONAM 2012*.

Papacharissi, Z., and Rubin, A. M. 2000. Predictors of Internet Use. *Journal of Broadcasting & Electronic Media* 44(2):175–196.

Preece, J., and Shneiderman, B. 2009. The Reader-to-Leader Framework: Motivating Technology-Mediated Social Participation. *AIS Transactions on Human-Computer Interaction* 1(1):13–32.

Preece, J.; Nonnecke, B.; and Andrews, D. 2004. The Top Five Reasons for Lurking: Improving Community Experiences for Everyone. *Computers in Human Behavior* 20(2):201–223.

Rafaeli, S.; Hayat, T.; and Ariel, Y. 2009. Knowledge Building and Motivations in Wikipedia: Participation as "Ba.". *Cyberculture and New Media* 51–68.

Rafaeli, S.; Ravid, G.; and Soroka, V. 2004. De-lurking in Virtual Communities: A Social Communication Network Approach to Measuring the Effects of Social and Cultural Capital. In *HICSS 2004*.

Rao, D.; Yarowsky, D.; Shreevats, A.; and Gupta, M. 2010. Classifying Latent User Attributes in Twitter. In *SMUC 2010*.

Soroka, V., and Rafaeli, S. 2006. Invisible Participants: How Cultural Capital Relates to Lurking Behavior. In *WWW 2006*.

Toubia, O., and Stephen, A. T. 2013. Intrinsic vs. Image-Related Utility in Social Media: Why Do People Contribute Content to Twitter? *Marketing Science* 32(3):368–392.

Uysal, I., and Croft, W. B. 2011. User Oriented Tweet Ranking: A Filtering Approach to Microblogs. In *CIKM 2011*.

West, R.; Weber, I.; and Castillo, C. 2012. Drawing a Data-driven Portrait of Wikipedia Editors. In *WikiSym 2012*.

Yang, S.-H.; Long, B.; Smola, A.; Sadagopan, N.; Zheng, Z.; and Zha, H. 2011. Like Like Alike: Joint Friendship and Interest Propagation in Social Networks. In *WWW 2011*.

Zamal, F. A.; Liu, W.; and Ruths, D. 2012. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. In *ICWSM 2011*.

Zhu, Y.; Zhong, E.; Pan, S. J.; Wang, X.; Zhou, M.; and Yang, Q. 2013. Predicting User Activity Level in Social Networks. In *CIKM 2013*.