Computer Communications xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

The rich and middle classes on Twitter: Are popular users indeed different from regular users?

Q2 Amit Ruhela [a,*], Amitabha Bagchi [a], Anirban Mahanti [b], Aaditeshwar Seth [a]

[a] CSE Department, IIT Delhi, India
[b] NICTA, Alexandria NSW, Sydney, Australia

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Twitter
Behavioral characteristics
Popular users
Event detection

ABSTRACT

Online social networking (OSN) websites such as Twitter and Facebook are known to have a wide heterogeneity in the popularity of their users, which is counted typically in terms of the number of followers or friends of the users. We add to the large body of work on information diffusion on online social networking websites, by studying how the behavior of the small minority of very popular users on Twitter differs from that of the bulk of the population of ordinary users, and how these differences may impact information diffusion. Our findings are somewhat counter intuitive. We find that on aggregate metrics such as the tweeting volume and degree of participation on different topics, popular users and ordinary users seem similar to each other. We also find that although popular users do seem to command an influential position in driving the popularity of topics on Twitter, in practice they do not affect growth rates of user participation and the causality of popular users driving event popularity is hard to establish. Our observations corroborate the findings of other researchers who show that user popularity in terms of number of followers does not translate into driving event popularity, but that event popularity may be driven by extraneous factors to do with the importance of the event.

© 2015 Published by Elsevier B.V.

1. Introduction

Online social networking (OSN) websites such as Twitter and Facebook have millions of users, and due to this sheer volume they have spawned entire new industries and research directions. They have become advertising properties by providing eyeballs that are arguably measurable in the number of impressions and clicks created, with tight targeting on voluntarily revealed personal user information. They have become barometers of user perception on topics ranging from news events to business, politics and products, by analyzing the subject matter and sentiment of user generated content shared on the OSN platforms. They provide an opportunity to sociologists and political scientists to understand the formation and propagation of public perception at large scales. Consequently, there is a large volume of research focused on building better algorithms for these applications.

We focus on a research question in this context to understand how the behavior of popular users on Twitter defined as those with many followers differs from the behavior of less popular users, specifically on aspects that may influence the spread of information. These popular users with thousands of followers are typically celebrities in real life and are considered to be highly influential in making topics popular by leveraging their large reach. First, studying three topics in the Indian context – politics, entertainment, and sports – we find out whether popular users defined at those in the top 0.1 percentile of the number of followers, tend to tweet more frequently or adopt topics earlier or engage for longer, than ordinary users who form the bulk of the user population. Second, we find out whether popular users show any preferences in retweeting tweets by other popular or ordinary users, and whether they seem to influence growth rates or the popularity of specific events that are discussed on Twitter. This can help understand if indeed popular users are influential in driving the popularity of events, and in uncovering events that may otherwise go unnoticed.

We make some surprising findings. On the first question, we find that popular and ordinary users do not differ much from each other in the volume of tweets, or the stage at which they become interested in events, and given that popular users end up participating in 90% of events, our conclusions therefore point towards the idea that just tracking popular users who are a small fraction of the overall Twitter userbase should be sufficient for most trend detection algorithms. On the second question, we find that being connected with popular users indeed gives an opportunity to less popular users to push information into the limelight, but popular users do not seem to have any influence on the event growth rate. Even causality on whether

* Corresponding author. Tel.: +91 9868840004.
E-mail addresses: aruhela@cse.iitd.ac.in, aruhela@gmail.com (A. Ruhela), bagchi@cse.iitd.ac.in (A. Bagchi), anirban.mahanti@nicta.com.au (A. Mahanti), aseth@cse.iitd.ac.in (A. Seth).

http://dx.doi.org/10.1016/j.comcom.2015.07.024
0140-3664/© 2015 Published by Elsevier B.V.

Table 1
Datasets attributes (M=Million).

Dataset	Seed user	Followees and followers (M)	Tweets (M)	Hashtags ≥ 10 K tweets	Topics specific hashtags
Entertainment	150	23	406	1568	119
Politics	55	7	115	558	182
Sports	40	9	129	580	59
Total	245	26	468	1631	360

popular users are critical in making an event popular is hard to establish, and rather it may be extraneous factors to do with the real life importance of the events themselves that may drive their popularity.

2. Related work

Understanding the mechanisms of information diffusion on online social media is an active area of research [1–6]. Among the millions of registered Twitter users, only a few known as elite users [7] or celebrities [8] and considered as being influential in affecting the diffusion of information. Various studies underpin the role of such users in the spread of news [9] and product marketing [10]. Researchers [11] show that a high number of followers and the volume of tweeting play a significant role in social advertisements that result in higher click rates. However, another study [12] challenges the role of these influential users, both as initiators of large cascades or as early adopters. Our study in many ways corroborates this latter finding that popular users do not differ much from ordinary users in aspects like tweeting volume or early adoption, and do not influence the growth rate of events either.

The approach of classifying Twitter users into multiple classes based on the popularity of the users is similar to [13], with subtle differences with regards to the thresholds that were chosen. Overall, this is a step in the direction of acknowledging user heterogeneity in Twitter, and studying the behavior of these disparate user classes.

Several studies focus on sampling strategies to crawl unbiased datasets from social networking websites to avoid having to process large amounts of data [14–16]. Our work is closest to [17] where the authors show that information collected from a few randomly selected individuals and their friends can detect contagious disease outbreaks in advance. Our findings similarly indicate that tracking only popular users may be sufficient for most purposes.

3. Datasets and definitions

Our goal was to obtain a dataset that could allow us to directly compare the activities of popular users with ordinary users. We chose to study this in the context of three common topics in India: Entertainment (specifically Bollywood), Politics, and Sports. For each of these topics, we first manually identified 245 seed users (Table 1) from among famous personalities mentioned on Forbes India [18] and other websites [19–21]. We only considered celebrity users who had a verified profile on Twitter. We then completely crawled the immediate neighbors of these seed users, both their followers as well as users they are following, and obtained all tweets within the last 95 days for these users. Overall, we were able to assemble a dataset of 26M Twitter users through this method. When we sampled the location of these users, we found that 40% of them were from India. These users in fact represent 60% of the entire Twitter userbase from India, according to statistics from [22] where India had 18M Twitter users in 2013. We therefore feel that this dataset conveys a good representation of the Indian Twitter userbase, and our method of starting with celebrity users to build a dataset may very well be a replicable method since it seems that a large fraction of users end up following some celebrity or the gff3w1.

We did not use the search API of Twitter to obtain tweets (and users) for certain keywords, because this API only returns a sample of tweets. Rather, by exhaustively listing all users, we were able to use the timeline API that returns the last 3200 tweets of a user. Out of all these tweets, we considered tweets in the last 95 days (roughly 3 months) from December 22 2013 to March 26 2014. We chose this threshold because a span of three months seemed sufficient to be able to witness the entire lifetime of events occurring within these topics, and only 0.006952% of users (1800 in count) seemed to have posted 3200 or more tweets in this period for whom we might miss some data. We are therefore confident that for each of the three topics, our datasets not only include a large proportion of twitter users interested in these topics, but also considers all tweets posted by most of these users during the study period.

Our next task was to prune the large number of tweets we crawled, to only consider tweets that belonged to one of the three topics of Entertainment, Politics, or Sports. To do this, we selected those hashtags which have received more than 10K tweets during the study period. For example, in the Entertainment dataset we found 1568 hashtags that have received at least 10K tweets. We manually went over these hashtags and removed non-entertainment related hashtags to come up with a list of 119 entertainment hashtags. Overall, we identified 360 hashtags under the topics of entertainment, politics, and sports, and these collectively represent 13% of the total number of Tweets in our dataset. This cascading selection method is shown in Table 1. Instead of using hashtags, we could have used other Natural Language Processing methods provided by APIs from OpenCalais, Alchemy, Yahoo! term extractor, etc., but due to API usage volume restrictions these methods would have been very time consuming and hence hashtags are a good substitute.

We understand that spam users on Twitter could affect our analysis. According to [23], 77% of spam accounts are deleted by Twitter itself on the first day of tweets by these users. Our crawlers for collecting user information started about one week later after building the social graph of seed users, so we feel that most spammers would already have been removed. Further, one year after our datasets collection, we again collected profile of all users of our datasets from Twitter. We found that 54,440 out of 816,626 expected spam accounts were deleted on Twitter. We observed that none of the celebrity users were among these users whose accounts did not exist one year after our datasets collection, and these users together posted less than 0.003% of the tweets. Considering this approach of building datasets, we feel that spam accounts are likely to have had an insignificant repercussion in our analysis.

We next outline a few definitions to build a vocabulary for our work before we present the actual analysis.

3.1. Defining the popularity of users

There is a wide diversity among the Twitter population in the number of followers of users and the volume of tweets done by them, as shown in Fig. 1. Much like how economic literature uses income classes to differentiate people between elite/upper/middle/poor classes, we used the number of followers and the volume of tweets to create four classes of users:

1. Popular users: the top 0.1 percentile of users based on the number of followers. This included 23,059 users in the Entertainment dataset, 6966 users in the Politics, and 9129 users in the Sports

UNCORRECTED PROOF

Please cite this article as: A. Ruhela et al., The rich and middle classes on Twitter: Are popular users indeed different from regular users? Computer Communications (2015), http://dx.doi.org/10.1016/j.comcom.2015.07.024

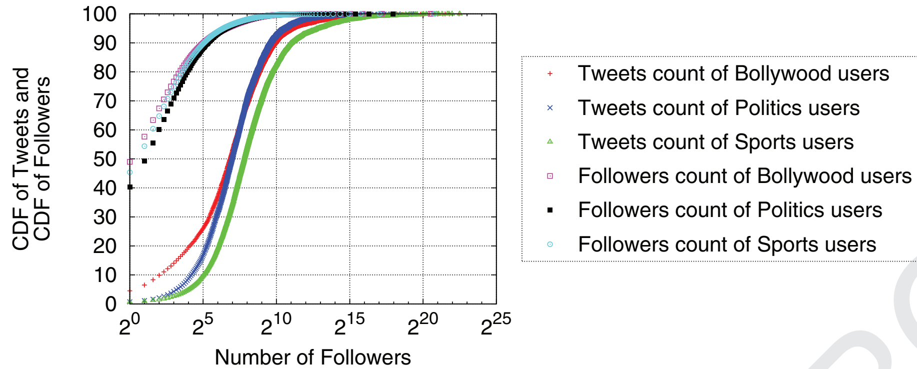


Fig. 1. Cumulative distribution of Tweets wrt followers count.

dataset. These users generated 1% of the tweets, and taking the Politics dataset as an example they had more than 6828 followers. This set of popular users contained 149 out of the original 150 seed users we had identified in the Bollywood dataset, 52 out of 55 original seed users we had identified in the Politics dataset and 39 out of 40 users we had identified in the Sports dataset. Thus, our definition of popular users is indeed a superset of the original seed set of celebrity users with a very little exception.

2. Medium popular users: the top 0.1 to 5 percentile of users based on the number of followers. This included 1.12M users in the Entertainment dataset, 0.33M users in Politics, and 0.44M users in the Sports dataset. This group produced 58% of tweets, and in the Politics dataset each user had 95 to 6828 followers. This set of medium popular users contained only 1.22% of seed users of all three datasets.
3. Ordinary users: the top 5–30 percentile of users. This included 5.65M users in the Entertainment dataset, 1.68M users in Politics, and 2.21M users in the Sports dataset. This group produced 37% of tweets, and in the Politics dataset each user had 8 to 94 followers.
4. Inactive users: the bottom 70 percentile of users. This included 16.25M users in the Entertainment dataset, 4.93M users in Politics, and 6.45M users in the Sports dataset. This group produced 4% of tweets, and in the Politics dataset each user had less than 8 followers. The large population of inactive users in our datasets is in-line with several previous studies [3,13,24,25].

This classification allowed us to compare the tweeting behavior characteristics of different groups of users in aggregate. For most comparisons, we compared the popular users category with the ordinary users category, which we feel allows us to understand whether the small minority of very popular users behaves differently from the large majority of regular Twitter users.

3.2. Defining events and event phases

Now that we are in a position to examine the tweeting patterns on a particular hashtag by different popularity groups of users, we wanted to go deeper to study these patterns within specific phases when events are occurring on the hashtags. We use the common definition of an event an occurrence sharply localized in a definite space and time instant. We say that an event occurs when the volume of tweets on a topic rises at a high rate. Each hashtag can thus contain several major and minor events during the study period. Periodic topics e.g. Follow Friday results in distinct events on each occasion of their occurrence. We use a threshold based event detection algorithm to find events for all hashtags under consideration. The algorithm is described in Appendix A. For each event, it outputs three phases: growth, peak, and decay. The growth phase is marked with a continuous increase in the tweet rate (barring minor fluctuations that are captured as hysteresis), the decay phase is similarly marked with a continuous decrease in the tweet rate, and the intermediate peak phase marks the period of highest tweet rate after which it starts dropping. Fig. 2 shows an example in which an event is detected for the topic “AbkiBaarModiSarkar”. The event corresponds to a rally by the prime-ministerial candidate Shri Narendra Modi (who went on to become the current prime minister of India) at Sambalpur in Odisha on the afternoon of 14th March 2014.

For some analysis in the next section, we also define the average growth rate of an event as the rate at which tweeting frequency increases within the growth phase:

$$GrowthRate = \frac{STA[PeakPhaseStartTime] - STA[EventStartTime]}{PeakPhaseStartTime - EventStartTime}$$

We further classify events as follows based on the growth rate.

1. Events with a low growth rate: Those with growth rate in the lowest 20 percentile
2. Events with a moderate growth rate: Those with growth rate between 20 and 80 percentile
3. Events with a high growth rate: Those with growth rate in the highest 20 percentile

4. Results and discussion

We next begin to analyze the tweeting behavior of different classes of users.

4.1. Volume of tweeting

The first question we answer is whether popular users tweet more or less or the same as ordinary users, where *popular* and *ordinary* users are the classes as defined in the previous section. We do not use simple methods like CDF of tweets/user to study tweeting volume because the popularity distribution of events is widely distributed between 0.6K and 250K tweets per event. Therefore, we performed this analysis on the basis of events and normalized the tweeting volume of users according to the following two ways. First, we determine the count of tweets produced by each user in an event, and rank order the users with a min–max normalization for each event on the number of tweets by the users. The normalization gives a score for each user within 0 and 1, and we look at the distribution of these scores for popular users and for ordinary users. Fig. 3 shows this cumulative distribution of rank scores of popular users and ordinary users for the politics dataset. Popular users rank only slightly higher than ordinary users in the events.

We also use another method to test this hypothesis: we find the average number of tweets by popular users and by ordinary users within each event, normalize this by the size of the event in terms of the total number of tweets, and then compare the distributions for various phases of the events. We find that popular users tweet slightly

UNCORRECTED PROOF

Please cite this article as: A. Ruhela et al., The rich and middle classes on Twitter: Are popular users indeed different from regular users? Computer Communications (2015), http://dx.doi.org/10.1016/j.comcom.2015.07.024

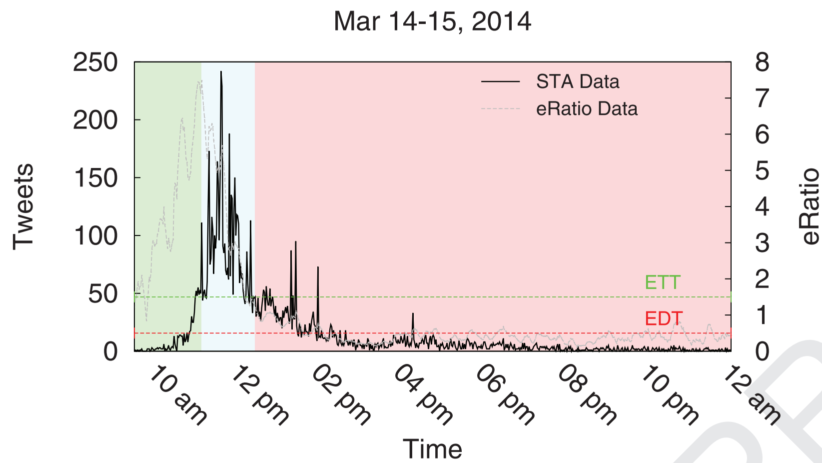


Fig. 2. An example on event detection for topic “AbkiBaarModiSarkar”.

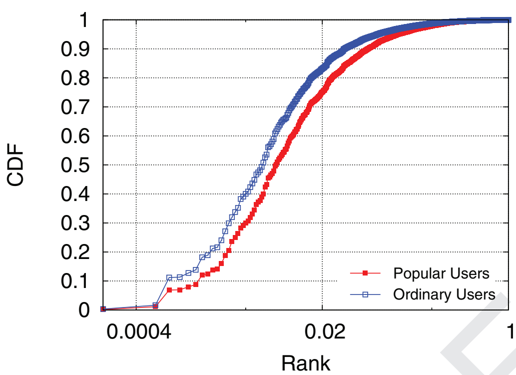


Fig. 3. CDF plot of rank scores on basis of tweets volume in politics dataset.

higher than ordinary users in the growth phase, but slightly lesser than ordinary users in the peak and decay phases. We also checked the volume of tweeting of medium popular users and found it to be similar. *This implies that the volume of tweeting by popular, medium popular and ordinary users is more or less the same.*

As an additional check to verify whether user popularity is a result of user activity, that those users who post frequently are the ones who end up becoming popular, we also checked the correlation between the number of followers of a user and the tweeting frequency of the user. We found a near zero correlation with all methods, including a correlation check for all users, a check weighted on the number of users in different popularity bins, and different binning strategies both linear and logarithmic. This therefore lends greater significance to our results, that *the cause of similar tweeting volume between different classes of users is not related to the activity of the users.*

4.2. Early adopters

We next answer the question of whether or not popular users are early adopters to begin tweeting on an event. To do this, we find out how much time after the event was triggered do popular and ordinary users post their first tweet in growth phase of the events. Fig. 4 shows the cumulative distribution for this time of posting for the politics dataset. Popular users started tweeting earlier on an average by 7 min than medium popular users and by 21 min than ordinary users. Relative to the average duration of the growth phase of events which is 223 min; we find that popular users start tweeting sooner than medium popular users by approximately 3% and ordinary users by approximately 10% of the growth phase duration. *This implies that popular and medium popular users have no significant difference in when they jump on to discussing a topic.*

4.3. Engagement with the event

Next we try to understand whether there is a difference in the degree to which popular and ordinary users engage with an event. We do this in two ways. One, we find the time difference between the first time and the last time an user tweets throughout the event lifetime, and compare the distribution for these time differences between the sets of popular and ordinary users. Fig. 5 shows the cumulative distribution for the politics dataset where we see that at

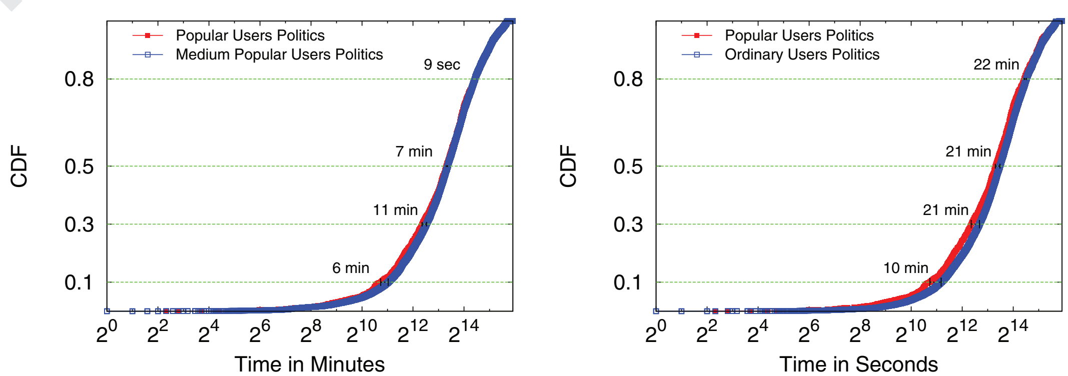


Fig. 4. (a) Relative time at which popular and medium users start tweeting on topics from start of the events. (b) Relative time at which popular and ordinary users start tweeting on topics from start of the events.

Please cite this article as: A. Ruhela et al., The rich and middle classes on Twitter: Are popular users indeed different from regular users? Computer Communications (2015), http://dx.doi.org/10.1016/j.comcom.2015.07.024

UNCORRECTED PROOF

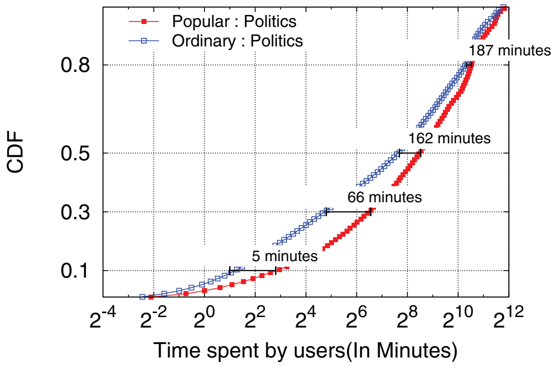


Fig. 5. CDF plot of time spent by popular and ordinary users in political events – popular users spend more time than ordinary users in the events.

the 30th percentile popular users stay active 103 min more than the ordinary users, and at the 50th percentile popular users stay active 207 min more than ordinary users. Relative to the average lifetime of events in the politics dataset which is 1548 min, this difference translates to 13% of the event lifetime.

This method has the obvious problem that it does not quantify the degree of intensity or the continuity of participation by the users. We therefore build another method, where we divide the entire event into time units of 15 min, close to the average session time of 12.51 min [26] (or 17 min [27]) of users on Twitter. We then count the number of slots in which users have posted at least one tweet, and normalize it by the total number of slots in the event. We find that the differences in cumulative distribution of this normalized slot count for popular and ordinary users are positive but much smaller indicating that *popular users have a much larger attention span because they engage for a longer amount of time, but there is not much noticeable difference in the intensity of participation.*

4.4. Participation across different event phases

We next go deeper to understand how the participation of popular and ordinary users carries forward along the different event phases of growth, peak, and decay. Fig 6 shows the breakdown for each event phase, of whether it was new popular or ordinary users who tweeted in this phase, or how much fraction of users who had participated in earlier event phases also participated in this phase. The values on the transition arrows are the median value across the events.

We find that 40% of popular users and 44% of ordinary users, who participate in the growth phase, do not participate in subsequent phases. Similarly, 69% of popular users and 79% of ordinary users who participate in the peak phase do not participate in the decay phase. Although these are large values, the interesting difference is that ordinary users tend to drop off between 4% and 14% more than popular users. *This is consistent with the results in the previous section, where we found that popular users participate longer in an event.*

4.5. Influence on growth rates

The next question we answer is whether popular users have an influence on the growth rate of events, i.e. does a higher participation by popular users lead to a faster growth of the event. We classify the events along two axes: on their growth rate as low/medium/high growth events (explained in Section 3.2), and on the participation by popular users as low/medium/high in the same way. A user is considered as having participated in the growth of an event if he/she tweets at least once during the growth phase of the event timeline. Fig. 7 shows for each (x, y) cell an example event timeline, and mention the proportion of events in each cell.

Looking at the column of high growth events, we can see that there are more high growth events with a low participation by popular users (3.08%) than events with a high participation by popular users (1.79%), which negates the hypothesis. Similarly, if we look at the low growth events column, we do see that there is a large proportion of events with low growth rates and low participation by popular users, but the trend is not consistent because there are more low growth events with high participation by popular users than medium participation by popular users. The hypothesis therefore seems weak, which indicates that *event growth rate is more likely to be dictated by extraneous phenomena related to the importance of the event itself, than driven by the participation of popular users on Twitter*. Note that our claim here is about event growth rates only, and not about a broader (and stronger) argument of whether or not popular users are necessary to make an event popular in the first place. We discuss this in more detail in the next two sections.

We also separately study if the event growth rate is correlated with other variables such as the sum of the follower count of participating users, retweet count, likes count, replies count, number of tweets, etc., but do not see any strong trends, again pointing

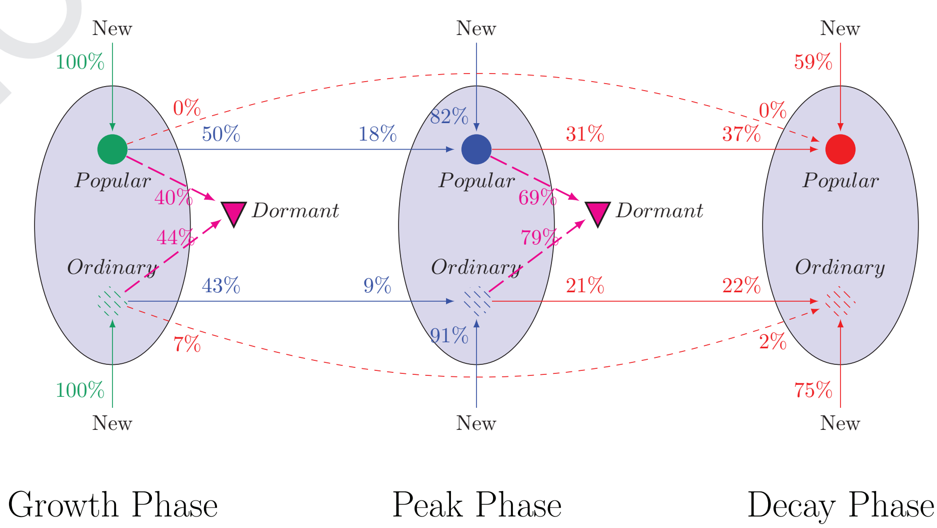


Fig. 6. Participation of users in subsequent phases of the political events: The nearest value on each edge for a user type display the percentage of that user type that goes out of the corresponding event phase to either other phase of the event or to the dormant state.

UNCORRECTED PROOF

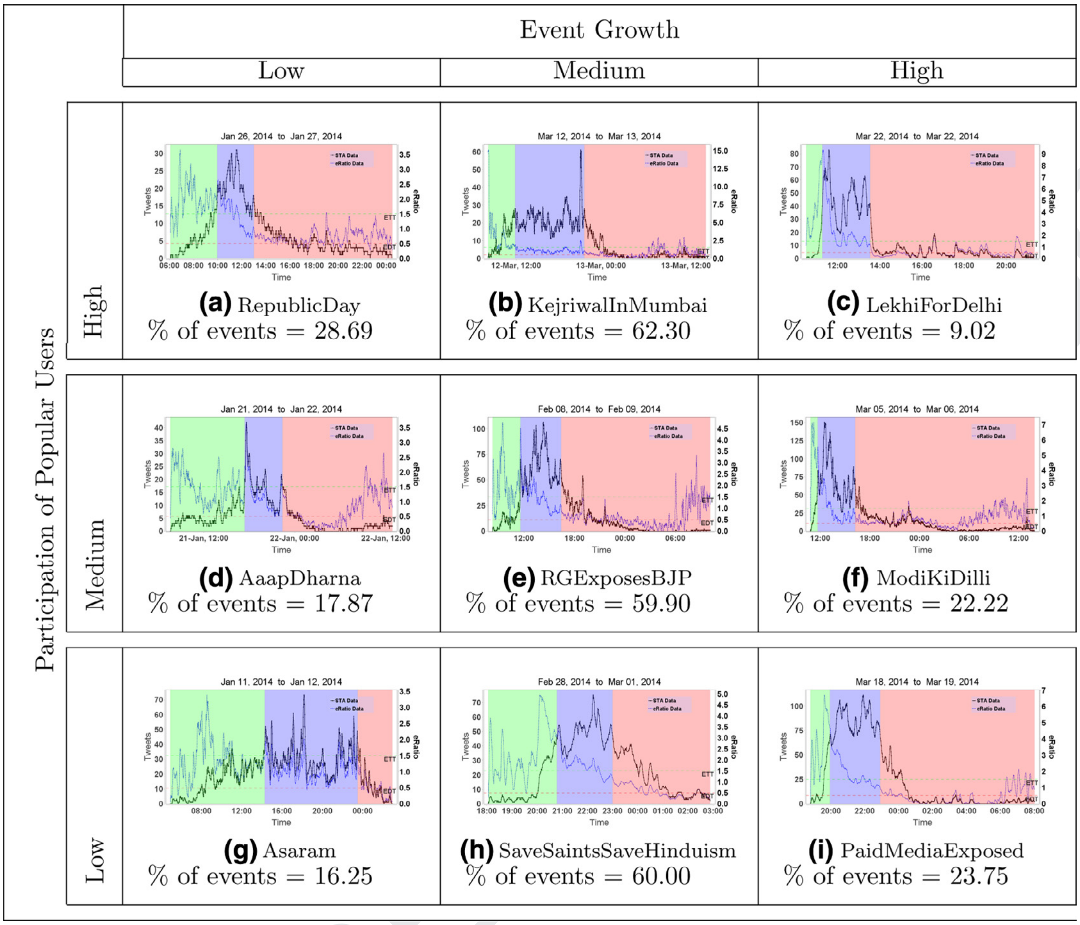


Fig. 7. Temporal plot of events on the basis of 'Growth Rate' and 'Participation of Popular users' for Political dataset. The growth, peak and decay phase of the event is colored green, blue, and red respectively. Each row in the plot indicates that growth rate of events is independent of popular user's participation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

towards extraneous phenomena being more important drivers of event growth.

4.6. Content copying characteristics

We next analyze retweeting characteristics of users: Are tweets by popular users retweeted more than tweets by others? Do popular users retweet more or write their own tweets on different topics? Fig. 8a shows the four classes of users based on their popularity, and labels each incoming arrow into a class with the proportion of tweets from other classes. The *original tweets* label indicates the percentage of non-retweets by that class of users.

We can make a couple of interesting observations. Across all user classes, popular users have the highest fraction of original tweets authored by them (61%), i.e. they do not retweet as much as other classes of users but prefer authoring their own tweets. Across all user classes, we also see that retweets of tweets authored by popular users are more or less of the same order as retweets of tweets authored by medium popular users; considering that overall only 1% of tweets are written by popular users while 58% of tweets are written by medium popular users, this indicates that the chance of a tweet by a popular users getting retweeted is much higher than the chances of retweets for tweets by less popular users. Calculating this specifically, we find that the probability of retweets of tweets by popular users is 0.77, while that for ordinary users is 0.13. Both these insights indicate that popular users are indeed influential in attracting a lot of retweets of their tweets, but we have also seen from the previous section that this influence does not necessarily translate into higher event growth rates.

Looking at the activity of medium popular users, we find that a high percentage (43%) of tweets are retweets by other medium popular users. A possible explanation could be the reciprocity of relations among this set of users, i.e. unlike popular users who seem to be followed because of their celebrity status, these users are likely to be followed by their friends whom they too follow. Hence, they tend to retweet tweets by their friends. We do not have reciprocity information in this dataset, but we used another dataset [3] to separately verify the correlation between the reciprocity of relations on Twitter and the number of followers. This is explained in more detail in Appendix B, and indeed we find that reciprocity is strongly related to the number of followers: Users with between 100 and 5000 followers reciprocate almost 40–60% of their follower relations, but reciprocity rapidly decreases for more popular users. The Twitter policy on aggressive following does not restrict the followers' count of a user, which means reciprocity is not affected by Twitter policy. A high reciprocity therefore seems to indicate that users retweet each other's tweets.

4.7. Time-delay characteristics of content copying

The final aspect we analyze is how soon do users copy content, and is there any preferential treatment given to popular vs. ordinary users. We draw a similar Fig. 8b as in the previous section, labeled with the median value of the retweet delay in minutes.

Our first observation is that popular users retweet more quickly than other users. This can be seen clearly from the median values around different user types, which range in single digits for popular users but are much larger for medium popular and ordinary users,

UNCORRECTED PROOF

Please cite this article as: A. Ruhela et al., The rich and middle classes on Twitter: Are popular users indeed different from regular users? Computer Communications (2015), http://dx.doi.org/10.1016/j.comcom.2015.07.024

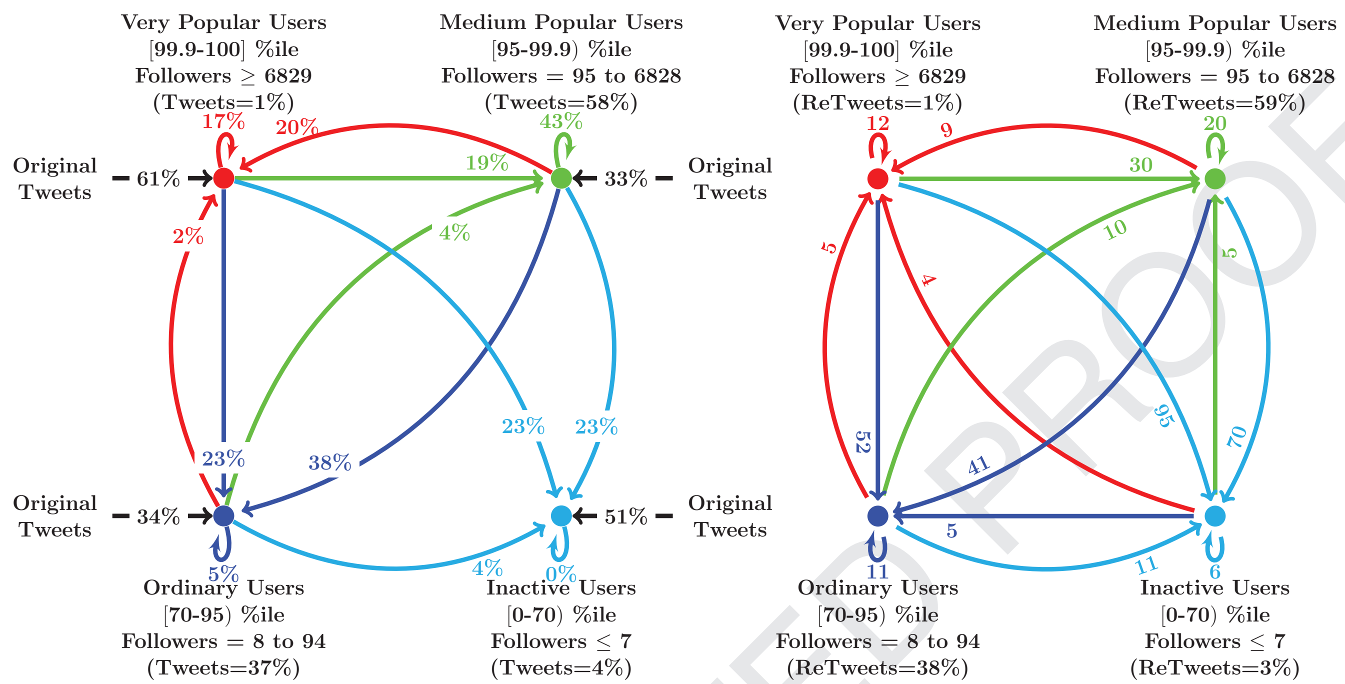


Fig. 8. (a) Production of tweets in political dataset by very popular [99.9–100] percentile, medium popular users [95–99.9) percentile, ordinary users [70–95) percentile and inactive users [0–70) percentile. The values above arrows directed towards the 'Original Tweets' state the percentage of tweets self-produced by the user. On all other arrows, the value on an incoming arrow state the percentage of tweets copied by the user from where the arrow originate. (b) Time-delay(in minutes) for 50th percentile of retweets between very popular, medium popular, ordinary and inactive users in political dataset.

indicating that popular users are more alert probably because they spend more time on Twitter. Overall for retweeting, popular users take 8 times less number of minutes than inactive users, 4.5 times less number of minutes than ordinary users, and 2.5 times less number of minutes than medium popular users.

Another interesting observation is that popular users retweet tweets by ordinary users and other less popular users faster than tweets by other popular users. This can be seen for popular users by looking at the retweet latency of 5 min for tweets by ordinary users, but of 9 min and 12 min for the retweet latencies of tweets by medium popular and popular users. This preferential trend holds true for medium popular users as well. This is potentially explained by the same reciprocity argument we used earlier in Section 4.6, that since popular users follow only a few users, therefore these few users are likely to be the friends of popular users, and hence popular users retweet tweets by their friends more conscientiously than tweets by other more popular users whom they follow.

Combining insights from this section and the previous section, it seems that popular users retweet more quickly than other users, their tweets tend to get retweeted more, and they also show a preference to retweeting tweets by less popular users. Popular users therefore certainly seem to command an influential position and could potentially drive the popularity of events, especially events initiated by less popular users. Section 4.5 however shows that in aggregate; at least the growth rate of events seems to not be dependent on the level of participation by popular users and is likely to be driven by entirely extraneous phenomena. What we cannot tell from the dataset though is whether the participation of popular users is critical to making an event popular – this causality can only be correctly understood by comparing the growth trajectories of otherwise identical events which have different degrees of participation by popular users. We studied this question indicatively as part of another research [5] on a different dataset, and run a similar test on this dataset, by correlating the popularity of an event (total number of users who engage with the event) with the number of popular users during the growth phase of the event. This is shown in Fig. 9a. A high correlation is indeed present (Pearson correlation = 0.74), although we can see there are events that become popular without much help from popular users, as well as events that do not become popular despite participation from popular users. It is hard to establish causality though, because more popular users may have participated in an event that ended up becoming popular just because the event itself was more important. Overall therefore, we are not able to establish whether or not, and in which ways, are popular users useful for the popularity of events.

5. Discussion and conclusion

The main insights we have gained through this study are outlined below:

1. Section 4.1: the volume of tweets by popular vs. ordinary users is not distinguishable from each other.
2. Section 4.2: within the growth phase, popular users tweet earlier than ordinary users by approximately 10% of the event growth duration.
3. Sections 4.3 and 4.4: popular users engage with an event for 13% of the event lifetime longer than ordinary users, and tend to drop off up to 14% less across different event phases than ordinary users. However, their intensity of participation is not very different.
4. Section 4.5: the participation of popular users does not seem to influence the event growth rate.
5. Section 4.6: popular users write more original tweets than retweets by a factor of 60:40, while for ordinary users this ratio is almost the inverse.
6. Section 4.6: the tweets by popular users are retweeted 6 times more than tweets by other users.
7. Section 4.7: popular users are the quickest to retweet tweets by a factor of 8 than other users.
8. Section 4.7: popular users show a preference to retweet tweets by less popular users sooner, probably those who are their friends.

UNCORRECTED PROOF

Please cite this article as: A. Ruhela et al., The rich and middle classes on Twitter: Are popular users indeed different from regular users? Computer Communications (2015), http://dx.doi.org/10.1016/j.comcom.2015.07.024

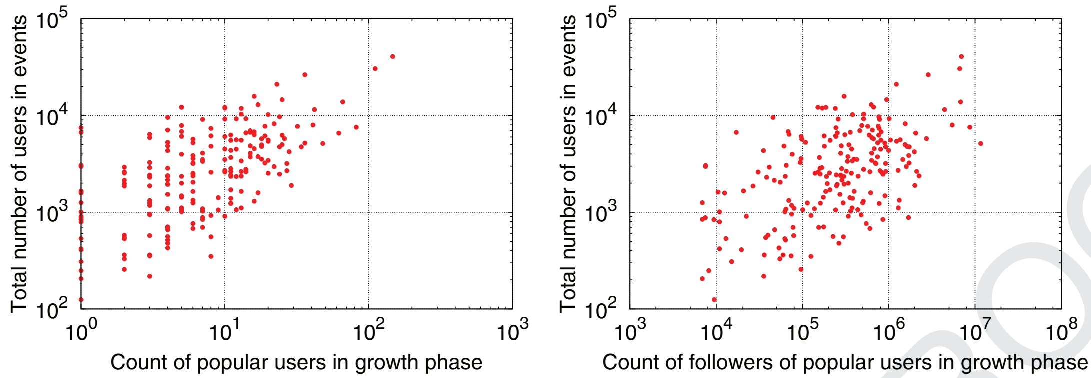


Fig. 9. Participation of popular users in growth phase of events vs. events' popularity.

These insights point towards some curious trends. First, it seems that aggregate characteristics such as tweeting volume, event participation, and early adoption do not differ much among popular and ordinary users. Any minor differences are in fact likely to be accentuated in the case of popular users. This can have important implications in the design of trend detection algorithms for various purposes – understanding the flow of information on social networks, targeted advertising, business intelligence, etc. Tracking a small set of popular users may be sufficient to capture most trends, instead of mining large volumes of tweets from across many users.

Second, it appears that popular users can be influential in driving event popularity given that their tweets are retweeted more, and that they retweet more quickly than other users. Furthermore, popular users seem to show a preference to retweeting less popular users, which can help bring attention to events that otherwise may not become popular. However, although there appears to be a correlation between event popularity and participation by popular users, we find that popular users are not able to influence the event growth rates. This indicates that event growth rates are more likely to be dictated by extraneous factors related to the importance of the event itself, or activities occurring outside of Twitter such as mass media interest in the event. Whatever be the direction of causality, the correlation can certainly be leveraged to make the job of detecting trends easier, by tracking only popular users instead of all Twitter users. Going forward, we plan to microscopically analyze individual events to get a better sense of the direction of causality.

Acknowledgments

This work is dedicated to Late Sh. Sumer Singh Ruhela (Father of Amit Ruhela) whose endless love, support and encouragement has always been an inspiration to the first author. The authors would like to thank Vipin Tyagi and Ravi Gupta from C-DOT India for their valuable support and suggestions. This work was supported by the Commonwealth of Australia and the Department of Science and Technology, India, under the Australia-India Strategic Research Fund, and the Department of Information Technology, India under research Project #RP02442 and #RP02355.

Appendix A. Event detection algorithm

For simplicity, we wanted to only use the tweet timings within a hashtag, to detect events for the hashtag. The tweet timeline of course suffers from sudden short term variations that we wanted our event detection algorithm to ignore, and only capture well defined events. After trying a number of different approaches, we found a variation of the method adopted by [28] to work best.

We start with maintaining two averages, given the samples $S(i)$ of tweet frequency at various timestamps:

1. Long term average, calculated as an exponentially weighted moving average: $LTA[i] = \alpha * S(i) + (1 - \alpha) * LTA(i-1)$
2. Short term average, calculated using the Average Loss Interval method [28]. This method calculates the mean over the last n samples, giving an equal weightage to the last $n/2$ samples and a progressively lesser weightage to older samples. $STA[i] = \sum_{j=0}^{n-1} (S(i-j) * w_i) / \sum_{j=0}^{j=n-1} w_i$

We then calculate the ratio of the short term average to the long term average, labeled *eRatio*, which rises with an increase in the tweet rate and drops as the event cools down. To clearly define the event phases, we use several thresholds:

1. When the *eRatio* exceeds an *"Event Trigger Threshold"-ETT*, we declare it as the start of the event.
2. When the *eRatio* subsequently drops below *"Event DeTrigger Threshold"-EDT*, we declare it as the end of the event .
3. In between, we find a $Threshold_{Peak}$ value, and declare the part of the event between ETT and Threshold during the rising period as the event growth phase, the part when the eRatio is larger than the threshold as the peak phase, and the part between the Threshold during the drop period and the EDT as the decay phase.

To further smooth out short term fluctuation in *eRatio*, we use a state transition diagram shown in Fig. 10, where despite *eRatio* changes the event is not exited (toggle between states S_1 and S_2) unless the short term average drops below what it was when the event had started. We further run post-processing to capture only significant events, defined as those with at least 1000 tweets. The choice of thresholds is made carefully and shown in Table 2, where

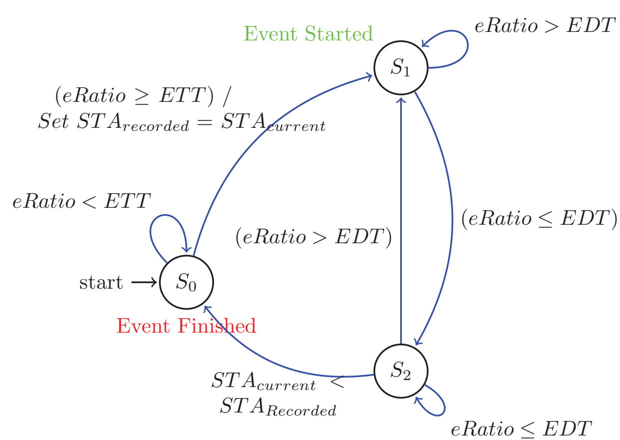
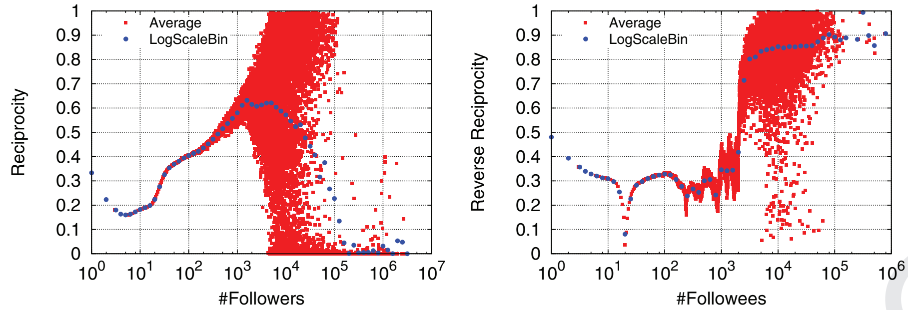


Fig. 10. State diagram ofevents detection.



(a) How many followers follow back the user? **(b)** How many friends follow back the user?

Fig. 11. Reciprocity and reverse reciprocity.

Table 2
Variables of events detection algorithm.

Const1(Time unit) = 20
Length of STA window = 9 time interval
$\alpha = 0.01$
ETT = 1.5 and EDT = 0.5
$Threshold_{Peak} = \mathrm{Min}((Mean + 1.2 * STD)$ or $(0.6 * STA_{Peak}))$
Max time gap to merge the adjacent events = 8 h

we manually reviewed the events detected in 75 out of 360 hashtags and chose values that produced the clearest defined events.

Appendix B. Reciprocity

We define the reciprocity of a user as the fraction of the user's own followers whom the user follows back. Measuring reciprocity requires the complete social graph of users, which we did not have in our current dataset. We therefore used a publicly available dataset [3] which contains the entire social graph of 40 million users from July 6, 2009 to July 31, 2009. The scatterplot and average reciprocity are shown in Fig. 11a. Reciprocity values seem to be positively correlated with the number of followers for up to 5000 followers. Beyond this threshold, the reciprocity rapidly decreases. This shows that very popular users, who attain celebrity status, do not follow their followers back, but less popular users do follow back and it seems these users are friends with each other and hence follow each other.

To confirm this hypothesis, we also define the reverse reciprocity as the fraction of the number of users a user is following, who follow him/her back. Fig. 11b shows the scatterplot and mean values for reverse reciprocity and we see a similar trend. Popular users have a high reverse reciprocity and are followed by users whom they are following. For users with less than 2000 followers however, the reverse reciprocity is much lower. A curious dip can also be seen at a follower count of 20, which seems to be because Twitter allows new users to follow 20 people in a single click and in fact throws up recommendations of popular users when new users join Twitter. Since the popular users are unlikely to follow these new users back, the dip is constituted of those users who recently joined Twitter and chose to follow several popular users according to the recommendations given by Twitter.

References

[1] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, A. Mahanti, Characterizing and modelling popularity of user-generated videos, Perform. Eval. 68 (11) (2011) 1037–1055.
[2] D.M. Romero, B. Meeder, J. Kleinberg, Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter, in: Proceedings of the 20th International ACM Conference on World Wide Web, New York, NY, USA, 2011, pp. 695–704.
[3] H. Kwak, C. Lee, H. Park, S. Moon, What is Twitter, a social network or a news media? in: Proceedings of the 19th International ACM Conference on World Wide Web, New York, NY, USA, 2010, pp. 591–600.
[4] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, S. Moon, Analyzing the video popularity characteristics of Large-Scale user generated content systems, IEEE/ACM Trans. Netw. 17 (5) (2009) 1357–1370.
[5] S. Ardon, A. Bagchi, A. Mahanti, A. Ruhela, A. Seth, R.M. Tripathy, S. Triukose, Spatio-temporal and events based analysis of topic popularity in Twitter, in: Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, CIKM, San Francisco, CA, USA, 2013, pp. 219–228.
[6] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, Z. Su, Understanding retweeting behaviors in social networks, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM, New York, NY, USA, 2010, pp. 1633–1636.
[7] S. Wu, J.M. Hofman, W.A. Mason, D.J. Watts, Who says what to whom on Twitter, in: Proceedings of the 20th ACM International Conference on World Wide Web, New York, NY, USA, 2011, pp. 705–714.
[8] M.S. Srinivasan, S. Srinivasa, S. Thulasidasan, Exploring celebrity dynamics on Twitter, in: Proceedings of the 5th ACM IBM Collaborative Academia Research Exchange Workshop, I-CARE, New York, NY, USA, 2013, pp. 13:1–13:4.
[9] M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, K.-L. Ma, Breaking news on Twitter, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI, ACM, New York, NY, USA, 2012, pp. 2751–2754.
[10] Y.-M. Li, Y.-L. Lee, N.-J. Lien, Online social advertising via influential endorsers, Int. J. Electron. Commer. 16 (3) (2012) 119–154.
[11] J.Y. Park, K.-W. Lee, S.Y. Kim, C.-W. Chung, Ads by whom? ads about what?: exploring user influence and contents in social advertising, in: Proceedings of the First ACM Conference on Online Social Networks, COSN, ACM, New York, NY, USA, 2013, pp. 155–164.
[12] D.J. Watts, P.S. Dodds, Influentials, networks, and public opinion formation, J. Consum. Res. 34 (4) (2007) 441–458.
[13] M. Cha, F. Benevenuto, H. Haddadi, K. Gummadi, The world of connections and information flow in twitter, IEEE Transactions Syst. Man Cybern. Part A: Syst. Hum. 42 (4) (2012) 991–998, doi:10.1109/TSMCA.2012.2183359.
[14] J. Leskovec, C. Faloutsos, Sampling from large graphs, in: Proceedings of Knowledge Discovery and Data Mining, KDD, 2006, pp. 631–636.
[15] B.F. Ribeiro, P. Wang, F. Murai, D. Towsley, Sampling directed graphs with random walks, in: Proceedings of INFOCOM, 2012, pp. 1692–1700.
[16] M.D. Choudhury, Y.-R. Lin, H. Sundaram, K.S. Candan, L. Xie, A. Kelliher, How does the data sampling strategy impact the discovery of information diffusion in social media? in: Proceedings of International Conference on Weblogs and Social Media, ICWSM, 2010, pp. 34–41.
[17] G. arcía Herranz, E.M. Egido, M. Cebrián, N.A. Christakis, J.H. Fowler, Using friends as sensors to detect global-scale contagious outbreaks, PLoS ONE abs/1211.6512 (4) (2014) e92413.
[18] F. India, 2013 Celebrity 100 List – Forbes India Magazine, 2013, URL http://forbesindia.com/lists/2013-celebrity-100/1439/1.
[19] Slideshare, Blogworks Most Mentioned Political Leaders Index January, www.slideshare.net/Blogworks/blogworks-most-mentioned-political-leaders-index-january-20142014.
[20] Bollywoodbuzz, Bollywood Celebrities on Twitter, http://www.bollywoodbuzz.in/bollywood-celebrities-on-twitter/2013.
[21] D. Mahapatra, @devi4u/bollywood on Twitter. URL https://twitter.com/devi4u/lists/bollywood/members. 2013
[22] EMarketer, Japan, India Boast Largest Twitter Audiences in APAC – Emarketer., 2015. URL http://www.emarketer.com/Article/Japan-India-Boast-Largest-Twitter-Audiences-APAC/1011917.
[23] K. Thomas, C. Grier, D. Song, V. Paxson, Suspended accounts in retrospect: an analysis of twitter spam, in: Proceedings of the ACM SIGCOMM Conference on Internet Measurement Conference, in: IMC, New York, NY, USA, 2011, pp. 243–258, doi:10.1145/2068816.2068840.
[24] M. Gabielkov, A. Legout, The complete picture of the twitter social graph, in: Proceedings of the 2012 ACM Conference on CoNEXT Student Workshop, in: CoNEXT Student, New York, NY, USA, 2012, pp. 19–20, doi:10.1145/2413247.2413260.

UNCORRECTED PROOF

[25] Businessinsider, Most People on Twitter Dont Actually tweet. 2015URL http://www.businessinsider.in/Most-People-On-Twitter-Dont-Actually-Tweet/articleshow/33621062.cms.
[26] Statista, Chart: 8 Reasons Why Twitter is no Second Facebook (yet), http://www.statista.com/chart/1598/twitter-compared-to-facebook/, [Online; accessed 11 Sept-2014] (Nov. 2013).
[27] W. Yichuan, L. Xin, C. David, L. Yunxin, Earlybird: Learning-Based Mobile Prefetching Through Content Preference and Usage Pattern, http://www.cs.ucdavis.edu/~liu/preprint/earlybird.pdf, [Online; accessed 11 Sept-2014] 2014.
[28] S. Floyd, M. Handley, J. Padhye, J. Widmer, Equation-based congestion control for unicast applications, in: Proceedings of the ACM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, SIGCOMM, New York, NY, USA, 2000, pp. 43–56.

UNCORRECTED PROOF

Please cite this article as: A. Ruhela et al., The rich and middle classes on Twitter: Are popular users indeed different from regular users? Computer Communications (2015), http://dx.doi.org/10.1016/j.comcom.2015.07.024