Circadian Patterns in Twitter

Marijn ten Thij, Sandjai Bhulai

VU University Amsterdam,
Faculty of Sciences,
Amsterdam, The Netherlands
Email: {m.c.ten.thij,s.bhulai}@vu.nl

Peter Kampstra

RTreporter,
Amsterdam, The Netherlands
Email: peter@rtreporter.com

***Abstract*—In this paper, we study activity on the microblogging platform Twitter. We analyse two separate aspects of activity on Twitter. First, we analyse the daily and weekly number of posts, through which we find clear circadian (daily) patterns emerging in the use of Twitter for multiple languages. We see that both the number of tweets and the daily and weekly activity patterns differ between languages. Second, we analyse the progression of individual tweets through retweets in the Twittersphere. We find that the size of these progressions follow a power-law distribution. Furthermore, we build an algorithm to analyse the actual structure of the progressions and use this algorithm on a limited set of tweets. We find that retweet trees show a star-like structure.**

Keywords*–*Data analytics; Twitter; Retweet graph; Language use; Daily pattern

I. Introduction

In the current digital age, many different (micro)blogging platforms have emerged, e.g., Twitter, Facebook and LinkedIn. Using these media, users can share everyday thoughts and activities. Through this sharing behaviour, large quantities of information are available to researchers who have distinguished many applications for the valorisation of this information. Within these applications, the focus is placed on predicting the future. The implementation of predictions using data from Twitter cover very different areas, e.g., predictions with respect to policital elections [1], the prediction of stock market prices [2], estimating the box-office revenue of a movie [3], and the detection of earthquakes [4].

In most social media, there is a notion of trending topics. With this notion, questions arise as to how and when these topics emerge and grow. Since these questions are far from trivial, we need more insight in the normal use of this social media to be able to answer these questions. Because if one understands the patterns of usage of a social medium, this insight will lead to understanding of the trending mechanism.

Therefore, in this paper, we study the patterns of use in Twitter. We focus on two aspects that provide a first insight in the usage of Twitter. First, we give the reader an overview of related research in Section II. Then, in Section III, we describe the process of gathering the data we used in the study. The analysis of circadian patterns in Twitter is presented in Section IV. Thereafter, in Section V, we display our analysis of the progression of retweets through the user network of Twitter. Finally, we draw our conclusions and discuss possible extensions of our work in Section VI.

II. Related Research

In this section, we provide the reader with a brief overview of related research in the two areas that we adress in this paper. The first of these fields is the circadian pattern that appears in social media usage. Secondly, we focus on the spread of information in social media.

A. Circadian patterns in Social Media

First, we consider social media activity. This has been researched for many different social media platforms. For instance, Kaltenbrunner et al. [5] analyse the activity pattern in Slashdot, a news site. They observe both daily and weekly activity patterns in the use of the site. Also, Gill et al. [6] study the activity of Youtube. They find clear circadian and weekly patterns, where the majority of the activity takes place at the end of the day during weekdays. Szabo and Huberman [7] examine the activity patterns on Digg and Youtube. They notice a weekly cycle of activity in Digg and investigate the popularity pattern of articles in Digg and videos in Youtube.

Noulas et al. [8] investigate the user activity pattern on Foursquare. They find clear geo-temporal rhythms in its activity, both for weekdays and weekends. Moreover, Grinberg et al. [9] use Foursquare data to extract real-life activity patterns. They observe that there are clear patterns for coarse categories, such as food or nightlife. They also notice that these patterns are present in Twitter.

Yasseri et al. [10] analyse circadian patterns in editorial activity on Wikipedia, an online encyclopedia. They find a clear daily pattern in activity per language. The only exception to these patterns is the English Wikipedia. For this language, the activity is more spread out over the day. Also, they find four weekly activity patterns for different groups of countries. Ten Thij et al. [11] study the page-view activity patterns for Wikipedia and observe circadian patterns in page-view activity.

Poblete et al. [12] investigate the user activity on Twitter for the top 10 countries in their sample. They perform an analysis of the activity based on sentiment and network properties. They find that the network and user properties can differ from country to country, from small connected networks to a large and more hierarchical structured network. Mocanu et al. [13] study GPS-tagged tweets by location and language. They analyse the heterogeneity of language use for many levels (e.g., global, country and city) and observe clear peaks in activity by tourists in some countries in the Mediterranian.

ISBN: 978-1-61208-358-2

B. *Information spread*

In the second part of our work, we focus on the spread of information through the network. Again, this type of activity analysis has been executed for multiple platforms. For instance, Jurgens and Lu [14] analyse temporal patterns in edits to Wikipedia articles. Their analysis reveals motif instances in the edit-patterns to pages.

Lerman and Gosh [15] analyse user activity on Digg and Twitter. They conclude that despite their different setup, both sites display similar patterns of information spread. Kamath et al. [16] analyse the geo-spacial progression of hashtags in Twitter using geo-tagged tweets. They use their analysis to find analytics techniques to characterize the relative impact of locations on spread dynamics of a topic. Yang and Leskovec [17] investigate temporal patterns arising in the popularity of online content. They formulate this as a time series clustering problem and formulate an algorithm to cluster these time series with respect to the patterns they exhibit. Finally, Bhamidi et al. [18] develop a random graph model that models the giant component of the retweet network induced by an event on Twitter. We aim to extend this insight to a message-based insight in the spread dynamics of Twitter.

III. Dataset

In this section, we describe how we obtained the tweets that were using in our analysis. The tweets were scraped by RTreporter, a company that uses the incoming stream of Dutch tweets to detect news in the Netherlands. These tweets are scraped using the filter stream of the Twitter Application Programming Interface (API) [19]. We set up four different streams, the first two streams (called sample and geo-located) are meant to give an overview of the stream of Twitter messages. The third and fourth stream (called “Netherlands (NL) general” and “NL specific”) are set up with the goal to scrape as many Dutch tweets as possible.

The first stream is the so-called sample stream. It outputs a sample of the complete Twitter Firehose. The sample that is given, contains roughly 1% of all tweets. The second stream uses the option **location** of the filter stream, in which a geo-location square is defined. All the tweets within this square are caught. We filter the stream on the geo-square induced by ((-179.99, -89.99), (179.99, 89.99)).

We call the third stream the “NL general” stream. In this stream, we use filter stream with the option **track**, where a list of words must be defined. All tweets containing one of these words are caught. We define a list of general Dutch words (e.g., *‘een, het, ik, niet, maar, heb, jij, nog, bij’*). In total, this list consists of 130 words. Lastly, the fourth stream, “NL specific” uses the filter stream combining **track** and **follow**. For this last option, we add a list of user IDs for which all tweets are caught. In total, we define a list of 1,303 users. Examples of accounts are *@NUnl*, *@TilburginBeeld*. The list of terms consists of 395 entries (e.g., *‘brandweer, politie, gewond, ambulance’*). Note that a specific tweet may be contained in multiple streams, we have not distinguished duplicates in our dataset.

Since we do not have access to the Twitter Firehose, we do not receive all tweets that we request due to rate limitations by Twitter [20]. An overview of the missed tweets is presented in Table I. Furthermore, in Figure 1, we display the number of tweets that were missed per stream on a daily scale. We see that the number of missed tweets in the ‘NL general’ stream is gradually decaying over time. In our experience, this decrease is probably caused by a decrease in the number of ’spam’ tweets made by Dutch teens (e.g., tweets like “Welterusten!”, which means “Good night!”). The geo-located stream follows an increasing trend. Furthermore, the sample stream has no missed tweets, which is logical, since the maximum number of tweets that one can receive is bounded by this number. With respect to the ‘NL specific’ stream, we see that the number of missed tweets is small, with the exception of some dates. We have not performed a more detailed analysis of the specific activity during these days.

TABLE I: NUMBER OF MISSED TWEETS PER STREAM FROM FEBRUARY 1ST 2013 TO FEBRUARY 1ST 2014.

‘NL general’	Sample	Geo-located	‘NL specific’
58,711,359	0	1,744,800,984	10,025

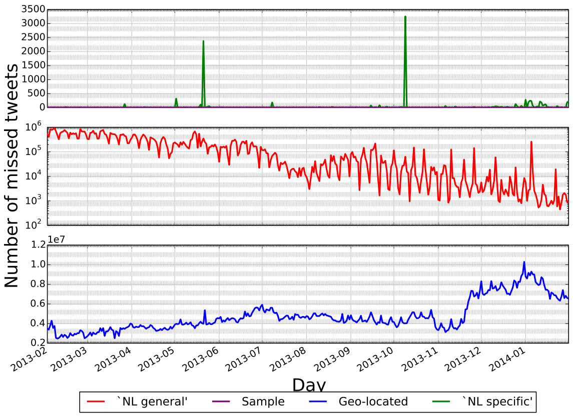


Figure 1: Number of missed tweets per day, indicated per stream.

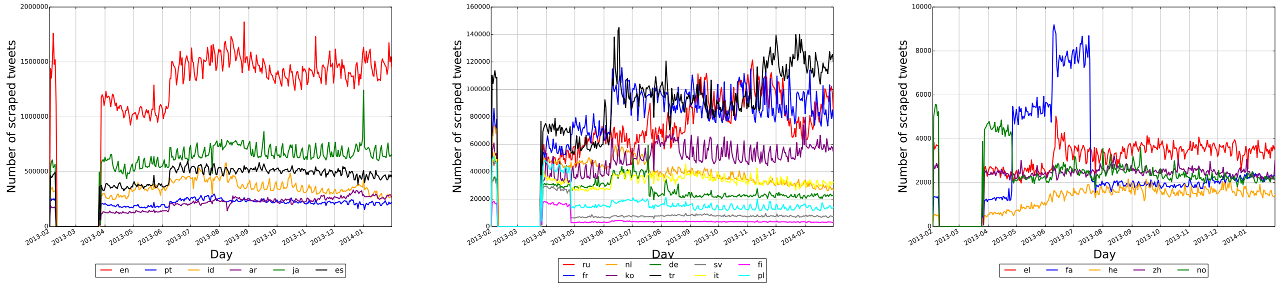
We process the tweets that have been obtained by these four streams from February 1st 2013 to February 1st 2014. The number of tweets, clustered by language, for each stream is indicated in Table II. For the sample stream, we also indicate which percentage of all tweets was posted in the given language (denoted in brackets behind the number of tweets). The timezone information displayed in Table II is used to correct the daily/weekly patterns for comparison. If a language is commonly used in multiple timezones, we search for a city that is located in the center of this area and base the correction on this city. These cities are also mentioned in Table II.

IV. Tweet Patterns

In this section, we discuss the temporal patterns that emerge in the data. We focus on two streams, namely the geo-located and sample streams, since these streams give a wide perspective on the traffic on Twitter. We present an analysis of these two streams on a daily scale for the complete year. Also, we present a more fine-grained analysis of the hourly patterns, both on a daily and a weekly basis.

TABLE II: NUMBER OF RECEIVED TWEETS PER LANGUAGE AND PER STREAM FROM FEBRUARY 1^{ST} 2013 TO FEBRUARY 1^{ST} 2014.

Language	Abbreviation	UTC Timezone	Sample	Geo-located	'NL general'	'NL specific'
English	en	-5	457,243,925 (33.85%)	503,979,412	26,590,575	8,805,644
Japanese	ja	9	214,383,682 (15.87%)	39,163,635	1,565,480	72,675
Spanish	es	-5	159,954,649 (11.84%)	145,905,496	3,162,450	476,042
Indonesian	id	7 (Jakarta)	116,797,591 (8.65%)	173,151,505	12,490,099	758,133
Portuguese	pt	-3 (Brasilia)	75,455,200 (5.59%)	137,408,037	1,840,142	250,478
Arabic	ar	2 (Egypt)	72,798,062 (5.39%)	40,181,252	64,144	11,973
Turkish	tr	2	31,035,914 (2.3%)	62,847,302	9,698,937	97,305
French	fr	1	28,284,488 (2.09%)	47,852,027	2,870,243	342,836
Russian	ru	4 (Moskow)	24,798,379 (1.84%)	29,639,926	602,388	22,434
Korean	ko	9	16,506,590 (1.22%)	5,474,115	65,661	10,385
Dutch	nl	1	13,270,846 (0.98%)	16,481,387	567,200,368	110,396,915
Italian	it	1	11,145,933 (0.83%)	14,450,810	316,434	79,364
German	de	1	8,919,771 (0.66%)	9,808,823	12,100,125	898,117
Polish	pl	1	6,044,235 (0.45%)	6,904,949	1,004,977	203,577
Swedish	sv	1	3,347,244 (0.25%)	6,247,073	2,131,957	168,306
Finnish	fi	2	1,687,673 (0.11%)	2,322,875	733,630	84,428
Greek	el	2	1,075,685 (0.08%)	921,053	18,737	770
Persian (Farsi)	fa	4	1,035,230 (0.08%)	746,396	2,537	1,081
Norwegian	no	1	870,515 (0.06%)	1,414,377	528,787	193,297
Chinese	zh	8	809,889 (0.06%)	1,188,903	14,126	2,101
Hebrew	he	2	460,390 (0.03%)	1,222,151	2,042	152
Other			104,715,420 (7,75%)	122,068,237	9,814,957	1,383,383
Total			1,350,641,311	1,369,379,741	652,818,796	124,259,396

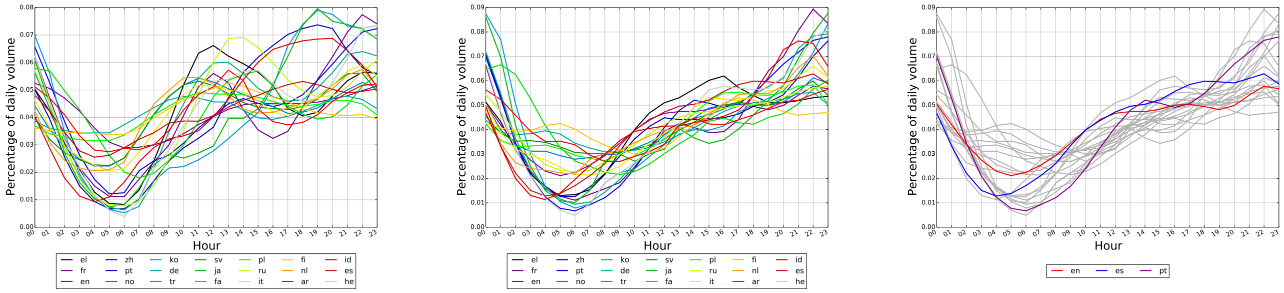


(a) max #tweets ≥ 200,000

(b) 10,000 ≤ max #tweets < 200,000

(c) max #tweets < 10,000

Figure 2: Yearly patterns for tweet volume in sample stream.



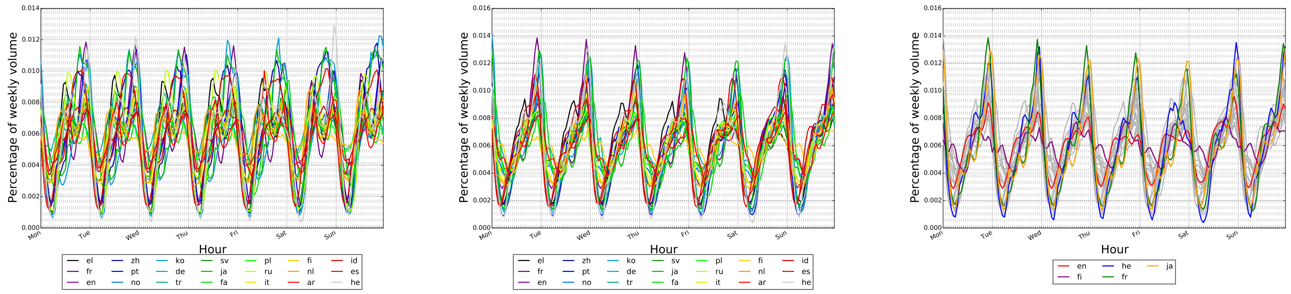
(a) Daily cycle geo-located stream

(b) Daily cycle sample stream

(c) Daily cycle sample stream for English, Spanish, and Portuguese

Figure 4: Average daily tweet patterns.

First, we analyse the trend in the daily number of tweets for the sample stream. We see large differences between languages. Since several plots greatly overlap, we display these plots in Figure 2 in three separate figures. The clear majority of the tweets we found are written in English (see Figure 2a), followed by the number of tweets in Japanese. Next we find Spanish, Indonesian, Portuguese, and Arabic. In Figure 2b, we select all languages that do not have more than 200,000 tweets per day in our dataset and in Figure 2c we choose this number to be 10,000. During February 2013, all plots in Figure 2



(a) Weekly cycle geo-located stream

(b) Weekly cycle sample stream

(c) Weekly cycle sample stream for English, Hebrew, Japanese, Finnish, and French

Figure 5: Average weekly tweet patterns.

indicate 0 tweets per language. This is because Twitter updated the language detection algorithms during this time [21]. After language detection is turned on, there are still clear jumps for certain languages. Most likely, these jumps correlate with updates of language detection algorithms in these languages. For the number of Dutch tweets in the sample stream, we see that this quantity is decreasing. We note that a decreasing line in Figure 2 does not have to imply that the number of tweets of that language is decreasing, since the quantities in the sample stream are relative to the total Twitter stream. However, Figure 3 shows the daily number of Dutch tweets that were received for all four streams. We see that all four streams show a decreasing pattern. Thus, we can conclude that the total volume of Dutch tweets is decreasing. This fact is also supported by Figure 1, in which the number of missed tweets for the 'NL general' stream is decreasing.

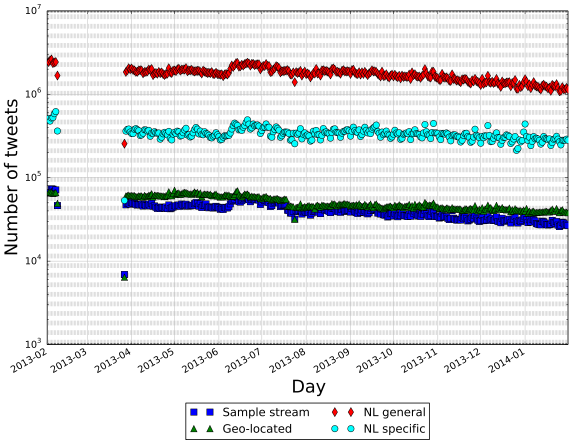


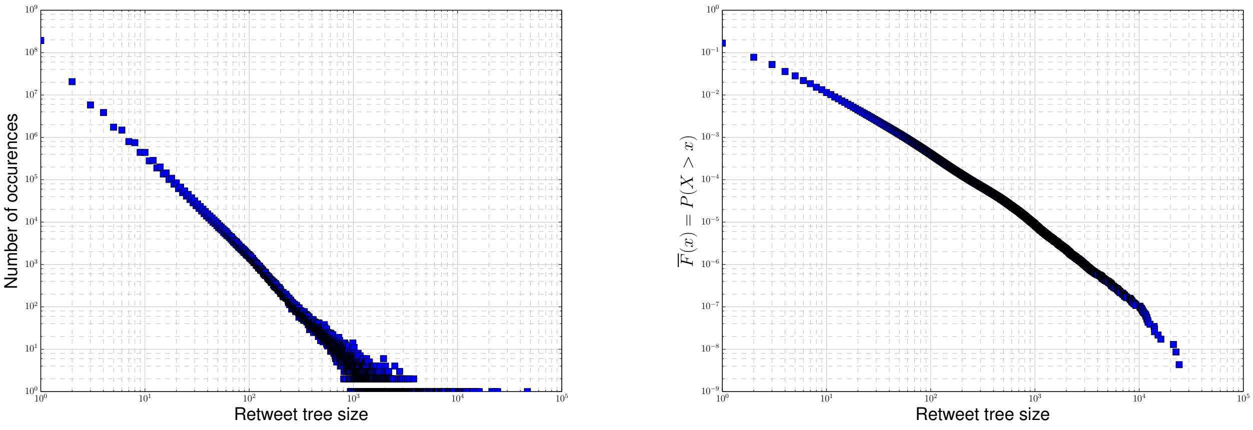
Figure 3: Number of Dutch tweets scraped daily.

Second, we focus on the daily patterns in the tweet volume. Since all languages are spoken in different timezones, we adjust the time-series to UTC time. The corrections that we used can be found in Table II. We analyse the daily patterns of two streams, namely the geo-located and the sample stream (Figures 4a and 4b, respectively). One striking difference between these figures is that while the sample stream displays a similar pattern for all languages, the geo-located stream differs strongly between languages. In the geo-located stream, we see two clear intervals where the activity peaks, namely during lunch hours and during the evening, which concurs with the findings of [9]. In Figure 4c, we highlight the daily patterns of three languages: English, Spanish, and Portuguese. These patterns are rescaled to American times, thus the majority of the usage of Twitter in these languages originates there. Furthermore, the amplitude of the English pattern is very low, therefore the activity in English is very spread out during the day. A large contrast to this is the pattern in Portuguese tweets. In this pattern, we clearly see a large decay in the number of tweets during the hours of the night.

Third, we regard weekly patterns in the tweet volume. Again, we focus on the geo-located and the sample stream. Similar to the daily patterns, we see that the sample stream patterns in Figure 5b are more consistent than the patterns in the geo-located streams in Figure 5a. For the majority of languages, we see a decreasing activity throughout the week in the sample stream patterns. Two good examples of this 'standard' pattern are Japanese and French, which are highlighted in Figure 5c. After a weekly decrease, the activity increases again on Sunday. Another language that follows this pattern is English; however, it has a very low amplitude with respect to the aforementioned languages. However, a clear exception to this pattern is the Hebrew pattern. In this language, the increase in activity happens on Saturday (since it is the Sabbat). Further, we find that for languages for which we have a small number of tweets (e.g., Finnish), there is only a decrease in activity during the night. The afternoon and evening activity for these languages are evenly distributed.

V. RETWEET TREES

In this section, we analyse the progression of tweets in our dataset through retweets. We define the progression of a message m as the retweet tree related to that message, which we denote by T_m. The graph T_m is a rooted tree, where all non-root vertices indicate people who retweeted the original message. If a user retweets an already retweeted message, this is shown as a new level in T_m. We use all four streams to determine the retweet trees. Figure 8 gives some examples of retweet trees.



(a) Distribution of the retweet tree size

(b) CCDF of the retweet tree size

Figure 6: Retweet tree size.

```
Require: $(t_{TS}, t_{MID}, t_{UID})\ \forall\ t \in V_{T_m}$ and root node $(r_{TS}, r_{MID}, r_{UID})$.
1: stop = False; $C = \{r_{MID}\}$;$L_c = \{(r_{MID}, r_{UID})\}$; $L_n = \emptyset$.
2: while stop = False do
3:     $T = V_{T_m} \setminus C$;
4:     $T^* = sort\_on\_time(T)$ {Time of posting}
5:     for $(t_{MID}, t_{UID}) \in T^*$ do
6:         for $(u_{MID}, u_{UID}) \in L_c$: do
7:             if $t_{UID}$ is followed by $u_{UID}$: then
8:                 $(t_{MID}, u_{MID}) \to E_{T_m}$
9:                 $u_{MID} \to C$
10:                $u_{MID} \to L_n$
11:            end if
12:        end for
13:    end for
14:    if $L_n = \emptyset$ then
15:        stop = True
16:    else
17:        $L_n \to L_c$
18:        $L_n = \emptyset$
19:    end if
20: end while
21: for $t \in V_{T_m}$: do
22:     if $t_{MID} \notin E_{T_m}$ then
23:         $(r_{MID}, t_{MID}) \to E_{T_m}$
24:     end if
25: end for
26: return $E_{T_m}$
```

Figure 7: Determine progression of message m.

We aim to derive the distribution function for the size of a retweet tree. Thus for each retweet tree in our dataset, we calculate the number of nodes in T_m, denoted by $|V_{T_m}|$, and use this to build the distribution function. This function is displayed in Figure 6a. Using this distribution function, we determine the Complementary Cumulative Density Function (CCDF) of the retweet tree size (see Figure 6b). From these plots, we find that the retweet tree size follows a power-law distribution.

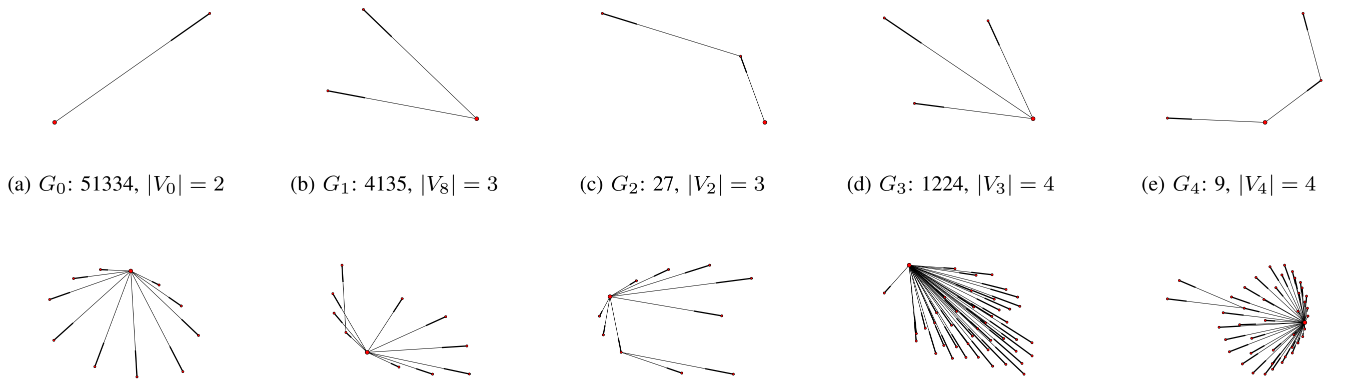
When a retweet is received in the Twitter API, one also receives the original message. However, the level at which this message lies in the retweet tree T_m is not given. Therefore, we propose the following algorithm to determine the progression of a retweet tree T_m. Given all retweets of a certain message, we start with this message. Then, for all retweets, we find out if the user that made that retweet is following the original poster of the message. If this is the case, this user retweeted the original poster. After we checked all users, we find the first level of the retweet tree. If we iterate this procedure until we cannot add new retweets to the tree, we are done. However, using this approach, it could be the case that some retweets have not been placed in the tree. Since these users do not follow any of the other users, we assume they found the tweet through search and thus retweet the original message. This algorithm is defined more formally in Figure 7. Here, we denote L_c as the current level, L_n as the new level and C is a list of checked messages. Furthermore, TS is short for timestamp, MID and UID are the message and user ID number, respectively, and E_{T_m} indicates the edge-set of tree T_m.

Hereafter, we studied the progression of retweet trees for January 13th 2014 from 18h to 19h. We chose a smaller dataset for this analysis due to the rate limitation of the Twitter API, which we need to check the follow-relation between two users in line 7 of Figure 7.

After retrieving the follow-relations and after processing the retweets, we find that the retweet trees tend to be wide and shallow. For instance the retweet tree consisting of three nodes of a star-shape (Figure 8b) occurs 4,135 times whereas the path-shape retweet tree of three nodes (Figure 8c) only occurs 27 times. Although part of this preference is caused by the last part of our algorithm, we find that we can allocate 67.71% of the 93,579 retweets of the timeframe by using the follow-relations in Figure 7.

VI. Conclusion and future work

In this paper, we analysed two aspects of user behaviour in Twitter. First, we analysed daily and weekly patterns that emerge from user activity in Twitter. We found that there



(a) G_0: 51334, $|V_0| = 2$ (b) G_1: 4135, $|V_8| = 3$ (c) G_2: 27, $|V_2| = 3$ (d) G_3: 1224, $|V_3| = 4$ (e) G_4: 9, $|V_4| = 4$

(f) G_5: 53, $|V_5| = 1$ (g) G_6: 5, $|V_6| = 11$ (h) G_7: 1, $|V_7| = 11$ (i) G_8: 1, $|V_8| = 44$ (j) G_9: 1, $|V_9| = 44$

Figure 8: Examples of retweet trees: number of occurences, retweet tree size.

are clear circadian patterns for every language we studied. Moreover, all studied languages show a similar daily pattern throughout the week. This concurs with studies done for other social media.

Also, we examined the number of daily tweets per language. Here we found no global patterns that hold for every language. However, through an analysis of the number of tweets that were not received through the streaming API, we see that the percentage of tweets that contains geo-locational data is increasing over time.

Furthermore, we studied the distribution of the size of a retweet tree. We found that these sizes follow a power-law distribution. Moreover, we extended this analysis to the actual progression of retweet trees in Twitter and found that retweet trees tend to be wide and shallow in their structure. For the algorithm in Figure 7, we need to know the network of relations within Twitter. Since this is a very time-consuming process, a possible extension of this work is to find a way to determine the progression of a retweet tree without knowing the complete graph.

REFERENCES

[1] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment." *ICWSM*, vol. 10, pp. 178–185, 2010.

[2] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[3] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, vol. 1. IEEE, 2010, pp. 492–499.

[4] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World Wide Web*. ACM, 2010, pp. 851–860.

[5] A. Kaltenbrunner, V. Gómez, A. Moghnieh, R. Meza, J. Blat, and V. López, "Homogeneous temporal activity patterns in a large online communication space," *IADIS International Journal on WWW/INTERNET*, vol. 6, no. 1, pp. 61–76, 2008.

[6] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "Youtube traffic characterization: A view from the edge," in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*. New York, NY, USA: ACM, 2007, pp. 15–28.

[7] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, vol. 53, no. 8, pp. 80–88, 2010.

[8] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "An empirical study of geographic user activity patterns in Foursquare." *ICWSM*, vol. 11, pp. 70–573, 2011.

[9] N. Grinberg, M. Naaman, B. Shaw, and G. Lotan, "Extracting diurnal patterns of real world activity from social media," in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM–13)*, 2013.

[10] T. Yasseri, R. Sumi, and J. Kertész, "Circadian patterns of Wikipedia editorial activity: A demographic analysis," *PloS one*, vol. 7, no. 1, p. e30091, 2012.

[11] M. ten Thij, Y. Volkovich, D. Laniado, and A. Kaltenbrunner, "Modeling and predicting page-view dynamics on Wikipedia," *arXiv preprint arXiv:1212.5943*, 2012.

[12] B. Poblete, R. Garcia, M. Mendoza, and A. Jaimes, "Do all birds tweet the same?: characterizing Twitter around the world," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 1025–1030.

[13] D. Mocanu, A. Baronchelli, N. Perra, B. Gonalves, Q. Zhang, and A. Vespignani, "The Twitter of babel: Mapping world languages through microblogging platforms," *PLoS ONE*, vol. 8, no. 4, 2013.

[14] D. Jurgens and T.-C. Lu, "Temporal motifs reveal the dynamics of editor interactions in wikipedia," in *International AAAI Conference on Weblogs and Social Media*, 2012.

[15] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on Digg and Twitter social networks." *ICWSM*, vol. 10, pp. 90–97, 2010.

[16] K. Y. Kamath, J. Caverlee, K. Lee, and Z. Cheng, "Spatio-temporal dynamics of online memes: A study of geo-tagged tweets," in *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2013, pp. 667–678.

[17] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, ser. WSDM '11. New York, NY, USA: ACM, 2011, pp. 177–186.

[18] S. Bhamidi, J. M. Steele, and T. Zaman, "Twitter event networks and the superstar model," *arXiv preprint arXiv:1211.3090*, 2012.

[19] Retrieved: June 27, 2014. [Online]. Available: https://dev.twitter.com/docs/api/1.1/post/statuses/filter

[20] Retrieved: June 27, 2014. [Online]. Available: https://dev.twitter.com/docs/faq#6861

[21] Retrieved: June 27, 2014. [Online]. Available: https://blog.twitter.com/2013/introducing-new-metadata-for-tweets