Who are my Audiences? A Study of the Evolution of Target Audiences in Microblogs

Ruth García-Gavilanes[1], Andreas Kaltenbrunner[2], Diego Sáez-Trumper[3],
Ricardo Baeza-Yates[3], Pablo Aragón[2], and David Laniado[2]

[1] Universitat Pompeu Fabra, Barcelona, Spain
[2] Fundación Barcelona Media, Barcelona, Spain
[3] Yahoo Labs, Barcelona, Spain

Abstract. User behavior in online social media is not static, it evolves through the years. In Twitter, we have witnessed a maturation of its platform and its users due to endogenous and exogenous reasons. While the research using Twitter data has expanded rapidly, little work has studied the change/evolution in the Twitter ecosystem itself. In this paper, we use a taxonomy of the types of tweets posted by around 4M users during 10 weeks in 2011 and 2013. We classify users according to their tweeting behavior, and find 5 clusters for which we can associate a different dominant tweeting type. Furthermore, we observe the evolution of users across groups between 2011 and 2013 and find interesting insights such as the decrease in conversations and increase in URLs sharing. Our findings suggest that mature users evolve to adopt Twitter as a *news media* rather than a social network.

1 Introduction

Online social networks like Twitter have become extremely popular. Twitter has grown from thousands of users in 2007 over millions in 2009 to hundreds of millions in 2013. Through the years, users have learned to use Twitter following certain conventions in their messages, limited to 140 characters. In certain occasions, these conventions help users to imagine a target audience or set a topic that goes along with what the community is talking about. For example, the use of the symbol @ (at) before a user name to mark a dyadic interaction between two users and the use of *re-tweets* for spreading the content of a tweet posted by someone else. Likewise, the use of URLs (often shortened) to share external information, etc.

As a consequence, Twitter is used in several contexts, for different audiences and with different purposes. In fact, scholars have argued that Twitter is used as an hybrid between a communication media and an online social network [6, 17]. Additionally, user behavior is not static, it changes through the years, the way the first Twitter users interacted with the platform when it started may differ from how they interact now. While the set of research using Twitter data has expanded rapidly, little work has studied the change/evolution in Twitter ecosystem itself.

In this paper, we propose a step towards understanding the evolution of user behavior focusing on *how* people tweet and their audiences. To this end, we carry out a longitudinal study of tweets posted during 10 weeks in 2011 and 10 weeks in 2013 by more than 4M users who have been active in Twitter in both of these periods.

First, we propose a taxonomy of messages based on Twitter conventions (mentions, links, re-tweets). In doing so we obtained 6 tweet formats. To identify models of behavior, we cluster users based on these types of tweets and study how users change their behavior in time. To present our results, we organize the paper as follows. Section 2 provides related work. Section 3 describes the data. In Section 4 we explain our methodology and the taxonomy given to the types of tweets. In Section 5 we report how user behavior changes in 2013 with respect to 2011. We finish with conclusions and next steps in Section 6.

2 Related Work

The goal of this work is to study the variation of tweeting behavior across time based on a taxonomy of tweet types and audiences. In a similar way, researchers have already analyzed how a variety of aspects change across time in Twitter and other online platforms. They have studied the following aspects:

The nature of Twitter. While most messages on Twitter are conversations and chatter, people also use it to share relevant information and to report news [4]. In fact, scholars have concluded that from the highly skewed nature of the distribution of followers and the low rate of reciprocated ties, Twitter more closely resembles an information sharing network than a social network [6].

Evolution of users and behavior. Liu *et al.* [8] studied the evolution of Twitter users and their behavior by using a large set of tweets between 2006 and 2013. They quantify a number of trends, including the spread of Twitter across the globe, the shift from a primarily-desktop to a primarily-mobile system, the rise of malicious behavior, and the changes in tweeting behavior. The main part of this study is based on the accumulative number of tweets. We address, instead, the evolution based on individual users' behavior.

Audiences. Marwick and boyd [10, 13] claim that users in Twitter *imagine* their target audiences since they do not know "which few" will read their tweets. They find that users do not have a fixed target audience and that having one would be a synonym of "inauthenticity".

Behavior and clusters. Naaman *et al.* [11] find 4 relevant categories of tweets based on the content of the messages. For each one of these categories, they cluster users and find two types of users: Meformers (talking about one self) and Informers (sharing news). Luo *et al.* [9] classify tweets based on language and syntactic structure and Huang *et al.* [3] show that tagging behavior (hashtags) has a conversational, rather than organizational nature.

Many attempts have been done to classify users according to their audiences and tweet content. However, most of these studies are language-dependent and need manual labeling. In this work, we categorize audiences and tweet types using a language-independent approach.

3 Data Set

For the results we present here, we crawled the profile information of users who posted tweets with the hashtag *#followfriday* or *#ff* on the first Friday of March, 2011 as in [2].

From this set, we randomly selected 55K users with a number of followers and followees in the range of [100, 1000] and crawled their corresponding followee network (for a user u, it contains all users who u is following).

We then proceeded to collect all of the tweets posted in English by the original 55K users as well as their followees during 10 weeks starting from the second half of March 2011. By crawling the information of the followees, we attempt to target the typical accounts twitterers like to follow. It is mostly on these users and the 55K seed set that all our results are concerned.

In total, we obtained 8M users who tweeted around 2.4B tweets. We then crawled Twitter during 10 weeks between October and December 2013 looking for the same users and found that around 4.3M users tweeted at least once also in 2013. After the end of the crawling period, we identify the language in which tweets are written. We then proceed to classify as *active users* those who tweeted at least 55 and less or equal than 1540 tweets in English during 10 weeks to exclude inactive or hyperactive users and bots. In total we found around 538K users tweeting within this range in both years. We chose this range as to set a threshold of 1 tweet per working day (5 per week) and a maximum of 22 per day. The maximum limit was chosen based on a marketing study by Zarrella [19], which argues that most users tweet an average of 22 times a day. With this we attempt to include users likely to be engaged with the platform excluding those with an abnormal activity (i.e, advertisers or bots). Appendix A describes details about the crawling process and Table A1 presents the summary of the dataset used for the experiments.

4 Methodology

As previously discussed in the related work section, some researchers argue that everybody has an *imagined audience* in a communicative act even if that act involves social media [10]. Given the various ways people consume and spread tweets, it is virtually impossible for Twitter users to account for their potential audience, although we often find users tweeting as if these audiences were bounded. For instance, the use of the @ sign before a user login name allows to “poke” that user which may trigger a reply and start dyadic conversations (through mentions) which are visible at the same time to others as well. In fact, Marwick and boyd [10] found, through interviews to twitterers, that sometimes

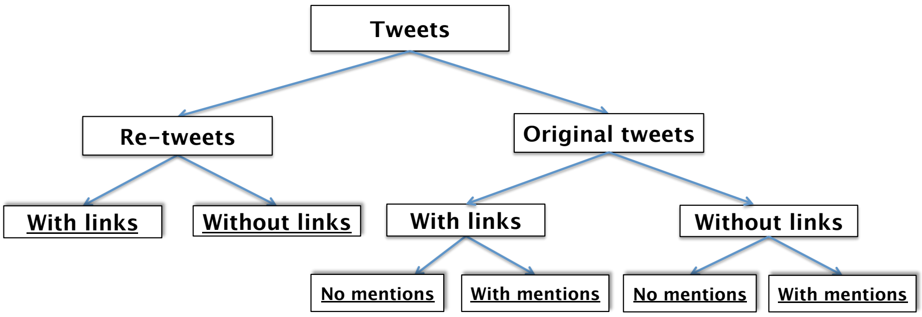
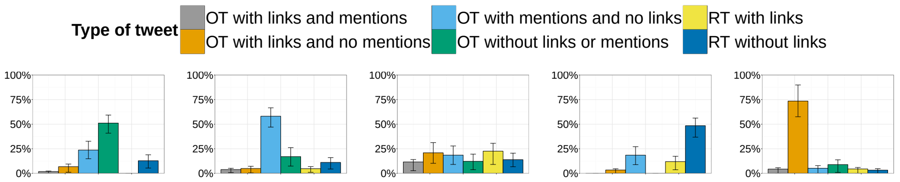


Fig. 1. This classification tree represents the tweet formats used to classify users in different groups. The top groups include the tweets in the subsequent levels. The underlined nodes (leaves of the tree) are used in the clustering process (6 types).

users are "conscious of potential overlap among their audiences (i.e, friends, family, co-workers, etc)." The authors report cases where users tweet to themselves, to fans, to fellow nerds, to super users, etc.

We propose a language-independent taxonomy of tweet types. The proposed types are based on the conventions established by Twitter such as the mention symbol @, the retweet flag and the URLs, *imagining* an audience through the combination of these symbols. Figure 1 shows these categories.

We start by classifying two main groups of tweets: retweets (RT) and original tweets (OT). Retweets refer to those tweets forwarded from other users. We hypothesize that a retweet targets the user who created the forwarded tweet and the followers of the user forwarding the tweet. Next, original tweets refer to tweets posted by users themselves and the audience could vary between the followers and the users themselves. For the RT and OT sets, we make two other distinctions: tweets with URLs and without URLs. We hypothesize that URLs target audiences who are willing to obtain information from the links posted and generally interested in exogenous stimuli. For tweets without URLS, users want to transmit a self-contained idea in maximum 140 characters. For the OT set we make yet another distinction, for the tweets with URLs and without URLs we divide them between tweets containing a mention (conversational) and those without a mention (textual). A OT containing a link with a mention implies that a user calls the attention of another user to open the link shared in the tweet. We do not make this last distinction (mention and link) for the RT set given than all retweets already refer to another user. In this study, we focus on the tweet types at the deepest level of each branch (6 in total): a) re-tweets with links, b) re-tweets without links, c) original tweets with links and no mentions, d) original tweets with links and mentions, e) original tweets without links and no mentions and finally f) original tweets without links and mentions.



(a) Endogen. (b) Conver. (c) Gener. (d) Echoers (e) Link F.

Fig. 2. Clustering based on 6 tweet types posted by *active* users during 10 weeks in 2011 and 2013. The clusters appear from left to right according to their size in descending order. Each bar shows the average percentage of that tweet type. Error bars represent the interquartile range. Clusters (a) and (d) do not contain tweets of all types.

Based on this scheme, we classify the tweets of the *active* set of users (538 K) in 2011 and 2013 and find a slight increase in tweets with URLs in 2013 (from 14.62% to 18.74%). Table B1 of Appendix B has the percentage of tweets in each category for *active* users.

Furthermore, for each *active* pair (user, year) we calculate the percentages of tweets belonging to each of the tweet types. Each pair (user, year) is represented by a 6-dimensional vector, *6* being the number of all numerical features (the percentages) used to describe the objects to be clustered. We use the well-known k-means algorithm for clustering. To decide the k points in that vector space, we used the so called *elbow method*. This is a visual standard method [12] that runs the k-means algorithm with different numbers of clusters and shows the results of the sum of the squared error. The value of k is chosen by starting with $k = 2$ and increasing it by 1 until the gain of the solution drops dramatically, which will be the bend or elbow of the graph. This is the k value we want and is chosen visually. We found that the *bend* lingered between 4 and 5 (see Figure B1 in Appendix B). We analyzed both cases and chose $k = 5$ because we observed that it best encapsulates interesting and distinctive patterns of tweeting behavior.

5 Results

We now proceed to the results and study how users have changed their tweeting behavior through time. Figure 2 shows the average composition of tweet type vectors in the clusters. The clusters are ordered by size and the bars indicate the interquartile range for each case. Note that we have abbreviated some of the names in the captions due to space concerns. We observe that each cluster has a dominant tweet type except for the third cluster (*Generalists*) that reports a balance among the tweet types.

We discuss now each of the identified patterns of tweeting behavior and relate them to the concept of the *imagined audiences* discussed in the previous section.

Endogenous: Users in this cluster mostly post and forward messages not linked to external information. Users in this cluster are supposed to use Twitter more as a social network than as a news media. The dominant type of tweets are self-contained posts created by the user herself without mentioning other users such as quotes, thoughts or even futile information. In second place we observe original tweets with mentions which is a sign of conversation with other users.

Conversationalists: Users following this pattern are characterized mostly by tweets containing mentions with no links. Similarly to the *Endogenous* type, users in this cluster are also supposed to use Twitter more as a social network but with an emphasis on interacting with other users more than sharing self-contained ideas.

Generalists: This cluster groups users who use Twitter without a distinctive tweet type. It is interesting to notice that in this cluster, retweets with links and original tweets with links are slightly above the rest which may suggest an inclination to audiences interested in obtaining external information.

Echoers: These are users characterized by forwarding other people's tweets with no links. These users are mostly inclined to read what others have to say, indicating in a way that they make part of the audience of other users‘s original ideas (being these informative or not). An example of such users are those who follow accounts posting jokes, positive thinking, quotes, etc. The second dominant category in this cluster involves tweets with mentions, which most likely mean that users reply or chat with others.

Link Feeders: This cluster involves all those accounts that mostly tweet messages containing external links. In 2011 [18] found that around 50% of URLs posted in tweets came from media producers. We expect then that the owners of these accounts are mainly news media, journalists, link builders, SEO specialists, etc. Since these are tweets that contain no mentions, the expected target audience is then a general public that aims to obtain information through these accounts (i.e, followers of news papers).

The clustering process was based on the tweets of active users in both 2011 and 2013. Figure B2 in Appendix B shows the number of users falling in one of the clusters for each year.

5.1 Change in Tweeting Behavior

Here we study how users have changed their tweeting behavior in 2013 with respect to 2011. Based on the active users only (those who remained active in 2011 and 2013), we plot these groups into a Sankey diagram in Figure 3 to observe the proportion of users moving from one cluster to another.

We observe that in general around half of these active users remain in the same cluster in both periods, except for the Echoers. On the other hand, we observe an increase in 2013 of the *Generalists* and *Link Feeders* cluster with respect to 2011. The increase in the *Generalists* cluster is expected since our dataset contains users who have remained in Twitter for more than two years. These users have matured with the platform and most likely learned to use it for multiple reasons (chat, share information, retweets, etc). Moreover, the increase in the *Link Feeders* cluster goes along with Table B1, which also shows an increase in the percentage of tweets with URLs. Nowadays, Twitter automatically shortens URLs using the t.co service [1] which makes it easier for users to share

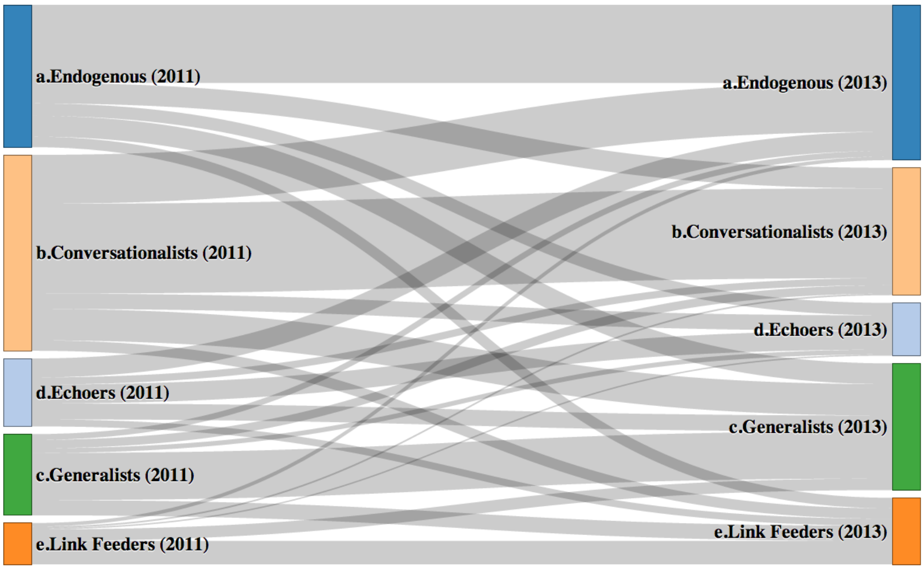


Fig. 3. The Sankey shows how active users have changed the way they tweet in 2013 with regard to 2011.

links without the need to visit other URL shortener sites. This was not the case in 2011. Additionally, an increasing number of external sites allow to automatically post on Twitter with their link included. It is expected then that by 2013 users share more URLs than before.

On the other hand, we see a decrease in 2013 of the *Conversationalists* type. It seems that some users who used to chat a lot are evolving to chat less and be more *Endogenous* (posting their own tweets with no links or mentions) and *Generalists*. Mature users would have quickly realized that it was hard to continue conversations once the chat channel has passed in Twitter. On top of that, cross-platform instant messaging services more oriented to conversation purposes (i.e.,WhatsApp) have become increasingly popular. Neverthless, in 2013 Twitter made it easier to follow conversations in the timeline [5]. Perhaps, we will witness an increase in conversations after 2013.

Finally, the decrease in the *Echoers* cluster from 2011 to 2013 shows that users who tend to forward other people's ideas most of the time have evolved to generate more content themselves, moving to the *Endogenous* or *Generalist* clusters.

For a better readability of the evolution of active users' behavior, we did not include in the Sankey diagram the proportion of users who were filtered out of the active set in 2011 and moved to any of the clusters in 2013. We include this information in Table 1 in percentages (of around 4.3 M users) and show in Table B2 (Appendix B) the corresponding absolute values. We observe that the majority of users from any cluster in 2011 become inactive in 2013. Similarly,

Table 1. Percentage of users who changed clusters from 2011 (rows) to 2013 (columns). Some users passed from inactive or hyperactive/bot to other clusters and vice versa.

2011/ 2013	Endogenous	Conver.	Gener.	Echoers	Link F.	Inactive	Hyper./Bots
Endogenous	22.38%	5.89%	5.96%	3.56%	3.02%	58.33%	0.86%
Conver.	11.33%	20.79%	7.26%	3.54%	2.41%	53.80%	0.87%
Generalists	2.67%	3.88%	21.78%	2.17%	7.02%	62.07%	0.41%
Echoers	9.93%	3.72%	8.31%	9.93%	3.65%	63.62%	0.84%
Link Feeders	3.38%	1.47%	11.11%	1.25%	22.59%	59.45%	0.75%
Inactive	6.64%	3.30%	3.12%	2.38%	2.31%	82.00%	0.26%
Hyper./Bots	28.13%	17.42%	8.15%	6.48%	4.91%	26.71%	8.19%

inactive users tend to remain as such even two years later. Interestingly, the majority of hyperactive users move to one of the clusters but we also observe a significant percentage (26.71%) becoming inactive in 2013.

These findings go along with Liu *et al.* [8], who found a massive percentage of inactive accounts by the end of 2013. As Twitter users mature, many also choose to move to other platforms and to be less active.

6 Conclusions

In this paper we have carried out a study in Twitter between 2011 and 2013. We propose a taxonomy of 6 tweet types and found that users fall into 5 clusters of behavior: Endogenous (those who mostly tweet without links or mentions), Conversationalists (those who mostly converse with others), Generalists (those who post different type of tweets), Echoers (those who re-tweet more) and Link Feeders (those who share URLs most of the time). We then observed the evolution of users across clusters between these years and noticed a general tendency to become inactive or maintain the same type of behavior over years, with the exception of *echoers* who show to be active in a year full of controversial events. We also observed a decrease of *conversationalists*, likely due to the maturation of users, the emergence of instant message services and the difficulty of chatting in Twitter before 2013. We also found more Link Feeders and Generalists in 2013. In the past, Twitter has been described as hybrid platform, being a social network and a news media at the same time [6]; our results, with the increase in news feeders and decrease in conversationalists, suggest that the main usage of the service by mature users is shifting towards the latter: a news media.

After completing this study, there are several complementary projects ahead. For instance, we plan to look closely at the behavior of the inactive and hyperactive users and bots. We also plan to study the lexical variation in dyadic conversations across time. Furthermore, it would be interesting to analyze if users tweeting in several languages differ in tweeting behavior for each language. Finally, we plan to compare this evolution to the change in user popularity.

References

1. T. H. Center. Posting links in a tweet. *https://support.twitter.com/entries/78124-how-to-shorten-links-URLs*.
2. R. García-Gavilanes, N. O'Hare, L. M. Aiello, and A. Jaimes. Follow my friends this friday! an analysis of human-generated friendship recommendations. In *The 5th International Conference on Social Informatics (SOCINFO)*, 2013.
3. J. Huang, K. Thornton, and E. Efthimiadis. Conversational tagging in twitter. In *Proceedings of the 21st ACM Conf. on Hypertext and Hypermedia*, 2010.
4. A. Java, X. Song, T. Finin, and B. Tseng. Why We Twitter: Understanding Microblogging Usage and Communities. In *Procedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007*, 2007.
5. J. Kamdar. Keep up with conversations on twitter. *https://blog.twitter.com/2013/keep-up-with-conversations-on-twitter*.
6. H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference companion on World Wide Web*, 2010.
7. K. Lee, B. Eoff, and J. Caverlee. Seven months with the devils: A long-term study of content polluters on twitter. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.
8. Y. Liu, C. Kliman-Silver, and A. Mislove. The tweets they are a-changin': Evolution of twitter users and behavior. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2014.
9. Z. Luo, M. Osborne, S. Petrovic, and T. Wang. Improving twitter retrieval by exploiting structural information. In *In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, JToronto*, 2012.
10. A. Marwick and danah boyd. I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media and Society*, Sept. 2010.
11. M. Naaman, J. Boase, and C.-H. Lai. Is it really about me?: Message content in social awareness streams. In *Proceedings of the 13th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW'10)*, 2010.
12. A. Ng. Machine learning. *Available at https://www.coursera.org*, 2014.
13. Z. Papacharissi. The presentation of self in virtual life: Characteristics of personal home page. *Journalism and Mass Communication Quarterly*, 2002.
14. S. Petrović, M. Osborne, and V. Lavrenko. Rt to win! predicting message propagation in twitter. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.
15. B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *The 2nd IEEE International Conference on Social Computing*, 2010.
16. K. Thomas, C. Grier, D. Song, and V. Paxson. Suspended accounts in retrospect: an analysis of twitter spam. In *ACM Proceedings of the 2011 ACM SIGCOMM on the Internet Measurement Conference*, 2011.
17. S. Wu, J. Hofman, W. Mason, and D. Watts. Who says what to whom on Twitter. In *Proceedings of the 20tt International Conference companion on World Wide Web*, 2011.
18. S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on twitter. In *Proceedings of the 20tt International Conference companion on World Wide Web*, 2011.
19. D. Zarrella. The science of timing. *Available at http://www.slideshare.net/HubSpot/the-science-of-timing*, 2011.

A Appendix : Detailed dataset description

The Follow Friday hashtag emerged in 2009 as a spontaneous convention from the Twitter user base: users post tweets with the *#followfriday* (or #ff) hashtag, and include the usernames of the users they wish to recommend on Fridays. Back in 2010 and 2011, this hashtag was one of the most used in Twitter [15, 14] and so we hypothesized that engaged twitter users would likely adopt this hashtag because they care about recommending users to follow.

We crawled 55K users with number of followers and followees in the range of [100, 1000] not exceed the limit of the API calls at that time. It also has the added benefit of filtering out less legitimate (e.g., spam) users, since, according to Lee *et al.* [7], the majority of spam users tend to have out-degree and in-degree outside the range of [100; 1000]. Also Kurt *et al.* [16] showed that 89% of users following spam accounts have fewer than 10 followers. So, while we cannot guarantee that our dataset does not contain spammers, previous studies indicate that our sample will indeed have a higher probability of containing mostly legitimate users.

The information collected includes the user id, the screen name, the information in the location field of the profile, the date stamp of the tweet, the number of followers and followees, the *id* and the text of the tweet. We continue by finding the geolocation of each user via the location field entered in their profiles and we kept those geolocated users as to add one additional anti-spam filter. We believed that users who specified a valid geolocation are less likely to be spammers.

There is a higher proportion of active users among those users who tweeted in *both* 2011 *and* 2013 (the 5th row of the 3rd and 4th column) than those who tweeted in 2011 but not necessarily in 2013 (the 5th row of the 2nd column).

Table A1. The second column shows the full data crawled in 2011. The 3rd and 4th column show information of users who tweeted in *both* 2011 and 2013. Rows 3 to 5 contains information about active and inactive users. Rows 7 to 9 contain information of the active users only. Active users are those considered to have tweeted in English more than 55 and less than 1540 times.

Active and inactive set

	Full Data Set 2011	**Users active in 2011 & 2013**	
	2011	**2011**	**2013**
Users	8,092,891	4,350,583	4,350,583
Tweets	2,280,707,094	1,527,675,950	679,507,450
English Tweets	1,086,233,182	768,940,902	369,452,361
Active set			
Active Users	1,868,150	1,315,313	1,125,968
Tweets	1,248,300,919	880,889,333	375,741,789
English Tweets	562,134,366	406,719,99	256,330,241

B Appendix : Complementary Material

Table B1. Tweets from active users in 2011 and 2013, and the corresponding percentage of tweets that belong to each type.

	Full DS 2011	2011	2013
	Tweets	Tweets	Tweets
Original tweets	77.30%	76.94%	74.77%
With URLs	14.93%	14.62%	18.74%
with mentions	6.39%	3.46%	4.16%
without mentions	11.36%	11.16%	14.58%
Without URLs	62.37%	62.32%	56.03%
with mentions	35.18%	35.36%	27.44%
without mentions	27.19%	26.96%	28.59%
Retweets	22.70%	23.06%	25.23%
With URLs	6.29%	6.75%	8.6%
Without URLs	16.41%	16.31%	16.63%

Table B2. The absolute number of users who moved across clusters from 2011 (rows) to 2013 (columns). Some users passed from inactive or hyperactive/bot to the other clusters and vice versa.

2011/2013	Endogenous	Conver.	Gener.	Echoers	Link F.	Inactive	Hyper./Bots
Endogenous	79,472	20,900	21,159	12,657	10,705	207,108	3,036
Conver.	49,832	91,429	31,945	15,570	10,616	236,624	3,807
Generalists	5,886	8,542	47,997	4,784	15,479	136,813	903
Echoers	19,308	7,235	16,149	19,306	7,105	123,704	1,640
Link Feeders	3,573	1,548	11,736	1,315	23,855	62,781	794
Inactive	194,636	96,641	91,391	69,684	67,769	2,403,596	7,481
Hyper./Bots	29,275	18,131	8,484	6,745	5,109	27,803	8,529

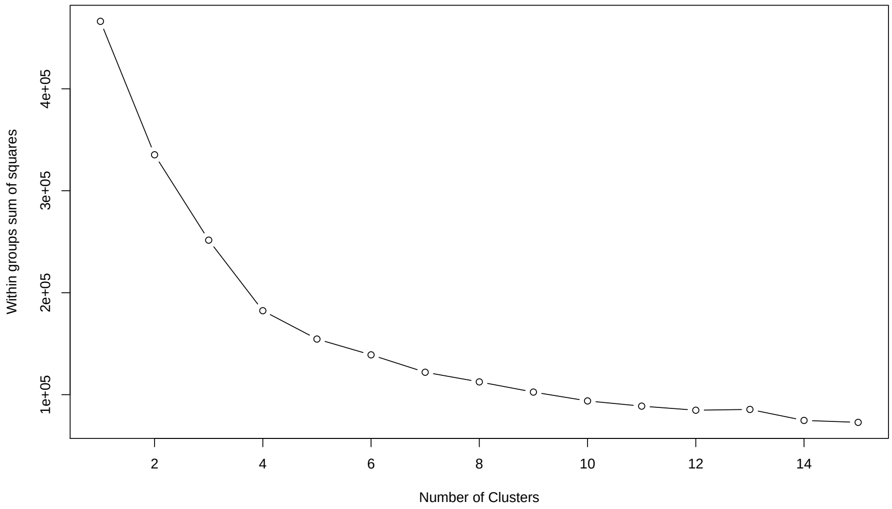
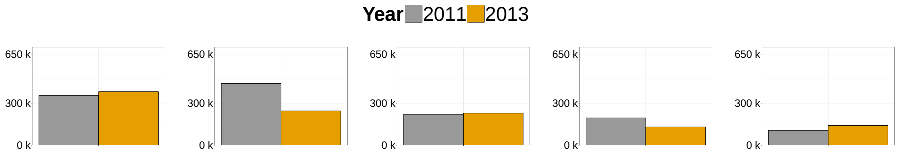


Fig. B1. Elbow method for clustering : the *bend* lingers between 4 and 5.



(a) Endog. (b) Conver. (c) Gener. (d) Echoers (e) Link F.

Fig. B2. Number of active users in each cluster for 2011 and 2013.