Soc. Netw. Anal. Min. (2016) 6:12
DOI 10.1007/s13278-016-0320-6

ORIGINAL ARTICLE

Hashtags and followers

An experimental study of the online social network Twitter

Eva García Martín[1] · Niklas Lavesson[1] · Mina Doroud[2]

Received: 18 August 2015 / Accepted: 22 February 2016 / Published online: 14 March 2016

Abstract We have conducted an analysis of data from 502,891 Twitter users and focused on investigating the potential correlation between hashtags and the increase of followers to determine whether the addition of hashtags to tweets produces new followers. We have designed an experiment with two groups of users: one tweeting with random hashtags and one tweeting without hashtags. The results showed that there is a correlation between hashtags and followers: on average, users tweeting with hashtags increased their followers by 2.88, while users tweeting without hashtags increased 0.88 followers. We present a simple, reproducible approach to extract and analyze Twitter user data for this and similar purposes.

Keywords Experimental study · Correlational analysis · Hashtags · Followers

1 Introduction

Twitter is a social network founded in 2006 that publishes around 1 billion tweets every 2 days (Terdiman 2012), with the goal of spreading world breaking news or opinions about certain events (Wang et al. 2011). The research conducted on Twitter has been centered in sentiment analysis, event prediction, retweet prediction and graph modeling.

Twitter has a key characteristic that was later adopted by other social networks, the hashtag symbol, commonly used to tag posts into different categories. Although there has been a lot of research conducted in Twitter in the past years, we believe that not enough studies have been published that try to discover how hashtags affect users in Twitter. Hashtags have gained relevance in the past years since companies use them to marketize events, tv shows and sport matches, to attract users engagement towards their brand. A way to increase this engagement is by gaining followers on Twitter on the companies account. It has to be noted that having more followers do not always mean being more influential. The reason is that retweets and mentions are not strongly correlated with the number of followers, being retweets and mentions two measures that capture the real audience reactions. This means that follower count is not enough to capture the influence of a user, since it only measures the size of the audience, not its reaction (Cha et al. 2010).

For these reasons and considering that there is still a need to understand how users react to posts with hashtags, we have conducted a large-scale exploration of the potential correlation between the use of hashtags and the increase of followers on Twitter. More specifically, a natural experiment has been created where the average increase of followers is compared between two groups of distinctive users, one tweeting with hashtags and another tweeting without hashtags. The results of the experiment show that users tweeting with hashtags gain more followers than users tweeting without hashtags. To avoid confounding factors on the results, users tweeting with and without

This work is part of the research project "Scalable resource-efficient systems for big data analytics" funded by the Knowledge Foundation (Grant: 20140032) in Sweden.

✉ Eva García Martín
eva.garcia.martin@bth.se

Niklas Lavesson
niklas.lavesson@bth.se

Mina Doroud
m.doroud@gmail.com

[1] Blekinge Institute of Technology, Karlskrona, Sweden

[2] Twitter Inc., San Francisco, USA

hashtags have been picked following the exact same random procedure, ending up with a random Twitter user population. Therefore, if confounding variables appear, they would be canceled out (Peirce et al. 1935).

Some analyses have been made in relation to popular users. Due to the fact that Twitter considers popular tweets as those that generate more engagements, we believe that users with a high number of followers could be considered popular users, since they have more probability to create engagement towards them. Our results show that popular users that tweet with hashtags gain more followers than popular users tweeting without hashtags. The main reason behind this is that the hashtag channel is attracting more followers to the already popular users. When a user clicks on a hashtag, popular tweets will appear first, thus increasing the probability for popular users to gain visibility.

Although there is a clear relationship between the use of hashtags and the increase of followers, this study does not go beyond than portraying a correlation between both variables. A cause-effect claim can not be made between using hashtags and increasing followers. However, we can recommend the use of hashtags to increase visibility in Twitter, as the results suggest.

The article is structured as follows: in Sect. 2, the background is presented with the Twitter-related terminology and the different articles related to social networks. In Sect. 3, the purpose statement is detailed. In Sect. 4, the research methodology, the experimental design, data collection and statistical analysis are described. In Sect. 5, the raw and analyzed results are presented. In Sect. 6, we summarize the conclusions. Finally, in Sect. 7, we provide further perspectives of future work.

2 Background

2.1 Terminology

Tweets are short 140 character messages. To tweet is the action of posting a tweet in Twitter. Twitter users tweet to show to their followers their thoughts about a specific matter, to post breaking news or to post information about topics they are interested in (Mathioudakis and Koudas 2010). In the following paragraphs we are going to detail all the Twitter glossary[1]. Mentions are used for connecting users. If user U_a wants to mention user U_b, U_a posts the character @ followed by U_b username, e.g. *@username*.

Retweet is a particular case of mentioning. A retweet is the action of a user tweeting the tweet of another user. Retweets are either created automatically, by clicking on the retweet button; or manually, by placing the text indicator *RT*. If the user has manually retweeted a message with modifications, we use the text indicator *MT*[2].

When users want to categorize their messages into specific topics, they add the hash symbol to the topic. For example *#computerscience*. The hash symbol plus the topic or word is called a hashtag (Wang et al. 2011).

Users connect with each other by following each other. If U_a follows U_b, then U_a receives in his or her timeline all the tweets from U_b. In this case, U_a is a follower of U_b and U_b is a friend of U_a.

For gathering data from Twitter, there is an open API available for developers[3]. This API makes it easy for the developer to send requests to Twitter to ask for specific information (Makice 2009). In Sect. 4.2, we detail how the data were gathered from the API. In order to allow Twitter to monitor the number of requests we make, we need to follow an authentication protocol, *Oauth*[4]. The information is then gathered by the authenticated user.

2.2 Related work

This section is organized in three parts. First, we briefly introduce studies related to online social networks in general. Second, we detail the three main focuses, mostly in Twitter, more closely related to this study. Finally, we end up detailing the two closest works to this one and explaining the main differences between them.

Online social networks are becoming more and more popular nowadays. Several models have been built that try to define the behavior of such networks (Kumar et al. 2010; Mislove 2009). The most common model is the preferential attachment, created by Barabási and Albert (1999) and tested with positive results by Newman (2001) and Jeong et al. (2003). Preferential attachment states that new links tend to form towards already popular links. Popular links are users that have a higher number of followers compared to the average Twitter user. On the other hand, Lang and Wu (2011) built a growth model of the social network *Buzznet* looking for evidence of preferential attachment. They unexpectedly discovered that *Buzznet* follows an anti-preferential attachment model; therefore, high-degree nodes create edges to low-degree nodes. Degree can be defined as the number of connections that a node creates towards other nodes. This means that users that we expect to have a higher number of followers than friends (high-degree nodes), such as celebrities, end up having a lower number of followers than friends (low-degree nodes).

[1] http://support.twitter.com/articles/166337-the-twitter-glossary.

[2] https://support.twitter.com/articles/166337-the-twitter-glossary#m.

[3] https://dev.twitter.com/.

[4] https://dev.twitter.com/docs/auth/oauth.

Moreover, several studies such as Sakaki et al. (2010), Ritterman et al. (2009) and Qiu et al. (2011) have been made on event prediction using Twitter as a source. Based on hashtags or trends they try to predict the appearance of future events. We have also observed that social networks have been used by many researchers to detect the sentiment from different users to certain products or events. This is called sentiment analysis and it used both to analyze users opinions individually and to analyze opinions of groups of users. Looking into users as a group creates a more complete solution when combined with traditional sentiment analysis approaches (Bifet and Frank 2010; Pak and Paroubek 2010; Wang et al. 2011; Thelwall et al. 2011; Wang et al. 2013; Go et al. 2009; Diakopoulos and Shamma 2010).

In relation to the research conducted in Twitter, we have identified three main focuses related to this study. The first, influence, tries to identify the reasons behind influential posts on Twitter. Influence stands for "the power to cause an effect", in this case, the power to cause users to take certain actions, for instance retweet a certain tweet. Originating from the traditional influential theories, it was widely believed that only a minority group of people, influencers, were capable of causing influence to other users (Katz and Lazarsfeld 1955; Rogers 2010). However, it was later discovered that ordinary users can also create influence, and that influence depends more on the type of posts, how society reacts to a specific trend and how close are the opinions of such users (Cha et al. 2010; Domingos and Richardson 2001). The discovery most related to this study, is that indegree, that is, the number of followers of a user, is not a strong measure of the influence of such user, however retweets and mentions are (Cha et al. 2010). A final remark from the same authors, is that users with active followers are more likely to be retweeted, and since retweet is an important measure of influence, we could conclude that having active followers increases the influence of a user.

The second focus is related to link prediction. Link prediction studies different factors that are related to a user gaining followers. A first study was conducted where researches discovered that links in Flickr were usually created by users who already had many links (Mislove 2009). In the Twitter area, it has been shown that the topic of the tweet is one of the indicators to predict whether users will gain or loose followers. Tweets with happy messages gain followers in contrast to tweet with sad messages (Kivran-Swaine and Naaman 2011; Quercia et al. 2011). Another important factor, stated by Hutto et al. (2013), is the shared interests between the users. As was noted in the influence aspect, there are more links between users that have similar interests. On the same line, popular users often remain popular, in contrast to ordinary users. Regarding the friend and follower network (Huberman et al. 2008), several studies focus on modeling this network into balanced or unbalanced, depending if two users have a friend in common whom they both follow, and its impact into maintaining followers in Twitter (Kivran-Swaine et al. 2011). Other studies focus on predicting future friend nodes in the network (Nia et al. 2013).

The third focus is on the hashtag topic. A discovery was made by Suh et al. (2010), where they state that tweets containing hashtags and URLs are more likely to be retweeted. They also mention that the amount of followers of a user does affect retweetability. This indirectly contradicts the claims by Cha et al. (2010), since they mention that influence is not related to the amount of followers but that retweets are. However, if followers and retweets are related, based on the claim by Suh et al. (2010), then followers should also be related to influence. In relation to the hashtag topic, Kong et al. (2014) have created a study where they predict trending topic hashtags in real time. Since the topic of the tweet is related to increase of followers, as was stated before, we believe that analyzing which hashtags attract more followers is a perfect follow-up work. Regarding prediction, Maruf et al. (2014) predicts the personality of a user analyzing their hashtags and She and Chen (2014), Otsuka et al. (2014), Yu and Shen (2014), Wang et al. (2014) have created several recommender systems that suggest hashtags to the Twitter user.

Finally, Hutto et al. (2013) and Jungselius et al. (2014) have created two studies related to ours. Jungselius et al. (2014) extract that users in Instagram do attract more followers and likes if such users adds hashtags to their publications. Hutto et al. (2013) have conducted a complete study on the reasons behind follower growth patters. They measure the relationship between many control variables and the increase or decrease of followers. They conclude that hashtags is one of the factors related to the increase of followers, matching with our discovery. The new perspective of our study is the ability to isolate hashtag usage from all other variables to ensure that the measured correlation is not affected by any other variable. Our complete experimental study is designed to study the correlation between hashtags and followers and avoid any confounding variables. Hutto et al. (2013) study is more a general analysis into active users.

3 Purpose statement

The purpose of this experimental study is to investigate the potential correlation between hashtag usage and the increase or decrease of followers; controlling for retweets, user characteristics and user popularity. The independent variable is hashtags. The dependent variable is the change in the number of followers. In this natural experiment we control for retweets, user popularity, user characteristics, *the million follower fallacy* (Cha et al. 2010) and new mentions. We take into account the effect of retweets since

one of the key reasons for new followers is that U_a starts to follow U_b because they had a friend in common that retweeted a tweet from U_b.

The million follower fallacy is a term used when U_a starts following U_b just for etiquette or for being polite to follow someone that already follows you. As for user characteristics, we sample random users to have a set of users that is representative of the whole population, with random ages, genders, nationalities, languages and popularity level. Popular users are users widely known to the public with many links towards them, such as celebrities and famous sport players. We also consider the fact that U_a mentions U_b in his or her tweet. The username of U_b is publicly being seen by all the followers of U_a, increasing the chance of U_b to get new followers.

Finally, as for confounding variables, we consider the following three:

- A user posting his or her user account on a public place in the Internet.
- Twitter suggesting to follow new users.
- A Twitter user tweeting with a specific hashtag when attending a concert, conference or other events that might attract new followers.

We are aware that other confounding variables might appear in the future. However, in relation to the third bullet point, we are not studying the type of hashtag that leads to an increase of followers, we are studying if the presence of any hashtag affects this increase.

At this point, we have disclosed which control and confounding variables we are going to take into consideration. The way these variables are going to affect the experiment design and how we are going to control them is explained in Sect. 4.

4 Research methodology

To investigate the possible relationship between the use of hashtags and the increase of followers, we propose a specific research question, namely; whether the addition of hashtags to tweets produces new followers.

We want to discover if hashtags are related to user visibility. In Twitter, if you click on any hashtag, trendy or not, you can see the users that are tweeting with that hashtag. Hence, we want to discover if users that search tweets based on a specific hashtag, actually start to follow the authors of those tweets. For that reason, we hypothesize that there is an increase in the number of followers for users tweeting with hashtags.

4.1 Experimental design

The aim is to perform a natural experiment to determine if hashtags are related to the increase of followers. The main characteristics that we want to achieve are randomization, non-biased choices and a clear distinction between users tweeting with hashtags and without hashtags.

We created two independent groups of users, an experimental group and a control group. All users from both groups are gathered following the same procedure, being the only difference between them the hashtag usage. Therefore, since the sample group is large enough, if any confounding variables emerge, they will appear in both groups and they will be canceled out. This is known as *random assignment* (Peirce et al. 1935), where the objective is to minimize the chance that a difference between both groups will be due to confounding factors. Users that have tweeted with a hashtag in the moment they were collected form part of the experimental group and users that have tweeted without a hashtag form part of the control group.

There is also a relevant design choice made, being, that tweets from the past are not considered. The main reason is that it is not possible to know if past tweets attracted new followers, since we can only know the number of followers in present time. We are aware that other factors might still influence the results, therefore, we plan to identify and address such factors in subsequent work. With this experimental design and after a statistical analysis of the data we are able to answer the research question.

4.2 Data collection

Tweets and users are obtained from Twitter using the Twitter API. We developed a Python script to make the different requests to the Twitter API, the code is described in Sect. 4.3. We used an open source package called Twython, from developer Ryan McGrath[5], to make the different requests to Twitter.

From the Twitter API two requests were used:

- GET statuses/sample
- GET users/lookup

The first request returns a small sample of random tweets from Twitter's public timeline independent of the location or language. For every tweet that does not contain a hashtag, its author becomes a member of the control group, and for every tweet that contains a hashtag, its author becomes a member of the experimental group. The following information is saved for each user: *Username*, *number of tweets*, *number of followers*, *number of friends*, *gathering date*, *number of times that tweet has been chosen as favorite*, *number of times that tweet has been retweeted*, *the tweet text*.

[5] https://github.com/ryanmcgrath/twython.

The second request returns the information from a maximum of 100 users. We use this request to update the information of each user to discover if the number of followers increased or decreased since they were collected. Every user is updated a total of 5 times, every 12 min. The reason we choose to update every 12 min in an hour span, is that based on a study conducted by Lardinois (2009), the lifespan of a tweet is roughly 1 h. Therefore, it is more likely that if an increase of followers occurs after 1 h, it is due to another tweet or external factor. Every update is made after 12 min to see which of the five updates have the highest increase of followers; to match this with the study by Bray (2012) which states that a tweet is more popular during the first 18 min after it was published.

There were a total of 502, 891 users acquired, 252, 957 users tweeting without hashtags and 249, 934 tweeting with hashtags. There are some special considerations relevant for this study. Twitter API has a limit in the total number of requests that the developer can make per hour, therefore we have gathered the total number of users based on this limit. The updates of all the users could not be made a total of 5 times in all cases, because on rare occasions Twitter's service was unavailable due to the server being overloaded with requests. To ensure that we make the correct calculations, we have removed all users that were not updated 5 times and whose updates where outside the interval between 1 and 2 h.

4.3 Python script

Script 1 portrays the code for obtaining all users and tweets from Twitter[6].

Script 1: To obtain a sample of tweets from Twitter with Twython.

```
users = [];
days = 7 ;
TOT−HOURS = 24 * days ;
for hour in range (TOT−HOURS):
    tweets = TwythonStreamer;
    users = getUsers(tweets);
    saveTweets(tweets);
    saveUsers(users);
    for update in range (5):
        sleep(12);
        udpt−users = twitter.LookupUser(users);
        db.users.update(udpt−users);
```

The first step is to obtain the random sample of tweets from Twitter with the functions already implemented in Twython. We save these tweets as a collection in MongoDB named *tweets*. The next step is to extract the authors of such tweets and save them in another collection on MongoDB named *users*.

Moreover, after the users have been obtained, we update their information (number of tweets, number of followers and number of friends) five times in 1 h, once every 12 min. We append this new information to each *document* from each user created in the second step. When the five updates are finished, new users are gathered and the process is repeated again for a complete week. In summary, we have a script that is going to be executed during one week, picking up new users every hour and updating their information every 12 min. Sometimes, this interval will be higher due to Twitter rate limits.

The next step is to preprocess the data to calculate the difference between the number of followers between the last and the first instance. Finally, we import the data into a statistical tool (Matlab in this case) to perform the statistical analysis.

4.4 Data analysis

We first perform the Kolmogorov–Smirnov normality test (Kolmogorov 1933) on the difference of followers for the control and the experimental group. In Fig. 1 we represent the histogram of such differences, having zoomed both axis for an easier display. The results of the normality test are shown in Table 1.

The null hypotheses are the following:

- "The values that measure the difference in the number of followers for the control group follow a normal distribution".
- "The values that measure the difference in the number of followers for the experimental group follow a normal distribution".

Both hypotheses are rejected, since the P value is less than 0.05, concluding that the difference of followers for both groups does not follow a normal distribution.

Since the data are tested negative for normality, we are going to perform a non-parametric test, namely, the Mann–Whitney U test (Mann and Whitney 1947).

The null hypothesis is: "The mean of the populations of the experimental group and the control group are equal".

The alternative hypothesis is the following: "The mean of the populations of the experimental group is significantly higher than the mean of the populations of the control group".

In this test, the absolute value of parameter z,

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \quad (1)$$

is compared with the critical value for a confidence level of 0.05 shown in Altman's Z score table (Altman 1968). This

[6] https://github.com/evek2/tw-hashtags.

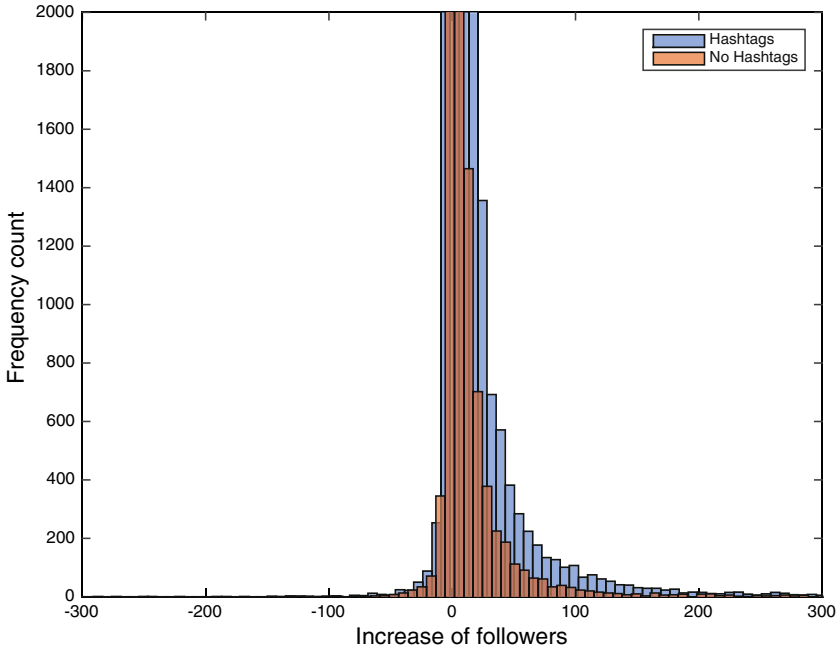

Fig. 1 Increase of followers. Approximation of the probability distribution of the increase of followers for the users who tweet with and without hashtags. Users tweeting with hashtags (*blue*) have a higher increase of followers than users tweeting without hashtags (*orange*)

Table 1 Kolmogorov–Smirnov test on the control and experimental group

	Control	Experimental
P value	<0.05	<0.05
Confidence level	0.05	0.05
Final result	Hypothesis rejected: the data do not follow a normal distribution	Hypothesis rejected: the data do not follow a normal distribution

table displays the values of the normal distribution parametrized for different confidence levels.

If the absolute value of z is bigger than the critical value, for a one-tailed test, then the null hypothesis is rejected. If the null hypothesis is rejected, then the mean of the populations, in this case the experimental and control group, are significantly different.

The way of collecting data, explained in Sect. 4.1, is known as simple random sampling with replacement. Each user is randomly chosen from a large data set, in this case Twitter's public timeline. Therefore, each user has the same probability of begin chosen, ensuring independent sample values (Moore and McCabe 2011; Cochran 2007; Starnes et al. 2010; Given 2008).

4.5 Validity evaluation

For this study we have considered four validity threats: internal, external, construct and statistical conclusion (Shadish et al. 2002).

Internal validity stands for the extent to which a causal conclusion can be made and how the variables in the experiment are manipulated. In this case, we have not made a cause-effect claim, we study a potential correlation between the dependent and independent variable. We designed the experiment to show that the increase of followers is only influenced by hashtags, by having a large and random Twitter population. Therefore, an appearance of a confounding variable will be canceled out. This experiment is a natural experiment, in contrast with a true experiment. The reason is that we are observing certain individuals, in this case users, and how their actions lead to a change in the number of followers. Since we are merely observers, we are categorizing users into one or other group, but we are not testing subjects based on some predefined actions.

External validity is defined as: "to what populations, settings and variables can this effect be generalized". In terms of this study, it means if the correlation between hashtags and followers could be generalized to all twitter users, or, on the other hand, only applies to certain users. Since we gather users from different periods of the day, during a complete week, and with a completely random procedure, we believe that the results indeed generalize for the complete Twitter population.

Construct validity is "the degree to which a test or program measures what it claims". Therefore, we need to ensure that the data which have been collected, and that the computations that have been made are correct. After gathering the data, we analyzed the data in search for some outliers. Several data points were found that seemed like a mistake, e.g., duplicate users or deleted profiles. We

cleaned the data after performing tests on all users, and we also performed some tests to check whether the computations (increase of followers, friends, etc.) were done correctly.

Finally, statistical conclusion validity stands for the degree to which conclusions about the relationship between variables match with the correct use of statistics. We used two statistical tests, first, the Kolmogorov–Smirnov test, which concluded that the data were not normally distributed. Then we chose to use a non-parametric test. The main reason is that parametric tests make no assumption of the probability distribution of the variables. We then chose the Mann–Whitney *U* test, with the main objective of testing whether two samples, in this case increase of followers with and without hashtag usage, are drawn from the same distribution (Sheskin 2003).

5 Results and analysis

5.1 Statistical characterization

Table 2 shows an overview of different statistics obtained from the dataset. Our goal was to portray a better understanding of the users presented in the dataset that were later analyzed in this study. In particular, we present statistics from the followers, friends and tweets. The top 10 % users means that the 10 % users with the highest value of some variable were analyzed. Popular users are users with a high number of followers. In this case, the ratio of popular users details how many users with highest increase of followers are actually the ones with the highest number of followers. We calculate this by getting the 10 % users with the highest increase of followers and the 10 % users with the highest number of followers and see how many users were in common from the population analyzed. In this case, the population analyzed is the users with positive increase of followers. Top 10 % users with high followers and tweets increase, represent the percentage of users that have both the highest 10 % increase of followers and the highest 10 % increase of tweets. Conclusions obtained from these values are explained in Sect. 5.2.

5.1.1 Mann–Whitney U test

Since the samples do not follow a normal distribution, the Mann–Whitney *U* test is performed to test the existence of

Table 2 Statistical summary of the data

	Non-hashtag group	Hashtag group
Users		
Total	252,957	249,934
Followers increase		
Maximum	6248	17,048
Mean	0.88	2.38
Users with 0 followers increase (%)	71 %	64 %
Users with 1 followers increase (%)	12.12 %	12.82 %
Users between 2–20 followers increase (%)	8.24 %	13.25 %
Users with more than 20 followers increase (%)	0.84 %	1.89 %
Top 10 %	36.22	73.86
Followers decrease		
Average	−2.5	−4.2
% of users	7.8 %	8.05 %
Total followers		
Average	1010	1776
Top 10 %	20,342	36,080
Popular users ratio	17.19 %	22.06 %
Tweets		
Average increase	10.35	12.18
Top 10 % users with high followers and tweets increase	29.12 %	31.12 %
% of users with positive increase of tweets	82.19 %	80.95 %
Friends		
Minimum of the difference of friends	−36,936	−3403
Maximum of the difference of friends	5712	16,815
Mean of the difference of friends	0.62	1.61

Table 3 Numerical values for the Mann–Whitney U test

Parameters	Control group	Experimental group
Group size	$n_1 = 252{,}957$	$n_2 = 249{,}934$
Group rank	$\sum R_1 = 4.23 \times 10^{10}$	$\sum R_2 = 8.42 \times 10^{10}$
U parameter	$U_1 = 5.29 \times 10^{10}$	$U_2 = 1.03 \times 10^{10}$

a significant difference between the control and the experimental group. The numerical values obtained for this test are displayed in Table 3.

We compute U as $U = \min\{U_1, U_2\}$, where the values of U_1 and U_2 are: $U_1 = 5.29 \times 10^{10}$ and $U_2 = 1.03 \times 10^{10}$.

Since $U_2 < U_1 \Rightarrow U = U_2 = 1.03 \times 10^{10}$. Then, we compute the value of z, obtaining: $|z| = 414.27$. Since $z_{0.05} = 1.65 \Rightarrow |z| > z_{0.05}$.

The absolute value of z is bigger than $z_{0.05}$; therefore, the null hypothesis is rejected and the alternative hypothesis is supported. The mean of the experimental group is significantly higher than the mean of the control group.

5.2 Data exploration

The goal of this section is to check the relationship between several control variables and the increase of followers, to discover if they are related. We also portray interesting findings from the users in the dataset.

Popular users are users with a high number of followers. In this study, we characterize these users as the ones with the 10 % highest number of followers. From the results shown in Table 2, we can see that, on average, users tweeting with hashtags have a higher number of followers, 36,080 against 20,342. This means that there are more popular users tweeting with hashtags that without. The next step is to check if there is a relationship between popular users and the increase of followers. For that reason, we calculate the percentage of users that have both a high increase of followers and a high number of followers. For the non-hashtag case, 17.19 % users are the top 10 % in both followers increase and number of followers. This number increases by a 5 % in the hashtag scenario. Based on these results, we can conclude that even though hashtag users have more popular users, those are increasing their number of followers only 5 % more than non-hashtag users, and they only represent a 22 % of the users increasing high the number of followers.

In relation to the increase of followers, we can observe how, on average, users tweeting with hashtags increase more than double their number of followers that users without hashtags. This matches with the results of our study, that defines a correlation between tweeting with hashtags and the increase of followers. To understand how this increase is distributed, we have computed how many users, as a percentage of the total number of users, actually increase zero followers, one, between 2 and 20 and more than 20. What the results show is that a 71 % of the users do not increase their followers in the non-hashtag group and a 64 % do not increase in the hashtag group. Then this number of users decreases while the number of followers increases, i.e., there are less users increasing a lot their number of followers, but those that tweet with hashtags have always a higher increase than users that tweet without. An interesting phenomenon is that users tweeting with hashtags are also decreasing their followers more, on average, than users tweeting without hashtags. This suggests *visibility*.

We believe that users tweeting with hashtags are more visible in Twitter. This claim is supported by two facts. First, the fact that they are both loosing and gaining more followers, which suggests that they have more presence in Twitter. Second, the fact that, on average, users tweeting with hashtags are also tweeting more than users tweeting without, 10.35 against 12.18 increase of tweets. Therefore hashtag users are more active than non-hashtag users. So the reason why they are more visible could be because of their activeness and their hashtag use. The question to answer, is then: *Do users tweeting with hashtags increase more their followers because they are tweeting more (12.18 vs 10.35) rather than their use of hashtags?* They way we controlled for this possible occurrence is by calculating how many users that have a high increase of followers are actually tweeting at a high rate. Since, if the claim were true, then users tweeting at a higher rate would have a higher increase of followers. We calculate the 10 % users with the highest increase of followers, the 10 % with the highest increase of tweets and see how many users are in common. The results show that 29.12 % of non-hashtag users have both a high increase of tweets and followers, and that 31.12 % of hashtag users have both a high increase of tweets and followers. The main conclusion extracted from these results is that there is no apparent relationship between tweeting at a high rate and having a high increase of followers. If it were, then the difference between hashtag and non-hashtag group would be higher than 2 %, because we already know that more users are increasing their followers for the hashtag group (74 % vs 36 %).

5.3 Trend analysis

This section illustrates analyses between certain features presented in the data set obtained from Twitter. The main general trend and the key finding in this study is portrayed in Fig. 1. This figure represents that on average, users that tweet with hashtags have a higher increase of followers than users that tweet without hashtags. This figure matches with the results shown in Table 2, where the average

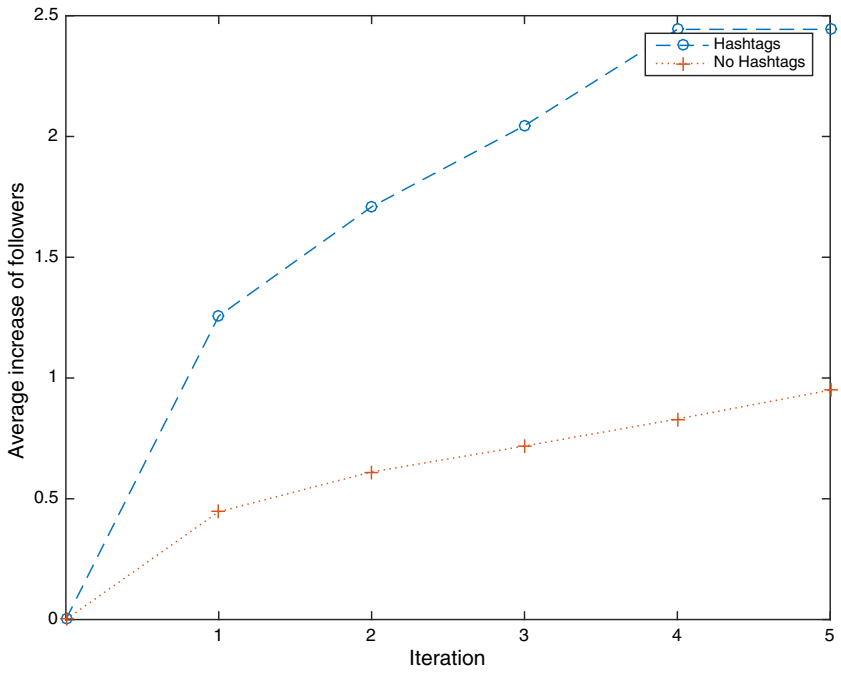

Fig. 2 Average increase of followers with and without hashtags. This figure shows a comparison between the increase of followers of users tweeting with hashtags and the increase of followers of users tweeting without hashtags. This comparison is per iteration, that is every time the information from each user is updated

increase of followers for users with hashtags is 2.38 and the average increase of followers for users without hashtags is 0.88. Figure 2 shows a comparison between the same increase of followers in the hashtag and non-hashtag usage, but separates the increase of followers into different iterations or updates. It can be observed that the highest increase occurs in the first iteration, during the first 12 min after the user has published the tweet. Bray (2012) stated that a tweet is popular during the first 18 min after it has been published. For that reason, it makes sense that the highest increase of followers happens during the first 12 min, since after the next iteration, in the minute 24, the tweet's popularity will have decreased. Also, looking at the blue line that represents the increase of followers for users with hashtags, we can observe that the last update is almost flat. This complements the study by Lardinois (2009), where they discovered that the lifespan of a tweet was 1 h. Moreover, Fig. 3 shows how users tweeting with and without hashtag decrease their number of followers. We analyzed the behavior of those users that loose followers and concluded that except from the fourth and fifth update, both types of users decrease their followers in the same way. However, users tweeting with hashtags decrease significantly more their number of followers in the fourth update.

Figure 4 shows the relationship between the total number of followers and the total number of friends. The axis are zoomed for a clearer view of the plot. In this figure we can identify two types of Twitter users, already explained in Sect. 3, differentiated between two clusters. The first cluster represents users with high number of followers and low number of friends. We can see this cluster in all the points that move around the X axis but have low values of Y, forming a horizontal line. These users are users that have a lot of followers but that do not follow other users. Users such as celebrities can follow this type of trend, since a lot of fans are following them but they only follow some friends back. This is called preferential attachment, and was discovered by Barabási and Albert (1999). The second cluster is formed by the points with a linear relationship between followers and friends. The cluster can be observed by looking at the points in the figure that form a diagonal line across the frame. These users could be matched to the *follower fallacy* (Cha et al. 2010). As explained before, users follow other users that follow them just for etiquette or being polite, ending up with similar number of followers and friends. Another interesting characteristic of this figure is the fact that there are seldom users with a high number of friends and a low number of followers. There are some outliers that become visible if we zoom out the figure. Finally, another group of users are the ones with a more random relationship between followers and friends that do not really follow any pattern.

Figures 5, 6 and 7 detail relationships between the increase or decrease of followers and the number of hashtags that the user tweeted with. As we can see from Fig. 5, most of the users that tweeted with hashtags tweeted with one hashtag. The number of users decreases as the number of hashtags increases. The goal with this figure is to portray how many hashtags were used by all users, to

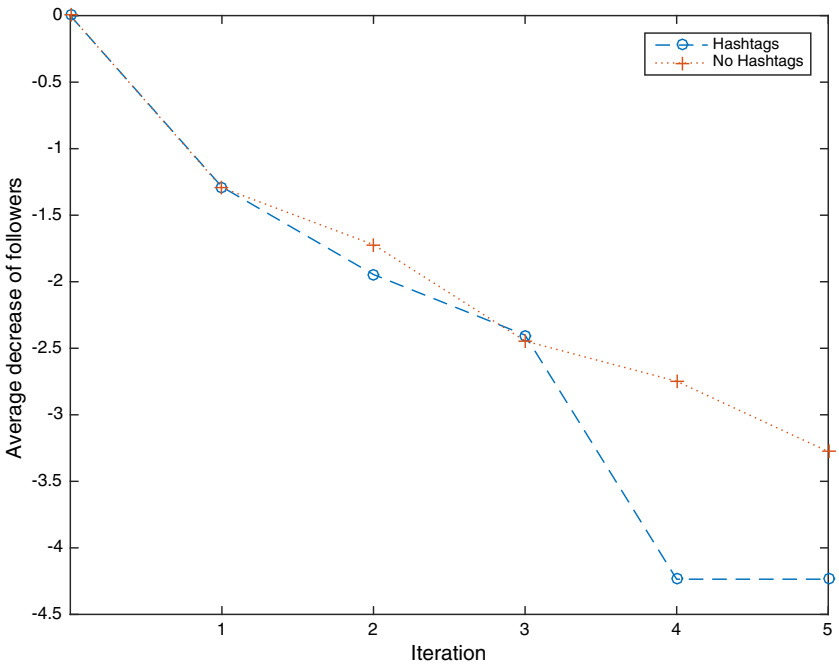

Fig. 3 Average decrease of followers with and without hashtags. This figure shows a comparison between the decrease of followers of users tweeting with hashtags and the increase of followers of users tweeting without hashtags. This comparison is per iteration, that is every time the information from each user is updated

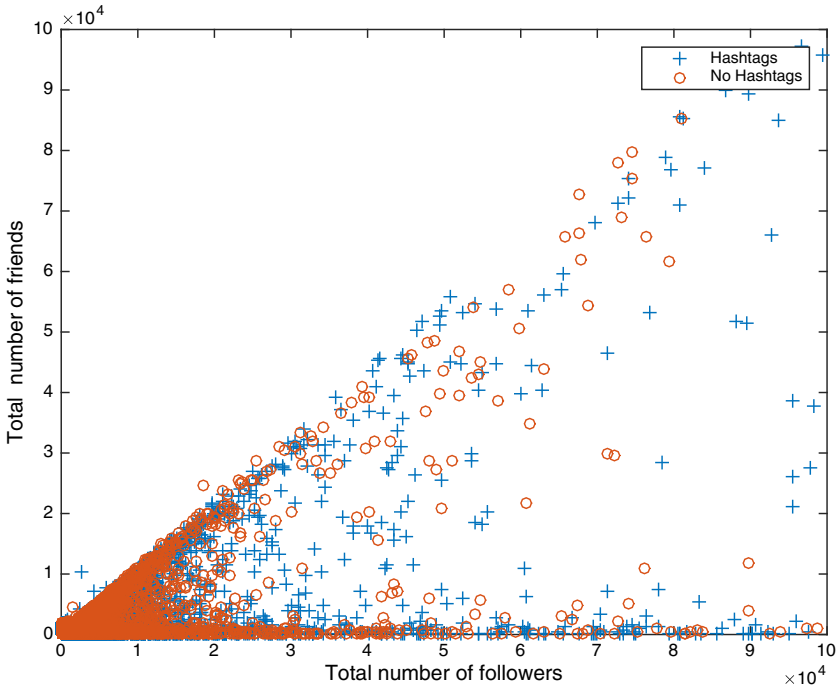

Fig. 4 Total number of followers against total number of friends. Comparison between the total number of followers and friends for users that tweet with and without hashtags. Both users with and without hashtags present a similar trend, in the form of two different clusters. The first cluster is formed by the points that create a *horizontal line*, and the second cluster by the points that create a *linear* and *diagonal line*

then analyze it together with Figs. 6 and 7. Having a lot of users tweeting with a specific number of hashtags does not mean that they increase a lot their followers, but if we check the average increase and decrease of followers per hashtag, it can gives us an overview on how these two measures are related. We can observe that tweeting with one hashtag had, on average, the second highest increase of followers and that users tweeting with six hashtags had the highest increase of followers, which is not expected. On average, the higher the number of hashtags the lower the number of followers. In an online study, it was suggested that using more than one or two hashtags will drop the user engagement by 17 %[7]. This claim was also presented in the

[7] https://blog.bufferapp.com/a-scientific-guide-to-hashtags-which-ones-work-when-and-how-many.

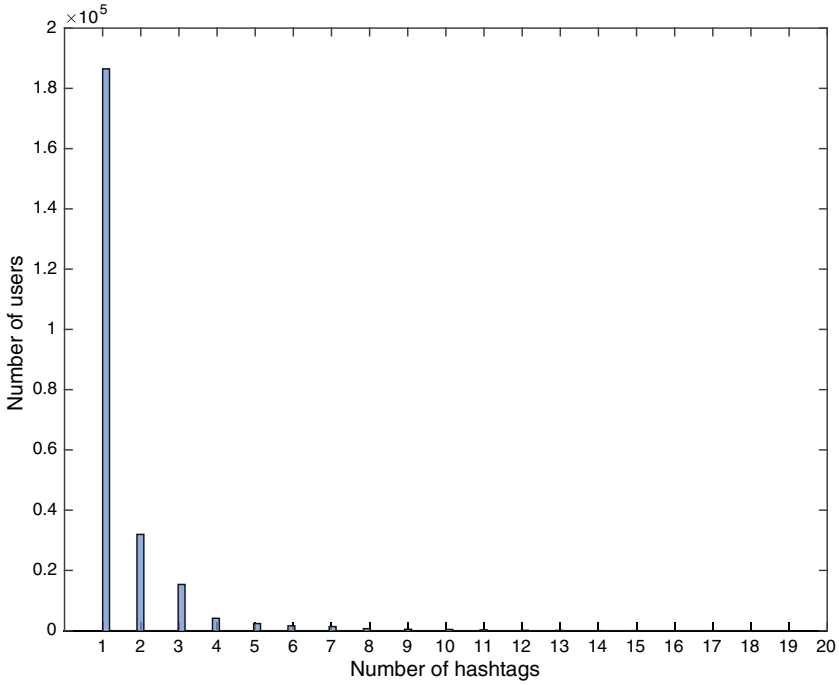

Fig. 5 Number of users with a positive, negative or zero increase of followers against the number of hashtags. This figure shows the number of users that tweeted with *X* number of hashtags

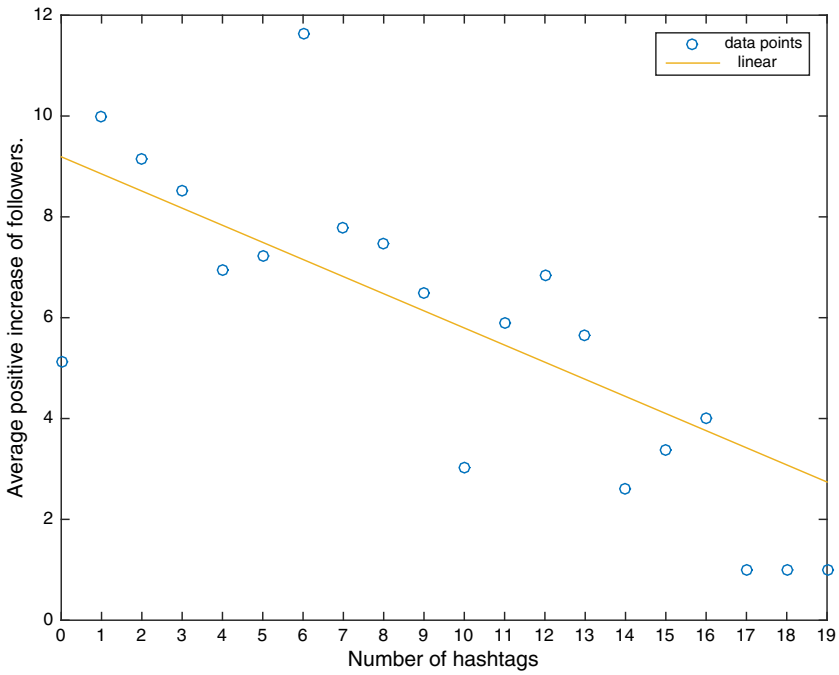

Fig. 6 Average increase of followers against the number of hashtags. The average increase of followers of users tweeting with a certain (x) number of hashtags. On average, the higher the number of hashtags the lower the increase of followers

work by Hutto et al. (2013). If we analyze the decrease of followers from Fig. 7, tweeting with one hashtag had a higher decrease than the baseline of not tweeting with any hashtag. The higher the usage of hashtags the lower the decrease, although since the change from one hashtag to another is low, the decrease does not seem to be dependent on the number of hashtags. Maybe users tweeting with a lot of hashtags do get less followed but they do not get unfollowed. The final remark is that tweeting with five hashtags get a significant decrease of followers. This could be related to some bot actions, were they are tweeting with these amount of hashtags and users unfollow them immediately.

In terms of connectivity, we have analyzed the tweets that are replies or mentions to other users. These tweets represent an 8.15 % of the total users that tweet with hashtags. From these 8.15 %, the aim was to understand the connection between them. For example, one finding

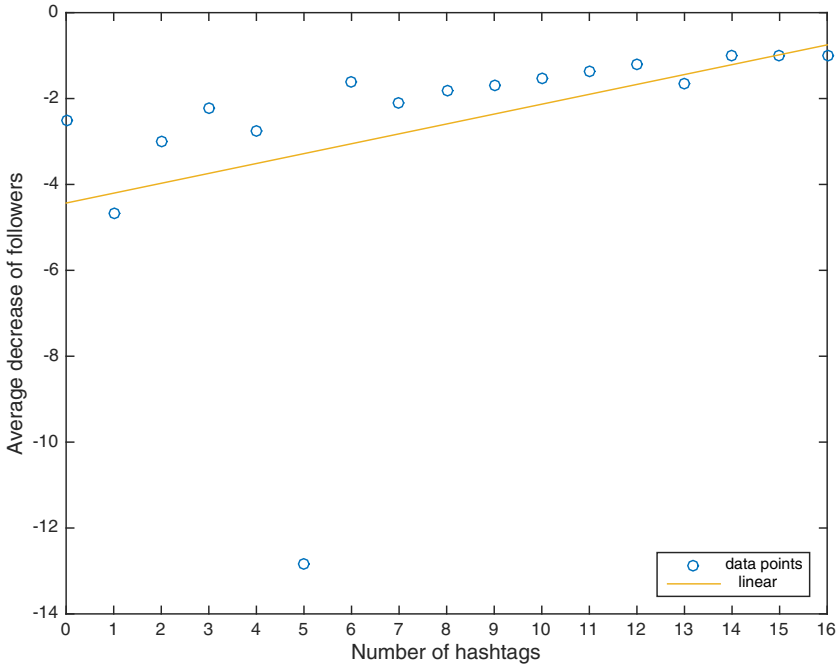

Fig. 7 Average decrease of followers against the number of hashtags. The average decrease of followers of users tweeting with a certain (x) number of hashtags. On average, there is no significant relationship between the number of hashtags and the decrease of followers

could be that almost all tweets were from the same community, therefore not being generalizable to the complete Twitter population. Figure 8 represents those users that tweeted a reply or a mention to another user in their tweets and their connection. The size of the node is proportional to the number of connections to that users. All users with less connections that four were removed from the graph to make it more understandable. What can be extracted is that there are two main groups of users tweeting to a specific user. Those two groups are represented by the blue and the red node. We expect these nodes to be famous users, being that the reason why they receive so many tweets to them. After analyzing them in the database, we discovered that the blue node is the Twitter account from the music band *5 Seconds of Summer* (@5SOS) and that the red node corresponds to one of the members of that band (@ashton5-sos). Apart from these nodes, there is not an apparent connection between the rest of the users.

A final discovery extracted from the data is that tweets that contain hashtags contain also more URLs than tweets without hashtags. This matches with the claim from Hutto et al. (2013), where they found a correlation between URLs and follower growth. Based on this relationship, hashtags and URLs could also be correlated, being this a potential study that we propose as future work.

In summary, the main conclusions extracted from these analyses are the following:

- Users that tweet with hashtags have a tendency to increase more their number of followers than users that tweet without hashtags.
- There is a relationship between the number of hashtags and the increase of followers.
- There is a relationship between the number of friends and the number of followers.
- There is a relationship between hashtags and URL usage.

6 Conclusions

The goal of this study was to determine whether the addition of hashtags to tweets produces new followers. For this reason, we performed a natural experiment in which we gathered random users that tweeted with and without hashtags for a period of 7 days, obtaining a total of 502, 891 users. The next step was to compute the difference in the number of followers in 1 h slots. Since the data were not normally distributed, we performed a non-parametric test to find out whether the increase of followers was significantly different for users tweeting with and without hashtags. The results showed that users tweeting with hashtags have a significant higher increase in the number of followers than users tweeting without hashtags.

Moreover, we extracted several conclusions after performing an analysis on the data. We showed that, on average, users increased more their number of followers during the first 12 min after they tweeted. At the same time, users tweeting with hashtags did not increase their number of followers after the minute 48, the 4th update. These numbers indicate the possible lifespan of a tweet, having its

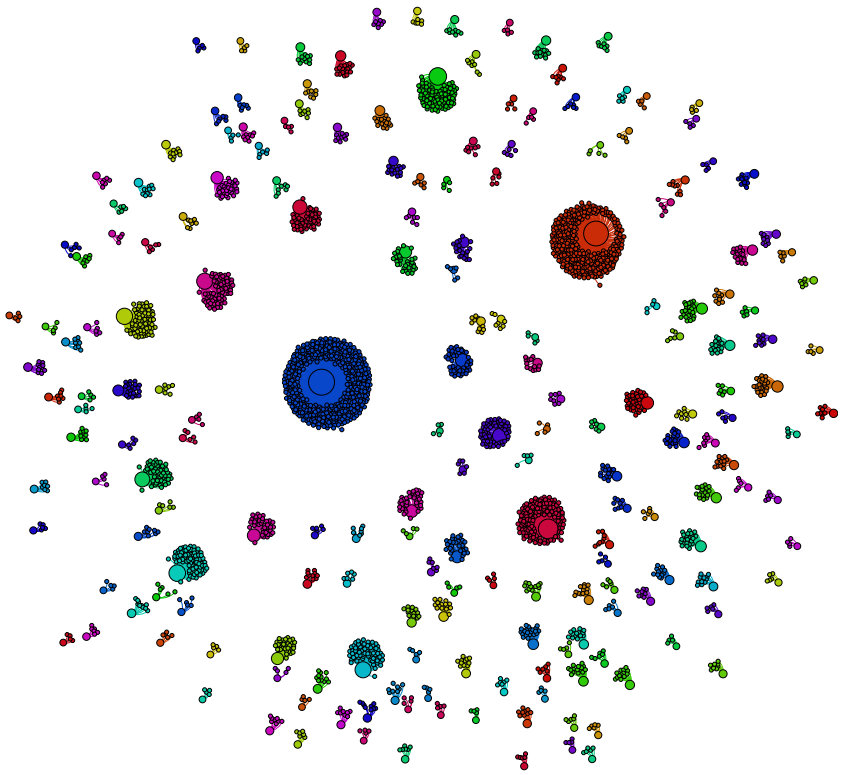

Fig. 8 Users connectivity. This network represents the connection between users tweets. Each connection represents a tweet that was mentioning or replying to another user. The node's size is proportional to the number of connections to such node

peak during the first 12–18 min, as explained in Sect. 5. Therefore, if companies want to target specific clients, they should be aware that the visibility of their tweets will significantly decrease after the first 12–18 min. In addition, by analyzing the connection between the number of followers and friends, we also discovered that apart from the standard average user, there are two more types of users. The first type could be a celebrity or a famous person, since they have a lot of followers but they do not follow that many users. The second type are users with a linear relationship between friends and followers, having these users a similar number of friends and followers. Lastly, we also discovered the possible relationship between the number of hashtags and the increase of followers. On average, the increase of followers decreases when the number of hashtags increases. Therefore, if a company wants to efficiently target their clients, the data suggest that they should use two or less hashtags in their tweets.

We believe that the presented discoveries give a better understanding of users behavior inside Twitter by portraying correlations between certain features. Companies could benefit from these results by building more efficient models to target clients. Thus, they can tweet about campaigns with the knowledge that tweeting with hashtags and the number of hashtags do matter for their impact in Twitter. Finally, several suggestions on how to continue this study are presented in the following and final section.

7 Future work

First of all, we suggest that an interesting work could be made to discover which hashtags attract new followers and which do not. Right now we know that hashtags and increase of followers are correlated, but we do not know precisely the type of hashtags that are responsible for this phenomena. For that reason, one option could be to apply machine learning techniques to group hashtags into different types, and discover if there exists specific type of hashtags that produces an increase of followers. Moreover, another option could be to apply natural language processing techniques in order to morphologically analyze each hashtag.

As a second future investigation, we can use the findings of this study to create a predictor model that is able to predict follower formation depending on the tweets and users characteristics. Lastly, as was mentioned in Sect. 5, we could make an experiment with different users to test for a possible correlation between hashtags and URLs in tweets within the message content.

References

Altman EI (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J Finance 23(4):589–609

Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512
Bifet A, Frank E (2010) Sentiment knowledge discovery in twitter streaming data. In: Discovery science. Springer, Berlin, pp 1–15
Bray P (2012) When is my tweet's prime of life? A brief statistical interlude. http://moz.com/blog/when-is-my-tweets-prime-of-life
Cha M, Haddadi H, Benevenuto F, Gummadi KP (2010) Measuring user influence in twitter: the million follower fallacy. In: 4th international AAAI conference on weblogs and social media (ICWSM), vol 14, p 8
Cochran WG (2007) Sampling techniques. Wiley, New York
Diakopoulos NA, Shamma DA (2010) Characterizing debate performance via aggregated twitter sentiment. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 1195–1198
Domingos P, Richardson M (2001) Mining the network value of customers. In: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 57–66
Given LM (2008) Qualitative research methods, vol 2. Sage, Chennai
Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. CS224N project report, Stanford, pp 1–12
Huberman B, Romero D, Wu F (2008) Social networks that matter: Twitter under the microscope. Available at SSRN 1313405
Hutto C, Yardi S, Gilbert E (2013) A longitudinal study of follow predictors on twitter. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 821–830
Jeong H, Néda Z, Barabási AL (2003) Measuring preferential attachment in evolving networks. EPL Europhys Lett 61(4):567
Jungselius B, Hilman T, Weilenmann A (2014) Fishing for followers: using hashtags as like bait in social media. Selected papers of internet
Katz E, Lazarsfeld PF (1955) Personal influence. In: The part played by people in the flow of mass communications. Transaction Publishers, Piscataway
Kivran-Swaine F, Naaman M (2011) Network properties and social sharing of emotions in social awareness streams. In: Proceedings of the ACM 2011 conference on computer supported cooperative work. ACM, New York, pp 379–382
Kivran-Swaine F, Govindan P, Naaman M (2011) The impact of network structure on breaking ties in online social networks: unfollowing on twitter. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 1101–1104
Kolmogorov AN (1933) Sulla determinazione empirica di una legge di distribuzione. Giornale dellIstituto Italiano degli Attuari 4(1):83–91
Kong S, Mei Q, Feng L, Zhao Z (2014) Real-time predicting bursting hashtags on twitter. In: Web-age information management. Springer, Berlin, pp 268–271
Kumar R, Novak J, Tomkins A (2010) Structure and evolution of online social networks. In: Link mining: models, algorithms, and applications. Springer, Berlin, pp 337–357
Lang J, Wu SF (2011) Anti-preferential attachment: if i follow you, will you follow me? In: Privacy, security, risk and trust (passat). In: 2011 IEEE third international conference on social computing (socialcom). IEEE, New York, pp 339–346
Lardinois F (2009) The short lifespan of a tweet: retweets only happen within the first hour. Read Write Web (September 2009). Accessed 20 Febr 2012. http://wwwreadwritewebcom/archives/the-short-lifespan-of-a-tweet-retweets-only-happenphp
Makice K (2009) Twitter API: up and running. Oreilly & Associates Incorporated, CA
Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 18(1):50–60
Maruf HA, Mahmud J, Ali ME (2014) Can hashtags bear the testimony of personality? Predicting personality from hashtag use
Mathioudakis M, Koudas N (2010) Twittermonitor: trend detection over the twitter stream. In: Proceedings of the 2010 ACM SIGMOD international conference on management of data. ACM, New York, pp 1155–1158
Mislove AE (2009) Online social networks: measurement, analysis, and applications to distributed information systems. ProQuest, Ann Arbor
Moore DS, McCabe GP (2011) Introduction to the practice of statistics. AMC 10:12
Newman ME (2001) Clustering and preferential attachment in growing networks. Phys Rev E 64(2):025102
Nia R, Erlandsson F, Johnson H, Wu SF (2013) Leveraging social interactions to suggest friends. In: 2013 IEEE 33rd international conference on distributed computing systems workshops (ICDCSW). IEEE, New York, pp 386–391
Otsuka E, Wallace SA, Chiu D (2014) Design and evaluation of a twitter hashtag recommendation system. In: Proceedings of the 18th international database engineering & applications symposium. ACM, New York, IDEAS '14, pp 330–333
Pak A, Paroubek P (2010) Twitter as a corpus for sentiment analysis and opinion mining. In: LREC
Peirce CS, Hartshorne C, Weiss P (1935) Collected papers of charles sanders peirce, vol 5. Harvard University Press, Massachusetts
Qiu L, Rui H, Whinston A (2011) A twitter-based prediction market: social network approach. In: ICIS 2011 proceedings
Quercia D, Ellis J, Capra L, Crowcroft J (2011) In the mood for being influential on twitter. In: Privacy, security, risk and trust (PASSAT) and 2011 IEEE third international conference on social computing (SocialCom). IEEE, New York, pp 307–314
Ritterman J, Osborne M, Klein E (2009) Using prediction markets and twitter to predict a swine flu pandemic. In: 1st international workshop on mining social media
Rogers EM (2010) Diffusion of innovations. Simon and Schuster, New York
Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World Wide Web. ACM, New York, pp 851–860
Shadish WR, Cook TD, Campbell DT (2002) Experimental and quasi-experimental designs for generalized causal inference. Wadsworth Cengage learning, Belmont
She J, Chen L (2014) Tomoha: topic model-based hashtag recommendation on twitter. In: Proceedings of the companion publication of the 23rd international conference on World wide web companion. International World Wide Web Conferences Steering Committee, Quebec, pp 371–372
Sheskin DJ (2003) Handbook of parametric and nonparametric statistical procedures. CRC Press, Boca Raton
Starnes DS, Yates D, Moore D (2010) The practice of statistics. Macmillan, London
Suh B, Hong L, Pirolli P, Chi EH (2010) Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network. In: 2010 IEEE second international conference on social computing (SocialCom). IEEE, New York, pp 177–184
Terdiman D (2012) Report: Twitter hits half a billion tweets a day. http://news.cnet.com/8301-1023_3-57541566-93/report-twitter-hits-half-a-billion-tweets-a-day/
Thelwall M, Buckley K, Paltoglou G (2011) Sentiment in twitter events. J Am Soc Inf Sci Technol 62(2):406–418

Wang T, Wang KC, Erlandsson F, Wu SF, Faris R (2013) The influence of feedback with different opinions on continued user participation in online newsgroups. In: 2013 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE, New York, pp 388–395

Wang X, Wei F, Liu X, Zhou M, Zhang M (2011) Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In: Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, New York, pp 1031–1040

Wang Y, Qu J, Liu J, Chen J, Huang Y (2014) What to tag your microblog: hashtag recommendation based on topic analysis and collaborative filtering. In: Web technologies and applications. Springer, Berlin, pp 610–618

Yu J, Shen Y (2014) Evolutionary personalized hashtag recommendation. In: Web-age information management. Springer, Berlin, pp 34–37