# Why Do I Retweet It? An Information Propagation Model for Microblogs

Fabio Pezzoni[1], Jisun An[2], Andrea Passarella[1],
Jon Crowcroft[2], and Marco Conti[1]

[1] CNR-IIT, via G. Moruzzi, 1 - 56124 Pisa, Italy
[2] Computer Laboratory, University of Cambridge, UK
f.pezzoni@iit.cnr.it,jisun.an@cl.cam.ac.uk,
a.passarella@iit.cnr.it,jon.crowcroft@cl.cam.ac.uk
m.conti@iit.cnr.it

**Abstract.** Microblogging platforms are Web 2.0 services that represent a suitable environment for studying how information is propagated in social networks and how users can become influential. In this work we analyse the impact of the network features and of the users' behaviour on the information diffusion. Our analysis highlights a strong relation between the level of visibility of a message in the flow of information seen by a user and the probability that the user further disseminates the message. In addition, we also highlight the existence of other latent factors that impact on the dissemination probability, correlated with the properties of the user that generated the message. Considering these results we define an information propagation model that generates information cascades (i.e. flows of messages propagated from user to user) whose statistical properties match empirical observations.

## 1 Introduction

Online Social Networks (OSNs) have become one of the most popular services in the Web 2.0. They allow people to communicate and share content with each other, playing a fundamental role for the spread on information, ideas, and influence. In recent years, the study of the information diffusion in OSNs have attracted the attention of many researchers. A better characterisation of the phenomenon, in fact, can lead to more effective and fair use of these systems, suggest focused marketing strategies and provide insights into the underlying sociology. The properties of information diffusion (i.e. how information spreads in the social network due to communication between users) have been studied in different types of OSNs such as microblogging platforms like Twitter [1–3] and Facebook [4] and other specific Web 2.0 services, e.g. Flickr [5], blogs [6], Digg [7] and YouTube [8].

The main goal of this paper is contributing to the characterisation of the information diffusion in microblogs, analysing the role of the users' activity. For this reason we define an agent-based model to reproduce the behaviour of the users, such that the impact of the various parameters on information diffusion

can be studied "in vitro". For example, one of the most important factors for the formation of information cascades is the *decaying visibility* of the content. In fact, different studies have demonstrated that the probability that a user forwards a received content decreases with time [1, 9, 10]. We believe that, for a better characterisation of the content visibility, it can not be measured only in terms of time and that the users' activity patterns should be considered too.

Focusing on Twitter, a more straightforward way for estimating the visibility of a *tweet* is considering its *position* in the tweet feed that is the result of the global users' activity. In fact, as empirically demonstrated in Sect. 2.2, the tweet's position in the feed is strongly correlated with its probability to be retweeted giving rise to *information cascades*. In addition to the position of a tweet in the user's feed, we also show that other parameters related to the user that originally generates a tweet can impact on the diffusion of information in Twitter. We collectively represent them with a unique parameter, that we call *user standing*. These properties are the base for the agent-based model we describe in Sect 3. In the model, agents simulate the users' activity in creating new messages and forwarding previously received messages. Basing on an underlying network structure, messages are dispatched to the connected agents and, based on their position in the tweet feed and the standing of the originating agent, they are probabilistically forwarded, simulating the formation of information cascades. In Sect. 5 we evaluate our model (simulating the user activity) in a network whose parameters are derived from a Twitter dataset (Sect. 4). Simulation results match empirical observations with high statistical confidence both in terms of information cascade properties and characterisation of the user influence.

## 2   Dataset Analysis

In this section we analyse the properties of the information diffusion as a function of Twitter users' properties using a dataset we have collected. This analysis highlights key features that determine information cascades, and it is thus the starting point for the agent-based propagation model we propose in Sect. 3.

For our analysis, we collected Twitter data from 17 October 2012 to 11 February 2013 using the Twitter REST API. Using the crawling agent described in [11] (where we also present an analysis, orthogonal to this study, of structural properties of the Twitter social network) we extract a Twitter subgraph of $2,029,143$ users. For each of them we downloaded his profile, the lists of his *followers* (people who follow the user) and *followings* (people followed by the user), and all his published tweets up to the limit of $3,200$ tweets (the maximum number of tweets that can be downloaded using the REST API). In total our dataset contains around $2,500M$ tweets that we divided in *"regular" tweets* ($63.2\%$), *replies* ($19.9\%$) and *retweets* ($16.9\%$). As replies have not an active role in the propagation of information, in our analysis we consider just "regular" tweets and their retweets.

### 2.1   Influence in Twitter

The influence can be defined as the ability of a user to spread information in a network. In Twitter, the propagation of a message can be measured in terms of *retweet count*, that is the number of times the message has been retweeted and that is included in the metadata of each downloaded tweet. Using this information we can define the influence of a user in Twitter as the average retweet count of all tweets he created. Figure 1 displays the Complementary Cumulative Distribution Functions (CCDFs) of the retweet count and of the user influence by the solid and dotted lines respectively. These results are inline with other analysis in literature that have shown that the size of information cascades and the user influence tend to be highly skewed [1, 2, 12].

Starting from the measure of influence, we can now examine what factors are related to it using our dataset. Literature says that the structural feature that best correlates with the user influence is the number of followers [2, 10, 13] that corresponds to the in-degree of the nodes in the underlying network topology. The reason behind is that a tweet from a user with many followers reaches immediately a large audience that, possibly, will retweet it to other users. In Fig. 2 we show the log-log plot of the number of followers against the user influence. The correlation (Pearson coefficient equal to 0.532) is remarkable, however, given the same number of followers, the influence value can vary significantly. In fact, as previously mentioned, structural features of the nodes alone are not sufficient to explain the actual influence of a user in the network. Others factors should be investigated.
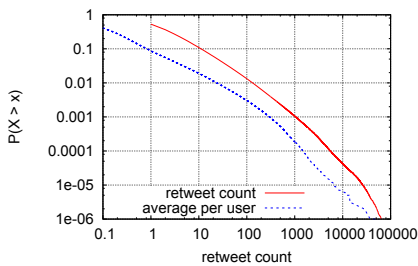


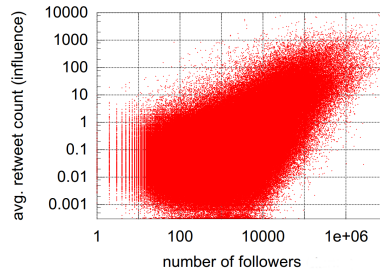Fig. 1: CCDFs of retweet count and average retweet count per user

Fig. 2: Relation between # of followers and influence.

### 2.2   Factors on Retweeting Behaviour

When a Twitter user accesses his tweet feed there are different factors that impact on his behaviour leading him to select a message to retweet. We perform our study by assuming that two main factors impact on the detailed retweeting behaviour of the users: the position of tweets in the feed, and an overall parameter describing all the properties of the creator of the tweet, that we call *user standing*.

**Position in the Tweet Feed.** Previous studies have inferred that visibility of the tweets is related to their probability to be retweeted [1, 9, 10]. A tweet has the maximum visibility immediately after it is received because it takes the least effort to be discovered at the top of the tweet feed. As soon as new tweets arrive in the feed, they push the old messages down in the queue reducing their visibility. We believe that the time span after receiving a tweet is a good estimator of its visibility however, it can be influenced by other factors like the temporal activity patterns of the users.

A more straightforward approach, is to analyse the actual position of the messages in the tweet feed. For this analysis we randomly selected a subset of 100, 000 users from our dataset. Then for each user we have recreated his message feed joining all the published tweets of the users he follows. Successively, comparing the timestamps, we have extracted for each retweeted message its position in the tweet feed at the time of the retweet. In our analysis we have considered only the first 1, 000 positions of the feed. Results in Fig. 3 show that the probability of retweetting a message in a certain position of the feed follows a power-law distribution with coefficient 1.433 estimated using the maximum-likelihood estimation (MLE).

It is worth noting that the position of the messages in a tweet feed is pretty much random, since it depends only on the time a user receives the messages and on the time he retweets. The relation between the position and the retweet probability, therefore, does not explain the variation on the user influence discussed at the end of Sect. 2.1. Visibility is, in fact, a general property of the tweets and doesn't depend on the influence or on the number of followers of the users.
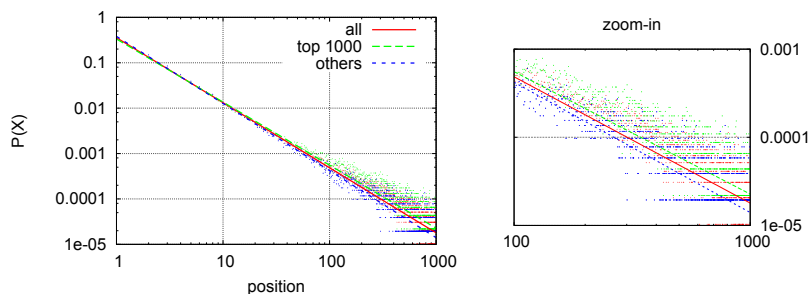


Fig. 3: Retweet probability given the position in the feed for all the tweets in the dataset ("all"), for the tweets created by the 1, 000 most influential users ("top 1000") and for the tweets created by all the other users ("others").

**User Standing.** In order to explain mentioned variations in the user influence we have to investigate the effect of the properties of the users on the retweeting behaviour. These properties are often qualitative and, therefore, hard to quantify (e.g. credibility, expertise, enthusiasm and popularity). For this reason, we use

a unique index called *user standing*, to take into consideration the joint effect of all of them.

The effect of the user standing can be observed as the variation of the retweet probability for different equally-positioned tweets. In this sense, the user standing can be considered as a sort of "favouritism in retweet" for the messages created by some users. In our case, we are interested in investigating if the tweets created by the most influential users are more likely to be retweeted than the tweets created by other users. For the analysis we ranked the users considered in the previous analysis based on their influence and then we selected the top $1,000$ influential users. In Fig. 3 we plot the retweet probability of their tweets compared with the retweet probability of tweets created by all the other users. The gap between the lines appears narrow, however the fit with a power-law function has coefficient 1.389 for the influential users and 1.478 for the others. This means that, considering the same position, the most influential users have a higher probability to get their messages retweeted.

## 3    Activity-Based Propagation Model

The model we present in this work describes the information propagation mechanism in a microblogging social network given the topology of the network and some features of the agents that represent the users. In the model any agent interacts with the network in two different ways: creating new messages and forwarding previously received messages. The frequency with which an agent $v$ is selected for creating and forwarding messages, is given by the parameters $f_v^{\mathrm{cr}}$ and $f_v^{\mathrm{fw}}$ respectively. Both in case of creation and forwarding, the messages are broadcast to other agents that "follow" the creator or forwarder. An agent $r$ follows the agent $v$ if, in the underlying network graph $G(V, E)$, a direct link between the nodes that represent agents $r$ and $v$ respectively exists. In this case the agent $r$ receives all the messages created or forwarded by agent $v$. If an agent receives multiple copies of the same message, it keeps in memory just the first received one and discharges the others.

Assuming that an agent $v$ is selected to perform a forwarding action at time $t$, the model takes the *message feed* list $F_{v,t}$ that includes all the messages received by $v$ before time $t$ sorted by reverse-chronological order. Then, for each message $w \in F_{v,t}$, it assigns the probability $P(w|v, t)$ to be forwarded such that $\sum_{w \in F_{v,t}} P(w|F_{v,t}) = 1$ where:

$$P(w|v, t) = \frac{\alpha_{o(w)} \varphi(\theta_{v,t}(w))}{\sum_{z \in F_{v,t}} \alpha_{o(z)} \varphi(\theta_{v,t}(z))}, \qquad w \in F_{v,t} \tag{1}$$

$\alpha_{o(w)}$ is the *standing* of the the agent $o(w)$, who is the creator of the message $w$, and $\varphi(\cdot)$ is a function called *position function* that takes as a parameter the position of $w$ in $F_{v,t}$ denoted as $\theta_{v,t}(w)$. According to (1), the probability of a message to be selected for the forward depends on: i) its position in the message feed and ii) the standing of its creator.

  i) The position of the message in the feed is considered in the model since, as we demonstrated in Sect. 2.2, there is evidence that last received messages (which are on top of the message feed) are more likely to be forwarded. For this reason the position function $\varphi(\cdot)$ has to be monotonically decreasing. For example, as our analysis suggests, it can be defined as a power-law function.

  ii) As discussed in Sect. 2.1, we introduced the concept of user standing that represents the joint effect of all the properties of the users that positively influence the forwarding probability of their messages. Each agent in the network $v$ is therefore characterised, in addition to the frequencies $f_v^{\mathrm{cr}}$ and $f_v^{\mathrm{fw}}$, also by a standing value $\alpha_v$. In the next section we discuss in detail how to model the user standing.

## 4  Deriving the Model's Parameters

In our simulation we implement the agent-based propagation model described in previous section in order to simulate the user activity and the information diffusion of a real social network. We used the Twitter dataset described in Sect. 2 to infer both the graph structure and the agents' properties.

### 4.1  Social graph

For computational reasons we selected a random subset of $100,000$ users among all the active users from our dataset. We considered a user to be active if he has at least 100 followers and if he has created at least 100 tweets. These constraints allow us to avoid low-active accounts that are not relevant for the propagation of information. From this set of users, we derived the social graph that consists of $5,756,450$ arcs and maintains well-known features of social networks' graphs such as high clustering coefficient and small average path length (small-world property) [14].

### 4.2  Position Function

As suggested in Sect. 3 we define the position function $\varphi(\cdot)$ as a power-law. In particular we use the result in Sect. 2.2 in which we have fit the retweet probability given the tweets' position with a power-law with coefficient 1.433. Considering that the position function is discrete, we define it as a ZipF Probability Mass Function with the given coefficient and limited to $N = 1,000$, which is the same number of positions we have used in our analysis.

### 4.3  Frequencies

For each user $v$ we extract, from the dataset, the frequency of creating messages per day $f_v^{\mathrm{cr}}$ and the frequency of forwarding messages per day $f_v^{\mathrm{fw}}$. Distributions of these frequencies are highly skewed since just few users have a very high activity.

### 4.4 User Standing

In Sect. 2.2, we defined the user standing as the joint effect of the latent factors that affect the forwarding of his messages. As previously discussed, these parameters of the model are not directly quantifiable. We could estimate them using a MLE estimator where the likelihood function is given by a sample of retweeting actions extracted from the dataset. Unfortunately, applying this method would have required to analyse the full propagation path of each and every tweet of all our users, which was not feasible due to the computational complexity and the fact that cascades can involve users not included in our dataset. Therefore, we use an approximate way to estimate the user standing, as follows.

The idea is to estimate the standing of a user as the average retweet probability of the tweets he has originated. This can be calculated as the ratio of his average retweet count (influence) to the average number of users who have received his tweets. However, the latter value is not derivable since it would require to track the full propagation trees. As approximation, we use the number of his followers instead. It is worth noting that, due to this approximation, the standing of the most influential users could be overvalued. This is because the number of followers can be significantly smaller than the number of users that received the tweets. In order to remove this bias we had to apply an exponent to the previously defined measure. As result of an extensive analysis, we set the exponent to $1/3$ as this value guarantees to obtain better performance of our model. Formally, the user standing values we considered in our simulation are defined as:

$$\alpha_r = \left( \frac{\sum_{w \in W_r} \pi(w)}{|W_r| \cdot k(r)} \right)^{1/3} \tag{2}$$

where $w$ is a message, $W_r$ is the set of messages created by user $r$, $\pi(w)$ is the number of times the message $w$ has been forwarded and $k(r)$ is the number of followers of the node $r$.

## 5 Simulations

Using the social graph and the user parameters described in Sect. 4, we simulated a period of 30 days of user activity. We run 10 independent simulations in order to calculate the 95% confidence intervals which are shown as error bars in the figures and between square brackets in the tables and in numerical data. The simulations produced an average of $24,026,886$ $[\pm 292]$ user interactions in that 77.1% $(18,515,225$ $[\pm 1,092])$ are related to the creation of new messages and the rest are forwarding messages. These proportions are consistent with those related to the dataset in Sect. 2 (excluding reply tweets). Among all created messages, 14.3% of them $(2,649,709$ $[\pm 1,128])$ have been forwarded originating cascades. In Fig. 4 we show the histogram of the depth of the cascades produced. As we can see, the trend is logarithmically decreasing with respect to the frequency. In fact, 78.7% of the forwarded messages are not propagated beyond the first

level of followers. This trend is exactly the same shown in several analysis in literature [1, 2].

As discussed in Sect. 2.1, we define, for each node $r$ in the simulations, the influence $\gamma_r$ as the average retweet count of the tweets $r$ has originated. In Fig. 5 we show the CCDFs of the number of forwards for each message as the solid line and the nodes' influence as the dashed line. Comparing these results with those in Fig. 1, we can see that the simulations replicated the presence of a small number of influential users located in the tail of the distribution[1].
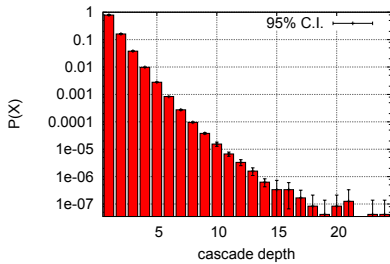


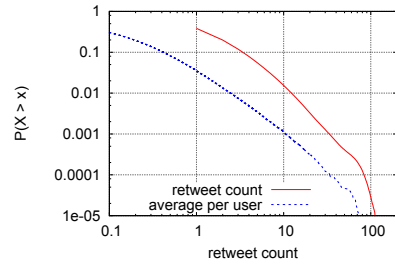Fig. 4: Cascade depth distribution.



Fig. 5: CCDFs of forwardings per message and user influence.

In the column "orig" of Table 1, we summarise the results of the simulations (upper part) and the correlation of the resulting influence with other variables (lower part). In the table we refer to the the vector of the nodes' influence as $\gamma$ while we use the symbol $k$ for the vector of the number of followers and $\alpha$ for the vector of the users' standing. Correlation values demonstrate that our model is able to replicate high correlation between the influence and both the number of followers and the user standing [2]. We also calculate the correlation between the simulated user influence and the influence $\gamma^*$ of the selected users in the dataset described in Sect. 2. Considering that the influence from the dataset refers to the actual influence of the users in the Twitter network and that in our simulations we consider just a small subset of this network, the correlation value is remarkable and proves the ability of our model to simulate the actual user influence distribution.

### 5.1   Message Positioning and User Standing Impact

In order to study the impact of the message positioning in our model we run 10 simulations with the same setting described in Sect. 4, excluding the position function $\varphi(\cdot)$ from the model. Results are shown in the column "no-pos" of

---

[1] Direct comparison between the two plots is not possible, due to the large difference of the number of users in the dataset and in the simulations.

[2] Note that, while in (2) the standing is clearly a function of the influence, the values of the user standing have been computed based on the information propagation in the dataset, while influence is measured based on the simulations' results.

Table 1: Summary of results. Column "orig" refers to the results obtained using the original model; columns "no-pos" and "no-sta" refers to the results obtained without considering the position function and the user standing respectively.

| | orig | no-pos | no-sta |
|---|---|---|---|
| max cascade depth | 19.0 [±2.1] | 121.6 [±6.0] | 10.0 [±0.5] |
| max msg forwards | 257.7 [±37.2] | 10,347.4 [±304.2] | 155.9 [±7.0] |
| max user influence | 79.1 [±0.6] | 1,436.0 [±187.6] | 131.6 [±5.2] |
| | [id:41801] | [id:98020] | [id:2019] |
| corr($\gamma$,$k$) | 0.544 [±0.010] | 0.100 [±0.009] | 0.646 [±0.003] |
| corr($\gamma$,$\alpha$) | 0.101 [±0.003] | 0.073 [±0.004] | 0.014 [±0.001] |
| corr($\gamma$,$\gamma^*$) | 0.595 [±0.003] | 0.126 [±0.011]] | 0.443 [±0.003] |

Table 1. The main consequence of such change is that some messages flood the entire network and some users become extremely influential. This indicates that the decreasing visibility of the messages in social networks is fundamental for limiting the size of the information cascades.

We also studied the impact of the user standing, running 10 simulations where we have excluded it from the probability of forwarding. In this case the main change in results, shown in column "no-sta" of Table 1, is an increase in the correlation between the number of followers and a decrease in the correlation between the influence and the standing values.

In both "no-pos" and "no-sta" cases, it is noticeable the sensible decrease of the correlation between the simulated influence and the actual influence registered in the our dataset. This demonstrates the importance of considering both parameters in our model.

## 6    Conclusions

In this work we analysed the properties of the information diffusion in Twitter, in particular the impact of the structural features of the users and their retweeting behaviour. Using a Twitter dataset we studied the relation between the probability of a message to be retweeted and its position in the tweet feed and we concluded that this relation is described by a power-law function. We also characterised the most influential users in the network discovering that, although their ability of spreading messages is mostly given by their large number of followers, other factors have to be considered. These factors, joint effect we called user standing, have effect at the forwarding behaviour level, scaling the retweet probability given by the position of the message.

Based on these observations we proposed an agent-based information propagation model able to generate cascades whose properties match empirical observations. Agents simulate the activity of the users in a network creating and forwarding messages independently. Received messages are organised in an ordered list for reproducing the effect of the position on the forward probability.

Through simulations, we show that our model is able to reproduce information cascades statistically similar those presented in the literature and that the generated user influence is strongly correlated with the actual influence measured in the dataset. These results demonstrated that our model can thus be used to realistically study how the user activity and the forwarding mechanism influence the propagation of information.

## Acknowledgements

## References

1. Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z., Kellerer, W.: Outtweeting the twitterers-predicting information cascades in microblogs. In: Proceedings of the 3rd conference on Online social networks, USENIX Association (2010) 3–3
2. Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J.: Everyone's an influencer: quantifying influence on twitter. In: Proceedings of the fourth ACM international conference on Web search and data mining, ACM (2011) 65–74
3. Ye, S., Wu, S.F.: Measuring message propagation and social influence on twitter.com. In: Social informatics, Springer (2010) 216–231
4. Sun, E., Rosenn, I., Marlow, C., Lento, T.M.: Gesundheit! modeling contagion through facebook news feed. In: ICWSM. (2009)
5. Cha, M., Mislove, A., Adams, B., Gummadi, K.P.: Characterizing social cascades in flickr. In: Proceedings of the first workshop on Online social networks, ACM (2008) 13–18
6. Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., Hurst, M.: Cascading behavior in large blog graphs. arXiv preprint arXiv:0704.2803 (2007)
7. Szabo, G., Huberman, B.A.: Predicting the popularity of online content. Communications of the ACM **53**(8) (2010) 80–88
8. Susarla, A., Oh, J.H., Tan, Y.: Social networks and the diffusion of user-generated content: Evidence from youtube. Information Systems Research **23**(1) (2012) 23–41
9. Oken Hodas, N., Lerman, K.: How visibility and divided attention constrain social contagion. (2012)
10. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web, ACM (2010) 591–600
11. Arnaboldi, V., Conti, M., Passarella, A., Pezzoni, F.: Ego networks in twitter: an experimental analysis. In: The Fifth IEEE International Workshop on Network Science for Communication Networks (NetSciCom 2013)
12. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P.K.: Measuring user influence in twitter: The million follower fallacy. ICWSM **10** (2010) 10–17
13. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: Social Computing (SocialCom), 2010 IEEE Second International Conference on, IEEE (2010) 177–184
14. Newman, M.E.: The structure and function of complex networks. SIAM review **45**(2) (2003) 167–256