

# How Far Does a Tweet Travel? Information Brokers in the Twitterverse

Diederik van Liere  
University of Toronto

Rotman School of Management  
105 St. George Street

diederik.vanliere@rotman.utoronto.ca

## ABSTRACT

In this paper, I present evidence on the geographic diffusion patterns of information of Twitter users. I identify three possible information diffusion patterns: random, local and information brokerage and show that the information brokerage pattern describes best how users of Twitter diffuse information through the act of retweeting.

## Categories and Subject Descriptors

J.4 [Computer Applications]: Social and Behavioral Sciences – sociology

## General Terms

Algorithms, Measurement.

## Keywords

Twitter, information diffusion, geography, influence, ranking, social networks, social media filter.

## 1. INTRODUCTION

Twitter is a microblogging Internet service that allows its users to share information and engage in conversations [7, 8]. In contrast to off-line conversations, the information shared is public<sup>1</sup>. Everybody can observe who is communicating with whom and what the exact topic of the conversation is.

Twitter makes conversations a public good, which creates new research opportunities for studying information diffusion as the act of information rebroadcasting becomes observable. Each time a person uses Twitter to share information, a tweet is broadcasted containing the information, date and time, sender and optionally the receiver and exact location. People will retweet some of the tweets they receive indicating that the information has utility to some people.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee

MSM'10, June 13, 2010 Toronto (ON, Canada)

Copyright © 2010 ACM 978-1-4503-0229-6/10/06 ...\$10.00.

Retweeting a tweet can be thought of as a vote for the information, just as a hyperlink can be thought of as a vote for a webpage [2].

In this paper, I will highlight three alternative geographic information diffusion patterns and using data from Twitter I show empirical evidence that the dominant information diffusion pattern is the one where people act as bridges between different geographic regions.

## 2. ALTERNATIVE GEOGRAPHIC INFORMATION DIFFUSION SCENARIOS

I define the *geographic distance* traveled by a tweet as the distance (in meters) between the sender of a tweet and the receiver who retweets that message. Retweeting is the act of rebroadcasting a tweet received from someone else and the 'RT' anchor in the tweet can identify such tweets. I define the *geographic information diffusion pattern* as the distribution of geographic distances between sender and the receiver who retweets a message. By looking at the geographic information diffusion pattern, we can induce the reason(s) why people use Twitter.

In this section, I will introduce three alternative geographic information diffusion patterns. The first information diffusion pattern, as a baseline, is the *random pattern*. The geographic distance between a sender (the person who tweets) and receiver (the person who is following the sender) is uniform distributed with minimum distance zero meters and maximum distance half of the earth's circumference, which equals to 20,037,510 meters.

Obviously, people are more likely to follow people who are close in terms of geographic proximity than to follow random people across the globe. Hence, the second geographic information diffusion pattern is the *local pattern*. This pattern is characterized by being left skewed, the average geographic distance between sender and receiver is short and long distances are negligible.

The local pattern is based on the assumption that Twitter conversations are mainly between people who are friends in the off-line world. The third pattern is based on following people with shared interests and not necessarily following friends, I will refer to this geographic information diffusion pattern as the *information broker pattern*. This pattern is right skewed, with most of the observations at the right hand of the distribution.

## 3. DATA COLLECTION

In this section I describe how I collected the data and how I calculated the distance a tweet travels

<sup>1</sup> With the notable exception of 'direct messages' which are private conversations between two persons.

First, for a period of 12 hours, I collected all tweets containing the RT identifier. boyd et al. [1] discuss the difficulties in recognizing a genuine retweet as there are multiple syntaxes to indicate that a tweet is a retweet. I use the RT moniker as this is the most often used syntax. Capital RT usually followed by the username signifies that this tweet was retweeted. For each retweeted tweet I went through the following sequences of steps:

- 1) Determine the topic being tweeted by stripping off Twitter usernames, URL's and hashtags. To identify all retweets of a particular tweet I used the following procedure:
  - I only include retweets that contain either the 'RT' or 'via' indication.
  - I only include retweets that contain a link to a website but I exclude social networks sites such as MySpace and Facebook as I cannot retrieve the actual page without logging in.
  - I visit each page mentioned in the (re)tweet and extract the title of the story.
- 2) Do an exact search on Twitter for the topic from step 1.
- 3) The results from step 2 are all the tweets and retweets about the topic identified in step 1. For each result, I download the user profile and store their location information if available.
- 4) I geocode the location and time zone information, if available, to longitude and latitude coordinates using the geonames.org website. Most people provide their location at the city level, although increasingly people are giving their latitude and longitude coordinates using their cell phones as well.
- 5) To create variance in distance for people from the same city, I randomly select latitude and longitude coordinates for that city. For example, if many of the Twitter users are from San Francisco then I randomly select one of the longitude – latitude pairs for San Francisco. This will assure that the distance travelled by a tweet is larger than zero meters<sup>2</sup>.
- 6) Using the haversine formula [11], I calculate the distance between two pairs of longitude and latitude coordinates. The haversine formula is defined as:
 
$$haversindR=haversin(\varphi1-\varphi2)+\cos\varphi1 \cos\varphi2haversin\Delta\Phi$$
 where  $d$  is the distance in kilometers between the two points,  $R$  is the radius of the earth (6371 kilometers) and  $\varphi1=latitude \text{ point } 1$ ,  $\varphi2=latitude \text{ point } 2$  and  $\Delta\Phi=\text{delta longitude}$ .
- 7) I excluded all users who have set their city to Tehran / Iran and Quito to prevent bias in the results. Many people using Twitter reset their location to Tehran during the demonstrations in June 2009 as there were indications / rumors that the Iranian government started tracking people from Iran who were using Twitter. To obfuscate their attempts, many people changed their location to Tehran and they might not have reset their location to their true location

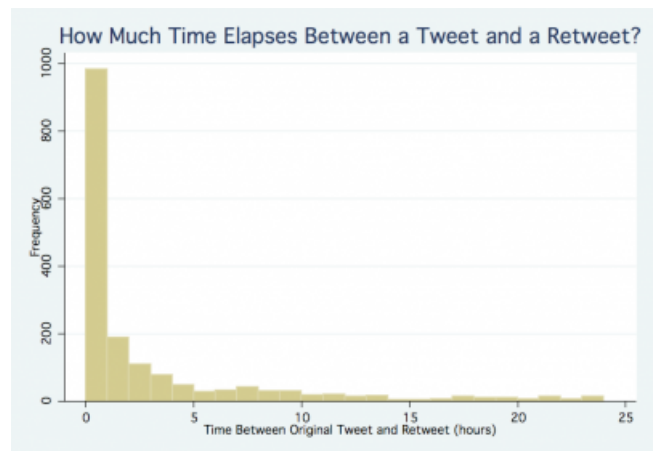
<sup>2</sup> Future research should use the extra information from geotagged tweets as those tweets contain exact longitude and latitude information from where they were sent.

[10]. In addition, for a while the default time zone in Twitter's user registration was set to Quito, Ecuador, and not all users set it to their true time zone.

#### 4. INFORMATION BROKERS AND INFLUENCE

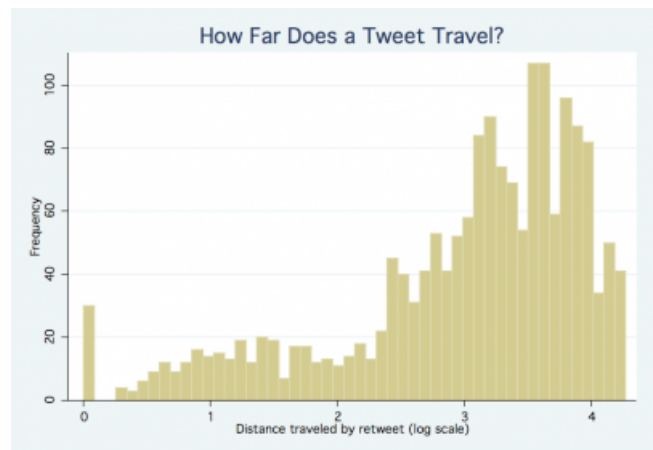
Of the 10,228 users that I collected, I was able to geocode 6,424 users. I assume that the missing geo-coordinate data is randomly distributed among Twitter users. This assumption means that regardless from the country where the person lives, the probability of sharing location information is the same. Whether this assumption is reasonable is a topic for future research.

My database consists of 13,399 (re)tweets comprising 285 original stories. I am able to create 1,758 dyads where I know both the geo-coordinates of the original sender and the person who retweeted it.



**Figure 1 Histogram of Time Elapsed Between Original Tweet and Retweet**

The first chart shows the histogram of elapsed time between the original tweet and the retweets. More than 60% of all retweets happen within the first hour, after that the probability of a tweet being retweeted quickly fades to almost zero and in the limited sample, nothing is retweeted after 24 hours.



**Figure 2 Histogram of the Distance Traveled by a Tweet**

The second chart shows the histogram of distance between the original sender and the person who retweets the story. To improve

the readability of the histogram, I take the log of the distance. An average retweet travels  $10^{2.98} = 955$  kilometers, while the median distance is  $10^{3.23} = 1698$  kilometers. These data suggest that Twitter makes the world smaller but not hugely. On the other hand, the average and median distance are too large to speak of local communication which suggest that the information broker pattern is the most appropriate pattern for this sample.

### 4.1 The Topic Retweet Network

The results from §4 suggest that information travels beyond the local proximity of a person’s friends. If we assume that in general people’s friends live in close geographical proximity then Figure 2 shows that information retweeted reaches new groups of people. The work from Burt [3-5] is important to understand why people would broker information between different groups of people. In his view, a broker is a person who bridges two otherwise unconnected (groups of) people. The broker benefits from this disconnection in the network: acting as the bridge, the broker can infuse new information to a group, bargain for better terms due to the information asymmetry between the unconnected groups and control the flow of information. Previous research has persistently demonstrated that brokers perform well [4, 6].

I construct a retweet network for each topic from the data as follows. The nodes in the network are people who either tweeted about the topic themselves or are people who retweeted somebody else. Two people are connected when the retweet acknowledges the person who sent the tweet first.

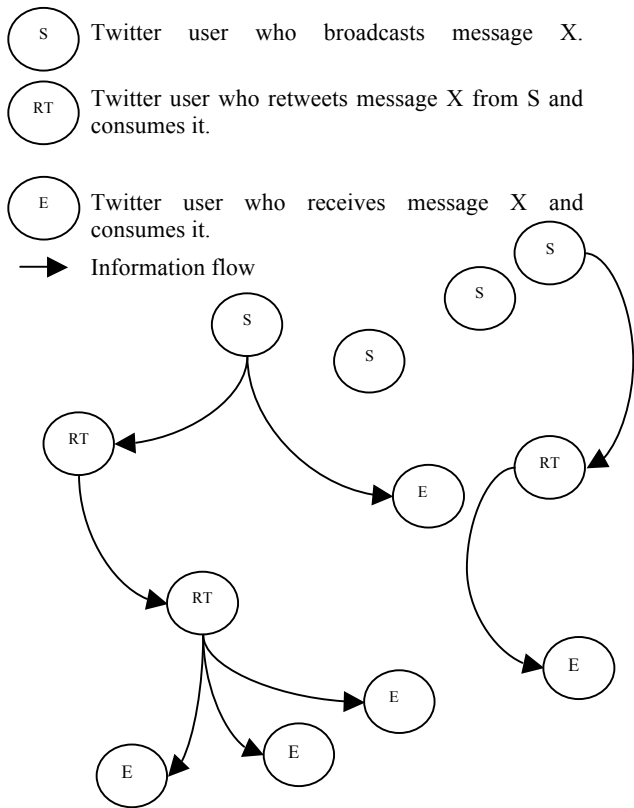


Figure 3 The Topic Retweet Network

This network can contain isolates: people did tweet about a particular topic but where never retweeted. Constructing such a

network allows us to derive a number of Twitter influence measures. Raw Twitter influence can be measured as follows:

1. Raw count of how often a message by a particular user is retweeted:

$$influence = \sum_{n=1}^n T_n$$

where  $n$  is a unique identifier for the tweet,  $T$  is 1 if tweet  $T_n$  for user  $i$  contains ‘RT’ and 0 otherwise. This measure will be biased towards Twitter users with many followers, hence we can correct for this by dividing by the number of followers.

2. Raw retweet count weighted by the number of followers:

$$weighted\ influence = \frac{\sum_{n=1}^n T_n}{i_f}$$

where  $i_f$  is the number of followers.

3. Average depth of a retweet:

$$retweet\ depth = \frac{\sum \max(d_{ij})}{\sum_{n=1}^n T_n}$$

where  $d_{ij}$  is the shortest path from user  $i$  and user  $j$  from the topic retweet network.

### 4.2 Discussion

Attention and influence are the scarce resources of social media. To overcome this attention deficit, people can choose to follow other people who act as social media filters [9] and retweet only those tweets that the social media filter thinks is valuable. Retweeting a tweet has three distinct functions. First, the act of retweeting is an attempt to vie for attention from unsubscribed Twitter users and to increase the follower count.

Second, retweeting is an attempt to gain influence by acting as an intelligent social media filter who specializes in a particular topic. The person who retweets has specifically chosen to retweet that tweet which can be seen as an endorsement of that particular piece of information. Retweeting can be thought of as a vote for the quality, novelty or timeliness of a piece of information.

Third, retweeting transfers information from one group of people to a second group of people as the results of this study shows. The person who retweets bridges distinct groups and injects new information and ideas to a new group. Retweeting within the same group is less functional as the information is continuously echoed among the same people. A random retweet connects different pockets in the Twitterverse possibly in a similar way as small world networks do: people bridging different regions of a network make the network smaller and the same seems to apply to Twitter: people who are part of different conversations re-broadcast this information to Twitter regions that have not heard the information yet. This suggests that there are opportunities for information arbitrage: people can capitalize on early access to information, re-broadcast this information and thereby developing a reputation as an expert on a particular topic.

### 4.3 Limitations

Two limitations are worth noting. First, the sample size is limited, especially when considering the volume of messages being exchanged on a daily basis. I intend to rerun the described analysis on a dataset consisting approx. 1,000,000 tweets. Second, I have not filtered out spam tweets which may bias the results. Although there seem little incentive to retweet spam, it might

happen especially when the person who retweets is promised some kind of reward. Future research should identify such spam tweets and handle them appropriately.

Concluding, conceptualizing the act of retweeting as a vote has two advantages. First, it allows for the identification of influential people. Second, it can be an effective strategy for building a social media filter that allows people to cope with the increasing volume of tweets.

## 5. REFERENCES

- [1] boyd, d., Golder, S. and Lotan, G. 2010. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. in *HICSS-43*, (Kauai, HI), IEEE.
- [2] Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30 (1-7). 107-117.
- [3] Burt, R.S. 2005. *Brokerage and Closure: an Introduction to Social Capital*. Oxford University Press, Oxford, UK.
- [4] Burt, R.S. 1992. *Structural Holes - The Social Structure of Competition*. Harvard University Press, Cambridge, MA.
- [5] Burt, R.S. 2004. Structural holes and good ideas. *American Journal of Sociology*, 110 (2). 349-399.
- [6] Burt, R.S. 2000. The network structure of social capital. in *Research in Organizational Behavior*, JAI-Elsevier Science, New York, 345-423.
- [7] Java, A., Song, X., Finin, T. and Tseng, B. 2007. Why we Twitter: Understanding Microblogging Usage and Communities. in *9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, (San Jose, CA).
- [8] Krishnamurthy, B., Gill, P. and Arlitt, M. 2008. A few chirps about Twitter. in *First Workshop on Online Social Networks*, (Seattle, WA).
- [9] Lerman, K. and Jones, L. 2006. Social Browsing on Flickr.
- [10] Morozov, E. 2009. Iran Elections: A Twitter Revolution? *The Washington Post*, The Washington Post Company, Washington.
- [11] Sinnott, R.W. 1984. Virtues of the Haversine. *Sky and Telescope*, 68 (2). 159.