Inform or Flood: Estimating When Retweets Duplicate

Amit Tiroshi[1], Tsvi Kuflik[1], and Shlomo Berkovsky[2]

[1] University of Haifa, Haifa, Israel
{atiroshi,tsvikak}@is.haifa.ac.il
[2] NICTA, Sydney, Australia
Shlomo.Berkovsky@nicta.com.au

Abstract. The social graphs of Twitter users often overlap, such that retweets may cause duplicate posts is a user's incoming stream of tweets. Hence, it is important for the retweets to strike the balance between sharing information and flooding the recipients with redundant tweets. In this work, we present an exploratory analysis that assesses the degree of duplication caused by a set of real retweets. The results of the analysis show that although the overall duplication is not severe, high degree of duplication is caused by tweets of users with a small number of followers, which are retweeted by users with a small number of followers. We discuss the limitations of this work and propose several enhancements that we intend to pursue in the future.

1 Introduction

The graphs of social network users represent the links established between the users. The entire graph can be decomposed into *ego-graphs* [9], representing the perspective of a single user on the network and containing only the links between the user and other users. In Twitter, user links are established through the 'Follow' feature, such that users have a set of users whom they follow (dubbed as *followees*) and a set of users who follow them (dubbed as *followers*). Previous works have shown that Twitter graph is a small-world network, i.e., most users can be reached within a small number of network hops [7]. As such, the degree of overlap between the ego-graphs of two users who established a link between them is high, and it increases over time, as users establish more links and their ego-graphs expand. This phenomenon is explained by the observed homophily of users: people often have mutual acquaintances and connect to like-minded people with similar interests [10].

Two popular ways of public communication in Twitter are to *tweet*, i.e., post new tweets to followers, and to *retweet*, i.e., re-post tweets from followees to followers. No official statistics of the number of followers of an average Twitter user are available. However, a 2009 data based on 56 million accounts, shows an average of 557 followers [1]. A 2012 dataset based on 80 million accounts that tweeted at least once, shows an average of 235 followers [2]. The expansion of the ego-graphs and the abundance of tweets/retweets may pose a significant information overload on the users. Furthermore, the high overlap between the ego-graphs of two linked users can potentially lead to a duplication of tweets reaching a user, as they may receive a tweet as

S. Carberry et al. (Eds.): UMAP 2013, LNCS 7899, pp. 267–273, 2013.

well as the retweets of the same tweet through a number of users. For example, consider three users – Alice, Bob, and Carol – such that Bob follows Alice, and Carol follows both Alice and Bob. In this setting, Alice's tweets will be duplicated in Carol's incoming stream, if retweeted by Bob. This may aggravate the information overload problem and make it even harder for users to identify tweets of interest and stay informed.

While the reasons for tweeting have been thoroughly studied [7,11], to the best of our knowledge the reasons for retweeting have received little attention so far. In [4], Boyd et al. survey the key motivating factors for retweeting, which can be split into three groups. The first includes *informative* factors, such as the desire to spread a tweet to followers because it matches their interests, will entertain them, or will make them aware of the tweet's topic. The second groups refers to *emotional* factors, such as endorsing the opinion expressed in the tweet, appealing to the user who posted the tweet, or trying to gain benefit from a tweet that might become popular. Finally, the last group includes *utility* factors, which virtually come to bookmark tweets of relevance. It should be noted that many emotional and utility factors can be fulfilled by other Twitter features, e.g., the 'Favorite' and the 'Reply' features.

As in many information sharing scenarios, retweets ought to strike the balance between the information need and the information overload (an example of how to aim for that balance in social news feeds is given at [3]). Although the value of the emotional and utility factors can hardly be quantified, the degree of duplication (as a proxy for information overload) caused by a retweet can be assessed and it can potentially affect the value of the informative motivating factor. For example, users may refrain from retweeting a tweet that has already been received by most of their followers or, conversely, be urged to retweet a tweet that a few followers have received so far. However, at the moment, Twitter users have no means for assessing the redundancy caused by their retweets.

Aiming to develop these means in the future, we present here an exploratory analysis of the degree of duplication created by more than 1000 real-life retweets. We computed the duplication caused by these tweets and found that, overall, 20% of Tweets caused duplicate posts for 20% or more of the recipients. The duplication depends on the number of followers of both the user who posted the original tweet and the user who retweeted it. We also discuss several interfaces that can communicate the duplication to users and potentially affect their retweeting decision.

2 Related Work

The problem of overlaps in social graphs and their effect on information propagation was examined by Boyd et al. [4], who focused on the retweeting phenomenon. It was noted that when Twitter users retweet posts, there may be an overlap between their followers and the followers of the user who posted the original tweet, but the retweeters are unlikely to be aware of this overlap. They also refer to a note made in [6] that in small-worlds, where people connect to each other seamlessly, the effort required for keeping tracks of who knows whom (or who follows whom) is immense.

Other works focused on various aspects of retweets, but provided possible explanations for the presence of overlaps in social graphs. For example, [10] evaluated several recency-, content-, and homophily-based computational models of retweeting. It was found that models that take user homophily into account, fit the observed retweeting behavior better than others. Besides being one of the key drivers for retweeting, homophily is also pivotal for establishing the follower/followee links [8,12]. In combination, these findings provide a strong evidence that user homophily may lead to a potential duplication of tweets. However, there is little evidence for the impact of the social graph overlaps on the duplication, which is the focus of this work.

3 Analysis of Overlapping Retweets

In this section we present an analysis of the overlap caused by post retweets on Twitter. Let us denote by *O* the user who posted the original tweet and by *R* the user who reposted the tweet. We consider the ego-graphs of Twitter users and denote by *fr(u)* and *fe(u)* the set of followers and followees of user *u*, respectively. Finally, we denote by |*s*| the cardinality of a set *s*. Given this notation, we quantify the degree of duplication caused by *R* retweeting a tweet posted by *O* as:

$$OL(O,R) = \frac{|fr(O) \cap fr(R)|}{|fr(R)|}$$

Note that the intersection of the two sets of followers is divided by the cardinality of the set of followers of *R*, in order to stress the duplications caused by *R*'s retweets.

In order to assess the overlap caused by retweets in the wild, we gathered in September 2012 a set of 1030 real-life retweets, as well as the ego-graphs of *O* and *R* for each. The data contains tweets posted by 1000 and retweeted by 1029 unique users. High-level statistics of the sets of followers[1] and their overlaps are shown in Table 1. Also, Figure 1-left shows the distribution of the *OL(O,R)* values across the collected retweets. Although most retweets have a low overlap, it can be seen that for 20% of them (205 retweets) there is an overlap of 20% or more between the followers of *O* and *R*. The distribution of overlap frequencies fits the long tail distribution.

Table 1. Followers and overlap statistics

	|fr(O)|	*|fr(R)|*	*OL(O,R)*
Minimum	4	0	0%
25th percentile	378	90	1%
Median	2,346	201	6%
Mean	210,492	986	11%
75th percentile	52,902	489.75	16%
Maximum	9,175,388	206,535	100%

[1] The difference between the mean and median of |*fr(O)*| is due to 60 retweets for posts of users who had more than a million followers each.

Since the number of followers distributes normally neither for *O* nor for *R*, we drill down to analyze how *OL(O,R)* varies across different users. For this, we split all the collected retweets into two equal-size bins of 515 retweets each, according to |*fr(O)*| and |*fr(R)*|. For the *O*-split we sort the collected retweets according to the number of *O*'s followers, such that 515 retweets where |*fr(O)*|≤2342 are considered as tweets of users with a low number of followers and the other 515 retweets where |*fr(O)*|>2342 are of users with a high number of followers. We repeat the same process for the *R*-split: 515 retweets where |*fr(R)*|≤115 are mapped to the bin of retweeters with a low number of followers and the other 515 to the bin with a high number of followers. Note that the cut-off point of the *O*-split is much greater than that of the *R*-split. This is due to the unbalanced distribution of retweets, which are often done to influential users and VIPs, and are less frequent for those using Twitter for everyday chat [4].

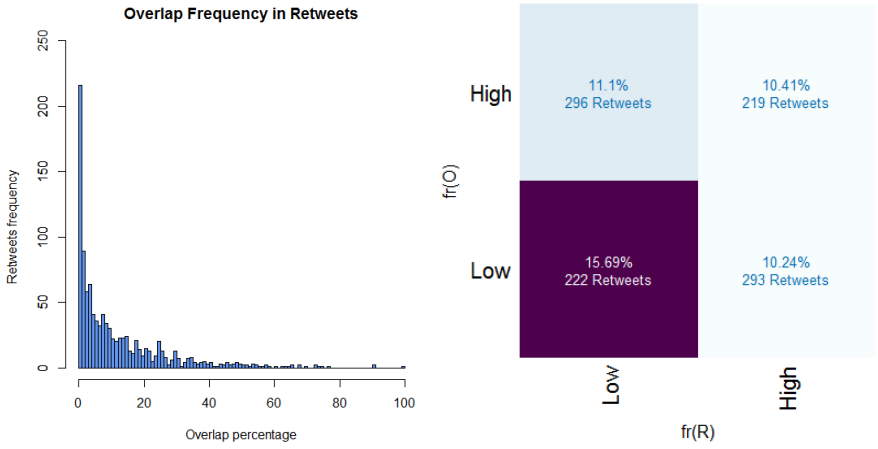

Fig. 1. Left – overall frequency of *OL(O,R)* and right – heat map of *OL(O,R)* for the combinations of *O*- and *R*-split

Figure 2 plots the cumulative distribution function (CDF) of *OL(O,R)* obtained for the 'low' and 'high' bins for the *O*-split (left) and *R*-split (right). We observe that most retweets have a small number of users who receive duplicate tweets and only a small number of retweets causes duplication for a high portion of recipients. For the *O*-split, about 16% of retweets of users in the 'high' bin cause duplication for 20% of users of more, whereas in the 'low' bin this ratio stands at about 25%. This gap between the two CDFs functions is lower for the *R*-split; about 18% of retweets in the 'high' bin cause duplication for 20% of users of more, whereas in the 'low' bin this ratio is about 22%. However, in both splits we observe the same trend: the duplication caused by retweets in the 'Low' bin exceeds the one in the 'High' bin. That is, retweets of users having a small number of followers cause more duplication than retweets of users having a large number of followers. This finding can be explained by the higher homophily observed for small communities [10]. Indeed, the larger a community of users is, the harder it is to maintain a high degree of similarity of users across the community. In case of followers, one intuitive argument supporting this is that for

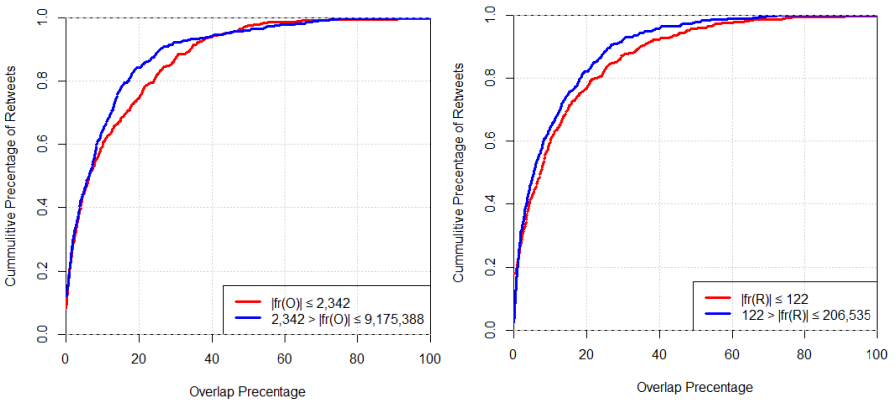

Fig. 2. CDF of *OL(O,R)* for a low/high number of followers; left – *O* and right – *R*

users with a large number of followers, these followers would normally fragment into more topics of interests (or sub-groups); hence, the overlap will decrease.

To better understand the combined dependency between |*fr(O)*|, |*fr(R)*|, and *OL(O,R)*, Figure 1-right shows the heat map of overlaps obtained for the various combinations of the number of followers of *O* and *R*. Every segment of the heat map shows the number of retweets matching this combination and their average *OL(O,R)*, whereas the background color of the segment communicates the degree of overlapping and duplication. The highest *OL(O,R)* is obtained for the segment with low |*fr(O)*| and low |*fr(R)*|. That is, retweets of users with a small number of followers done for tweets or users with a small number of followers cause the highest duplication. A decrease in *OL(O,R)* is observed when either |*fr(O)*| or |*fr(R)*| increases. The impact of the decrease in |*fr(R)*| on *OL(O,R)* is slightly higher than that of the decrease in |*fr(O)*| and the observed overlap is lower.

4 Discussion

In this work we conducted an exploratory analysis of the information overload and duplication caused by retweets. We gathered a corpus of retweets and measured the degree of overlapping between the set of followers of a tweet originator and of the user who retweeted it. We discovered that the overlapping and duplication decrease as the number of followers in user ego-graphs increases. On the first look, the observed degree of duplication was not high: overall, around 20% of retweets caused duplicate posts for 20% or more of the recepients. While not looking severe, this rate may increase in small communities with dense graphs. For example, consider a group of followers of an academic conference Twitter account. Many users in this group may also follow each other, such that retweets of the conference announcements may be duplicated many times. Hence, the work into the discovery of these duplicate retweets is needed and timely.

We would like to highlight two limitations of this work. The first refers to the fact that our approximation of information overload through tweet duplication addresses only the informative factor of retweeting. Indeed, many users consider also the emotional and utility factors when retweeting, and may retweet despite the information

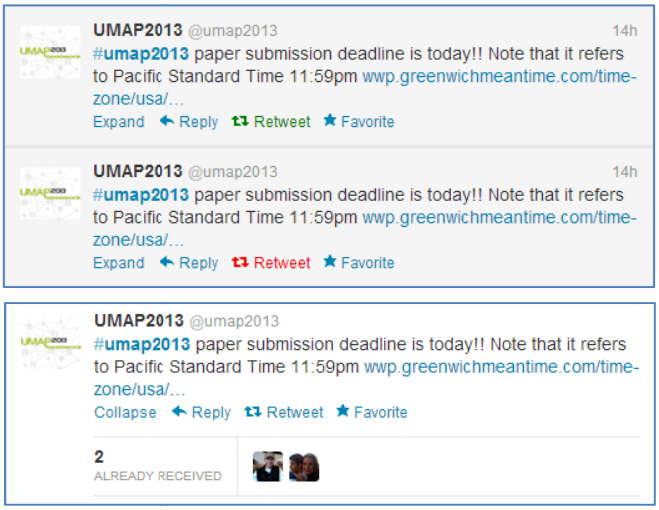

Fig. 3. Two mock-up interfaces to alert tweet duplication: top – traffic light rating; bottom – thumbnails of the recipients of duplicate tweets

overload they cause. Our overlapping computation cannot account for these factors and needs to incorporate techniques similar to the tie-strength model [5]. The second limitation refers to the small scale of the analysis. The size of the gathered data was constrained primarily by the REST API that prolongs the time needed to gather Twitter ego-graphs, especially when the set of followees and followers is large. There is no easy solution to this problem, but to substantially extend the duration of the data collection phase. As the density of the gathered ego-graphs resembles the density of those gathered in other works [1,2], we posit that our results reflect the results that could have been obtained in a larger-scale analysis.

Another issue that needs to be addressed is how to alert users to the potential overload of their retweets. We demonstrate two practical implementations. The first (see Figure 3–top) uses the traffic light rating system to color the 'Retweet' button according to the computed degree of duplication. This, however, is not a straightforward threshold-based coloring, but should rather consider the followers of the retweeters, the density of their ego-graph, and the nature of the followers (consider general interest users vs. professionals). Similarly, a different treatment should be given to different users. For example, a user may agree to cause duplicate tweets for some users and be reluctant to do this for others. Hence, we propose another implementation that shows the thumbnails of users, who will receive duplicate tweets due to the retweet (see Figure 3–bottom). The retweeter can then easily identify who will be affected by the retweet and take more informed retweeting decisions.

In the future we plan to conduct a user study with a cohort of users and a large corpus of tweets. We will expose the users to different alert visualizations and measure the persuasive impact of the alerts on their retweeting. In the study we will also be able to evaluate several personalized ways to compute the information overload,

which will take into account the structure of the followers' ego-graphs as well as the strength of the links between the retweeters and their followers. This study is necessary to ascertain the uptake of the retweeting decision support tool by real users.

Acknowledgement. This work is supported by ISF grant 226/2010.

References

1. 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 (WSDM 2011), pp. 65–74. ACM, New York (2011)
2. Basch, D.: Some Fresh Twitter Stats (as of July 2012, Dataset Included), http://diego-basch.com, http://diegobasch.com/some-fresh-twitter-stats-as-of-july-2012 (accessed July 31, 2012)
3. Berkovsky, S., Freyne, J., Kimani, S., Smith, G.: Selecting items of relevance in social network feeds. In: Konstan, J.A., Conejo, R., Marzo, J.L., Oliver, N. (eds.) UMAP 2011. LNCS, vol. 6787, pp. 329–334. Springer, Heidelberg (2011)
4. Boyd, D., Golder, S., Lotan, G.: Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In: Proceedings of the 2010 43rd Hawaii International Conference on System Sciences (HICSS 2010), pp. 1–10. IEEE Computer Society, Washington, DC (2010)
5. Gilbert, E., Karahalios, K.: Predicting tie strength with social media. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2009), pp. 211–220. ACM, New York (2009)
6. Granovetter, M.: Ignorance, knowledge, and outcomes in a small world. Science 301(5634), 773–774 (2003)
7. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microb-logging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (WebKDD/SNA-KDD 2007), pp. 56–65. ACM, New York (2007)
8. 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 (WWW 2010), pp. 591–600. ACM, New York (2010)
9. Lacaze, A., Moscovitz, Y., DeClaris, N., Murphy, K.: Path planning for autonomous vehi-cles driving over rough terrain. In: Proceedings of Intelligent Control (ISIC), 1998. Held jointly with IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA), Intelligent Systems and Semiotics (ISAS), September 14-17, pp. 50–55 (1998)
10. Macskassy, S., Michelson, M.: Why do People Retweet? Anti-Homophily Wins the Day! In: International AAAI Conference on Weblogs and Social Media, North America (July 2011)
11. Marwick, A., Boyd, D.: Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience. New Media and Society 13, 96–113 (2011)
12. Weng, J., Lim, E.P., Jiang, J., He, Q.: TwitterRank: finding topic-sensitive influ-ential twitterers. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM 2010), pp. 261–270. ACM, New York (2010)
13. Wilson, C., Boe, B., Sala, A., Puttaswamy, K.P.N., Zhao, B.Y.: User interactions in social networks and their implications. In: Proceedings of the 4th ACM European Conference on Computer Systems (EuroSys 2009), pp. 205–218. ACM, New York (2009)