High Correlation between Incoming and Outgoing Activity: A Distinctive Property of Online Social Networks?

Diego Saez-Trumper, David Nettleton
Universitat Pompeu Fabra
Barcelona, Spain
{*diego.saez,david.nettleton*}*@upf.edu*

Ricardo Baeza-Yates
Yahoo! Research
Barcelona, Spain
rbaeza@acm.org

Abstract

User influence is an important topic of research for online social networks. Recent work has shown that a users influence is not directly related with node in-degree. However, the definition of what is an influential or relevant user is still an open subject. In general, we can say that an influential user has the ability to produce incoming activity in an interaction graph. For this reason we have focused our attention on the user's incoming activity and the search for which factors are related with this indicator. We have studied a Facebook dataset and have found that outgoing activity is highly correlated with incoming activity. This characteristic is valid not only for users with a low level of activity, but it is also valid for high activity users. This result hasn't been reported before and appears to be an important factor of user behavior. To contrast this finding, we have compared it against a popular e-mail dataset and a Twitter dataset. The result was that we found a similar behavior for the Twitter Dataset, but for the e-mail dataset the correlation was lower. Hence, we conjecture that the high correlation may be a distinctive property of social networks. In our future work, we propose to extend this study to other social network platforms.

Introduction

The importance of a user in an online social network (OSN) has been widely studied using different approaches. Recent studies have shown that the node degree (for example the number of followers in Twitter) is not correlated with user influence (Cha et al. 2010; Kwak et al. 2010). However, several different ways exist to define what is a relevant or influential user. To mention some examples, Goyal *et al*. (Goyal, Bonchi, and Lakshmanan 2008) define "leaders" (influential users) as being the first to tag some given urls in del.icio.us which were subsequently tagged by other users. Goyal also measured the same behavior with respect to the rating of Songs in Yahoo! Music. Other definitions of "influential users" have been made in studies about Twitter, for example: an influential or relevant user is one who obtains more *Retweets* [1] (Cha et al. 2010). In this case *retweets* is the metric for relevance. In order to use a more specific concept than influence or relevance, we focused on a user's incoming activity. We represent the social network as a graph where the node degree represents the number of contacts that the user has, for example, the number of friends in Facebook, the number of followers in Twitter or number of contacts in an e-mail network. The activity is represented by the weight of the graph's edges. Incoming activity can easy be defined as actions that a given user receives from other users.

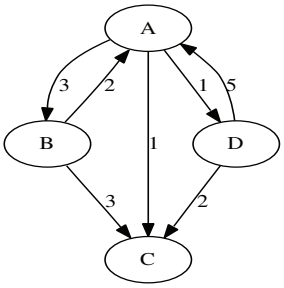

Figure 1: Activity example: user A has 4 friends, with an outgoing activity of 5, incoming activity of 7, and total activity equal to 12. User C has incoming activity but no outgoing activity, and therefore we do not consider user C as an active user.

For example, in an e-mail network, incoming activity is defined as a received mail. In our current work we studied an Email Network and a Facebook Wall Posts Network. A Wall post is one way of communicating in Facebook where one Friend (contact in Facebook) can write a public message to another person. If user A posts a message in user B's wall we consider that User A has a Wall post Done and User B has a Wall Post Received. So, we studied which factors are related with incoming activity in both these networks. Our results confirm previous work which showed that the number of friends is not highly correlated with incoming activity. We also found that in the Facebook Network, a user's outgoing activity is highly correlated with his/her incoming activity. In the e-mails network the correlation was lower, but still remained over 0.5 when we considered all the active users. These results suggest that one important factor which influences becoming a relevant user is to generate outgoing

[1] Retweets is when a second user copies a post from the first user.

activity. We also extended this analysis to a Twitter sample. As mentioned before, previous work showed that there is very little correlation between Followers and Mentions, therefore we decided to take another approach considering the number of updates (micro-post) as an outgoing activity, and trying with two different kinds of incoming activity: the number of replies and the number of followers. Our results show that the group of users with most updates usually has more replies and followers than users with a low level of outgoing activity. However, these results are preliminary and need to be tested with other datasets, adding new and more complex parameters that consider, for example, temporal analysis.

Datasets

We have used three different datasets, summarizing their characteristics in Table 1, as follows:

Enron's Emails Enron's e-mails dataset it is a well known dataset. From the version that we obtained (Shetty and Adibi 2005) we processed information from 250,483 users.We consider each unique e-mail address as an user. Because we are interested in active users we filtered considering only users that have at least one e-mail sent and one e-mail received. Applying this filter we obtained 11,254 active users.

Facebook Wall Posts The dataset used corresponds to New Orleans Facebook's Regional Network[2] (Viswanath et al. 2009) containing information about 58016 different users, who have been anonymized. Specifically we have two lists, one containing user-to-user links *(Friends)*, and a second list with user-to-user *Wall Posts*, most of them with a third column containing a Timestamp. The information covers a time-span from September 2006 to January 2009. From these lists we can obtain the number of friends for each user and his/her outgoing and incoming activity. Because we are interested in users interaction we only consider users that have at least one Wall Post done and one Wall Post Received, defined as "active users". Applying this filter we obtained 34,277 active users.

Twitter In Twitter every user has a unique ID. Users information can be accessed by the Twitter API (Benevenuto et al. 2010) using this ID. We randomly selected 250,000 numbers between 0 an 150,000,000 being able to download approximately 55% of these ids, corresponding to 136,662 users. The information about users contains, among other things, their number of Tweets (statuses), followers, friends (also called followees), profile age (date of profile creation) and if there exists a URL associated to the profile. We also downloaded the last 200 tweets from each user.

Our dataset is consistent with similar and more detailed studies of Twitter (Kwak et al. 2010), which have demonstrated power law distributions for user's followers and friends. The time of the profile creation (profile's age) covers June 2006 to July 2010, the number of followers ranges from 0 to over 600,000 and the number of tweets from 0 to over 300,000 per user.

[2]Regional Networks has been deprecated by Facebook since August 2009

Dataset	Users	Activity
Enron	11,254	1,277,214 emails
Facebook	34,277	836,576 wall posts
Twitter	136,662	54,764,095 tweets

Table 1: Summary of datasets used.

Experiments

We are interested in finding factors related with incoming activity. Related work shows that a relation does not exist between node degree in the social graph (friends, followers, etc) and the incoming activity (Cha et al. 2010) (wall posts received, mentions, etc). For this reason, we have focused our attention on the relation between incoming activity and outgoing activity, the latter in our case, corresponding to "wall posts done". In other words, we compare the out-degree and in-degree of nodes, taking into consideration the weight of the edge. In the case of the Facebook and E-mail datasets, we have compared the same kind of outgoing and incoming activity. That is, we compare posts done versus posts received and mails sent versus mails received, respectively. Then we apply a simple correlation parameter. In the case of Twitter, the nature of the outgoing (micro-post) and the incoming activity (followers) is different. In this case, we group users by their level of activity and then we compare them.

Facebook and Emails Analysis

The results in Table 2 confirm previous work in the literature which demonstrates that a strong correlation does not exist between the number of friends and the number of posts received. However, the correlation between posts done and posts received appears very strong (0.91) in the Facebook dataset. We recalculated the correlation coefficients against the users activity, in order to test if the correlations hold true when we remove the less active users. In Figure 2 we can observe that the correlation remains over 0.90 for 90 percent of the users (users whose activity is below 200 actions), and remain strong (over 0.70), if we consider only the most active users.

In the e-mail dataset, the correlation will depend of how we compute e-mails carbon copies "CC:"[3] for the outgoing activity. Suppose that user A sent an e-mail to user B with copies to users C and D. To compute user A outgoing activity we can consider that *(a)* she have 3 outgoing-actions, one per recipient, or we can consider *(b)* only 1 action, one per sent mail. In the first case, the correlation between outgoing and incoming activity results low (see Figure 2). In the second case, the correlation is higher, starting from 0.69 and remaining over 0.5 for the most active users (see Figure 3). However, this correlation is significant lower than in the Facebook graph.

[3]Note that we use "CC:" for e-mails copies and CC for Clustering Coefficient.

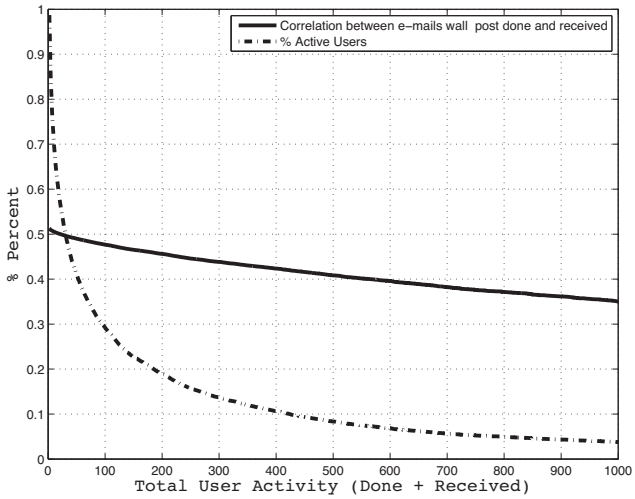

Figure 2: Enron's e-mails with Carbon Copy*(a)*: Correlation between out-going (e-mail's sent) and incoming (e-mails received) compared with users total activity. The correlation is under 0.5 if we consider the top 50 percent of most active users.

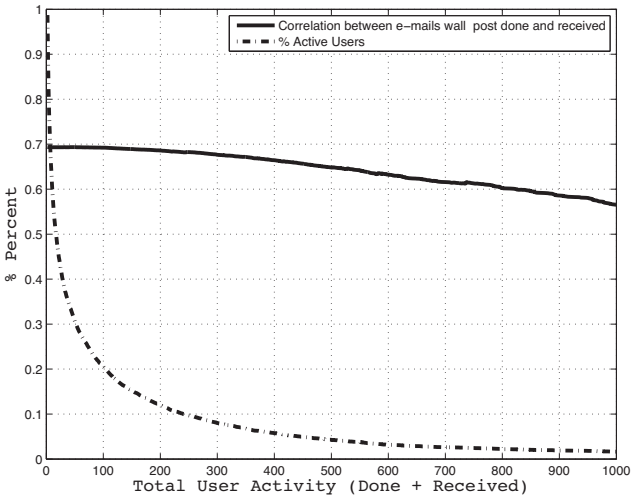

Figure 3: Enron's e-mails without Carbon Copy *(b)*: Using this configuration the correlation increase, going over 0.6. This correlation is significant, but is lower than in Facebook Dataset

Variable	Num. F.	CC.	P. Re.	P. D.
Number of Friends	1	-0.11	0.47	0.43
Clustering Coeff.	-0.11	1	-0.11	-0.11
Posts Received	0.47	-0.11	1	0.91
Posts Done	0.43	-0.11	0.91	1

Table 2: Facebook's Correlation Matrix.

Twitter Analysis

In this section, we consider two different kinds of incoming activity: first we have used the number of replies, meaning that user A gets one reply when user B posts with the prefix @A. This is a standard in Twitter, and intuitively is similar to the e-mail reply we used in previous section. However, we repeated our study considering the number of followers as an incoming activity. We have use followers as outgoing activity for three reasons: firstly, as we mentioned before, in the literature there are authors who have already studied different kinds of incoming activity such as Retweets or Mentions, secondly because this information (the number of followers and updates) is public and can be obtained directly from the Twitter statistics, avoiding any type of noise or errors, and thirdly because we consider that obtaining followers could be considered as an important goal for Twitter users, moreover, previous work has consider the relation between number of post and followers to construct influence rankings (Weng et al. 2010).

We started studying the relation between post an replies. The number of replies for each user is not contained in the information given by Twitter API. Therefore, was necessary to use the Twitter Search API to try to estimate this number. To face this challenge, we divided our sample by outgoing activity level (number of post) in five categories(where category 1 are inactive users and category 5 are the very active users) and created an stratified sub-sample, selecting randomly 2,000 users for each category. Next, we estimate the number of replies for these 10,000 users. We found that around the 80% of users does not have at least one mention, and only the 5% users have over 15 replies. Moreover, in the 3 categories with lowest outgoing activity (that is users with 300 post or less) only the 10% users have at least one reply. Users in the most active groups (category 4 and 5), over 300 and 1200 post, has 22% and 47% of users with at least one reply. However, an small fraction of users have at least 15 replies, and they are concentrated in the category of most active users (see Table 3. The small number of replies does not allow to make conclusions, but we can see that there exists a relation between outgoing activity and Twitter replies.

Outgoing Activity	Over 1 reply	Over 15 replies
Very Low	0%	0%
Low	2%	0%
Medium	7%	1%
High	19%	9%
Very High	48%	17%

Table 3: Twitter: Percent of users with replies (incoming activity)

The information about the number of followers is provided by the Twitter, then we can work with our sample of 136,662 users. We have divided our sample in 5 bins using a equal frequency discretization for each feature. In this way, each user can be defined as an instance with two features, thus: U(outgoing,incoming). Considering that we have 5 possible values for each feature, it is possible to have 25 different types of users. As an example, a user type T could be characterized by a low outgoing activity and a very high incoming activity. Figure 4 shows that when the level of out-

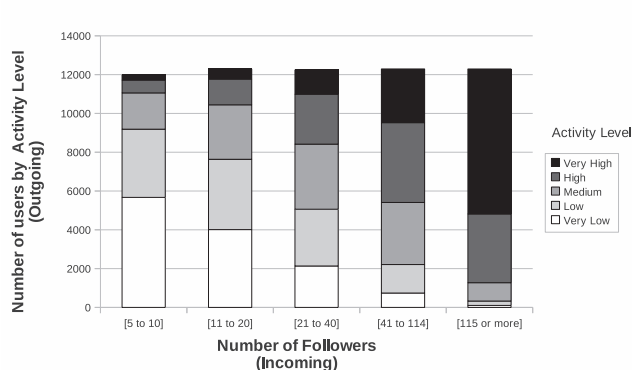

Figure 4: Twitter: Relation of Followers and Number of Posts. On the x-axis we have divided users by their number of followers (incoming activity), and on the y-axis we show the number of users with different levels of post activity (outgoing).

going activity increases, incoming activity also increases. As a consequence, users with a high number of followers have the biggest number of post. With this simple analysis we cannot infer which is the dependent variable, but is clear that a relation exists.

Discussion and Future Work

We have found a strong correlation between outgoing and incoming activity in the Facebook Dataset. Our analysis of Twitter confirms this relation. The results for the e-mail dataset shows that the correlation will depend of how outgoing activity is define. However, we have studied three different networks with different communications patterns: Facebook Wall posts has an one-to-one pattern, e-mails could be consider as one to n pattern, and Twitter has a kind of broadcast pattern, and all of them shows a this outgoing/incoming correlation . Moreover, the strong correlation persists in Facebook users with a very high outgoing activity, and is also present for users with the same characteristics in Twitter, suggesting that this relation could be a distinctive property of Online Social Networks and Social Media.

However, these conclusions are preliminary and it will be necessary to confirm them comparing with other datasets. It is also important to analyze the dynamic behavior of users, considering the time that they use to perform their actions. It is reasonable to think that the results obtained will change if a high activity is concentrated over a short time span - which could indicate spammer behavior- compared with an user whose activity is periodical and more spread over time. In our future work we will analyze this temporal behavior and its implication for incoming activity.

References

Benevenuto, F.; Magno, G.; Rodrigues, T.; and Almeida, V. 2010. Detecting spammers on twitter. In *Proceedings of the 7th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. P. 2010. Measuring user influence in twitter: The million follower fallacy. In *in ICWSM 10: Proceedings of international AAAI Conference on Weblogs and Social*.

Goyal, A.; Bonchi, F.; and Lakshmanan, L. V. S. 2008. Discovering leaders from community actions. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, 499–508. New York, NY, USA: ACM.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is Twitter, a Social Network or a News Media? In *World Wide Web Conference*. ACM Press.

Shetty, J., and Adibi, J. 2005. Discovering important nodes through graph entropy the case of enron email database. In *Proceedings of the 3rd international workshop on Link discovery*, LinkKDD '05, 74–81. New York, NY, USA: ACM.

Viswanath, B.; Mislove, A.; Cha, M.; and Gummadi, K. P. 2009. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09)*.

Weng, J.; Lim, E.-P.; Jiang, J.; and He, Q. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, 261–270. New York, NY, USA: ACM.