

The Complete Picture of the Twitter Social Graph

Maksym Gabielkov, Arnaud Legout

▶ **To cite this version:**

Maksym Gabielkov, Arnaud Legout. The Complete Picture of the Twitter Social Graph. ACM CoNEXT 2012 Student Workshop, Dec 2012, Nice, France. hal-00752934v3

HAL Id: hal-00752934

https://inria.hal.science/hal-00752934v3

Submitted on 3 Dec 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The Complete Picture of the Twitter Social Graph

Maksym Gabielkov
INRIA Sophia Antipolis – Méditerranée
Sophia Antipolis, France
maksym.gabielkov@inria.fr

Arnaud Legout
INRIA Sophia Antipolis – Méditerranée
Sophia Antipolis, France
arnaud.legout@inria.fr

ABSTRACT

In this work, we collected the entire Twitter social graph that consists of 537 million Twitter accounts connected by 23.95 billion links, and performed a preliminary analysis of the collected data. In order to collect the social graph, we implemented a distributed crawler on the PlanetLab infrastructure that collected all information in 4 months.

Our preliminary analysis already revealed some interesting properties. Whereas there are 537 million Twitter accounts, only 268 million already sent at least one tweet and no more than 54 million have been recently active. In addition, 40% of the accounts are not followed by anybody and 25% do not follow anybody. Finally, we found that the Twitter policies, but also social conventions (like the follow-back convention) have a huge impact on the structure of the Twitter social graph.

Categories and Subject Descriptors

C.2.m [**Computer-communication Networks**]: Miscellaneous

Keywords

Twitter, social networks, data mining.

1. INTRODUCTION

Twitter[1] is the most popular micro-blogging service in the world. It allows its users to exchange short messages (tweets) that are limited to 140 characters. It was created to enable people to find out what is currently happening with people and organizations they are interested in.

As a very popular information propagation system, Twitter is attracting the interest of scholars, politicians, and advertisers. Also, unlike classical social networks (e.g., Facebook), the relation between Twitter users is unidirectional, which makes information propagation in Twitter much closer to how information propagates in real life.

[1] `http://twitter.com`

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.
CoNEXT Student'12, December 10, 2012, Nice, France.
Copyright 2012 ACM 978-1-4503-1779-5/12/12 ...$15.00.

In this paper, we make two major contributions. First, we collected the entire Twitter social graph that consists of 537 million accounts connected by 23.95 billion links. For each account, we collected all public information (including user name, account creation date, number of published tweets, number of followings and followers) and the list of all followings. In order to deal with the rate limit in the number of requests made with the Twitter API that we used to collect information, we implemented a distributed crawler that we ran from 550 PlanetLab nodes. Second, we performed a preliminary analysis of the collected information (our future work is on digging deeper into the data) focusing on the correlation among the number of followers, the number of followings and the number of tweets.

To the best of our knowledge, this study is the first one to provide a complete picture of the Twitter social graph. For instance, Kwak et al. [3] have used a breadth-first search in the follower graph starting from Perez Hilton who had over a million followers at that time. Also, they have collected profiles of users who referred to popular topics in their tweets during the crawl. Therefore, their dataset has a bias towards well connected accounts and they have a partial view of the social graph. Several other studies also worked on a partial view of the social graph, [2] or focused on tweets instead of the social graph [4, 1].

This paper has the following organization. Section 2 describes the methodology of data collection. The preliminary analysis of our dataset is presented in Section 3. Finally, we conclude and present future work in Section 4.

2. METHODOLOGY

Twitter provides access to its data via a website, SMS and a Twitter Application Programming Interface (API)[2]. The information about user profiles and links between users is accessible through a REST API that we used to create our dataset. However, requests made with this API are rate-limited. Unauthenticated host can make at most 150 requests per hour with that API.

Given this limit, it would take approximately 13 years to crawl all user accounts on Twitter from one host. One way to speed up the crawl is to distribute it on several machines. We used PlanetLab[3] to deploy our crawler on 550 machines. We discovered that four PlanetLab machines have been whitelisted, two machines with a rate limit of 20,000 requests per hour and two with 100,000 requests per hour. Twitter has discontinued whitelisting new machines since

[2] `https://dev.twitter.com/`
[3] `https://www.planet-lab.org/`

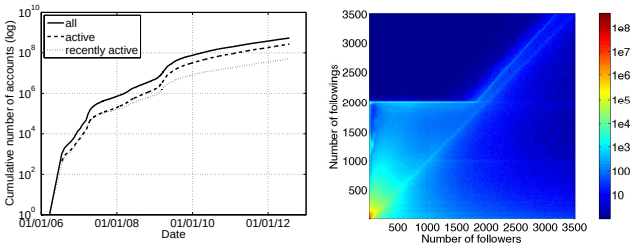


Figure 1: (left) Evolution of the number of user accounts on Twitter. (right) Number of Twitter accounts with a given number of followers and followings.

February 2011, but existing whitelisted machines can still be used.

Twitter assigns IDs for new accounts sequentially [2]. Therefore, we first determined using a random pooling that there is no ID assigned above 800 million, then we divided the range from 1 to 800 million into chunks of 10,000 IDs. We selected an upper bound (800 million) much larger than the claimed number of Twitter accounts by the time of our crawl to be sure to do not miss any account. Finally, the crawler distributed each chunk to one of the crawling machine to check each ID in the chunk for a valid account. If an ID corresponds to a valid account, all public account information[4] is retrieved along with the complete list of followings for this account. Each node reserved on PlanetLab for this crawl has a crawling script and uses an API wrapper described by M. Russell [5]. We performed our crawl from March 20, 2012 to July 24th, 2012.

3. RESULTS

We collected the public profile information for 537.5 million accounts, that is all Twitter accounts by the end date of the crawl. We show in Fig. 1 (left) the cumulative number of created Twitter accounts (based on the account creation date) with time since Twitter's creation in 2006. We first notice a huge growth in the number of created accounts almost two orders of magnitude within the last 3 years. However, even if the number of created accounts is very large (537 million), only 268 million already sent at least one tweet (*active* in Fig. 1), that is 50% of the accounts never sent any tweet. Moreover, 54 million accounts have sent at least one tweet within the week before the crawl date of each account (*recently active* in Fig. 1). Therefore, only 10% of the accounts can be considered active.

We now focus on the Twitter social graph. 94% of the users have a public account (that is all information about such accounts is public, including tweets). For each of these accounts, we collected the list of followings, which we used to build the follower graph for all public accounts. We note that we removed all links to protected accounts. The public follower graph contains 23.95 billion edges. We show in Fig. 1 (right) the number of accounts with a given number of followings and followers. We group the accounts within bins of 10 followings and 10 followers, and for each bin the color map gives the number of accounts within this bin (due to the density of bins, each bin appears like a dot in the figure instead of a square).

We notice in Fig. 1 (right) that 99.8952% of the Twitter accounts have less than 3,500 followers and followings (we do not show the 563,481 accounts outside of this range in Fig. 1 (right)). We observe three interesting properties in this figure. First, we observe that there are approximately 322 million accounts with less than 10 followers and followings, that is almost 60% of all Twitter accounts[5]. Second, we see a diagonal on the figure. It is made of users who have chosen to follow back everyone who follows them. It is a social trend to ask followed accounts to follow back. This trend is strong enough to have a noticeable impact on the overall social graph. Third, there is a horizontal line at 2,000 followings that starts growing after 1,800 followers. This is the limit on the number of followings set by Twitter to prevent users monitoring too many accounts whereas they have no active role in Twitter. However, we can see some accounts above the followings limit. After manual inspection, we found no specific correlation between these accounts. Our guess is that these users followed a large number of accounts *before* the limit on the number of followings was introduced. Finally, we found that 40% of accounts have no followers, 25% have no followings.

4. CONCLUSIONS AND FUTURE WORK

We have crawled the entire Twitter social graph and all Twitter users profiles. We presented in this paper a preliminary analysis of the collected data, but we plan a much deeper analysis. Our future plans are to: i) perform the analysis of the Twitter follower graph; ii) study information propagation on Twitter and understand the propagation speed of tweets; iii) identify the most influential users on Twitter and find a strategy for increasing the influence of any user; iv) find private communities disconnected from the main connected component; v) assess how discriminant social relations are, that is estimate the probability that two different users have the same followers or followings list.

5. ACKNOWLEDGMENTS

This work has been partially funded by EIT ICTLabs.

6. REFERENCES

[1] D. Boyd, S. Golder, and G. Lotan. Tweet, Tweet, Retweet: Conversational aspects of retweeting on Twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference*, pages 1–10, Los Alamitos, CA, USA, Jan. 2010. IEEE.

[2] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about Twitter. In *Proceedings of the first workshop on Online social networks*, WOSN'08, pages 19–24, New York, NY, USA, 2008. ACM.

[3] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World Wide Web*, WWW'10, pages 591–600, New York, NY, USA, 2010. ACM.

[4] H. Mao, X. Shuai, and A. Kapadia. Loose tweets: an analysis of privacy leaks on Twitter. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, WPES'11, pages 1–12, New York, NY, USA, Oct. 2011. ACM.

[5] M. Russell. *21 Recipes for Mining Twitter*. Real Time Bks. O'Reilly Media, Inc., 2011.

[4] `https://dev.twitter.com/docs/platform-objects/users`

[5] This sentence contains typos in the camera ready version.