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 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.

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). 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 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.

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.

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.

1http://twitter.com

2https://dev.twitter.com/

3https://www.planet-lab.org/
Twitter has discontinued whitelisting new machines since 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 (500 million) much larger than the claimed number of Twitter accounts by the time of our crawl to be sure to 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 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 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 (500 million) much larger than the claimed number of Twitter accounts by the time of our crawl to be sure to not miss any account. Finally, the crawler distributed each chunk to one of the crawling machines to check each ID in the chunk for a valid account. If an ID corresponds to a valid account, all public account information 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 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 (500 million) much larger than the claimed number of Twitter accounts by the time of our crawl to be sure to not miss any account. Finally, the crawler distributed each chunk to one of the crawling machines to check each ID in the chunk for a valid account. If an ID corresponds to a valid account, all public account information 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 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 (500 million) much larger than the claimed number of Twitter accounts by the time of our crawl to be sure to not miss any account. Finally, the crawler distributed each chunk to one of the crawling machines to check each ID in the chunk for a valid account. If an ID corresponds to a valid account, all public account information 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 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 (500 million) much larger than the claimed number of Twitter accounts by the time of our crawl to be sure to not miss any account. Finally, the crawler distributed each chunk to one of the crawling machines to check each ID in the chunk for a valid account. If an ID corresponds to a valid account, all public account information 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 February 2011, but existing whitelisted machines can still be used.