How Information Propagates on Twitter?

Maksym Gabielkov

UNIVERSITE NICE SOPHIA ANTIPOLIS

ECOLE DOCTORALE STIC

SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THESE

pour l'obtention du grade de

Docteur en Sciences

de l'Université Nice Sophia Antipolis

Mention : INFORMATIQUE

présentée et soutenue par

Maksym GABIELKOV

Comment se propagent les informations sur Twitter ?

Thèse dirigée par Arnaud LEGOUT

et préparé au sein du laboratoire Inria, équipe DIANA

soutenue le 15 juin 2016

Jury :

Directeur :	Arnaud LEGOUT	-	Inria (DIANA)
Rapporteurs :	Ashish GOEL	-	Université Stanford
	Krishna GUMMADI	-	MPI-SWS
President :	Guillaume URVOY-KELLER	-	Université Nice Sophia Antipolis
Examinateurs :	Laurent MASSOULIÉ	-	Centre de Recherche Commun Inria - Microsoft Research
	Laurent VIENNOT	-	Inria (GANG)

Comment se propagent les informations sur Twitter ?

Résumé: Cette thèse présente une étude sur la mesure des réseaux sociaux en ligne avec un accent particulier sur Twitter qui est l'un des plus grands réseaux sociaux. Twitter utilise exclusivement des liens dirigés entre les comptes. Cela rend le graphe social de Twitter beaucoup plus proche que Facebook du graphe social représentant les communications dans la vie réelle. Par conséquent, la compréhension de la structure du graphe social Twitter et de la manière dont les informations se propagent dans le graph est intéressant non seulement pour les informaticiens, mais aussi pour les chercheurs dans d'autres domaines, tels que la sociologie. Cependant, on sait peu de choses sur la propagation de l'information sur Twitter.

Dans la première partie, nous présentons une étude approfondie de la structure macroscopique du graphe social de Twitter dévoilant les routes sur lesquelles les tweets se propagent. Pour cette étude, nous avons crawlé Twitter pour récupérer tous les comptes et toutes les relations sociales (liens de following et follower) entre les comptes. Nous présentons une méthodologie pour dévoiler la structure macroscopique du graphe social de Twitter qui se compose de 8 composants définis par leurs caractéristiques de connectivité. Nous avons découvert que chaque composant regroupe les utilisateurs avec un usage spécifique de Twitter. Enfin, nous présentons une méthode pour explorer la structure macroscopique du graphe social de Twitter dans le passé, nous validons cette méthode en utilisant des anciens ensembles de données, et nous discutons l'évolution de la structure macroscopique du graphe social de Twitter durant les 6 dernières années.

Dans la deuxième partie, nous étudions la propagation de l'information sur Twitter en étudiant les articles de presse partagés sur Twitter. Les médias en ligne comptent de plus en plus sur les médias sociaux pour générer du trafic vers leur site Web. Pourtant, nous savons étonnamment peu de choses sur la façon dont les conversations sur les médias sociaux mentionnant un article en ligne génèrent un clic "social" vers cet article. Nous présentons une étude validée et reproductible des clics sociaux en collectant un mois de clics vers des articles mentionnés dans Twitter vers 5 grands journaux en ligne. Nous montrons que les clics et les clics par follower impactent plusieurs aspects de la diffusion de l'information, tous jusque-là inconnus. Par exemple, les ressources secondaires (non promues dans les gros titres des journaux) génèrent plus de clics que les gros titres. De plus, alors que l'attention des utilisateurs des médias sociaux est courte en ce qui concerne les postes, elle est étonnamment longue lorsque l'on regarde les clics. Pour finir, on montre que l'influence réelle d'un intermédiaire ou d'une ressource est mal prédite par le comportement d'envoi, et nous montrons comment cette prédiction peut être rendu plus précise.

Dans la troisième partie, nous présentons une étude expérimentale d'échantillonnage du graphe social de Twitter. Les réseaux sociaux en ligne (RSL) sont une source importante d'information pour les scientifiques dans différents domaines tels que l'informatique, la sociologie, ou l'économie. Cependant, il est difficile d'étudier les RSL car ils sont très grands. En outre, les entreprises prennent des mesures pour prévenir les analyses de leurs RSL et s'abstiennent de partager leurs données avec la communauté des chercheurs. Pour ces raisons, nous affirmons que les techniques d'échantillonnage sont une option efficace pour étudier les RSL à l'avenir. Dans cette dernière partie, nous prenons une approche expérimentale pour étudier les caractéristiques des techniques d'échantillonnage bien connues sur un graphe social complet de Twitter nous avons crawlé en 2012.

Mots clé: Twitter, réseaux sociaux, propagation de l'information, analyse de graphes, clics sociaux, échantillonnage de graphes

How Information Propagates on Twitter?

Abstract: This thesis presents the measurement study of Online Social Networks focusing on Twitter. Twitter is one of the largest social networks using exclusively directed links among accounts. This makes the Twitter social graph much closer to the social graph supporting real life communications than, for instance, Facebook. Therefore, understanding the structure of the Twitter social graph and the way information propagates through it is interesting not only for computer scientists, but also for researchers in other fields, such as sociologists. However, little is known about the information propagation in Twitter.

In the first part, we present an in-depth study of the macroscopic structure of the Twitter social graph unveiling the highways on which tweets propagate. For this study, we crawled Twitter to retrieve all accounts and all social relationships (follow links) among accounts. We present a methodology to unveil the macroscopic structure of the Twitter social graph that consists of 8 components defined by their connectivity characteristics. We found that each component group users with a specific usage of Twitter. Finally, we present a method to approximate the macroscopic structure of the Twitter social graph in the past, validate this method using old datasets, and discuss the evolution of the macroscopic structure of the Twitter social graph during the past 6 years.

In the second part, we study the information propagation in Twitter by looking at the news media articles shared on Twitter. Online news domains increasingly rely on social media to drive traffic to their websites. Yet we know surprisingly little about how social media conversation mentioning an online article actually generates a click to it. We present a large scale, validated and reproducible study of social clicks by gathering a month of web visits to online resources that are located in 5 leading news domains and that are mentioned in Twitter. As we prove, properties of clicks and social media Click-Per-Follower rate impact multiple aspects of information diffusion, all previously unknown. Secondary resources, that are not promoted through headlines and are responsible for the long tail of content popularity, generate more clicks both in absolute and relative terms. Social media attention is actually long-lived, in contrast with temporal evolution estimated from posts or receptions. The actual influence of an intermediary or a resource is poorly predicted by their posting behavior, but we show how that prediction can be made more precise.

In the third part we present an experimental study of graph sampling. Online social networks (OSNs) are an important source of information for scientists in different fields such as computer science, sociology, economics, etc. However, it is hard to study OSNs as they are very large. Also, companies take measures to prevent crawls of their OSNs and refrain from sharing their data with the research community. For these reasons, we argue that sampling techniques will be the best technique to study OSNs in the future. In this part, we take an experimental approach to study the characteristics of well-known sampling techniques on a full social graph of Twitter we crawled in 2012.

Keywords: Twitter, social networks, information propagation, graph analysis, social clicks, graph sampling

Contents

Chapter 1

Introduction

Social network can be defined as a set of social entities (*e.g.*, individuals, groups, organizations, etc.) connected among each other with different types of relationships. The analysis of social networks is an interdisciplinary academic field at the intersection of sociology, psychology, mathematics and computer science.

The idea of social networks has its roots in the works of two sociologists, Émile Durkheim and Ferdinand Tönnies, published in early 1890s. These works on social groups foreshadowed the idea of social networks. Throughout the 20th century there have been major developments by several groups of scientists in different fields working independently. At that time the systematic recording and analysis of social interactions were performed on small groups, *e.g.*, work groups or classrooms, due to the natural difficulty of doing large scientific studies with real people. By the 1970s, different tracks and traditions of social network analysis were combined together. In 1969 Travers and Milgram [Travers 1969] set up their well-known experiment in which they asked 196 arbitrarily chosen individuals in Nebraska and Boston to deliver a letter to a target person in Massachusetts via an acquaintance chain. This experiment was groundbreaking as it has suggested that human society is a small-world network with a short path-length, the mean number of intermediary between the starts and the target was 5.2. It played an important role in the development of the concept of "six degrees of separation" that suggests that every person in the world is six or less steps away from any other person. This concept was highly popularized not only in academia, but also in popular culture.

Due to the fast development of the Internet in the late 1990s, online social networks (OSNs) emerged. OSN, also known as social-networking site, is an Internet-based service that allows individuals to (a) create a public or semi-public profile in the service, (b) establish relationships with other users of the service, (c) view all or some of the relationship between other users. The functionality of the service may also include the ability to exchange messages, multimedia content or express reactions, but these features may vary from one OSN to another.

In 1997 the first social-networking site `SixDegrees.com` was launched. SixDegrees was ahead of its time, users didn't know what to after registering and connecting to their friends (the service got closed in 2013). Starting from 2003 many OSNs were launched, and social networking in the Internet became mainstream (see survey on social networking sites [Boyd 2007]). Nowadays OSNs can be split into two categories, *general* and *specialized*[1]. Notable examples of general purpose OSNs are Facebook, Twitter and Google+. Popular specialized OSNs include `Instagram.com` (focuses on sharing photos and videos) and `LinkedIn.com` (focuses of professional and business networking). The popularity of social-networking sites varies with time and geographical region. For instance, due to government

[1] `https://en.wikipedia.org/wiki/List_of_social_networking_websites`

regulations, multiple websites such as Facebook, Twitter, and Google+ are unavailable on the territory of China, which led to the development of local social-networking sites such as Qzone, RenRen, and Sina Weibo. Another example is a general purpose OSN VK.com which is popular in post-Soviet republics.

General purpose OSNs are normally used to keep in touch with offline friend and not to establish new relationships. It implies that we can see OSNs as a model of social relationships of an individual. We can conclude that OSNs do not only facilitate communications between their users, but also create a valuable source of information for researchers and business. With the help of OSNs researchers can make studies of planetary scale.

1.1 Why Study OSNs?

With the number of registered users in OSNs reaching billions, it is hard to deny that OSNs are playing an important role in lives of people. In the following, we outline the reasons why we believe OSNs should be studied.

Traffic analysis. According to various sources, traffic from social networks is responsible for 30% of all referral traffic to the content published online (see Section 4.1). However, little is known about how this traffic is formed. Advertising on the Internet is the major income for many companies such as Google, Facebook, Twitter[2]. Moreover it is one of the ways to keep the content freely accessible for such online resources as news media. Studying the traffic originating from the OSNs is crucial for the business relying on the revenue from advertising.

Public opinion measurement. Internet is becoming widespread, according to the estimates in 2014, 40% of the world population is using internet[3], making it one of the best way to know the opinion of the people on events such as political elections or product releases. OSNs are a valuable source of information for business analysts and researchers, but it can also impact of the financial market. For example, a single tweet can cause major fluctuation on the stock market[4].

Recommender systems have become extremely common in recent years, they are used to suggest products to the consumer based on their interest. Previous studies show that users who have relationship in OSNs are more likely to share interests [Zhang 2014], *e.g.*, in movie or music genres. Thus, exploiting data from OSNs can improve search results or the quality of recommendations. Also it can be used to improve connectivity in peer-to-peer systems by creating additional connection between the users having social ties on OSNs [Zhang 2014].

Social trust and spam. In the modern world the amount of information people receive every day is huge and keeps increasing, however it is hard to know which information is trustworthy and which is not. Another closely related topic is spam detection. While it is hard to understand from the content of a particular message if it is unsolicited,

[2] http://www.statista.com/statistics/460687/digital-ad-revenue-select-companies/
[3] https://en.wikipedia.org/wiki/Global_Internet_usage
[4] http://j.mp/20P5vUc

undesired or apocryphal, we can benefit from the knowledge of user's social ties, *e.g.*, by considering the content originating from the "friends" on OSNs more trustworthy.

Modeling systems. Many parameters such as popularity distributions emerge from people behavior. These parameters can be revealed by measuring the OSNs and can be used for modeling. For example, the data about the number of followers on Twitter can be used to model the popularity distribution for information-centric networks.

1.2 Why Twitter?

In this thesis, we chose Twitter as a case study. Twitter is a micro-blogging service that allows its users to send short messages (tweets) of up to 140 characters, also users can subscribe for the messages of others. The main Twitter page of a registered user is called a "timeline", it shows the list of tweets of the people the user is subscribed to in a reverse chronological order.

Twitter was launched in 2006 and has now more than 332 million active users (as of January 2016), making it the third largest OSNs in the world. While Twitter is not the most popular OSN, it has some features that make it different from other OSNs and more interesting to study from the research point of view; these features as described in the following.

Earliest OSNs allowed their users to create "friendship" relationships among their users. "Friendship" requires a confirmation from both parties of the relationship and is implemented as a major type of relationship in OSNs such as SixDegrees, Facebook or LinkedIn. However friendship is not the only type of social relationship we encounter in the world. In a traditional setting, when people watch TV or read newspapers, they are normally interested in the content, however the publisher of the newspaper or the producer of the TV program has no direct way to communicate with people consuming the product. This publisher-subscriber relationship can be modeled by introducing a "follow" relationship when a user can follow another user to receive their updates without the confirmation of the latter. 'Follow'-like relationships in OSNs were first introduced on LiveJournal in 1999, but the terms 'follow' and 'follower' are more known in the context of Twitter. The concept of such unidirectional relationship is much more flexible, because it can both model a publisher-subscriber relationship, *e.g.*, when a user follows a celebrity, and a friendship relationship, *e.g.*, when two users follow each other. That is why we believe Twitter is one of the best sources of information about social relationship among people.

Another interesting feature of Twitter is the limitation of 140 characters on the length of the post. Initially introduced for compatibility with SMS messaging, the 140-character limit played a huge role in forming the culture of Twitter. Due to the short nature of tweets, Twitter is used to send updates to people. That is why Twitter is widely used during natural disasters or public gatherings to connect people and serve as an additional communication channel. We can say that Twitter became the word-of-mouth of the Internet. We believe that a short message is a better representation of what is on the people's mind than a multipage opus on Facebook, short tweets give us the opportunity to observe the instant reaction of people to the events.

Also, this limit increased the usage of URL shorteners such as `goo.gl` and `bit.ly` so that URLs would not consume the majority of the valuable 140-character limit. The vast usage

of shorteners on Twitter opens unprecedented possibilities of studying content dissemination though Twitter as URL shorteners provide the detailed statistics on the actual number of clicks made on the URLs (see Chapter 4).

Given the constrained length of the tweet, Twitter users developed a series of conventions that allow users to add structure to their tweets [Boyd 2010]. These conventions emerged from the Twitter crowd and got so popular that they were implemented as features of Twitter.

Mentions. Users began to use the `@username` syntax to mention or to address specific users in their tweets. Now users get notified if they are mentioned in some conversation.

Re-tweets. To repost a message of another Twitter user to their followers, a user may copy the content of the message and post it with preceding `RT @username` or `via @username`. Now it is implemented as a re-tweet button.

Hashtags. Users can group messages together by topic or type using a hashtag, that is a word prefixed with a "#" sign. A click on a hashtag gives the messages containing this hashtag in a reverse chronological order.

All these conventions do not only help users to better navigate through Twitter, but also provide researchers and analysts an easier way to interpret the information without involving heavy natural language processing tools. Also, in late 2009, *Twitter lists* were introduced, allowing users to create and follow ad hoc lists of authors instead of individual authors.

In July 2015, Twitter has extended the limit for the private messages to 10,000 characters. Later, in January 2016, Jack Dorsey (current CEO of Twitter) revealed that Twitter is planning to expand the character limit for tweets as well; this limit would also be 10,000 characters, however users will be required to click to see anything beyond 140 characters. He said that while Twitter would "never lose that feeling" of speed, users could do more with the text.

Twitter does not impose real-name policy on its users. On the one hand, it implies more freedom of expression. Some OSNs, *e.g.*, SixDegees.com, lost many users after introducing the real-name policy and suspending the accounts that did not look real. Also, it is important to remain anonymous when the conversations touch topics such as politics, or when disclosure of the information can cause harm to the sender. On the other hand, it creates more opportunities for spammers and opinion manipulation, which is an interesting research topic of its own.

1.3 Challenges and Contributions

In this section, we describe the approach we took in the analysis of Twitter, the challenges we faced, and the contributions we made. Our analysis consists of three steps. (i) we want to understand the medium where information propagates, that is the Twitter social graph. (ii) we study the information propagation in Twitter focusing on news media articles. (iii) we study the problem of using graph sampling for estimation of various metrics when access to the data is constrained.

1.3.1 Social Graph Structure

To understand information propagation in Twitter, we first need to understand the structure of the medium where it propagates. In Twitter this medium is the *social graph* that has users as vertices and follow relationships as directed edges (arcs). Information can propagate in this graph from a user to his followers, then it may be retweeted and reach the followers of the followers of a user, etc. While the will of users to retweet something strongly depends on the content and the users themselves, there is no doubt that there is no way for information to flow between the users that are disconnected in the social graph. We can say that the structure of the social graph constraints the information propagation, and naturally a first step in understanding how information propagates on Twitter is to study the structure of its social graph. We faced the following challenges during this study.

1.3.1.1 Challenges

Data collection. To analyse the structure of the Twitter social graph, we first need to crawl the graph. Twitter does not provide access to its social graph, moreover they use a distributed infrastructure to support the operation of Twitter, and even with a full access to this infrastructure, it would be quite challenging to extract the social graph.

The most reliable way to get the data is the Twitter application programming interface (API). However, access to this API is subject to strict rate-limits, at the lime we collected the data (2012) this rate-limits were applied per IP address. We had to implement a distributed crawler that was using 550 machines spread across the globe. It took us 4 month to collect the *complete* Twitter social graph, details are presented in Section 3.2.

Data processing. Another challenge is to process the data of a big size, the graph we collected consists of 537 million users and 24 billion arcs and requires 74GB of RAM to be stored in the format of an adjacency list.

We have tested multiple state of the art tools including Hadoop[5], NetworkX[6], SNAP[7], and GraphChi [Kyrola 2012]. Map-reduce (Hadoop) and vertex-centric approaches (GraphChi) showed to be inefficient in performing breadth-first-search (BFS) that is essential for computation of strongly connected components (SCCs). NetworkX and SNAP appeared to be unusable for large graphs or graphs with high degree nodes. At first, we developed our own solution that made the computations on the graph in a divide-and-conquer manner, *e.g.*, we split the Twitter social graph into chunks that fit into RAM, computed the SCCs in each chunk, then merged the results. Afterwards, when we gained access to machines with more RAM, we used a combination of Python with a module written in C that stored the graph in memory. It helps us avoid the Python memory overhead while preserving the advantage of Python in terms of fast implementation and efficient data manipulation[8].

[5] http://hadoop.apache.org/
[6] https://networkx.github.io/
[7] http://snap.stanford.edu/
[8] Python is great tool for fast development, however it has overhead in term of CPU utilization and in terms of memory consumption (in Python everything is represented as objects that have headers, *e.g.*, one integer requires at least 12 bytes in memory, whereas in C it would be 4 bytes).

Interpretation of the results. We have computed various statistics on the Twitter social graph, *e.g.*, distribution of degrees, weakly connected components, SCCs, and we have designed and applied the generalized version of the graph decomposition of Broder *et al.* [Broder 2000]. We obtained the graph decomposition in 8 components and we sought the physical meaning of those components. This step required the computation of numerous metrics per component together with manual inspection of user accounts belonging to different components, which is time consuming. The results of these activities are presented in Section 3.4.

1.3.1.2 Contributions

We have collected a large scale dataset on the Twitter social graph in 2012 that may be the last dataset of such scale. We went through an ethical process and shared an anonymized version of our dataset[9] that was IRB approved. As of April 2016, we got 32 access requests from researchers all over the world, 18 of which signed the license agreement and obtained the dataset for various purposes including studying unbiased sampling, graph generation, influence propagation, node credibility, testing graph processing software (*e.g.*, graph partitioning problem for Apache Spark[10]) or algorithms (*e.g.*, personalized PageRank, community detection, or graph diameter estimation).

We improved and applied the decomposition methodology of Broder *et al.* [Broder 2000] to the Twitter social graph, which allowed us to map different components of the decomposition to different types of users, hence, to see how Twitter is used today (see Section 3.4) and to observe the evolution of Twitter usage with time (see Section 3.5).

1.3.2 Information Propagation

We studied the information propagation focusing on news media URL dissemination on Twitter. According to Mitchell *et al.* [Mitchell 2014], more than half of American adults are using OSNs as their primary source of political news. Also, these are two technical reasons why we chose news media URLs. First, URLs are easy to track because they have a particular syntax, also news media URLs point to a reputable content that was prepared by the publishers and categorized. Second, we can use the number of clicks on those URLs to externally assess their popularity. During this analysis we faced the following challenges.

1.3.2.1 Challenges

Data collection. The daily volume of tweets is measured in hundreds of millions. It is practically impossible for a person outside Twitter, Inc. to obtain the full data on these tweets. However, we can build a methodology combining different endpoints of the Twitter API to obtain a *consistent* sample of data. For example, we used the 1% sample of tweets, provided by Twitter without subscription fees, together with the search API of Twitter.

Twitter can give us information about who shared what and when, but it cannot tell us exactly who saw this information on Twitter. We can approximate the number of people who saw it by estimating the number of *receptions*, *e.g.*, by summing up the numbers of followers

[9] http://j.mp/soTweet
[10] http://spark.apache.org/

of people who shared the information (we call them posters). This approximation is quite naive because it doesn't take into account the overlap between the followers of the posters, and we have no way to validate that people actually saw the information in their Twitter timeline; they may not be checking their timeline regularly or the amount of information in it is overwhelming. Luckily, there is a way to know if a particular news media article shared on Twitter engaged users. We monitored the 1% sample of tweets and discovered that from 70% to 90% of news media URLs are *shortened* using services such as `bit.ly`, that provides an API to retrieve the statistics on the number of clicks made on its shortened URLs. Moreover, we can distinguish between clicks originating from Twitter and clicks coming from other websites by looking at the referrer. More details on the data collection process can be found in Section 4.2.1.

We are the first ones to combine data on sharing behavior of users and their clicking behavior; these data unveils multiple aspects of information propagation, all previously unknown (see Chapter 4).

Bias correction. Due to the fact that our data is incomplete, we face the danger of introducing multiple biases in our data. For example, we use the 1% random sample of tweets to discover news media URLs. This sample yields a random subset of tweets, but in terms of URLs we are highly biased towards popular ones. The same URL may be contained in multiple tweets; if we sample tweets at a constant rate of 1%, URL shared on Twitter once has 1% of chance to appear in our dataset, whereas a URL shared 100 times has $1-(1-0.1)^{100} \approx 63\%$ of chance to appear in the dataset. We cannot recover the URLs we missed, but we can correct the statistics in our study by giving more weight to unpopular URLs. Note that without the correction of this bias, we would observe strikingly different results. There is also a bias in the estimation of the number of receptions that we mentioned in the previous paragraph. More details are presented in Section 4.2.3.

User influence analysis. Another question we want to study is the influence of users who share content on Twitter. We can do that by looking at how good they are in posting popular content. However, we don't have the attribution between the users who share some content and the clicks made on this content; we only have the number of clicks per contents, which is good from a privacy point of view, but harmful in our research. We take an approach of accessing the user influence by looking at the success of content they *participated* in sharing. We were able to validate our approach on a subset of users in our dataset, more on that in Section 4.5.2.

1.3.2.2 Contributions

We present a large scale, unbiased study of news articles shared on Twitter. We gathered a month of web visits to online resources that are located in 5 leading news domains and that are mentioned in Twitter. That is the first data that combines sharing activities with clicks. Our dataset amounts to 2.8 million shares, together responsible for 75 billion potential views on this social media, and 9.6 million actual clicks to 59,088 unique resources. We design a reproducible methodology and carefully correct its biases. We are planing to share our

dataset with the research community after it gets IRB approved (details will be available at http://j.mp/soTweet).

The analysis of sharing activities together with clicks revealed multiple aspects of information diffusion, all previously unknown. First, news media article that are not promoted through headlines and are responsible for the long tail of content popularity, generate more clicks both in absolute and relative terms. Second, social media attention is actually long-lived, in contrast with temporal evolution estimated from shares or receptions. Third, the actual influence of a user or an article is poorly predicted by their share count, which is unreasonably used nowadays as a metric of impact in online resources.

1.3.3 Graph Sampling

We studied the problem of social graph sampling. We have collected the complete social graph of Twitter in 2012; soon after our crawl was completed, Twitter introduced a new version of the Twitter API with mandatory authentication for each request. The rate-limit in the new API is applied per user, whereas before it was applied per IP address. The new API made the crawl of the Twitter social graph much harder because it is much harder to create user accounts than use multiple IP addresses (*e.g.*, by using PlanetLab, or other distributed platform). Moreover, automated creation of multiple user and app accounts violates the Twitter terms of use and all accounts created can be suspended.

We believe that our 2012 dataset is the last crawl of Twitter of such scale. Companies managing the OSNs are introducing measures to prevent large-scale crawls of their data because nowadays information is a valuable resource for business. That is why we believe that the only way to do social network measurements, except in case of direct collaboration with an OSN provider, is to use *sampling*.

However, inconsiderate use of sampling may be harmful. For example, when one decides to use BFS to make a complete crawl of the graph, one assumes that the graph has a giant connected component (that can be reached with BFS) and the remaining components are of negligible size. Depending on the goal of the study it can lead to incorrect results, *e.g.*, we discovered that 20% of the Twitter accounts are disconnected, also if one used only one direction of the links on Twitter to make the crawl, they would miss 25% to 50% of the graph. Also, the result of such a crawl would highly depend on the source to start the BFS from, which is often made blindly.

Another problem appears when one use random walk (RW) to make a measurement of the graph. Most of the theory behind RW assumes that the graph is connected and has a *fast mixing* property. However, this property has never been validated on real social graphs.

We aim to study the properties of well-known sampling techniques (*e.g.*, BFS, RW) on the complete social graph of Twitter collected in 2012. In particular, we want to see how efficient these techniques are in estimating of metrics related to activities of Twitter, *e.g.*, number of active users, or number of tweets sent. These metrics are rarely published by OSN providers and often cannot be verified.

1.3.3.1 Challenges

Data processing. In order to perform tests of sampling techniques on our Twitter dataset, we need to build a tool that would emulate the Twitter API. This step is not straightforward

due to two facts. First, to make a trustworthy computation on the graph, we need to repeat the experiments multiple times, *e.g.*, we cannot conclude anything on one RW made from a particular node in the graph. Here is when the second fact comes into play, due to the size of the dataset, it is important to have a fast access to the data. For example, if we decide to store the data on the hard drive, a RW will require numerous reads from random places of the file which is known to be extremely slow; then, since we need to repeat the computation multiple times (*e.g.*, to get the confidence interval), we will end up with a computation that can take weeks or month, which is often impractical. Actually, most of the time we need to keep two copies of the Twitter social graph in memory: the adjacency list of followers and the adjacency list of followings. We overcome this problem by putting the graph into memory as a read-only structure and by making multiple computation threads to benefit from the multi-core architecture of the server.

Moreover, some of the sampling techniques (*e.g.*, the ones derived from RW) require computations to be done on each step of the sampling. For example, RW are naturally biased toward high-degree nodes, hence, one needs to unbias the results in either online (Metropolis-Hasting RW) or offline fashion (re-weighted RW), which requires some computation. Such computations usually take few time, but since we aim to see the evolution of different metrics with the size of the sample, we have to repeat these computations numerous times, which in turn results in a significant increase of the computation time. The optimization of such computation is done by profiling and re-factoring the code.

Crawling cost. Another interesting problem is the cost of the crawl. Here, we propose to consider the cost in terms of time required to retrieve the information from the OSNs or in terms of number of requests we need to make to OSNs servers. In theory, we often represent social relationships among users as a graph $G(V, E)$ where V is a set of users and E is a set of relationship interconnecting the users. A naive approach would be to consider a sample $S \subset V$ to have cost of $|S|$ or 1 unit (*e.g.*, request) per user. However, in practice that is not what we observe. Indeed, the way information is presented in OSN may affect the cost of the crawl. For example, due to the huge amount of information, the Twitter API paginates the results of some requests, let's consider that we need to retrieve the list of followers of Lady Gaga, she had 22 million followers in 2012, and the API returns at most 5,000 followers per page, that makes $\frac{22\times10^6}{5000} = 4400$ requests $= 73.3$ hours to get the information; whereas for an ordinary user we would need only one minute. Moreover, some techniques (*e.g.*, Metropolis-Hasting RW) relies on some additional knowledge (*e.g.*, the degree of an adjacent node) during the crawl to decide if we jump to a new node or stay in the current one; this can increase the cost of the crawl by a huge factor (50x in case of Twitter).

In summary, we aim to evaluate the performance of sampling techniques to estimate various metrics of the social graph taking into account the real cost of sampling.

1.3.3.2 Contributions

We show that it is important to properly address the sampling bias, because classical sampling techniques, such as RW and BFS, are biased towards high degree nodes. Also, we argue that one needs to carefully account for the practical cost of sampling when designing sampling algorithms (*e.g.*, the cost of Metropolis-Hasting RW is up to two orders of magnitude higher

than the cost of BFS with the same number of nodes sampled). We show that we can easily estimate such metrics as number of active accounts or number of tweets sent within the time-frame of one day.

1.4 Thesis Outline

This thesis has the following structure. Chapter 2 contains the description of related work. In Chapter 3 we present the study of the structure of Twitter social graph, we identify eight different components base on the connectivity of the graph and map this components to a particular usage of Twitter. In Chapter 4 we study the dissemination of news media articles through Twitter by monitoring the URLs of five popular news media posted by both the news media and the ordinary Twitter users. In Chapter 5 we acknowledge that provided the rapid growth of OSNs and the constraints put by the OSN provides on the access to their data, it will be difficult if not impossible to collect large scale datasets, we discuss the problem of sampling of OSNs and address the bias of these samples. Chapter 6 concludes the thesis.

Chapter 2

Related work

Twitter has been widely studied for years, first works on Twitter were published one year after its launch in 2006 [Java 2007]. We describe in the following works related to this thesis, and we position our contributions to the state of the art.

Twitter Crawl. Some of them crawled partially the graph before 2009 [Java 2007, Krishnamurthy 2008, Huberman 2008], so before the wide adoption of Twitter. Two studies made a large crawl of the Twitter social graph. Kwak *et al.* used a technique close to a BFS and reverse BFS from a popular account and also collected accounts referring to trending topics. This crawling methodology cannot capture some users that are not connected to the LSC component, and that do not tweet about trending topic, thus a partial view of the Twitter social graph. Cha *et al.* [Cha 2010] used a crawl by account ID, that is close to what we did. Both of these studies made their dataset publicly available and others built on it [Lee 2010, Wu 2011, Cha 2012, Sharma 2012], but the datasets were collected in 2009 during the main change in the Twitter social graph we discuss in Section 3.5.2.

To the best of our knowledge, the dataset we present is the most up-to-date and complete description of the Twitter social graph.

Graph Macrostructure. None of the prior studies explored the macrostructure of a social graph. Broder *et al.* [Broder 2000] introduced first the notion of macrostructure for a directed graph in the context of the Web, but we significantly improved it, and we are the first ones to apply it to Twitter. Unlike what Broder *et al.* proposed, we present a methodology to compute the exhaustive macrostructure of any large directed social graph, along with the categorization of each account in the identified component, which is a significant methodological step.

Influence and intermediaries. Information diffusion naturally shapes collective opinion and consensus [Katz 1957, Degroot 1974, Lord 1979], hence the role of OSNs in redistributing influence online has been under scrutiny ever since their prehistoric forms, *e.g.*, blogs and email chains [Adamic 2005, Liben-Nowell 2008]. Information on traditional mass media follows a unidirectional channel in which pre-established institutions concentrate all decisions. Although the emergence of opinion leaders digesting news content to reach the public at large was pre-established long time ago [Katz 1957], OSNs presents an extreme case. They challenge the above vision with a distributed form of influence: OSNs allow *in theory* any content item to be tomorrow's *headline* and any user to become an *influencer*. This could be either by gathering direct followers, or by seeing her content spreading faster through a set of intermediary nodes.

Previous studies show that almost all online users exhibit an uncommon taste at least in a part of their online consumption [Goel 2010], while others point to possible bottlenecks in information discovery that limits the content accessed [Cha 2009]. To fully leverage opportunities open by OSNs, works propose typically to leverage either a distributed social curation process (*e.g.*, [Zadeh 2013, Hegde 2013, Wong 2015, May 2014]) or some underlying interest clusters to feed a recommender system (*e.g.*, [Xu 2014, Massoulié 2015]).

Prior works demonstrated that news content exposure benefits from a set of information intermediaries [Wu 2011, May 2014], proposed multiple metrics to quantify influence on OSNs like Twitter [Cha 2010, Bakshy 2011], proposed models to predict its long term effect [Kleinberg 2007, Lelarge 2012], and designed algorithms to leverage inherent influence to maximize the success of targeted promotion campaign [Kempe 2003, Ok 2014] or prevent it [Lelarge 2009]. So far, those influence metrics, models, and algorithms have been validated assuming that observing a large number of receptions is a reliable predictor of actual success, hence reducing influence to the ability to generate receptions. We turn to a new definition in which influence is measured by actual clicks, which are more directly related to revenue through online advertising, and also denote a stronger interaction with content.

Temporal Evolution of Diffusion. Studying the temporal evolution of diffusion on OSNs can be a powerful tool, either to interpret the attention received online as the result of an exogenous or endogenous process [Crane 2008], to locate the original source of a rumor [Pinto 2012], or to infer *a posteriori* the edges on which the diffusion spreads from the timing of events seen [Gomez-Rodriguez 2012]. More generally, examining the temporal evolution of a diffusion process allows to confirm or invalidate simple model of information propagation based on epidemic or cascading dynamics. One of the most important limitation so far is that prior studies focus only on the evolution of the collective volume of attention (*e.g.*, hourly volumes of clicks [Szabo 2010], views [Crane 2008, Cha 2009]), hence capturing the *implicit* activity of the audience, while ignoring the process by which the information has been propagated. Alternatively, other studies focus on *explicit* information propagation only (*e.g.*, tweets [Yang 2011], URLs shorteners, diggs [Wu 2007]) ignoring which part of those content exposure leads to actual clicks and content being read. Here for the first time we combine explicit shares of news with the implicit web visits that they generate.

Temporal patterns of online consumption was gathered using videos popularity on YouTube [Crane 2008, Cha 2009] and concluded to some evolution over days and weeks. However, this study considered clicks originating from any sources, including YouTube own recommendation systems, search engine, and other mentions online. One hint of the short attention span of OSNs was obtained through URLs shorteners[1]. Using generic `bit.ly` links of large popularity, this study concludes that URLs exchanged using `bit.ly` on Twitter today typically fades after a very short amount of time (within half an hour). Here we can study jointly for the first time the two processes of OSN sharing and consumption. Prior work [Abisheva 2014] dealt with very different content (*i.e.*, videos on YouTube), only measured the overall popularity generated from all sources, and only studied temporal patterns as a user feature to determine their earliness. Since the two processes are necessarily related, the most important unanswered question we address is whether the temporal property of one

[1] `http://bitly.is/1IoOqkU`

process allows to draw conclusion on the properties of the others, and how considering them jointly shed a new light on the diffusion of information in the network.

Spammers and Bots on Twitter. Twitter is very open, it does not impose real name police and has profiles set to be public by default. While having obvious benefits, this approach has a drawback, it is fairly easy to create fake accounts that are often used for spam, creation of fake followers, or opinion manipulation. Twitter suspends accounts that are violating its terms of use, however, to the best of our knowledge, it is done manually. Many studies are focused on studying malicious activity on Twitter and on automatic ways of its detention [Benevenuto 2010, Wang 2010a, Ghosh 2011, Chu 2012, Yang 2012].

Kurt et al. [Thomas 2011] studied the lifetime, properties, and behavior of spam accounts on Twitter based on the large collection of tweets aggregated during seven month. Authors identified an emerging marketplace of social network spam-as-a-service and analyzed its underlying infrastructure including an in-depth analysis of five of the largest spam campaigns on Twitter. Zhang et al. [Zhang 2011] performed an analysis of automated activity on Twitter and discovered that 16% of active accounts exhibit a high degree of automation. Authors present a method for detecting the automated Twitter accounts. Ghosh et al. [Ghosh 2012] studied link farming on Twitter, when users try to acquire large number of followers. Authors describe different types of users who farm links and propose a ranking scheme, where users are penalized for following spammers.

In this work, we identified the DISCONNECTED component that previous studies missed, for example, because they used BFS from a popular node to crawl the graph. This component contains a lot of spammer accounts that are not connected to anyone, but are used to send tweets that will appear in the search results.

Twitter Privacy and Anonymity. Another implication of Twitter being open is that people send sensitive information in their tweets forgetting about the consequences [Humphreys 2010, Mao 2011]. Public tweets may contain such sensitive information as user's health state, confessions about driving drunk, or the periods of time when people leave their houses. This information can be used by insurance companies, law enforcement, and robbers respectively; it can have drastic effect on the lives of people who post it. This problem looks even more severe because 70% of Twitter users are identifiable, *e.g.*, they disclose their full name in their Twitter profile [Peddinti 2014].

Even without disclosing sensitive information by themselves, users of OSNs can be deanonymized by websites. For example, Ramachandran *et al.* show that it is possible to map silent web visitors to their Twitter profiles by matching the list of webpages they visited with the list of tweets containing URLs they received on Twitter [Ramachandran 2014].

In our study, we share the data we collected from Twitter, and we take the following precautions to ensure that we do not disclose any private information of Twitter users. First, we anonymized the Twitter social graph we make publicly available. Second, the URL shorteners that we use to study the information propagation on Twitter were merged to make sure that users cannot be deanonymized using the data from our dataset (see Section 4.2.2).

Twitter Users Categorization. A Twitter user profile does not provide rich information about the user, it only includes basic fields such as name, free-form description, location, and

personal website URL. However, to understand information propagation in Twitter, one needs to understand the landscape of Twitter users. This set of studies is focused on categorizing Twitter accounts into several categories based on their functions, influence, behavior, or origin. Cha et al.[Cha 2010] study the influence of Twitter users based on the complete snapshot of Twitter graph and the entire collection of tweets in 2009[2]. Later, Cha et al. [Cha 2012] categorized Twitter users and study their patterns of information spread. Some researches [Sharma 2012, Wu 2011] have crawled *Twitter lists*, that are user-created lists meant to group together people of similar domain of interest. Lists were introduced to better organize Twitter subscriptions, users can follow the entire list instead of following all its members. Also, the data from the lists can be used to infer the field of expertise for the users listed.

We used the list of experts identified by Sharma *et al.* [Sharma 2012] to understand the roles of users in different components of graph macrostructure (see Chapter 3.4).

Graph sampling. Graph sampling is a set of techniques to pick a subset of vertices or edges from the original graph. Sampling has a wide spectrum of applications including sociological surveys, graph visualization and estimation of various metrics on the graphs with limited access [Hu 2013]. Classical graph sampling techniques are based on graph traversal algorithms, *e.g.*, BFS, RW, or multiple RWs [Ribeiro 2010]. RW and BFS are known to have bias towards high-degree nodes, this bias should be corrected to obtain correct results, it is often done using theoretical results on Markov chains for undirected social graphs, *e.g.*, Facebook [Gjoka 2010]. One simple approach to access the sampling bias in a directed graph is to ignore the directional of the links and apply previous results [Wang 2010b].

We are the first ones to take into account the practical cost of sampling and to validate the sampling techniques on a real Twitter social graph.

[2] http://twitter.mpi-sws.org/

Chapter 3

Twitter Social Graph

Contents

Twitter is one of the largest social networks using exclusively directed links among accounts. This makes the Twitter social graph much closer to the social graph supporting real life communications than, for instance, Facebook. Therefore, understanding the structure of the Twitter social graph is interesting not only for computer scientists, but also for researchers in other fields, such as sociologists. However, little is known about how the information propagation in Twitter is constrained by its inner structure.

In this chapter, we present an in-depth study of the macroscopic structure of the Twitter social graph unveiling the highways on which tweets propagate, the specific user activity associated with each component of this macroscopic structure, and the evolution of this macroscopic structure with time for the past 6 years. For this study, we crawled Twitter

to retrieve all accounts and all social relationships (follow links) among accounts; the crawl completed in July 2012 with 505 million accounts interconnected by 23 billion links[1]. Then, we present a methodology to unveil the macroscopic structure of the Twitter social graph. This macroscopic structure consists of 8 components defined by their connectivity characteristics. Each component group users with a specific usage of Twitter. For instance, we identified components gathering together spammers, or celebrities. Finally, we present a method to approximate the macroscopic structure of the Twitter social graph in the past, validate this method using old datasets, and discuss the evolution of the macroscopic structure of the Twitter social graph during the past 6 years.

This work was accepted and presented at ACM SIGMETRICS 2014 in Austin, TX, USA [Gabielkov 2014b].

3.1 Motivation

Twitter is one of the largest social networks with more than 500 million registered accounts. However, it differs from other large social networks, such as Facebook and Google+, because it uses exclusively arcs among accounts[2]. Therefore, the way information propagates on Twitter is close to how information propagates in real life. Indeed, real life communications are characterized by a high asymmetry between information producers (such as media, celebrities, etc.) and content consumers. Consequently, understanding how information propagates on Twitter has implications beyond computer science.

However, studying information propagation on a large social network is a complex task. Indeed, information propagation is a combination of two phenomena. First, the content of the messages sent on the social network will determine its chance to be relayed. Second, the structure of the social graph will constrain the propagation of messages. In this chapter, we specifically focus on how the structure of the Twitter social graph constrains the propagation of information. This problem is important because its answer will unveil the highways used by the flows of information. To achieve this goal, we need to overcome two challenges. First, we need an up-to-date and complete social graph. The most recent publicly available Twitter datasets are from 2009 [Kwak 2010, Cha 2010], at that time Twitter was 10 times smaller than in July 2012. Moreover, these datasets are not exhaustive, thus some subtle properties may not be visible. Second, we need a methodology revealing the underlying social relationships among users, a methodology that scales for hundreds of millions of accounts and tens of billions of arcs. Standard aggregate graph metrics such as degree distribution are of no help because we need to identify the highways of the graph followed by messages. Therefore, we need a methodology to both reduce the size of the social graph and keep its main structure.

In this chapter, we overcome these challenges and make the following specific contributions.

1. We collected the entire Twitter social graph, representing 505 million accounts connected with 23 billion arcs. To the best of our knowledge, this is the largest *complete* social graph ever collected.

[1] http://j.mp/soTweet

[2] Arcs—that are directed edges—represent the follow relationship in Twitter. If A follows B, A receives tweets from B, but B will not receive tweets from A, unless B follows A.

2. We unveil a macroscopic structure in the Twitter social graph that preserves the highways of information propagation. Our method extends the one of Broder *et al.* [Broder 2000] and can be applied to any kind of directed graph.

3. We show that not only the macroscopic structure of the Twitter social graph constrains information propagation, but that each component of the macrostructure corresponds to group of users with a specific usage of Twitter. In particular, we show that regular, abandoned, and malicious accounts are not uniformly spread on the components of the macroscopic structure of the Twitter social graph. This result is important to understand how Twitter is used, where users with a specific usage are, and how to sample Twitter without a significant bias.

4. We present a simple methodology to explore the evolution of the macroscopic structure of Twitter with time, we validate this methodology, and show that old datasets from 2009 do not represent the current structure of the Twitter social graph. We explore this time evolution to understand the changes in the usage of Twitter since its creation.

The remainder of this chapter is structured as follows. In Section 3.2, we present our methodology to crawl Twitter and discuss the dataset we collected. We present and discuss, in Section 3.3, the notion of macroscopic structure, then we describe a methodology to unveil this macroscopic structure. We present the result of applying this methodology to our dataset in Section 3.4. In Section 3.5, we propose a simple approach to estimate the evolution of the macroscopic structure of the graph with time, validate this approach, and discuss the evolution of the Twitter social graph from 2007 to 2012. Finally, we present the related work in Section 3.6, and conclude in Section 3.7.

3.2 Measuring Twitter at Scale

In this section, we describe the methodology used to crawl the Twitter social graph, some high level characteristics of the dataset, the limitations of our crawl, and the ethical issues.

3.2.1 Crawling Methodology

In order to collect our dataset, we used the Twitter REST API[3] version 1.0 to crawl the information about user accounts and arcs between users. The main challenge of the crawl is that API requests are rate-limited; an unauthenticated host could make at most 150 requests per hour with that API. However this limit could be overcome by using a whitelisted machine. Twitter used to whitelist the servers of research teams and data-intensive services upon request, this service has been discontinued since February 2011, but existing whitelisted machines could still be used. We used four whitelisted machines to perform our crawl, two machines with a rate limit of 20,000 requests per hour and two with 100,000 requests per hour.

We also implemented and deployed a distributed crawler on 550 machines of PlanetLab[4], doubling the crawling rate compared to whitelisted machines only.

[3]Twitter REST API, https://dev.twitter.com/overview/documentation
[4]PlanetLab, https://www.planet-lab.org/

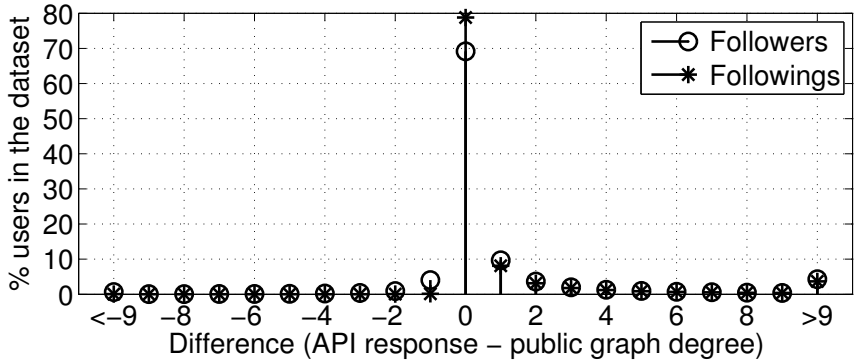


Figure 3.1 – **The difference in number of followers and followings between the data from user accounts and the public social graph reconstructed from our dataset.**

We crawled Twitter by user ID, such numeric IDs are assigned for new accounts sequentially, but with gaps [Krishnamurthy 2008]. Therefore, we first determined using a random polling that the largest assigned ID is lower than 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 largest observed ID to be sure to do not miss any account.

We performed our crawl from March 20, 2012 to July 24, 2012. We implemented a crawler that assigns chunks of 10,000 IDs to each crawling machine. Then, for a given chunk, each crawling machine performs two steps. First, the machine makes 100 requests for 100 IDs, the maximum number of IDs the lookup method of the API accepts, using an API call [Russell 2011a]. When an ID corresponds to a valid account, we retrieve all public numerical, boolean and date information[5]. Second, the machine collects the list of followings for all non-protected and valid accounts with at least one following.

We now define the notions of following, followers, and protected accounts that we use in this chapter. Each Twitter account can have *followings* and *followers*. An account receives all published tweets from its followings, and all its followers receive its tweets. Tweets, and list of followers and followings, are by default visible to everyone. However, users can make their account *protected* which makes this information visible only to its followers. Furthermore, following a *protected* account requires manual approval from its owner. For more information see `https://support.twitter.com/articles/14016`.

3.2.2 Limitations of the Crawl

There are some accounts that we could not crawl, representing 6.33% of the entire Twitter social graph. We explain in the following the reasons why some accounts are not present in our dataset.

1. 32,112,668 accounts (5.97% of the accounts in our dataset) are protected, so we cannot get their list of followings. The degrees of nodes in the graph we analyzed do not take into account arcs to and from protected accounts.

2. 1,855,945 accounts were referenced in the list of followings of other accounts, but the

[5]The public information returned by the API call we make is described in this URL `https://dev.twitter.com/overview/api/users`. We note that the history of the published tweets is not part of it.

API lookup did not return any profile information for these referenced accounts. Then we tried to perform further API lookups for these referenced accounts, and we obtained profile information for only 137,899 (7.43%) of them. For the rest, the API lookups did not return any profile information. We guess that these accounts were either deactivated [6] during the crawl or suspended by Twitter because these accounts violated Twitter's terms of use. Users can reactivate their account at any time during 30 days after deactivation, so we guess that the observed 7.43% have reactivated their accounts.

3. For 5,938 accounts, we did not crawl the list of followings because the API consistently returned an error code. We counted the number of followings for such accounts as 0.

4. 1,180 user accounts were lost because our archives with data were partially corrupted due to a system bug on two crawling machines.

The number of followings and followers for each account can be obtained in two ways. Either we get these values from an API call, or we compute them based only on the list of followings for each account. We use the latter to build our social graph, so we cross-validated the number of following and followers using the latter method with the former one. We see in Figure 3.1 that there is no difference between the numbers of followers (resp. followings) returned by the API and the number of followers (resp. followings) in the social graph we computed for 69.14% (resp. 78.79%) of the collected accounts. The difference observed for the other accounts is due to three different reasons. First, our graph does not include protected accounts and their incoming and outgoing arcs, so the number of following and followers in the computed graph is smaller than from the API, which explains that we observe a higher number of positive differences in Figure 3.1. Second, there is a delay between the time the account information was crawled and the time the list of followings was crawled because of the implementation of the crawler described in Section 3.2.1. This delay of 9 hours on average (9.5 minutes median) causes a difference in the number of followings reported by the API and the number of followings obtained by computing the social graph, because some arcs might be added or removed during this delay. Third, we crawled all accounts during a four months period. So a given account crawled at time T might be followed (resp. unfollowed) by accounts after time T, accounts that we crawled after they added (resp. removed) the follow links. Thus, there is a larger (resp. smaller) number of followers for this given account in the computed social graph than returned by the API.

3.2.3 Measured Twitter Social Graph

We collected all Twitter accounts, consisting of 537 million accounts at the end date of our crawl in July 2012, and accounts' public information (including account creation date, number of published tweets, number of followings and followers, etc.) We remind that there are 5.97% of all accounts (32 million) that are protected, which means one needs their approval to get the lists of their followings. So we collected the list of all followings for non-protected accounts only, resulting in a social graph with 505 million nodes and 23 billion arcs. The average node in-degree of this graph is 45.6, the median is 1, and the 90th percentile is 33.

[6] How to deactivate your account. https://support.twitter.com/articles/15358

Our dataset is, to the best of our knowledge, the largest and most complete dataset of a social network available today. We also believe that it will be harder in the future to collect such a large and complete dataset. Indeed, companies are taking measures to prevent large crawls of their social networks. For instance, Twitter is no more whitelisting machines. Moreover it has discontinued on June 11, 2013 the API 1.0 that supported anonymous requests and use of already whitelisted machines. The new API 1.1 requires user authentication for each request making crawls harder and longer to perform. For these reasons, we acknowledge that our dataset has value to communities interested in social graphs, and we publicly release it (with precautions described in Section 3.2.4)[7].

3.2.4 Ethical Issues

There are two main ethical issues with large scale measurement studies. First, we need to take care of users privacy. All data collected in this study are publicly available through the Twitter API, the Twitter applications, and the Twitter Web site. In particular, we did not collect any data that is not publicly available, or did not work around any protection mechanisms.

Second, we need to respect Twitter terms of use. We used the regular Twitter API to perform our crawl. We made half of our crawl using machines whitelisted by Twitter, and half of the crawl using a distributed crawler which used the regular Twitter API and conformed to its rate constraint. On average, we generated from the distributed crawler around 20 requests per second to the API, a rate of requests we believe to be negligible for the Twitter infrastructure.

We release our dataset that consists of the Twitter social graph in the format of an adjacency list. In the released dataset each account ID is anonymized.

3.3 Graph Analysis Methodology

We start discussing the motivation and insights behind the analysis of the macroscopic structure—henceforth called the macrostructure—of the Twitter social graph. There is a fundamental difference between directed social graphs such as Twitter and other directed graphs such as the Web. In a directed social graph, not only the links among accounts show the influence of accounts, but they also constrain the propagation of information. Therefore, unveiling the macrostructure of a social graph sheds light on the highways of information propagation.

However, it is a challenge to extract a macrostructure on a social graph of the size of Twitter. The intuition behind our macrostructure analysis is the following. We want to understand how the Twitter graph constrains the flow of information. Therefore, we start by identifying all the strongly connected components (SCCs) that are components with a directed path between any two nodes. In such components, the information can freely circulate, so we abstract each of these components by a single node. After this stage, we obtain a directed acyclic graph (DAG) that is half of the size of the original graph (in terms of number of nodes), still too large to be analyzed. Consequently, the next stage is to group nodes in this DAG based on their connectivity to the largest SCC. As discussed in the following, the

[7] http://j.mp/soTweet

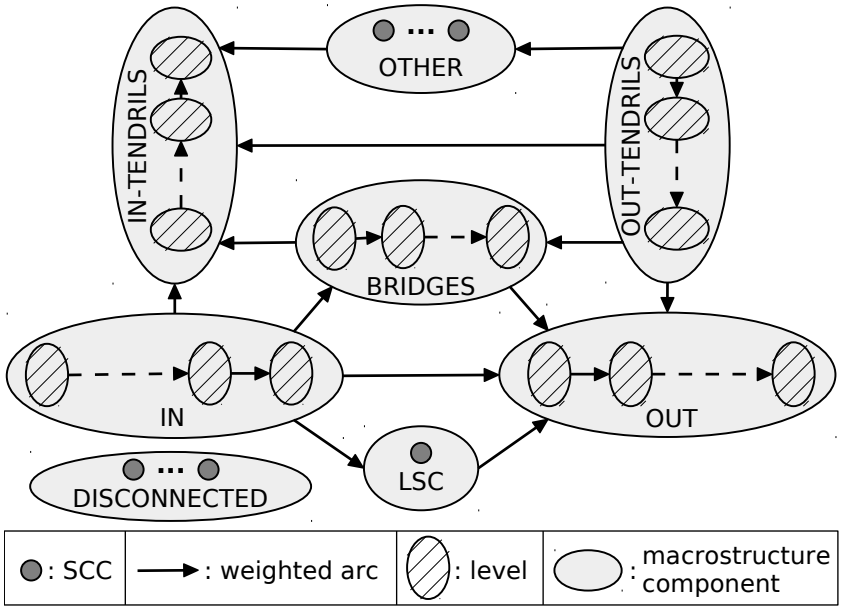


Figure 3.2 – **Macrostructure of any directed graph.**

largest SCC represents roughly half of the nodes. This is large and there is undoubtedly an interesting analysis to make on this component, but we keep this analysis for future work and focus in this chapter on the macrostructure. After this stage, we have 8 components representing a tractable graph. We now describe the details of this process.

We compute the macrostructure of the Twitter social graph in two stages. In the first stage, we use the Tarjan algorithm [Tarjan 1971] to compute the SCCs of the Twitter social graph. Then, we replace each SCC with a single vertex, and the multiple arcs between any two vertices with a weighted arc of weight equal to the number of arcs it replaces. As a result, we obtain a directed acyclic graph.

In the second stage, to uncover the macrostructure of the directed acyclic graph shown in Figure 3.2, we use the following procedure. We first identify the Largest Strongly Connected (LSC) component, the component with the largest number of original nodes. From this LSC component, we run a breadth first search (BFS). We define the set of vertices we find to be the OUT component, that is the set of nodes with a directed path from the LSC component. Inside the OUT component we distinguish *levels* (shown as hatched ellipses on Figure 3.2). Each level is a bin of SCCs that have the same distance from the LSC component. Then we run a reverse BFS from the LSC component and define the set of vertices we find to be the IN component which is a set of nodes with a directed path to the LSC component. Similarly to OUT we distinguish levels inside the IN component based on the distance to the LSC component. Next, we perform a BFS starting from the IN component and a reverse BFS from the OUT component, reachable nodes that were not yet in the LSC, IN or OUT components were identified as IN-TENDRILS and OUT-TENDRILS respectively. Inside the tendrils we can also identify levels depending on the distance to the components these tendrils are growing from. We separated nodes that were identified as both IN-TENDRILS and OUT-TENDRILS into the BRIDGES category that consist of accounts connecting the IN and OUT bypassing the LSC component, we can also distinguish levels based on the distance to OUT

and distance to IN. After that we put the nodes that were not categorized on previous steps into the OTHER category when there is an undirected path from them to categorized nodes or to DISCONNECTED category otherwise. All the possible arcs between the components of the macrostructure are shown on Figure 3.2.

The methodology we describe and the macrostructure representation is inspired from the work of Broder *et al.* [Broder 2000] in the context of the Web for 203 million Web pages. However, our methodology is significantly different from the one presented by Broder *et al.* Indeed, unlike our methodology that is exhaustive, they used a small random sample of 570 nodes from the LSC component to find other components. This difference in methodology has two important consequences. First, we perform a complete and accurate classification of all accounts, which is not possible with the methodology of Broder *et al.*, a methodology only intended to show the macrostructure, but not to accurately classify accounts. Second, the macrostructure we describe is more detailed and accurate. In particular, unlike Broder *et al.*, we identified a new component called OTHER, the structure of levels within components, links between components, and the exact number of such links.

In addition, insight we can get from unveiling the macrostructure of the Web is very different from the insight we can get from unveiling the one of a directed social graph such as Twitter. Indeed, the Web forms a directed graph and the arcs among Web pages are hypertext links. Therefore, the directed graph of the Web represents the paths to access Web pages, but no information propagates along the arcs of the graph. On the contrary, the directed graph of Twitter consists of the follow relationship among accounts. Each tweet published can only propagate along the paths of this graph. Therefore, whereas the notion of content propagation is irrelevant in the context of the Web graph, it is central in the context of the Twitter graph.

In summary, we present a method to compute the macrostructure of any directed graph. Figure 3.2 is not specific to Twitter and can be applied to any directed graph, and in particular to social graphs, where the components group together accounts with different roles in the social graph. This representation is, to the best of our knowledge, the first attempt to extract *exhaustively* a macrostructure of a large social graph, such as the one of Twitter, taking into account the connectivity of accounts in this graph. In Section 3.4, we will discuss the role of the Twitter accounts, depending on the component they belong too.

3.4 The Macrostructure of Twitter in July 2012

Exploring the macrostructure of the Twitter social graph is interesting because it sheds light on how information propagation is constrained. However, this macrostructure would be even more interesting if we can map specific usages of Twitter to components in this macrostructure. Unraveling a correlation between accounts usages and the macrostructure will improve the understanding of how Twitter is used.

In this section, we dissect the macrostructure of the Twitter social graph, focusing on regular, abandoned, and suspicious accounts. i) **Regular accounts** are by definition accounts that are neither abandoned nor suspended. Such accounts show the regular activity on Twitter. ii) **Abandoned accounts** are accounts with few followers and followings, and no recent tweet activity. Such accounts are important to understand Twitter adoption and to

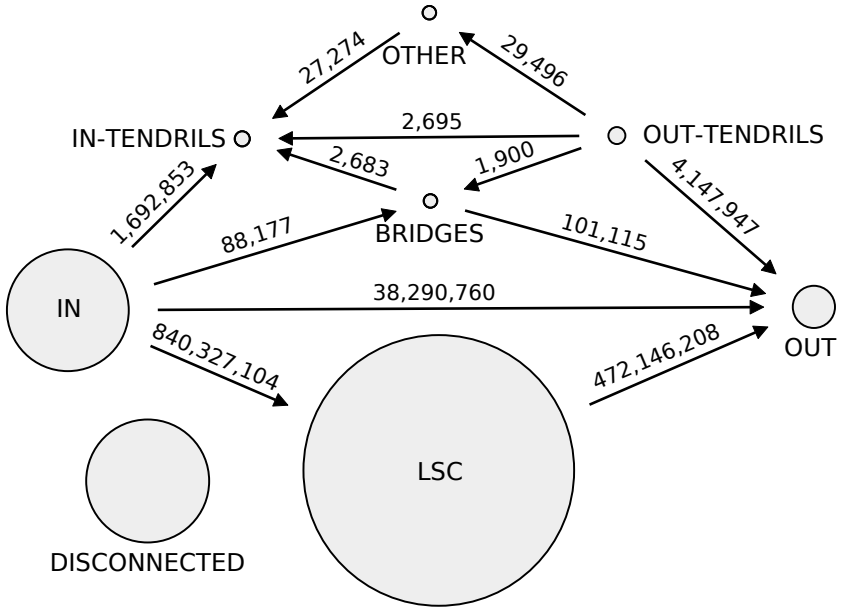


Figure 3.3 – **Macrostructure of Twitter in July 2012.** *The size of the circles is proportional to the number of accounts in components. The labels on arrows give the number of arcs between components.*

accurately quantify the bias when analyzing Twitter, bias due to these accounts that do not take part in any social activity. iii) **Suspicious accounts** are often suspended by Twitter because they infringed its terms of use. We checked that most suspended accounts show evident signs of malicious activity (bunch of sequentially generated accounts, accounts' user name generated with automatic patterns, etc.). There is no ground truth for the malicious activity, but the notion of suspended accounts is a reasonable metric to detect (in retrospect) malicious accounts [Thomas 2011]. For the purposes of our study we have recrawled a set of 1 million random users from our dataset on May 6, 2013 to check if they are still active. In the rest of the chapter, we refer to the number of suspended accounts as the number of accounts for which Twitter returned the 'suspended' status during this recrawl.

Figure 3.3 shows the macrostructure of Twitter computed with the methodology presented in Section 3.3. We identify 8 components in this graph, with 4 of them (LSC, OUT, IN, DISCONNECTED) representing 98.96% of all Twitter accounts; so we focus on them.

The LSC (Largest Strongly Connected) component is the core of the regular Twitter activity. Indeed, according to Table 3.1, the LSC component contains 96.95% of the 10,000 most followed accounts, 100% of the 10,000 accounts that follow the most, 88.66% of the 10,000 accounts that tweet the most, 94.28% of the 2.91 million experts identified by Sharma *et al.* [Sharma 2012] as influential accounts in their field, and 97.01% of the verified accounts [8] that are accounts of highly sought users (in music, acting, politics, etc.) that Twitter verified to be authentic. In addition, Table 3.2 shows that more than 96% of the following and follower links, and 98.05% of the tweets are for accounts in the LSC.

However, it is wrong to believe that the LSC component is the only one that matters when studying Twitter, other components contain a lot of accounts with specific roles in the

[8] FAQs about verified accounts. https://support.twitter.com/groups/31-twitter-basics/topics/111-features/articles/119135-about-verified-accounts

Component	Top followed (%)	Top following (%)	Top tweeting (%)	Experts (%)	Verified (%)	Suspended (%)
LSC	96.95	100	88.66	94.28	97.01	1.17
OUT	3.05	0	10.79	1.33	2.99	0.43
IN	0	0	0.07	0.01	0	1.77
DISC.	0	0	0.47	0.01	0	5.11
OUT-T.	0	0	0	0	0	0.18
IN-T.	0	0	0	0	0	0.49
BRID.	0	0	0	0	0	0
OTHER	0	0	0.01	0	0	1.25

Table 3.1 – **Distribution of noteworthy accounts among components.** *The first three columns represent the 10,000 accounts with the largest number of followers, followings, and tweets for the entire Twitter social graph. The fourth column represents the 2.91 million experts identified by Sharma et al. [Sharma 2012] as influential users in their field (the sum of this column is not 100% because 4.37% of the experts are not present in our dataset, most likely because they closed their account, or have been suspended). The fifth column represents the accounts verified by Twitter. The last column represents the percentage of suspended accounts.*

Twitter ecosystem. We see in Table 3.2 that the LSC contains only 50.71% of all accounts. This is surprising because it is easy to be part of the LSC component, an account only needs one following and one follower already in the LSC component. Also, we observe that a large fraction of the suspicious activity in Twitter is outside of the LSC component, as we see in Table 3.1 (last column). Finally, when looking at the percentage of accounts with no follower, no following, or no tweet, we see in Table 3.3 that each of the four main components has fundamentally different characteristics. Indeed 92.97% of the accounts in the OUT component have no following, 96.13% of the accounts in the IN component have no follower, and almost all accounts in the DISCONNECTED component have no following and no follower. Moreover, at least 60% of the accounts in these three components never sent any tweet, whereas it is only 23.87% for the LSC.

In summary, we see that even if most of the regular Twitter activity is in the LSC component, other components contain half of the Twitter accounts and present characteristics worth studying. In the following, we dig into each component to discuss its main characteristics.

3.4.1 LSC Component

We have seen that most of the regular Twitter activity is in the LSC component. However, due to the simplicity to belong to the LSC component, many abandoned and suspicious accounts also belong to it.

	Arcs (%)		Tweets (%)	Accounts (%)
	followers	followings		
LSC	98.01	96.13	98.05	50.71
OUT	1.96	0.02	1.49	5.30
IN	0.02	3.83	0.25	21.36
DISC.	<0.01	<0.01	0.21	21.60
Others	<0.01	0.02	<0.01	1.03
Total		23×10^9	127×10^9	505×10^6

Table 3.2 – **Distribution of the arcs, tweets and accounts per component.** *At the scale of the entire Twitter social graph, there is the same number of followings and followers, because they represent the same notion of arc. But, for each component, the number of followings and followers might be different due to the ingress and egress arcs, so we make a distinction between followings and followers for each component.*

Component	No follower (%)	No following (%)	No tweet (%)
LSC	0	0	23.87
OUT	0	92.97	61.82
IN	96.13	0	60.10
DISCONNECTED	99.63	99.63	79.31
OUT-TENDRILS	99.13	0	73.20
IN-TENDRILS	0	98.78	70.40
BRIDGES	0	0	67.34
OTHER	51.39	46.67	67.56

Table 3.3 – **Percentage of accounts with no follower, no following or no tweet per component.**

3.4.1.1 Abandoned Accounts

Most accounts with one following and one follower in the LSC are abandoned accounts. We see in Figure 3.4 (top, solid line) that there are 4.18% of accounts in the LSC component with one following and one follower. In addition, out of the accounts with one following and one follower in the LSC component, 86.34% are more than 6 months old and 59.57% never sent any tweet.

In summary, a large fraction of accounts in the LSC component with one following and one follower did not have any change in their number of followings and followers for months and did not send tweets recently. Considering that it is unlikely that such accounts will actively follow a single other account for month (so no serious follow activity) without tweeting anything (so no publishing activity), it is reasonable to believe that these accounts are abandoned.

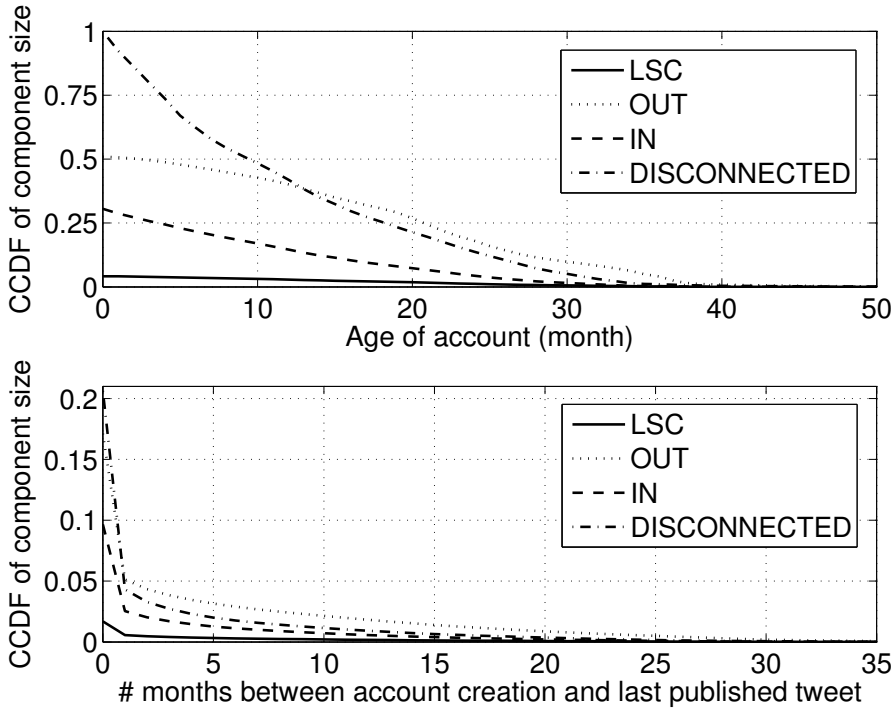


Figure 3.4 – **Characterization of abandoned accounts.** *(top) Identification of old abandoned accounts. CCDF of accounts with at most one follower and one following in a component according to the account creation date. (bottom) Characterization of accounts who published at least one tweet. CCDF of the duration between the creation date of an account and the date of its last published tweet for accounts with at most one follower and one following.*

3.4.1.2 Suspicious Accounts

The LSC component also contains suspicious accounts. We present in Table 3.4 the percentage of suspended accounts per component for five outlier categories. An *outlier* account is followed, following, or tweeting much more than a regular account, thus it is a good candidate for suspicious activity. The first three columns represent accounts with the largest number of followers, largest number of followings, and largest number of tweets. The fourth and fifth columns are for the accounts with the largest number of followings and tweets, but with at most one follower. We consider this notion of outliers because following a lot of accounts is a known technique used by spammers [Thomas 2011]. To reduce the impact of spammers, we remind that Twitter imposes a limit of 2,000 followings for accounts with no follower, and then a linear increase with the number of followers. Accounts close to this limit and with at most one follower are more likely to be spammers. The last column is for accounts that send the largest number of tweets, but with at most one follower. This is also a suspicious behavior, because it is strange to send a lot of tweets if nobody (or a single other account) follows them. Spammers can send a lot of tweets to interfere with trending topics or the Twitter search functionality, and to direct messages to a specific user using *@mentions* [Thomas 2011].

Considering the huge number of suspicious accounts, we cannot afford to manually inspect all of them. Therefore, we consider a suspicious account to be malicious if it was suspended by Twitter, see Table 3.4.

Component	Top followed (%)	Top following (%)	Top tweeting (%)	Top following with ≤ 1 follower (%)	Top tweet with ≤ 1 follower (%)
LSC	0.33	1.15	1.99	97.83	3.02
OUT	1.15	10.30	5.20	0.45	5.26
IN	2.78	96.87	3.87	96.87	3.89
DISC.	1.38	1.33	7.43	2.84	7.48

Table 3.4 – **Percentage of suspended accounts (on the 6th of May 2013) per component for 5 outlier categories.** *The first three columns represent the 10,000 accounts with the largest number of followers, followings, and tweets for the entire Twitter social graph. The fourth column is for the 10,000 accounts with the largest number of followings and at most one follower. The last column is for the 10,000 accounts with the largest number of tweets and at most one follower.*

As expected, the top followed accounts in the LSC component are regular, only 0.33% have been suspended. Indeed, it is complex to manipulate the number of followers, because it requires to either manipulate other accounts in order to incite them to follow, or to create fake accounts whose only one goal is to follow. More surprising, the top following accounts are also regular for Twitter, only 1.15% have been suspended. We expect accounts that follow a lot of other accounts to be spammers, but, according to Figure 3.5 (bottom), the LSC component is the only one to have accounts that break the limit of 2,000 followings. So the top following in the LSC component also have a lot of followers, thus the low number of suspended accounts.

Then we observe in Table 3.4 two important behaviors that characterize well the outlier activity in the LSC component. First, 97.83% of the top following with at most 1 follower have been suspended. This means that most of the accounts in the LSC component close to the limit of 2,000 followings are malicious. Second, only 3.02% of the top tweeting accounts, but with at most a single follower have been suspended. The rest looks like regular for Twitter. By manually inspecting these accounts that looks regular for Twitter, we found bots used as an interface to job forums, news site, Yahoo!Answers, YouTube published videos, etc. So, it seems that Twitter is used by developers to generate a stream of data collected from third party Web sites. As these accounts have only one follower, we guess that they are either used for tests only, or that the developers are using a Twitter widget to embed their account timeline into a Web site.

3.4.2 OUT Component

The OUT component represents all Twitter accounts with a directed path from the LSC component. In addition, these accounts can also have directed paths from other components, but no account in OUT can have a directed path to any other component (directed paths among OUT accounts are possible, so if an OUT account has following links, they necessarily

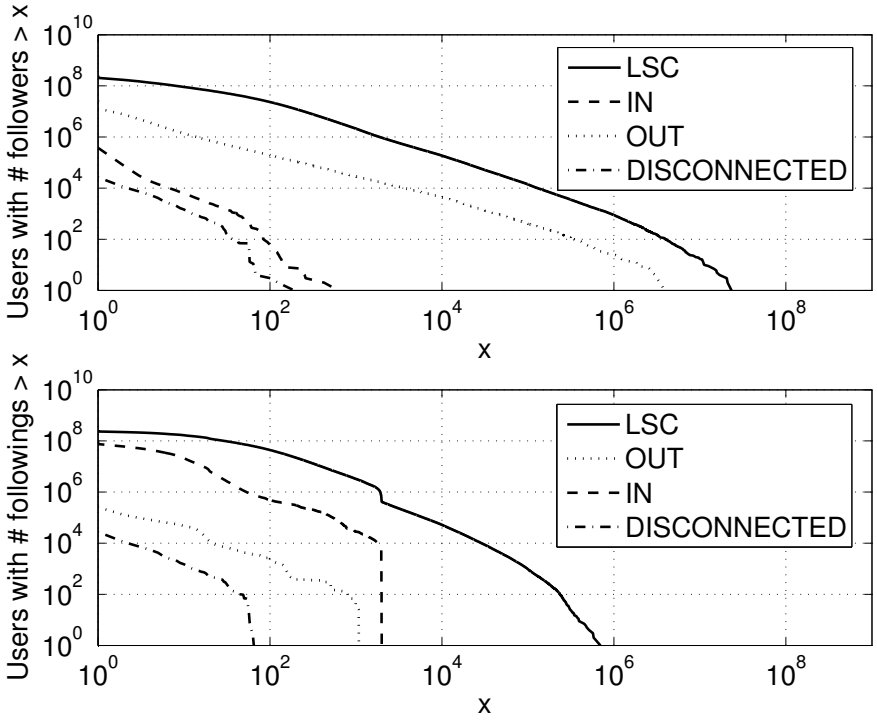


Figure 3.5 – **Distribution of followers (top) and followings (bottom) by category.** *Accounts with no follower (top) and no following (bottom) are filtered out (see Table 3.3)*

come to other OUT accounts).

3.4.2.1 Regular Accounts

A specificity of the OUT component is that a small set of accounts (belonging to celebrities) attract most of the follower links for this component. These are regular OUT accounts. We see in Figure 3.3 that more than 500 million links between components are directed to OUT, 37.93% of all inter-components links, whereas the OUT component represents only 5.30% of all accounts. Also, we see in Table 3.2 that accounts in OUT presents 1.96% of all follower links, which make it the second component with the largest number of follower links (we sum all follower links for all accounts in a given component). Among the 100 accounts that have the largest number of followers, we found that there are 35 verified accounts representing 12% of the arcs from the LSC to OUT. These accounts are owned by celebrities that belong to the OUT component because they do not follow any other account.

We observe another interesting specificity of the OUT component in Figures 3.4 (top) and Figures 3.6. The OUT component is the only one to show an inflection point for both curves around 20 months, meaning that the proportion of recent accounts in the component is lower than for other components. To explain this inflection point, we need to characterize the kind of accounts that stay in the OUT component. According to Table 3.3, 92,97% of the OUT accounts have no followings, but they all have at least one follower because they belong to the OUT component. These accounts are what we call selfish (they are not interested in tweets from other accounts), a decreasing trend in Twitter in the past two years. We will discuss further this trend in Section 3.5.3.

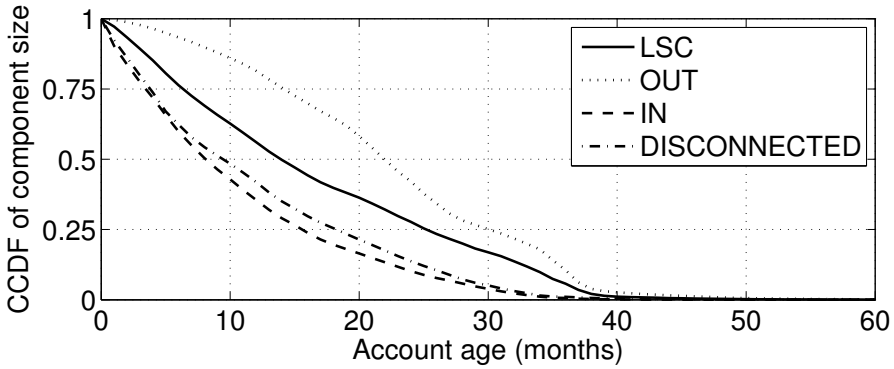


Figure 3.6 – **Age of accounts in each component.** *CCDF of accounts in a component according to the account creation date.*

3.4.2.2 Abandoned Accounts

As we discussed in Section 3.4.1, most accounts with at most one following and one follower are also abandoned accounts for the OUT component. We see in Figure 3.4 (top) that 50.94% of the accounts have at most one following and one follower, and 40.11% are more than 1 year old. We see in Figure 3.4 (bottom) that 82.89% of the accounts with at most one following and one follower never sent a tweet, and that only 5.13% of the accounts with at least one following and one follower sent a tweet more that 1 month after their creation date. This is a consequence of the *Find friends* feature available in Twitter that allows users to search their entire contact lists for Twitter accounts. By default, once the search is done, all accounts are checked to be followed. As a consequence, we observe many accounts in the LSC component that followed abandoned accounts in the DISCONNECTED component, making these abandoned accounts move to the OUT component.

3.4.2.3 Suspicious Accounts

There are fewer malicious accounts in the OUT component than for other components. We see in Table 3.4 that the percentage of suspended accounts for outlier accounts is low for the OUT component. We explain the low number of suspended accounts for the top followings because no account reaches the limit of 2,000 followings, and that few accounts have more than a hundred followings, see Figure 3.5 (bottom). As long as an account in OUT follows an account in the LSC, it belongs to the LSC, so spammers using following links to spam are likely to escape in the LSC component. We explain the low number of suspended accounts for the top tweeting with at most one follower because (as for the LSC component) most of these accounts are operated by bots. Finally, we see in Table 3.1 that out of the 4 main components, OUT is the component with the smallest number of malicious accounts.

3.4.3 IN Component

The IN component is much different from the two previous ones because accounts in this component have few followers (see Figure 3.5, top) and the distribution of the number of tweets is very different (see Figure 3.7) from the ones of the LSC and OUT components. The IN component contains the second largest fraction of abandoned and suspicious accounts,

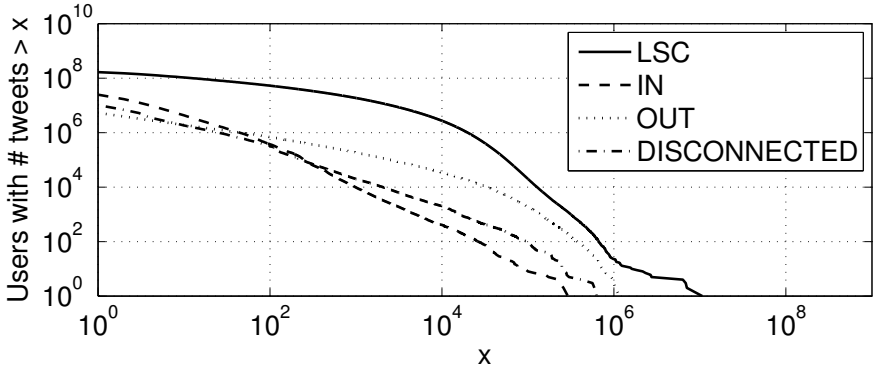


Figure 3.7 – **The distribution of number of tweets by component.** *Accounts with no tweets are filtered out (see Table 3.3).*

after the DISCONNECTED component.

3.4.3.1 Regular Accounts

The regular users for the IN component are passive followers, that are accounts who follow accounts in the LSC, but never publish tweets and are not followed. Indeed, in Tables 3.2 and 3.3 we see that the IN component contains 21.36% of all Twitter accounts, but 96.13% of them have no follower, and 60.10% of them published no tweet (we remind that accounts with followers in IN are followed by other accounts in IN only). This component consists of accounts who follow accounts in the LSC (99.6%) or an account in IN (0.4%). We will see in Section 3.5.3 that the trend of accounts to be passive followers on Twitter (that is, belong to IN component) has been growing since 2009.

Many accounts belonging to the IN component move to the LSC component. We see in Figure 3.4 (top) that 30.56% of the accounts in the IN component have at most one following and one follower, but that only 14.61% are more than one year old. So even if few of them have been tweeting close to the creation date of their accounts (see Figure 3.4, bottom), it is likely that they moved to the LSC component and tweeted from it. Indeed, we see in Table 3.1 that only 1.77% of the accounts in the IN component have been suspended, but that accounts are much younger in the IN component than in the LSC and OUT components, see Figure 3.6.

3.4.3.2 Abandoned Accounts

Whereas 96.13% of the accounts in the IN component never published any tweet (see Table 3.3), the fraction of abandoned accounts is much lower in this component than in the OUT and DISCONNECTED components. Indeed, we see in Figure 3.4 (top) that only 30.56% of the accounts in the IN component have at most one following and one follower, and 20.88% have at most one following, one follower, and never published any tweet. Moreover, according to Figure 3.5 (bottom), 23.04% of the accounts follow at least 10 other accounts, thus a passive follower activity.

3.4.3.3 Suspicious Accounts

The IN component contains many malicious accounts among the outliers. We see in Table 3.4 that 96.87% of the accounts with the largest number of followings are suspended. We note that all top followings have at most 1 follower in this component. There is also 3.87% of the accounts that tweeted the most that were suspended. For the rest, after manual inspection, we also found, as for the two previous components, that they are used by bots.

Finally, the IN component has a very interesting property for people looking for a reliable metric to assess influencers. Cha *et al.* [Cha 2010] show that the number of followers is not a reliable metric, because users perform link farming [Ghosh 2012] to increase their number of followers. However, this is a rare problem in the IN component. Indeed, accounts in the IN are clearly not interested in increasing their number of followers (see Figure 3.5, top) thus the accounts they follow will not be biased by this problem. Evaluating the benefit of considering accounts in the IN to assess influencers is an interesting problem for future work.

3.4.4 DISCONNECTED Component

Accounts in the DISCONNECTED component, like in the IN one, have few followers (see Figure 3.5, top) and the distribution of their number of tweets is very different (see Figure 3.7) from the ones of the LSC and OUT. The DISCONNECTED component contains the largest fraction of abandoned and suspicious accounts. There are almost no regular users in this component.

3.4.4.1 Abandoned Accounts

A specificity of the DISCONNECTED component is that is contains a lot of abandoned accounts. In spite of being the second largest component with 21.6% of all accounts (see Table 3.2), 78.94% of accounts in the DISCONNECTED component have no followers and no followings, and never published any tweet. Furthermore, 72.44% of accounts in DISCONNECTED component are older than one month. Therefore, we can conclude that the DISCONNECTED component has, by far, the largest number of abandoned accounts. We see in Fig. 3.4 (top) that 99.97% of its accounts have at most one following and one follower, but only 41.93% of them are older than 12 months. Like for the IN component, many account in the DISCONNECTED component are recent (see Figure 3.6), thus some accounts in this component have moved to another component.

3.4.4.2 Suspicious Accounts

Finally, we see in Table 3.1 that the DISCONNECTED component contains the largest fraction of malicious accounts, but we don't observe in Table 3.4 an outlier category grouping them. Indeed, most accounts have no followings, no followers and no tweets, so the number of outliers is much smaller than our sample size.

In summary, the DISCONNECTED component hosts a lot of abandoned accounts and a large fraction of the malicious activity on Twitter, it is also a transitional place for new accounts before they migrate to another component.

3.4.5 Other Components

The smallest components, IN-TENDRILS, OUT-TENDRILS, BRIDGES, and OTHER represent 1.03% of all accounts. Most accounts in these components are either accounts created for test, or new accounts that will migrate to another component after some time. We do not discuss deeper these components as their impact on the Twitter social graph is small compared to the 4 main components.

3.4.6 Discussion

We can draw several important lessons from the results discussed in this section.

First, the macrostructure of the Twitter social graph significantly constrains the propagation of information. Therefore, models of information propagation in social networks might lead to wrong results when abstracting the underlying social graph. This work sheds light on how to correctly abstract the social graph, and because the macrostructure is reasonably simple, with 3 main components with active accounts, we believe it is possible to model the underlying graph constraint.

Second, we identify a correlation between components in the macrostructure and the usage of accounts in these components. This result challenges the sampling techniques that follow arcs (such as random walks or bi-directional breadth first search) because the statistical validity of the sample might be low. For instance, all sampling techniques following arcs that start from well connected (or active) accounts will miss all of the malicious activity located in the DISCONNECTED component.

Last, the identification of the role of accounts in each components is important to understand who are the influencers in Twitter. For instance, as discussed in Section 3.4.3.3, users try to increase their popularity in Twitter by either offering reciprocation to the users that accept to follow them, or by buying follower links on the black market. Therefore, we can identify real influencers by focusing exclusively on the followers in the IN component, removing suspicious accounts by filtering out all accounts younger than, e.g., six months.

3.5 Evolution of the Macrostructure of the Twitter Social Graph with Time

In this section, we discuss the evolution of the macrostructure of the Twitter social graph with time from January 2007 to July 2012. To present this evolution, we first describe the estimation technique we use to estimate the Twitter social graph in the past. Then we validate our technique using two public datasets collected in 2009 [Kwak 2010, Cha 2010]. We discuss the evolution of the macrostructure of the Twitter social graph with time and explain how new accounts led to the time evolution we observed, shedding light on the evolution of the usage of Twitter in the past 6 years.

3.5.1 Methodology to Estimate the Macrostructure

The evolution with time of the macrostructure of the Twitter social graph is interesting, because it shows the evolution of the Twitter usage. We have seen in Section 3.4 that

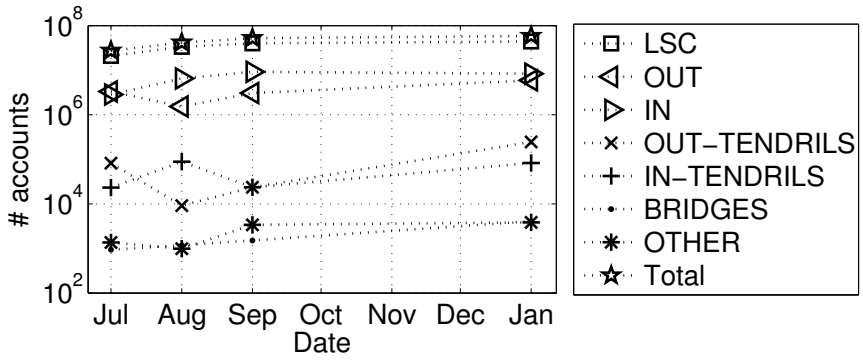


Figure 3.8 – **Comparison of our estimated graphs of 2009 (labeled Jul and Jan) with two existing Twitter datasets made in August [Kwak 2010] and September [Cha 2010] 2009.** *Our simple methodology gives an approximation of the macrostructure of the Twitter social graph that is consistent with existing datasets.*

components represent specific categories of usage. However, the Twitter API does not give access to the past social graph of Twitter.

We propose a simple approach to approximate the macrostructure of the Twitter social graph. The dataset we describe in Section 3.2.3 covers all Twitter accounts in July 2012 (with the limitation described in Section 3.2.2), and for each account we have the creation date. To approximate the macrostructure of the Twitter social graph at date D, we remove from our dataset all accounts created after this date, and all arcs to and from these accounts. Then, we use the methodology described in Section 3.3 to compute the macrostructure of the resulting graph at date D.

This simple methodology has two important limitations. First, we do not have any suspended and deactivated accounts in our dataset. Accounts are suspended by Twitter because they infringed the terms of use, most of the time they are spammers. Deactivated accounts have been closed by users themselves. We believe such accounts, when they were still active, had a small impact of the Twitter social graph. Second, the Twitter API does not give access to the arc creation date[9]. Therefore, we assume that all arcs between any two accounts in July 2012 existed at date D as long as the two accounts existed at this date; equivalently we assume that if there is an arc between two accounts, it is created close to the creation date of the youngest account. We are aware that, as reported by Kwak *et al.*, the creation of arcs among accounts is more complex than our simple approximation [Kwak 2011a]. However, our goal is to understand the evolution of the macrostructure of the Twitter social graph with time, not the fine grain evolution of arcs between accounts. For this reason, we believe that our approximation is reasonable.

Moreover, to validate this approximation on creation dates of arcs, we compare our approximation with two datasets collected in 2009 [Kwak 2010, Cha 2010]. Kwak *et al.* [Kwak 2010] and Cha *et al.* [Cha 2010] independently collected two Twitter datasets in August 2009 and September 2009 respectively and used different methodology. Kwak *et al.* used a technique close to a BFS and reverse BFS from a popular account and also collected

[9]For a given account, Meeder *et al.* observed that the 1.0 Twitter API returned the arcs in an order that was the reverse order of creation of the arcs for this account [Meeder 2011]. Our recent experiments with the Twitter API have shown that it is no more possible to rely on this ordering property.

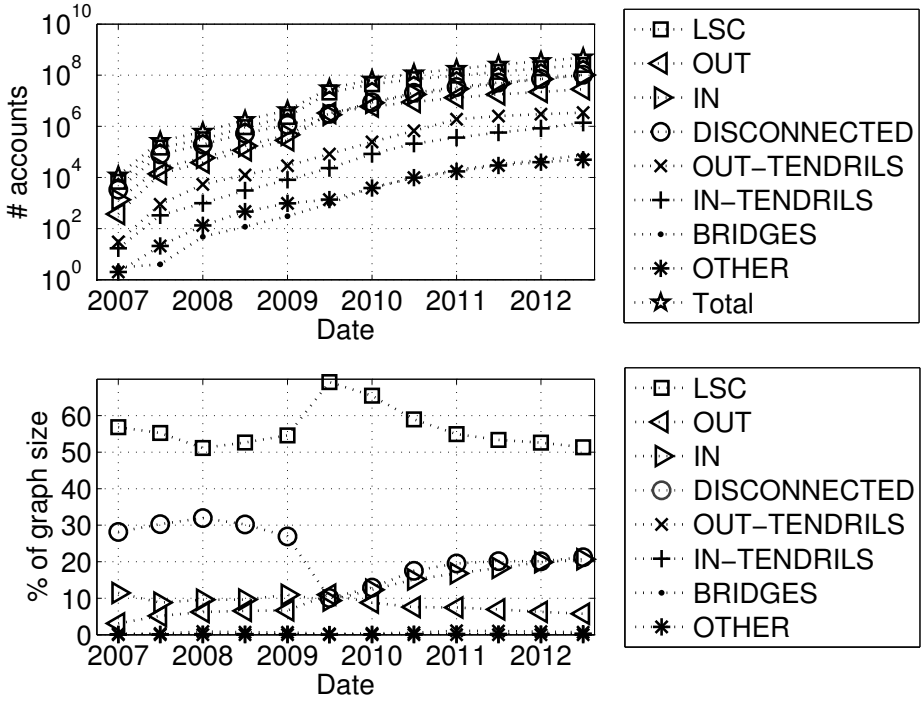


Figure 3.9 – **The estimated evolution of the macrostructure of Twitter with time.** *(top) Sizes of components in log scale. (bottom) Sizes of the same components as a percentage of the size of the graph.*

accounts referring to trending topics (so active accounts only), and Cha *et al.* used a crawl by account ID (as we did). For each of the two datasets we computed the Twitter macrostructure according to the methodology described in Section 3.3, and we approximated the macrostructure of Twitter using our dataset in July 2009 and January 2010. We show in Figure 3.8 the result of this validation: the order of the size of each components is consistent between the two validation datasets and our dataset. In addition, we have compared the dataset by Kwak *et al.* with our closest estimation (July 2009). We found that 88.25% of the users common to both datasets belong to the same components in both datasets. We cannot make such a validation for the second dataset because Cha *et al.* have anonymized it. In summary, the dynamics of the creation and deletion of arcs is complex [Kwak 2011a], but we have shown that our simple approximation is reliable enough for the purpose of our macrostructure study.

There is no DISCONNECTED component in Figure 3.8, because this component is missing in the two validation datasets. Either the methodology did not permit to crawl accounts in this component [Kwak 2010], or these accounts were filtered out in the published dataset [Cha 2010]. We observe in Figure 3.8 some small variations for the OUT, IN-TENDRILS, and OUT-TENDRILS components between the two validation datasets and our dataset. These variations can be explained by a major change in the Twitter macrostructure that happened in 2009. We discuss further this change in the next section.

3.5.2 Evolution of the Macrostructure

To observe the evolution of the Twitter social graph with time, we approximate its macrostructure using the simple methodology discussed in Section 3.5.1 every six months from January

1, 2007 to July 1, 2012. The first account on Twitter was created on March 21, 2006, but due to the small number of accounts created between March and July 2006, we decided to skip the macrostructure of the Twitter social graph in July 2006 and start our analysis in January 2007.

We see in Figure 3.9 (top) the evolution of the size of each component with time, confirming that the LSC, OUT, IN, and DISCONNECTED have always been the largest components in Twitter. However, by looking at the size of each component normalized with the graph size in Figure 3.9 (bottom), we observe an interesting change in proportion of macrostructure components in 2009.

Before 2009, the proportion of the DISCONNECTED component was around 30%, the IN component was stable in size, and the size of the OUT component was increasing. The real public adoption of Twitter started in 2009 where the total number of accounts went from 4.265 million in January 2009 to 67.487 million in January 2010. Several events contributed to attract new users on Twitter during that period: the terrorist attacks in Mumbai was one of the first event followed on Twitter in November 2008, attracting the attention of other news media such as CNN; some influential celebrities started to use Twitter such as Oprah Winfrey, and, for the first time, some accounts reached one million followers.

We see in Figure 3.9 (bottom) that the large adoption of Twitter in 2009 led to changes in the macrostructure of its social graph. The proportion of the DISCONNECTED component dropped to 10% while the LSC jumped to 70%. We have seen in Section 3.4 that the DISCONNECTED component corresponds to abandoned accounts, so during such a large adoption phase, the proportion of abandoned accounts is much lower. However, this proportion increased in 2010 and 2011 to reach a stable value, with the DISCONNECTED component representing around 20% of all accounts.

We also observe in Figure 3.9 (bottom) that the proportion of the OUT component has been decreasing since 2009. The reason is that a large fraction of celebrities joined Twitter in 2009 and 2010. Some of these celebrities created an account to increase their visibility, but never intended to follow other accounts, thus they joined the OUT component. The fraction of such celebrities is decreasing compared to regular accounts, and also joining Twitter without following anyone in the LSC component is a decreasing trend. Indeed, the proportion of the IN component has been increasing since 2009, showing that it is an increasing trend to follow accounts in the LSC component without tweeting and being followed.

It is worth noticing that the two most popular Twitter datasets [Kwak 2010, Cha 2010] have been collected in 2009. We have seen that the Twitter social graph macrostructure has significantly changed during the period 2009/2010, calling for a newer dataset such as the one we collected, which is more representative of the actual Twitter social graph. (In fact it may be the last dataset of Twitter social graph of such scale because Twitter hardened the rate-limits to its APIs in 2013, and it is no longer possible to crawl the entire Twitter graph through the API.) We also note that the two datasets of 2009 do not contain accounts belonging to the DISCONNECTED component, unlike our dataset, which is an issue for researchers focusing on malicious activities and abandoned accounts on Twitter.

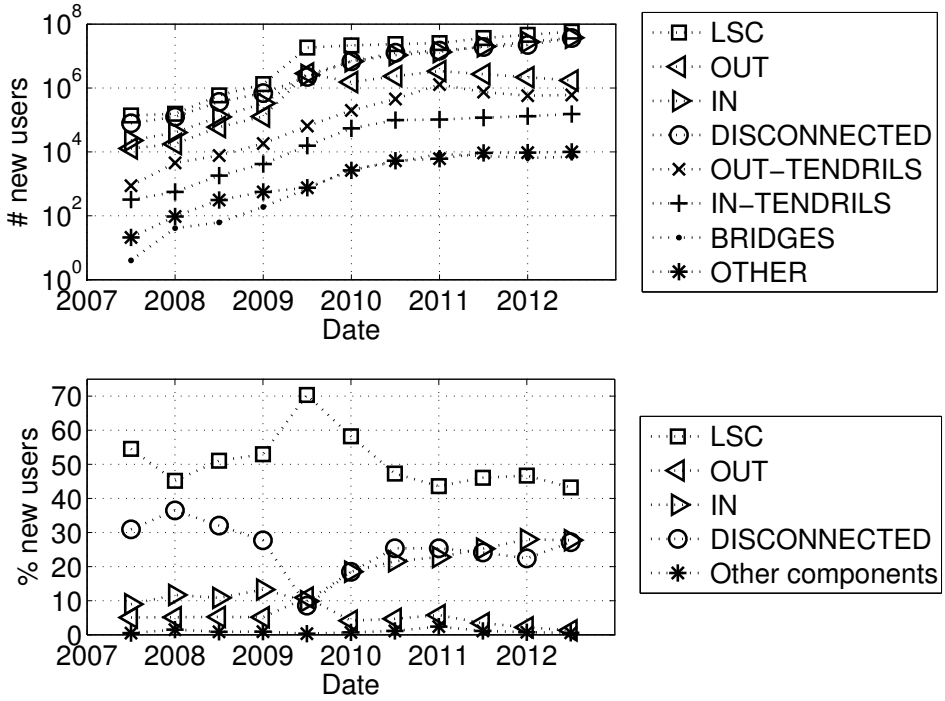


Figure 3.10 – **Distribution of new accounts per components with time.** *(top) Number of new accounts per component. (bottom) Fraction of the total number of new accounts per component.*

3.5.3 Distribution of New Accounts in Components

In this section, we evaluate to which component the new accounts created during each 6 months period belong to. To find this distribution, we use the approximations of the Twitter social graph macrostructure described in Section 3.5.2. Then, for each pair of contiguous approximations in time (e.g., July 2008 and January 2009), we remove all accounts already present in the oldest one to the newest one. This way, we obtain the evolution with time of the distribution of the new accounts in components, see Figure 3.10.

We observe in Figure 3.10 (top) that the total number of new accounts increases with time for the LSC, IN, and DISCONNECTED components, but not for OUT. This decrease confirms our discussion in Section 3.4.2 on the OUT component, explaining that new selfish accounts are decreasing in Twitter.

Figure 3.10 (bottom) shows the fraction of the total number of new accounts per component. We observe that new accounts join most the LSC component, but this is a decreasing trend at the benefit of the IN and DISCONNECTED components. We explain this trend by two changes in the usage of Twitter initiated in 2009. First, passive followers are taking an increasing role in Twitter; passive followers are accounts that follow other accounts, but that are not followed and never publish tweets, as described in Section 3.4.3. This increasing role of passive followers shows that Twitter is more and more used as a regular information media in which people receive information, but do not produce any. However, more than 40% of new accounts are still joining the LSC component, making Twitter the largest and most participative information media. Second, as Twitter is very popular, it is attracting a large fraction of users that are just creating a Twitter account out of curiosity, but never effectively

use it. Most of these accounts end up in the DISCONNECTED component.

3.6 Related Work

Twitter has been widely studied for years. A large fraction of the literature is on the identification of malicious behavior on Twitter [Thomas 2011, Zhang 2011, Ghosh 2012], on the study of tweets propagation [Sadikov 2009, Ye 2010], and on privacy [Humphreys 2010, Mao 2011]. All these studies are not directly related to our work as they do not crawl the Twitter social graph and do not explore its properties.

Closer to our work, several studies focused on the Twitter social graph. Some of them crawled partially the graph before 2009 [Java 2007, Krishnamurthy 2008, Huberman 2008], so before the wide adoption of Twitter. Two studies made a large crawl of the Twitter social graph. Kwak *et al.* used a technique close to a BFS and reverse BFS from a popular account and also collected accounts referring to trending topics. This crawling methodology cannot capture some users that are not connected to the LSC component, and that do not tweet about trending topic, thus a partial view of the Twitter social graph. Cha *et al.* [Cha 2010] used a crawl by account ID, that is close to what we did. Both of these studies made their dataset publicly available and others built on it [Lee 2010, Wu 2011, Cha 2012, Sharma 2012], but the datasets were collected in 2009 during the main change in the Twitter social graph we discussed in Section 3.5.2.

To the best of our knowledge, the dataset we present is the most up-to-date and the most complete description of the Twitter social graph. Moreover, none of these studies explores the macrostructure of the Twitter social graph, a new way to represent directed social graphs. Broder *et al.* [Broder 2000] introduced first the notion of macrostructure for a directed graph in the context of the Web, but we significantly improved it, and we are the first ones to apply it to Twitter. Unlike what Broder *et al.* proposed, we present a methodology to compute the exhaustive macrostructure of any large directed social graph, along with the categorization of each account in the identified component, which is a significant methodological step.

3.7 Discussion

In this chapter, we present the largest, most complete, and most up-to-date crawl of the Twitter social graph. This graph contains 505 million accounts connected with 23 billion arcs. In addition, we present a methodology to practically compute the macrostructure of any directed social graph and to exhaustively classify each account to one of the identified components. We applied this methodology to the Twitter social graph and found that only 50.71% of the accounts belong to the LSC component, and that 21.60% of the accounts (in the DISCONNECTED component) have no path to the other accounts.

We show that the main components of the macrostructure of the Twitter social graph correspond to specific usages. For instance, the LSC component hold most of the regular Twitter activity, and the IN component holds passive followers. Finally, we present a simple methodology to explore the evolution of the macrostructure of Twitter with time, we validate this methodology, and we show that the public datasets crawled in 2009 do not represent the current macrostructure of the Twitter social graph.

We believe that our collected dataset is a gold mine for any researcher working on social graphs and that the macrostructure analysis sheds a new light on the Twitter social graph that will be useful for both theoreticians and experimenters.

3.8 Acknowledgements

We thank Krishna P. Gummadi (MPI-SWS) for insights on the analysis of the dataset we collected and valuable feedback from the early stages of this work. We also thank him for sharing the list of influential Twitter users identified by Sharma *et al.* [Sharma 2012].

Chapter 4

Social Clicks

Contents

Online news domains increasingly rely on social media to drive traffic to their websites. Yet we know surprisingly little about how a social media conversation mentioning an online article actually generates clicks. Sharing behaviors, in contrast, have been fully or partially available and scrutinized over the years. While this has led to multiple assumptions on the diffusion of information, each assumption was designed or validated while ignoring actual clicks.

We present a large scale, unbiased study of *social clicks*—that is also the first data of its kind—gathering a month of web visits to online resources that are located in 5 leading news domains and that are mentioned in the third largest social media by web referral (Twitter). Our dataset amounts to 2.8 million shares, together responsible for 75 billion potential views on this social media, and 9.6 million actual clicks to 59,088 unique resources. We design a reproducible methodology and carefully correct its biases. As we prove, properties of clicks

impact multiple aspects of information diffusion, all previously unknown. (i) Secondary resources, that are not promoted through headlines and are responsible for the long tail of content popularity, generate more clicks both in absolute and relative terms. (ii) Social media attention is actually long-lived, in contrast with temporal evolution estimated from shares or receptions. (iii) The actual influence of an intermediary or a resource is poorly predicted by their share count, but we show how that prediction can be made more precise.

This work was accepted and presented at ACM SIGMETRICS / IFIP Performance 2016 in Antibes Juan-les-Pins, France [Gabielkov 2016].

4.1 Motivation

In spite of being almost a decade old, social media continue to grow and are dramatically changing the way we access Web resources. Indeed, it was estimated for the first time in 2014 that the most common way to reach a Web site is from URLs cited in social media[1]. Social media account for 30% of the overall visits to Web sites, which is higher than visits due to organic search results from search engines. However, the context and dynamics of social media referral remain surprisingly unexplored.

Related works on the click prediction of results of search engines [McMahan 2013] do not apply to social media referral because they are very different in nature. To be exposed to a URL on a search engine, a user needs to make an explicit request, and the answer will be tailored to the user profile using behavioral analysis, text processing, or personalization algorithms. On the contrary, on social media, a user just needs to create a social relationship with other users, then he will automatically receive contents produced by these users. At a first approximation, a web search service provides pull based information filtered by algorithms and social media provide a push based information filtered by humans. In fact, our results confirm that the temporal dynamics of clicks in this case is very different.

For a long time, studying clicks on social media was hindered by unavailability of data, but this chapter proves that today, this can be overcome[2]. However, no sensitive individual information is disclosed in the data we present. Using multiple data collection techniques, we are able to jointly study Twitter conversations *and clicks* for URLs from five reputable news domains during a month of summer 2015. Note that we do not have complete data on clicks, but we carefully analyze this selection bias and found that we collected around 6% of all URLs, and observed 33% of all clicks. We chose to study news domains for multiple reasons described below.

First, news are a primary topic of conversation on social media[3], as expected due to their real time nature. In particular, Twitter is ranked third behind Facebook and Pinterest in total volume of web referral traffic, and often appears second after Facebook when web users discuss their exposure to news[1]. Knowing which social media generates traffic to news site is important to understand how users are exposed to news.

[1] http://j.mp/1qHkuzi

[2] Like any social media studies, we rely on open APIs and search functions that can be rendered obsolete after policy changes, but all the collection we present for this chapter follows the terms of use as of today, and the data will remain persistently available. http://j.mp/soTweet

[3] For instance, more than one in two American adult reports using social media as the top source of political news, which is more than traditional network such as CNN [Mitchell 2014].

Second, diffusion of news are generally considered highly influential. Political opinion is shaped by various editorial messages, but also by intermediaries that succeed in generating interest in such messages. This is also true for any kind of public opinion on subjects ranging from a natural disaster to the next movie blockbusters. Knowing who relays information and who generates traffic to news site is important to identify true influencers.

Last, news exhibit multiple forms of diffusion that can vary from a traditional top-down promotion through headlines to a word-of-mouth effect of content originally shared by ordinary users. Knowing how news are relayed on social media is important to understand the mechanisms behind influence.

This chapter offers the first study of social web referral at a large scale. We now present the following contributions.

- We validate a practical and unbiased method to study web referral from social media at scale. It leverages URL shorteners, their associated APIs, in addition to multiple social media search APIs. We identify four sources of biases and run specific validation experiments to prove that one can carefully minimize or correct their impact to obtain a representative joint sample of shares/receptions/clicks for all URLs in various online domains. As an example, a selection bias leads to collecting only 6.41% of the URLs, but we validate experimentally that this bias can be entirely corrected. (Section 4.2)

- We show the limitations of previous studies that ignored click conversion. First, we analyze the long-tail content popularity primarily caused by the large number of URLs mentioned that are not going through official headline promotions and show their effect to be grossly underestimated: those typically receive a minority of the receptions[4], indeed almost all of them are shared only a handful of times, and we even show a large fraction of them generate no clicks at all, sometimes even after thousands of receptions. In spite of this, they generate a *majority* of the clicks, highlighting an efficient curation process. In fact, we found evidence that the number of shares, ubiquitously displayed today on media's web presence to indicate popularity, appears an inaccurate measure of actual readership. (Section 4.3)

- We show that that clicks dynamics reveal the attention of social media to be long-lived, a significant fraction of clicks by social media referrals are produced over the following days of a share. This stands in sharp contrast with social media *sharing behavior*, previously shown to be almost entirely concentrated within the first hours, as we confirm with new results. An extended analysis of the tail reveals that popular content tend to attract many clicks for a long period of time, this effect has no equivalent in users' sharing behavior. (Section 4.4)

- Finally we leverage the above facts to build the first analysis of URL or user influence based on clicks and Clicks-Per-Follower. We first validate our estimation, show it reveals a simple classification of information intermediaries, and propose a refined influence score motivated by statistical analysis. URLs and users influence can be leveraged to predict a significant fraction of future social media referral, with multiple applications to the performance of online web services. (Section 4.5)

[4]See Section 4.2.1 for definition.

4.2 Measuring Social Media Clicks

Our first contribution is a new method leveraging features of social media shorteners to study for the first time in a joint manner how URLs get promoted on a social media and how users decide to click on those. This method is scalable using no privileged accounts, and can in principle apply to different web domains and social media than those we consider here. After presenting how to reproduce it, we analyze multiples biases, run validation experiment, and propose correction measures. As a motivation for this section, we focus on estimating the distribution of Click-Per-Follower (CPF), which is the probability that a link shared on a social media generates a click from one of the followers receiving that share.

Background on datasets involving clicks As a general rule, datasets involving clicks of users are highly sensitive and rarely described let alone shared in a public manner, for two reasons.

First, they raise ethical issues as the behavior of individual users online (*e.g.*, what they read) can cause them real harm. Users share a lot of content online, sometimes with dramatic consequence as this spreads, but they also have an expectation of privacy when it comes to what they read. This important condition makes it perilous to study clicks by collecting individual behaviors at scale. We carefully analyze the potential privacy threat of our data collection method and show that it merely comes from minutiae of social web referral that can be accounted for and removed at no expense.

Second, clicks on various URLs have important commercial value to a news media, a company, or a brand: if a movie orchestrates a social media campaign, the producing company, the marketing company, and the social media itself might not necessarily like to disclose its success in gathering attention, especially when it turns out to be underwhelming. These issues have hindered the availability of any such information beyond very large seasonal aggregates. Moreover due to the inherent incentives in disclosing this information (social media buzz, or the lack thereof), one may be tempted to take any of those claims with a grain of salt. In fact, *prior to our work, no result can be found to broadly estimate the Click-Per-Follower of a link shared on social media.*

4.2.1 Obtaining Raw Data & Terminology

For our study, we consider five domains of news media chosen among the most popular on Twitter: 3 news media channels BBC, CNN, and Fox News, one newspaper, The New York Times, and one strictly-online source, The Huffington Post. Our goal is to understand how URLs from these news media appear and evolve inside Twitter, as well as how those URLs are clicked, hence bringing web traffic revenue to each of those medias.

To build our dataset, we crawled two sources: Twitter to get the information on the number of shares and receptions (defined below), and `bit.ly` to get the number of clicks on the shortened URLs. At the beginning of our study, we monitored all URLs referencing one of the five news media using the Twitter 1% sample, and we found that 80% of the URLs of these five domains shared on Twitter are shortened with `bit.ly`, using a professional domain name such as `bbc.in` or `fxn.ws`. In the following, we only focus on the URLs shortened by `bit.ly`.

Twitter Crawl. We use two methods to get tweets from the Twitter API. The first method is to connect to the Streaming API[5] to receive tweets in real time. One may specify to access a 1% sample of tweets. This is the way we discover URLs for this study[6].

The second method to get tweets, and *the one we use after a URL is discovered*, is to use the search API[7]. This is an important step to gather all tweets associated with a given URL, providing a holistic view of its popularity, as opposed to only tweets appearing in the 1% sample. However, this API call has some limitations: (i) we cannot search for tweets older than 7 days (related to the time the API is called), (ii) the search API is rate-limited at 30 requests per minute for an application-only authentication and 10 requests per minute for a user authentication. This method proved sufficient to obtain during a month a comprehensive view of the sharing of each URL we preliminary discovered.

bit.ly Crawl. The `bit.ly` API provides the hourly statistics for the number of clicks on any of its URLs, including professional domains. The number of API calls are rate-limited, but the rate limit is not disclosed[8]. During our crawl, we found that we cannot exceed 200 requests per hour. Moreover, as the `bit.ly` API gives the number of clicks originating from each web referrer, we can distinguish between clicks made from `twitter.com`, `t.co`, and others, which allows us to remove the effect of traffic coming from other sources.

In the rest of this chapter, we will use the following terms to describe a given URL or online article.

Shares. Number of times a URL has been published in tweets. An original tweet containing the URL or a retweet of this tweet are both considered as a new share.

Receptions. Number of Twitter users who are potentially exposed to a tweet (*i.e.*, who follow an account that shared the URLs). Note that those may not have necessarily seen or read the tweet in the end. As an example, if a tweet is published by a single user with N followers, the number of receptions for this tweet is N. This metric is related but different from the number of "impressions" (number of people who actually saw the tweets in their feed). Impressions are typically impossible to crawl without being the originator of the tweet (see a study by Wang*et al.* [Wang 2016] comparing those two metrics).

Clicks. Number of times a URL has been clicked by a visitor originating from `twitter.com` or `t.co`.

Click-Per-follower (CPF). For a given URL, the CPF is formally defined as the ratio of the clicks to the receptions for this URL. For example, if absolutely all followers of accounts sharing a URL go visit the article, the CPF is equal to 1.

Limitations The rest of this section will demonstrate the merit of our method, but let us first list a serie of limitations. (i) *It only monitors searchable public shares in social*

[5] https://dev.twitter.com/streaming/overview

[6] One can also specify to the Streaming API a filter to receive tweets containing particular keywords (up to 400 keywords) or sent by particular people (up to 5000 userIDs) or from particular locations. For the volume of tweets that we gathered, this method can be overwhelmed by rate limit, leading to real time information loss, especially during peak hours, that are hard to recover from.

[7] https://dev.twitter.com/rest/reference/get/search/tweets

[8] http://dev.bitly.com/rate_limiting.html

media. Facebook, an important source of traffic online, does not offer the same search API[9]. Everything we present in this chapter can only be representative of public shares made on Twitter. (ii) *It only deals with web resource exchanged using* `bit.ly` *or one of its professional accounts.* This enables to study leading news domain which are primarily using this tool to direct traffic to them. Our observations are also subject to the particular choice of domains (5 news channel, in English, primarily from North America). (iii) *It is subject to rate limits.* `bit.ly` API caps the number of request to an undisclosed amounts and Twitter implements a 1 week window on any information requested. This could be harmful to study even larger domains (*e.g.*, all links in `bit.ly`) as this is impractical, but as we will see it was not a limitation for those cases considered in this study. (iv) *It measures attention on social media only through clicks.* While in practice merely viewing a tweet can already convey some information, we consider that no interest was shown by the user unless they are actively visiting the article mentioned.

4.2.2 Ensuring Users' Privacy

Since we only collect information on URLs shared, viewed, and visited, we do not *a priori* collect individual data that are particularly sensitive in nature. Note also, that all those data are publicly available today. However, the scale of our data might *a posteriori* contains sufficient information to derive those individual behaviors from our aggregates. As an extreme example, if a user is the *only* user on that social media receiving an article, and that this article is clicked, that information could be used to infer that this user most likely read that webpage.

We found a few instance of cases like the above, but in practice those particular cases can be addressed. First, before we analyze the raw data we apply a preprocessing *merging* step in which all URLs shorteners that lead to the same developed URLs were merged together (ignoring information contained after the symbols "?" and "#" which may be used in URL to inform personalization). This also helps our analysis further as it allows to consider the importance of a given article independently of the multiplication of URLs by which it is shortened. After considering the data after this processing step, we found that case of clicked URLs with less than 10 receptions never occurred. Moreover, we found only a handful of URLs with less than 50 receptions, showing that in all cases we considered, all but a extremely small number of users are guaranteed to be k-anonymous with $k = 50$. Equivalently, it means that their behavior is indistinguishable from at least 49 other individuals. In practice, for most users, the value of k will be much larger. Finally, we observe that no explicit information connect clicks from multiple URLs together. Clicks are measured at a hourly rate, it is therefore particularly difficult to use this data to infer that the same user has clicked on multiple URLs and use it for reidentification.

Finally, we computed the probability for a discovered URLs to receive at least one click as a function of its number of receptions, and found that observed URLs with less than 500 receptions are a minority (3.96%), and *among those* 9.80% actually receive clicks. This shows that, while our study dealing with links from popular sources online raised little privacy issue, one could in practice extend this method for more sensitive domains simply by ignoring clicks

[9] Although, very recently as of October 2015 Facebook provides search over public posts http://www.wired.com/2015/10/facebook-search-privacy/.

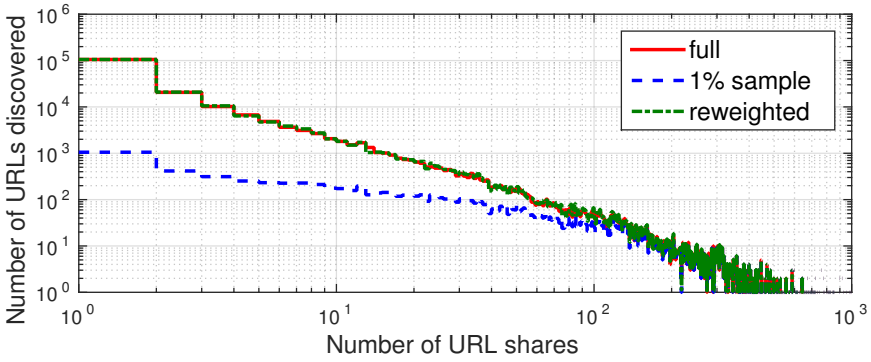


Figure 4.1 – **1% sample bias and correction.** *The bias due to the 1% sample is significant, but can be fully corrected using reweighting.*

from those URLs if a larger value of k-anonymity is required (such as $k = 500$) to protect users.

Removing clicks from all URLs with less than k receptions would in practice affect the results very marginally.

4.2.3 Selection Bias and a Validated Correction

The most important source of bias in our experiment comes from the URL discovery process. As we rely on a 1% sample of tweets from Twitter, there are necessarily URLs that we miss, and those we obtain are affected by a *selection bias*. First, Twitter does not always document how this sample is computed, although it is arguably random. (Our experiments confirmed that it a uniform random sample.) Second, the 1% sample yields a uniform sample of tweets, but not the URLs contained in these tweets. We are much more likely to observe a URLs shared 100 times than one shared only in a few tweets. We cannot recover the data about the missed URLs, but we can correct it by giving more weight to unpopular URLs in our dataset when we compute statistics. We note that this selection bias was present in multiple studies before ours, but given our focus on estimating the shape of online news, it is particularly damaging.

To understand and cope with this bias, we conduct a validation experiment in which we collected all tweets for one of the news sources (*i.e.*, `nytimes.com`) using the Twitter search API. Note that `bit.ly` rate limits would not allow us to conduct a joint analysis of audience for such volumes of URLs. This collection was used to compare our sample, and validate a correction metric.

Figure 4.1 compares distributions of the number of shares among URLs in our sample and in the validation experiment for the same time period. We see that 1% sample underestimates unpopular URLs that are shared less often. However, this bias can be systematically corrected by re-weighting each URL in the 1% sample using the coefficient $\frac{1}{1-(1-\alpha)^s}$, where $\alpha = 0.01$ is the probability of a tweet to get into the sample and s is the number of shares that we observed for that URL from the search API. We then obtain a partial but representative view of the URLs that are found online, which also indirectly validates that Twitter 1% sampling indeed resembles a random choice. All the results in this chapter take into account this correction.

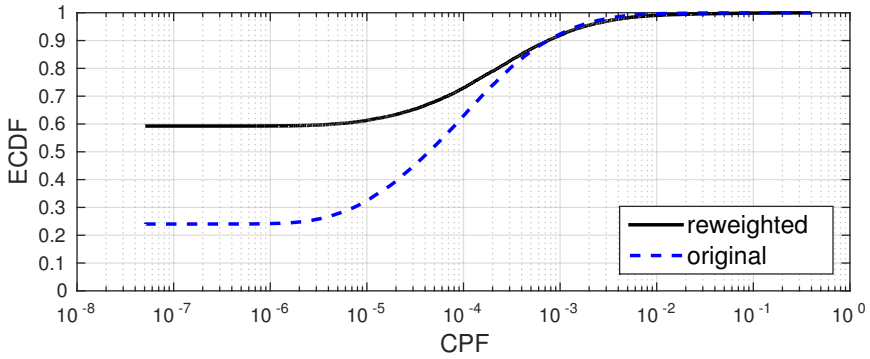


Figure 4.2 – **CPF distribution with and without selection bias.** *The number of URLs with low CPF is significantly underestimated if the bias due to the 1% sample is not corrected.*

Finally, we need to account for one particular effect of our data: URLs shorteners were first discovered separately, and crawled each for a complete view of their shares, but shorteners are subsequently merged (as explained in §4.2.2) into articles leading to the same developed URL. It is not clear which convention applies to best assign weight among merged URLs that have been partially observed, so we relied on simple case analysis to decide on the best choice. Our analysis showed that not to overwhelm the correction in favor of either a small or a large article, the best is to sum all shares of shorteners leading to the same article before applying the re-weighting exactly as done above.

Social media CPF and the effect of selection bias To illustrate the effect of the selection bias, Figure 4.2 presents the two empirical distributions of CPF obtained from the Twitter crawl by dividing for each article its sum of clicks by its number of receptions (all measured after 24 h), and after re-weighting each value according to the above correction.

We highlight multiple observations: First, CPF overall is low (as one could expect) given that what counts in our data as a *reception* does not guarantee that the user even has seen the tweet, for instance if she only accesses Twitter occasionally this tweet might have been hidden by more recent information. In fact, we estimate that a *majority* (59%) of the URLs mentioned on Twitter are not clicked at all. Note that this would not appear in the data without re-weighting, where only 22% of URLs are in that case, which illustrates how the selection bias present in the Twitter 1% sample could be misleading. Finally, for most of the URLs that do generate clicks, the CPF appear to lie within the $[10^{-5}; 10^{-3}]$ range. It is interesting to observe that removing the selection bias also slightly reinforce the distribution for high values of CPF. This is due to a minority of niche URLs shared only to a limited audience but that happens to generate in comparison a larger number of clicks. This effect will be analyzed further in the next section.

4.2.4 Other Forms of Biases

Limited time window On Figure 4.3, we present the time conventions and definitions we are using to describe the lifespan of a URL. We monitored the 1% tweet sample, and each time t_d we found a tweet containing a new URL (shortened by `bit.ly`) from one of the five news media we consider, we schedule a crawl of this URL at time $t_d + 24$h, the crawl ends at

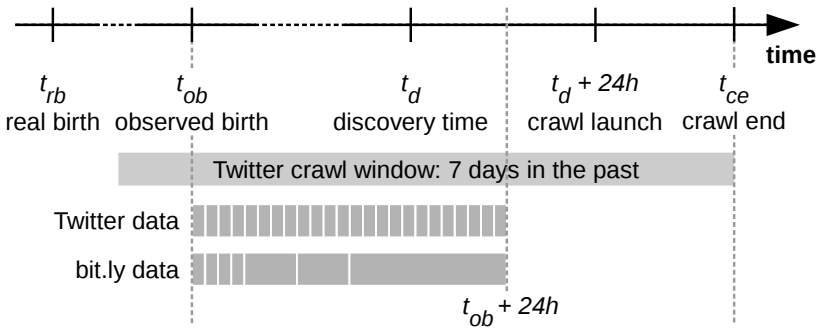


Figure 4.3 – **Time conventions and definitions for URL life description.**

time t_{ce}. The rational to delay by 24 hours the crawl is to make sure that each URL we crawl has a minimum history of 24 hours. The crawl of each URL consists of the three following steps.

1. We query the Twitter search API for all tweets containing this URL. We define the time t_{ob} of observed birth of the URL as the timestamp of the earliest tweet found containing this URL.

2. We query the `bit.ly` API to get the number of clicks on the URL. Due to the low limit on the number of queries to the `bit.ly` API, we make 7 calls asking for the number of clicks in the 1st, 2nd, 3rd, 4th, 5th through 8th, 9th through 12th, and 13th through 24th hours after t_{ob}.

3. After this collection was completed, we crawled `bit.ly` a second time to obtain information on clicks completed after a day (up to 2 weeks).

Note that this technique inherently leverages a partial view on both side of the temporal axis. On the one hand, some old information could have been ignored, due to the limited time window of our search. In fact, we cannot be sure that no tweet mentioned that URL a week before. On the other hand, we empirically measured that for a majority of URLs $t_d - t_{ob} \leq 1$ hour which means that the oldest observation of that URLs in the week was immediately before our discovery. Moreover, for an overwhelming majority of the URLs (97.6%) $t_d - t_{ob} \leq 5$ days. This implies that for all of those our earliest tweets were preceded by at least two days where this URLs was never mentioned (see Figure 4.3). Given the nature of news content, we deduce that earlier tweets are non-existent (we conclude that t_{rb} is equal to t_{ob}), or even if they are missed, not creating any more clicks at the time of our observations. Finally, we note that recent information occurring after our observation time is missing, especially as we cannot retroactively collect tweets after 24 h. While this could in theory be continuously done in a sliding window experiment, we observe that an overwhelming majority of tweets mentioning a URLs occurred immediately after our discovery (within a few hours). We also note that the effect of those tweets would not affect the clicks seen before 24 h.

The effect of multiple receptions An online article may be shown multiple times, sometimes using the same shorteners, to the same user. This comes from multiple reasons. Two

different accounts that a user follows may share the same link, or a given Twitter account may share multiple times the same URL, which is not so uncommon. This necessarily affects how one should count receptions of a given URLs, and how to estimate its CPF although it's not clear whether receiving the same article multiple times impacts its chance to be seen. Note finally, that even if we know the list of Twitter users who shared a URL, we cannot accurately compute how many saw the URLs at least once without gathering the list of followers of those who share, in order to remove duplicates, *i.e.*, overlaps between sets of followers. Because of the low rate limit of the Twitter API, it would take us approximately 25 years with a single account to crawl the followers of users who shared the URLs in our dataset. However, we can compute an upper bound of the number of receptions as a sum of the number of followers. This is an upperbound because we don't consider the overlap between the followers, that is any share from a different user counts as a new reception. Note that this already removes multiplicity due to several receptions coming from the same account.

To assess the bias introduced by this estimation, we consider a dataset that represents the full Twitter graph as of July 2012 (see Chapter 3). This graph provides not only the number of followers, but also the full list of followers for each account, at a time where the API permitted such crawl. Therefore, this graph allows to compute both an estimate and the ground truth of the number of receptions. We computed our estimated reception and the ground truth reception for each URLs on the 2012 dataset. We note that some users that exist today might not have been present in the 2012 dataset. Therefore, we extracted from our current dataset all Twitter users that published one of the monitored URLs and found that 85% of these users are present in the 2012 dataset, which is large enough for our purpose.

Figure 4.4 shows the difference between our estimation of the number of receptions and the real number of receptions based on the 2012 dataset. We observe that for an overwhelming majority (more than 97% of URLs) the two values are of the same order (within a factor 2). We also observe that for 75% of them, the difference is less than 20%.

This implies that the CPF values we obtain are conservative (they may occasionally underestimate the actual probability by about 20-50%) but within a small factor of values computed using other conventions (*i.e.*, for instance, if multiple receptions to the same user are counted only once). For the sake of validating this claim, and measuring how overlaps affect CPF estimation, we ran a "mock" experiment where we draw the distribution among URLs of the CPF assuming that the audience is identical to the followers found in the 2012 dataset. We compare the CPF when multiple receptions are counted (as done in the rest of this chapter), and when duplicate receptions to the same user are removed. Note that since 2012 the number of followers have evolved (typically increased) as Twitter expanded, hence receptions are actually underestimated. The absolute values we observe are hence not accurate (typically ten times larger). However, we found that the CPF for multiple conventions always are within a factor two of each other. It confirms that the minority of URLs where overlaps affect receptions are URLs shared to a small audience that typically do not impact the distribution.

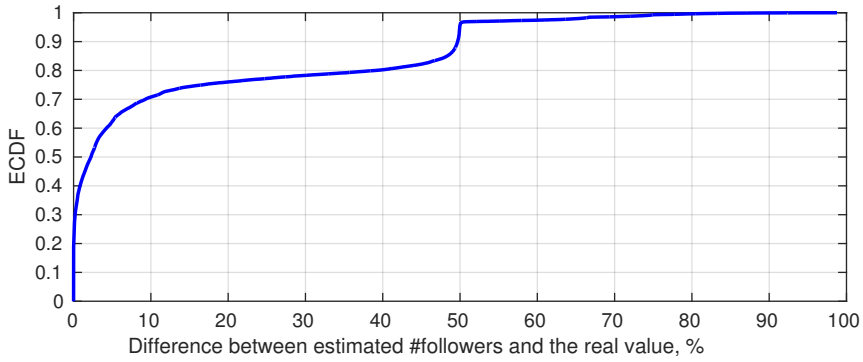


Figure 4.4 – **Bias of the estimation of the number of receptions.** The figure present the ECDF of the relative difference of the real number of receptions and our estimate. The lists of followers are taken from the 2012 Twitter dataset (see Chapter 3). *For 75% of the receptions, the estimation error is less than 20%, which is good enough for this study.*

4.3 Long Tail & Social Media

For news producers and consumers, social media create new opportunities by multiplying articles available and exposed, and enabling personalization. However, with no analysis of consuming behaviors that is, clicks, which we show vastly differ from sharing behaviors, it remains uncertain how the users of social media today take advantage of such features.

4.3.1 Background

Social media such as Twitter allow news consumption to expand beyond the constraints of a fixed number of headlines and printed articles. Today's landscape of news is *in theory* only limited by the expanding interests of the online audience. Previous studies show that almost all online users exhibit an uncommon taste at least in a part of their online consumption [Goel 2010], while others point to possible bottlenecks in information discovery that limits the content accessed [Cha 2009]. To fully leverage opportunities open by social media, works propose typically to leverage either a distributed social curation process (*e.g.*, [Zadeh 2013, Hegde 2013, Wong 2015, May 2014]) or some underlying interest clusters to feed a recommender system (*e.g.*, [Xu 2014, Massoulié 2015]). However, with no evidence on the actual news consumed by users, it is hard to validate whether todays' information has benefited from those opportunities. For instance, online media like those in our study continue to select a few set of headlines articles to be promoted via their official Twitter account, followed by tens of millions of users or more, effectively reproducing traditional curation under a new format. Those are usually leading to large cascades of information. What remains to measure is *how much* web traffic comes from such promoted content, and how much from different type of content reaching users organically through the distributed process of online news sharing. To answer these questions, we rely on a detailed study of the properties of the long tail and the effect of promotion on content accessed.

Long tail of online news: What to expect? Pareto's principle, or the law of diminishing return, states that a majority of the traffic (say, for instance 80%) should primarily comes from a restricted, possibly very small, fraction of the URLs. Those are typically referred to

as the *blockbusters*. This prediction is however complemented with the fact that most web traffic exhibits a long tail effect. The later focuses on the set of URLs that are requested infrequently, also referred to as *niche*. It predicts that, when it comes to generating traffic, whereas the contribution of each niche URLs is small, they are in such great numbers that they collectively contribute a significant fraction of the overall visits. Those properties are not contradictory and usually coexist, but a particular one might prevail. When it is the case, it pays off for a content producer who has limited resource and wishes to maximize its audience to focus either on promoting the most popular content, or, on the contrary, on maintaining a broad catalog of available information catering to multiple needs.

And what questions remains to answer? Beyond the above qualitative trends, we wish here to answer a set of precise questions by leveraging the consumption of online news. What level of sharing defines a blockbuster URL that generates an important volume of clicks, or a niche one; for instance, should we consider a moderate success a URL that is shared 20 times or clicked 5 times? Does it mean that niche URLs have overall a negligible effect in bringing audience to an online site? Since the headlines selected by media to feature in their official Twitter feed benefit from a large exposure early on, could it be that they account for almost all blockbuster URLs? More generally, what fraction of traffic is governed by blockbuster and niche clicks? Moreover, do any of those property vary significantly among different online news sources, or is it the format of the social media itself (here, Twitter) that governs users behaviors, in the end, determining how content gets accessed?

4.3.2 Traditional vs. Social Media Curation

To answer the above questions, we first introduce a few definitions:

Primary URL. A primary URL is a URL contained in a tweet sent by an official account of the 5 news media we picked for this study. Such URLs are spread through the traditional curation process because they are selected by new media to appear in the headlines.

Secondary URL. All URLs that are not primary are secondary URLs. Although they refer to official and authenticated content from the same domain, none benefited from the broad exposure that the official Twitter of this source offers. Such URLs are spread trough the social media curation process.

We see from the sharing activities (tweet and retweet) that primary URLs accounts for 60.47%, hence a majority, of the receptions in the network. This comes from the fact that, although primary URLs only account for 17.43% of the shares overall, non-primary URLs are typically shared to less followers.

If we assume that clicks are linearly dependent on the receptions then we will wrongly conclude that clicks are predominantly coming from primary or promoted content. In fact, one could go even further and expect primarily URLs to generate an ever larger fraction of the clicks. Those URLs, after all, are the headlines that are carefully selected to appeal to most, they also contain important breaking news, and are disseminated through a famous trusted channel (a verified Twitter account with a strong brand name). It is however, the

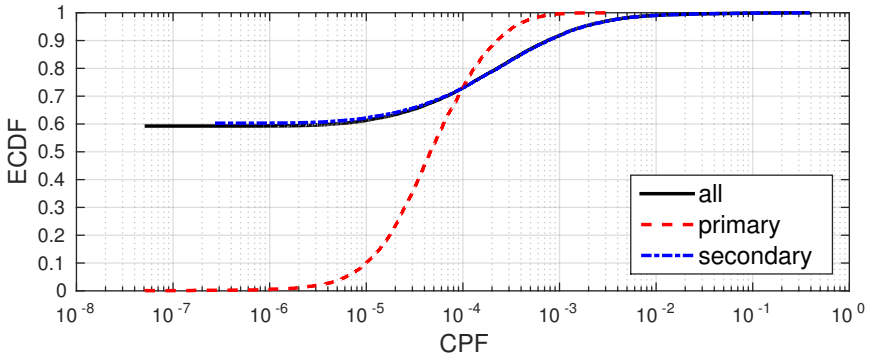


Figure 4.5 – **The empirical CDF of the CPF for primary and secondary URLs.** Primary URLs account for 2% of all URLs after selection bias is removed, are always receiving clicks, whereas secondary URLs account for 98% URLs, and 60% of them are never clicked.

opposite that holds. As our new data permits to observe for the first time, *secondary URLs, who receive a minority of receptions, generate a significant majority of the clicks (60.66% of them).* Note that our methodology plays a key role in proving this result, as without correcting the sampling bias, this trend is less pronounced (non-primary URLs are estimated in the raw data to receive 52.01% of the clicks).

We present in Figure 4.5 the CPF distribution observed among primary and secondary URLs separately. This allows to compare for the first time the effective performance of traditional curation (primary URLs) with the one performed through a social media (secondary URLs). Primary URLs account for less than 2% of the URLs overall and all of them are clicked at least once. Secondary URLs, in contrast, very often fail to generate any interest (60% of them are never clicked), and are typically received by a smaller audience. However, secondary URLs that are getting clicked outperform primary URLs in terms of CPF.

This major finding has several consequences. First, it suggests that social media like Twitter already play an important role, both to personalize the news exposed to generate more clicks, and also to broaden the audience of a particular online domain by considering a much larger set of articles. Consequently, serving traffic for less promoted or even non-promoted URLs is critical for online media, it even creates a majority of the visits and hence advertising revenue. Simple curating strategies focusing on headlines and traditional curation would leave important opportunities unfulfilled (more on that immediately below). In addition, naive heuristics on how clicks are generated in social media are too simplistic, and we need new models to understand how clicks occur. Beyond carefully removing selection bias due to sampling, there is a need to design CPFs model which accounts for various types of sharing dynamics.

Variation across news domain One could formulate a simple hypothesis in which the channel, or in this case the social network, that is used to share information governs how its users decide to share and click URLs, somewhat independently of the sources. But we found that the numbers reported above for the overall data vary significantly between online news media (as shown in Table 4.1 where they are sorted from the most to the least popular), proving that multiple audiences use Twitter differently.

Most of the qualitative observations remain: a majority of URLs (50-70%) are not clicked,

	bbc.in	huff.to	cnn.it	nyti.ms	fxn.ws
# URLs	13.14k	12.38k	6.39k	5.60k	1.82k
# clicks (million)	3.05	2.29	2.01	1.53	0.79
% primary shares	7.24%	10.34%	22.94%	31.86%	41.06%
% primary receptions	31.16%	61.78%	75.37%	78.97%	84.76%
% primary clicks	15.01%	30.81%	61.29%	60.31%	89.79%
% unclicked URLs	51.19%	59.85%	70.16%	64.94%	62.54%
Clicks from all URLs shared <100 times	51.10%	75.70%	33.67%	22.59%	34.89%
Clicks from secondary URLs shared <100 times	46.11%	55.28%	24.73%	14.83%	6.62%
Threshold share to get 90% clicks	8	6	22	34	90
Threshold receptions to get 90% clicks	215k	110k	400k	560k	10,000k
Threshold clicks to get 90% clicks	70	62	180	170	2,000

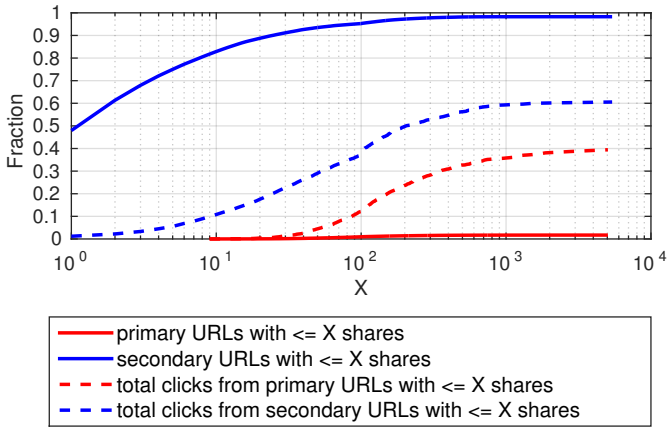
Table 4.1 – **Variation of metrics across different newssource.**

and primarily URLs always generate overall more receptions than they generate clicks. But, we also observe striking differences. First, domains that are less popular online (in terms of mentions and clicks) shown on the right in Table 4.1 typically rely more on traditional headline promotion to receive clicks: `nyti.ms` and `fxn.ws` stand as the two extreme examples with 60% and 90% of their clicks coming from headlines, whereas the Huffington Post and the BBC which receive more clicks, present opposite profiles. These numbers suggest that most famous domains like those receive the same absolute order of clicks from primary URLs (roughly between 0.5 and 1.2 million), but that some differ in successfully fostering broader interest to their content beyond those. Although those domain cater to different demographics and audience, Twitter users represent a relatively young crowd and this may be interpreted as an argument that editorial strategy based on a breadth of topics covered might be effective as far as generating clicks from this population is concerned.

4.3.3 Blockbusters and the Share Button

Our results show that information sharing on social media plays a critical role for online news; it complements curation offered traditionally through headlines. These two forms of information selection appear strikingly similar: in traditional curation 2% of articles are chosen and produce a disproportionate fraction (near 39%) of the clicks; the sharing process of a social media also results in a minority of secondary URLs (those clicked at least 90 times, about 7% of them) receiving almost 50% of the click traffic. Together, those blockbuster URLs capture about 90% of the traffic. Each typically benefits from a very large amount of receptions (from Figure 4.6, we estimate that 90% of the clicks come from URLs with at least 150,000 receptions). More details on each metric distribution and how those URLs generate clicks can be seen in Figure 4.6.

Before our study, most analysis of blockbusters content made the following assumptions: since the primary ways to promote a URL is to share or retweets this article to your followers, one could identify blockbusters using the number of times a URL was shared. This motivates today's ubiquitous use of the "share number" shown next to an online article with various button in Twitter, Facebook, and other social media. One could even envision that, since a fraction of the users clicking an article on a social media decide to retweet it, the share number is a somewhat reduced estimate of the actual readership of an article. However, our



(a) Shares.

Figure 4.6 – **Fraction of Primary/Secondary URLs (divided by all URLs) with less than X shares, receptions and clicks, shown along with the cumulative fraction of all clicks that those URLs generate (dashed lines).**

joint analysis of shares and clicks reveal limitations of the share number.

First, 59% of the shared URLs are never clicked or, as we call them, *silent*. Note that we merged URLs pointing to the same article, so out of 10 articles mentioned on Twitter, 6 typically on niche topics are never clicked[10].

Because silent URLs are so common, they actually account for a significant fraction (15%) of the whole shares we collected, more than one out of seven. An interesting paradox is that there seems to be vastly more niche content that users are willing to mention in Twitter than the content that they are actually willing to click on. We later observe another form of a similar paradox.

Second, when it comes to popular content, sharing activity is not trivially correlated with clicks. For instance, URLs receiving 90 shares or more (primary or secondary) comprises the top 6% most shared, so about the same number of URLs as the blockbusters, but they generate way less clicks (only 45%). Conversely we find that blockbuster URLs are shared much less. Figure 4.6 shows that 90% of the clicks are distributed among all URLs shared 9 times and more.

The 90% click threshold of any metric (clicks, shares, receptions) is the value X of this metric such that the subset of URLs with metric *above* X receives 90% of the clicks. We present values of the 90% click threshold for different domains in Table 4.1. We observe that the threshold for share is always significantly smaller than other metrics, which means that any URLs shared even just a few times (such as 6 or 8) may be one generating clicks significantly.

The values of 90% click thresholds, when compared across domains, reveals another trend: most popular domains (shown on the left) differ not because their most successful URLs receive more clicks, but because clicks gets generated even for URLs shared a few times only. For instance, in `fxn.ws`, URLs with less than 90 mentions are simply not contributing much to the click traffic; they could entirely disappear and only 10% of the clicks to that domain would be lost. In contrast, for `bbc.in` and `huff.to` the bulk of clicks are generated even

[10]They could, however, be accessed in other ways: inside the domain's website, via a search engine etc.

among URLs shared about 6-8 times. Similarly, for Fox News 90% of clicks belong to URLs generating at least 10 million receptions on Twitter, this goes down by a factor 50x to 100x when it comes to more popular domains receiving 3 or 4 times more clicks overall. In other words, although massive promotion remains an effective medium to get your content known, the large popularity of the most successful domains is truly due to a long-tail effect of niche URLs.

In summary, the pattern of clicks created from social media clearly favors a blockbuster model: infrequent niche URLs—whatever numerous they are—are generally seen by few users overall and hence have little to no effect on clicks, while successful URLs are seen by a large number (at least 100,000) of the users. But those blockbusters are not easy to identify: massive promotion of articles through official Twitter accounts certainly work, but it appeals less to users as far as clicks are concerned, and in the end, most of the web visits come from other URLs. All evidence suggest that for a domain to expand and grow, the role of other URLs, sometimes shared not frequently, is critical. We note also that relying on the number of shares to characterize the success of an article, as commonly done today for readers of online news, is imprecise (more on that in Section 4.5).

4.4 Social Media Attention Span

We now analyze how our data on social media attention reveals different temporal dynamics than those previously observed and analyzed in related literature. Indeed, prior to our work, most of the temporal analysis of social media relied on sharing behavior, and concluded that their users have a *short attention span*, as most of the activity related to an item occurs within a small time of its first appearance. We now present contrasting evidence when clicks are taken into account.

4.4.1 Background

Studying the temporal evolution of diffusion on social media can be a powerful tool, either to interpret the attention received online as the result of an exogenous or endogenous process [Crane 2008], to locate the original source of a rumor [Pinto 2012], or to infer *a posteriori* the edges on which the diffusion spreads from the timing of events seen [Gomez-Rodriguez 2012]. More generally, examining the temporal evolution of a diffusion process allows to confirm or invalidate simple model of information propagation based on epidemic or cascading dynamics. One of the most important limitation so far is that prior studies focus only on the evolution of the collective volume of attention (*e.g.*, hourly volumes of clicks [Szabo 2010], views [Crane 2008, Cha 2009]), hence capturing the *implicit* activity of the audience, while ignoring the process by which the information has been propagated. Alternatively, other studies focus on *explicit* information propagation only (*e.g.*, tweets [Yang 2011], URLs shorteners, diggs [Wu 2007]) ignoring which part of those content exposure leads to actual clicks and content being read. Here for the first time we combine explicit shares of news with the implicit web visits that they generate.

Temporal patterns of online consumption was gathered using videos popularity on YouTube [Crane 2008, Cha 2009] and concluded to some evolution over days and weeks. However, this study considered clicks originating from any sources, including YouTube own

recommendation systems, search engine and other mentions online. One hint of the short attention span of social media was obtained through URLs shorteners[11]. Using generic `bit.ly` links of large popularity, this study concludes that URLs exchanged using `bit.ly` on Twitter today typically fades after a very short amount of time (within half an hour). Here we can study jointly for the first time the two processes of social media sharing and consumption. Prior work [Abisheva 2014] dealt with very different content (*i.e.*, videos on YouTube), only measured overall popularity generated from all sources, and only studied temporal patterns as a user feature to determine their earliness. Since the two processes are necessarily related, the most important unanswered question we address is whether the temporal property of one process allows to draw conclusion on the properties of the others, and how considering them jointly shed a new light on the diffusion of information in the network.

4.4.2 Contrast of Shares and Clicks Dynamics

For a given URL and type of events (*e.g.*, clicks), we define this URL half-life as the time elapsed between its birth and half of the event that we collected in the entire data (*e.g.*, the time at which it received half of all the clicks we measured). Intuitively, this denotes a maturity age at which we can expect interest for this URL to start fading away. Prior report[1] predicted that the click half-life for generic `bit.ly` links using Twitter is within two hours. Here we study in addition half-life based on other events such as shares and receptions, and we consider a different domain of URLs (online news). We found, for instance that for a majority of URLs, the half-life for any metric is below one hour (for 52% of the URLs when counting clicks, 63% for shares, and 76% for receptions). This offers no surprise; we already proved that most URLs gather a very small attention overall, it is therefore expected that this process is also ephemeral.

But this metric misrepresents the dynamic of online traffic as it hides the fact that most traffic comes from unusual URLs, those that are more likely to gather a large audience over time. We found for instance that only 30% of the overall clicks gathered in the first day are made during the first hour. This fraction drops to 20% if it is computed for clicks gathered over the first two weeks; overall the number of clicks made in the second week (11.12%) is only twice smaller than in the first hour. Share and receptions dynamics are, on the contrary, much more short lived. While we have no data beyond 24 hours for those metrics due to Twitter limitations, we observe that 53% of shares and 82% of all receptions on the first day occur during the first hour, with 91% and 97% of those created during the first half of the day. By the time that we reach the limit of our observation window, it appears that the shares and receptions are close to zero. In fact, the joint correlation between half-life defined on shares and clicks (see Figure 4.7) revealed that those gathering most of the attention are heavily skewed towards longer clicks lifetime. Note that we are reporting results aggregated for all domains but each of them follow the same trend with minor variations.

4.4.3 Dynamics & Long Tail

To understand the effect of temporal dynamics on the distribution of online attention, we draw in Figure 4.8 for shares, receptions and clicks the evolution with time. Each plot presents,

[11] `http://bitly.is/1IoOqkU`

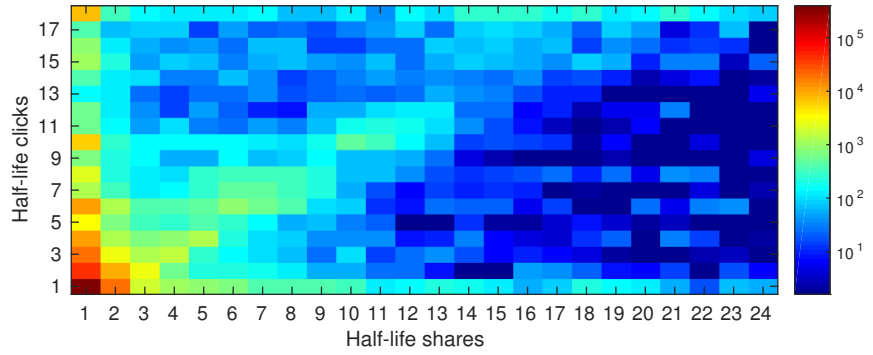


Figure 4.7 – **Joint distribution of the half-life defined on shares and clicks for the URLs in the dataset.**

using a dashed line, the distribution of events observed after an hour. In solid line we present how the distribution increases cumulatively as time passes, whereas the dashed-dotted line shows the diminishing rate of hourly events observed at a later point. Most strikingly, we observe that shares and especially receptions dropped by order of magnitude at the end of 24 h. Accordingly, the cumulative distribution of receptions seen at 24 h is only marginally different from the one observed in the first hour. Clicks, and to some extent shares, present a different evolution. Clicks drop more slowly, but more importantly it does not drop uniformly. The more popular a tweet or a URL the more likely it will be shared or clicked after a big period of time.

We highlight a few consequences of those results. Social media have often been described as entirely governed by fast successive series of flash crowds. Although that applies to shares and receptions, it misrepresents the dynamics of clicks which, on the contrary, appear to follow some long term dynamics. This opens opportunities for accurate traffic prediction that we analyze next. In addition, our dynamics motivate to further investigate the origins and properties of the reinforcing effect of time on popularity, as our results suggest that gathering clicks for news on social media rewards URLs that are able to maintain a sustained attention.

4.5 Click-Producing Influence

In this section, we propose a new definition of influence based on the ability to generate clicks, and we show how it differs from previous metrics measuring mere receptions. We then show how influence and properties of sharing on social media can be leveraged for accurate clicks prediction.

4.5.1 Background

Information diffusion naturally shapes collective opinion and consensus [Katz 1957, Degroot 1974, Lord 1979], hence the role of social media in redistributing influence online has been under scrutiny ever since blogs and email chains [Adamic 2005, Liben-Nowell 2008]. Information on traditional mass media follows a unidirectional channel in which pre-established institutions concentrate all decisions. Although the emergence of opinion leaders digesting

news content to reach the public at large was pre-established long time ago [Katz 1957], social media presents an extreme case. They challenge the above vision with a distributed form of influence: social media allow *in theory* any content item to be tomorrow's *headline* and any user to become an *influencer*. This could be either by gathering direct followers, or by seeing her content spreading faster through a set of intermediary node.

Prior works demonstrated that news content exposure benefits from a set of information intermediaries [Wu 2011, May 2014], proposed multiple metrics to quantify influence on social media like Twitter [Cha 2010, Bakshy 2011], proposed models to predict its long term effect [Kleinberg 2007, Lelarge 2012], and designed algorithms to leverage inherent influence to maximize the success of targeted promotion campaign [Kempe 2003, Ok 2014] or prevent it [Lelarge 2009]. So far, those influence metrics, models, and algorithms have been validated assuming that observing a large number of receptions is a reliable predictor of actual success, hence reducing influence to the ability to generate receptions. We turn to a new definition in which influence is measured by actual clicks, which are more directly related to revenue through online advertising, and also denote a stronger interaction with content.

4.5.2 A New Metric and its Validation

A natural way to measure the influence (or the lack thereof) for a URL is to measure its ability to generate clicks. We propose to measure a URL's influence by its CPF. To illustrate how it differs from previous metrics, in Figure 4.9 we present the joint distribution of shares (used today to define influence online) and clicks among URLs in our data using a color-coded heatmap. We observe that the two metrics loosely correlate but also present important differences. URLs can be divided into three subsets: the bottom left corner comprising URLs who have gathered neither shares nor clicks (cluster 1 in green), and two cluster of URLs separated by a straight line (cluster 2 in cyan and cluster 3 in red). Immediately below the figure we present how much share events are created by each cluster. Not shown here is the same distribution for clicks, which confirms that those are entirely produced by URLs in the cluster 2 shown in cyan. This produces another evidence that relying on shares, which classify all those URLs as influential, can be strikingly misleading. While about 40% to 50% of the shares belong to URLs in the cluster 3 in red, those are collectively generating a negligible amount of clicks (1%).

Ideally, we would like to create a similar metric to quantify the influence of a user. Unfortunately, it is not straightforward. One can define the *participatory influence* of a user as the median CPF of URLs that she decides to share. However, this metric is indirect since clicks generated by these URLs are aggregated with many others. In fact, even if this metric is very high, we cannot be sure that the user was responsible for the clicks we observe on those URLs. To understand the bias that this introduces we compare the CPF of the URLs that users were the only ones to send with the CPF of the other URLs they sent. There are 1,358 users in our dataset that are the only ones tweeting at least one of their URL shorteners (note that for validation we did not merge the URL shorteners). For each user, using those URLs whose transmission on Twitter can be attributed to them, we can compare the CPF of attributed URLs, and ones they participated in sharing without being the only ones. Figure 4.10 presents the relative CPF difference between these two metric computation. We see that for 90% of users with attributed clicks the difference between the

participating CPF and attributed CPF is below 4×10^{-4}. We also see that the difference is more frequently negative. Overall using the participating CPF is a conservative estimate of their true influence.

However, one limitation of the above metric is that a very large fraction of users share a small number of URLs. If that or those few URLs happen to have a large CPF, we assign a large influence although we gather small evidence for it. We therefore propose a refined metric based on the same principles as a statistical confidence test. For each user sharing k URLs, we compute a CPF threshold based on the top 95% percentile of the overall CPF distribution, and we count how many (l) of those URLs among k have a CPF above the threshold. Since choosing a URL at random typically succeeds with probability $p = 0.05$, we can calculate the probability that a user obtain at least l such URLs purely due to chance:

$$\sum_{n=l}^{k} \binom{k}{n} p^n (1-p)^{k-n} = 1 - \sum_{n=0}^{l-1} \binom{k}{n} p^n (1-p)^{k-n} \, .$$

When this metric is very small, we conclude that the user must be influential. In fact, we are guaranteed that those URLs are chosen among the most influential in a way that is statistically better than a random choice. It also naturally assigns low influence to users for which we have little evidence, as the probability in this case cannot be very small. After computing this metric and manually looking at the results, we found that it clearly distinguish between users, with a minority of successful information intermediaries receiving the highest value of influence.

4.5.3 Influence and Click Prediction

So far, our results highlight a promising open problem: as most clicks are generated by social media in the long term, based on early sharing activity, it should be possible to predict, from all the current URLs available on a domain, the click patterns until the end of the day. Moreover, equipped with influence at the user level, we ought to leverage context beyond simple count of events for URLs to make that prediction more precise.

To focus in this chapter on the most relevant outcomes, we omit the details of the machine learning methodology and validation that we use to compute this prediction. They are however, presented in appendix to our paper [Gabielkov 2016] for references. We now present a summary of our main findings:

- One can leverage early information on a URL influence in order to predict its future clicks. For instance, a simple linear regression is shown, based on the number of clicks received by each URL during its first hour, to correctly predict its clicks at the end of the day, with a Pearson R^2 correlation between the real and predicted values being 0.83.

- We found that this result was robust across different classifications methods (such as SVMs) but it varies with information used. Being able to observe clicks at 4 h, for instance, increased the correlation coefficient of the prediction to 0.87, while information on share after 4 h leads to lower correlations of the prediction of 0.65.

- Including various features for the first 5 users sharing the URLs (in terms of followers,

or the new influence score) is not sufficient to obtain prediction of the same quality. However, this information can be used in complement to the 1 h of shares to reach the same precision as with clicks only.

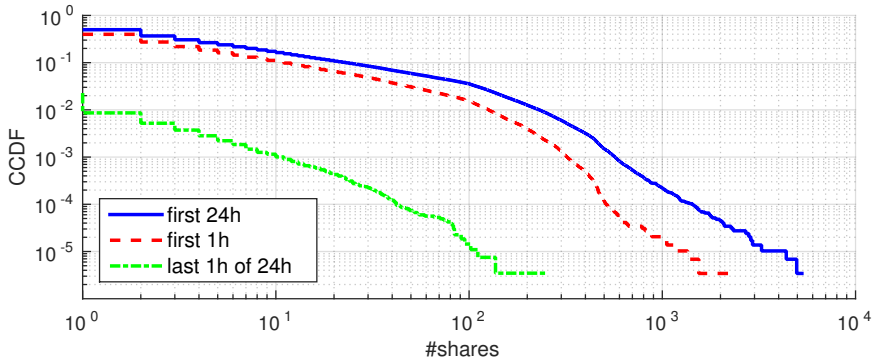
4.6 Conclusion

As we have demonstrated, multiple aspects of the analysis of social media are transformed by the dynamics of clicks. We provide the first publicly available dataset[12] to jointly analyze sharing and reading behavior online. We examined the multiple ways in which this information affects previous hypotheses and inform future research. Our analysis of social clicks showed the ability of social media to cater to the myriad taste of a large audience. Our research also highlights future area that require immediate attention. Chiefly among those, predictive models that leverage temporal property and user influence to predict clicks have been shown to be particularly promising. We hope that our methodology, the data collection effort that we provide, and those new observations will help foster a better understanding of how to best address future users' information needs.

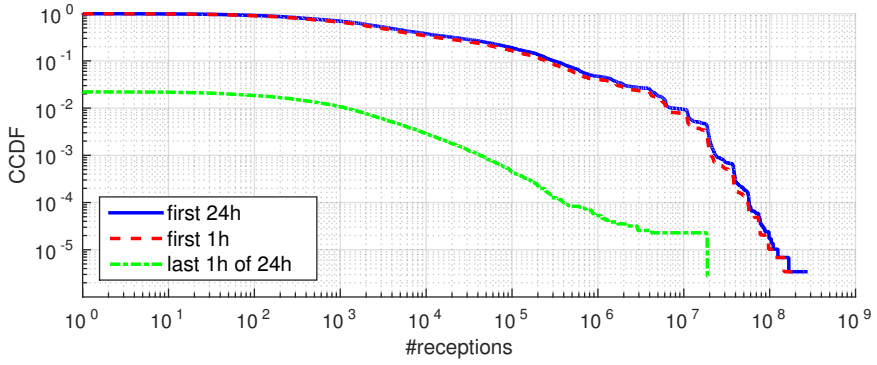
4.7 Acknowledgments

This material is based upon work supported by the National Science Foundation under grant no. CNS-1254035.

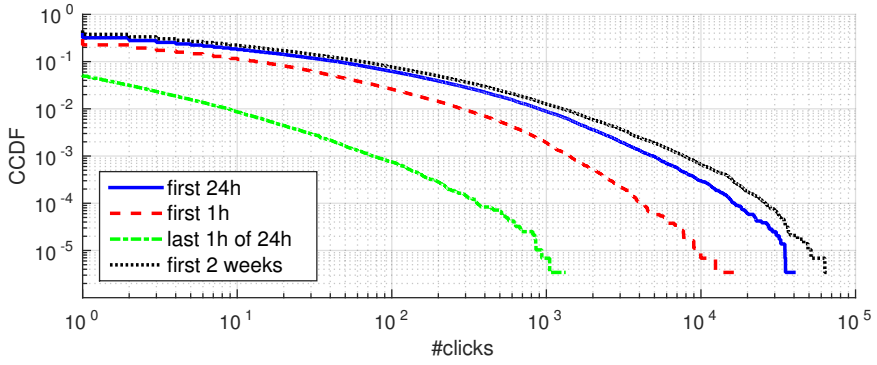
[12] http://j.mp/soTweet



(a) Shares.



(b) Receptions.



(c) Clicks.

Figure 4.8 – **Evolution of the empirical CCDF with time for three metrics.**

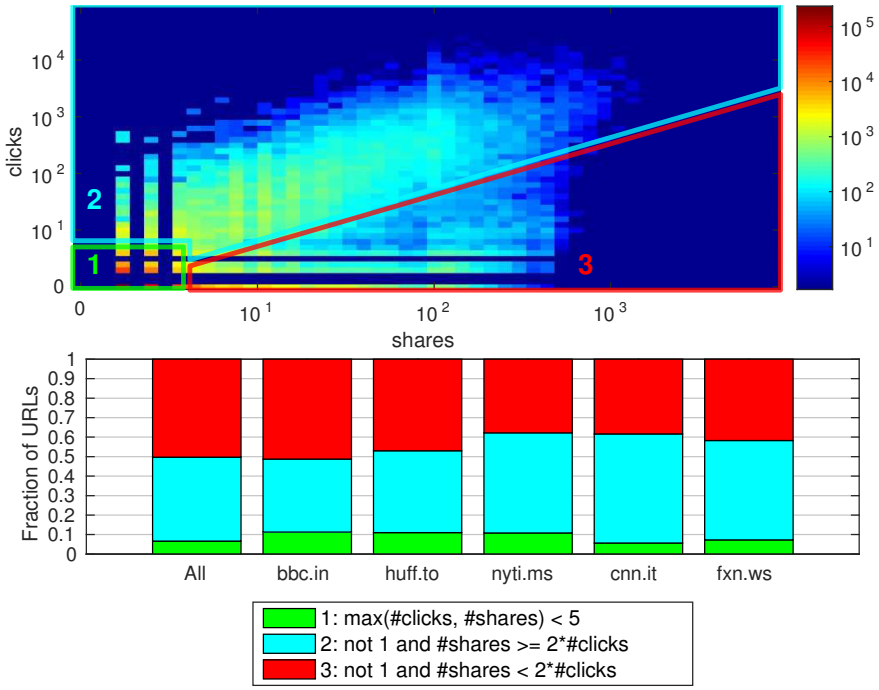


Figure 4.9 – **Joint distribution of shares and clicks, and volume of shares created by different subset of URLs.**

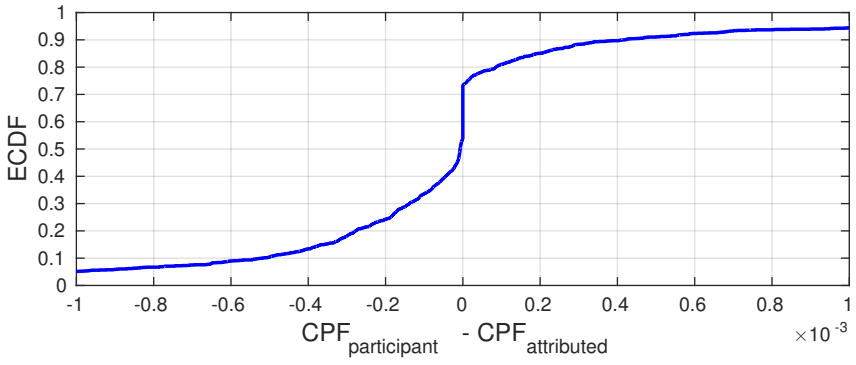


Figure 4.10 – **Bias of the estimation of CPF.**

Chapter 5

Sampling Twitter

Contents

Online social networks (OSNs) are an important source of information for scientists in different fields such as computer science, sociology or economics. However, it is hard to study OSNs as they are very large. For instance, Facebook has 1.28 billion active users in March 2014 and Twitter claims 255 million active users in April 2014. Also, companies take measures to prevent crawls of their OSNs and refrain from sharing their data with the research community. For these reasons, we argue that sampling techniques will be the best technique to study OSNs in the future.

In this work, we take an experimental approach to study the characteristics of well-known sampling techniques on a full social graph of Twitter crawled in 2012 (see Chapter 3). Our contribution is to evaluate the behavior of these techniques on a real directed graph.

5.1 Motivation

Online Social Networks (OSNs) fundamentally changed the way people communicate. The main reason of this success is that OSNs enable the creation of social links among users, mimicking in a virtual world the social relationships existing in the real life, and OSNs break the distance among users, providing seamless communications.

As a consequence, OSNs are an interesting subject of study for numerous scientific fields such as sociology, or computer science; the corner stone of such studies is the access to data such as user profiles, and user social links. However, accessing such data is often a challenge. The number of users of OSNs is constantly increasing. It will be harder to study OSNs as the data grows bigger. Users might set privacy settings that prevent access for any third party to specific information (such as the social links), and even if OSNs provide public API to access data, they rate limit this access. Indeed, companies take measures to prevent the

crawls of their social networks, *e.g.*, Twitter has discontinued the API 1.0 that supported anonymous requests and the use of already whitelisted machines; the new API 1.1 requires user authentication for each request making crawls harder and longer to perform.

One solution to work around these limitations is to sample the OSNs, but it is a challenge to assess the bias of a sample and its implications on the property we want to observe. In 2012 we collected the full graph of Twitter, resulting in a graph with 505 million nodes and 23 billion arcs (see Chapter 3). We use this graph to see how the classical sampling techniques are working with a limited sampling budget, that is a limited number of nodes that can be sampled.

5.2 Sampling Techniques

The social relationships between users are traditionally represented as a graph $G(V, E)$ with a set of vertices V and a set of edges E. In social graphs vertices represent users and edges represent 'friendship'.

The first step in the analysis of social networks is the data collection. There are multiple ways of collecting information on social networks including (i) collaboration with OSN providers, (ii) scraping and parsing the web pages of the OSN, or (iii) using the application programming interface (API) of the OSN. Collaborating with an OSN provider is the ideal but often unreachable case. OSN providers either have their own research department (*e.g.*, Facebook) or often do not show interest in collaboration with researchers. Nowadays collection of the data by scraping the web pages of an OSN is a daunting task, because modern OSNs become interactive web applications that include a lot of dynamically loading content. Scraping such OSNs require emulating the browser which is a task that is hard to set up and scale. Also web scraping can create additional load on the OSN servers. The classical way to obtain the data from an OSN is to use its API [Gabielkov 2014b, Kwak 2010, Cha 2010]; in this study we consider the latter.

The first way to crawl OSNs relies on the graph traversals, *e.g.*, breadth-first search (BFS) and random walk (RW). These methods are well known and were used in many areas, however previous studies [Gjoka 2010] show that they are biased towards high degree nodes. These methods are often used without proper characterization and correction of the bias. The bias of BFS is not formally characterized, whereas the bias of RW can be accessed via Markov Chain analysis and corrected by re-weighting the sample in either online or offline fashion. There has been multiple studies focusing on methods to obtain a sample with the original node degree distribution, *e.g.*, Metropolis-Hastings Random Walk [Gjoka 2010], unbiased sampling for directed social graphs [Wang 2010b], or Frontier Sampling [Ribeiro 2010]. Other metrics apart from the degree distributions were rarely studied.

The second way to crawl OSN is uniform sampling by ID which consists of randomly probing the ID space of the OSN for existing accounts. It is often inefficient on real OSNs, that is why it is often used *only* to obtain the 'ground truth' in studies focusing on sampling (see Section 5.3). Uniform sampling by ID is a case of rejection sampling [Leon-Garcia 2008] and is proved to yield a uniform sample regardless of the distribution of IDs in the ID space.

In the following, we provide the description of graph sampling techniques we use in this study.

BFS. *Breadth-first search* is a way to traverse a graph by sequentially visiting its nodes in the increasing order of distance from a *source* node. The algorithm of BFS with limited *budget* (size of the sample) is presented in Algorithm 1.

Algorithm 1 Breadth-First-Search

Input: Graph $G(V, E)$, source node s, budget b
Output: Sample of nodes $S \subset V$
create empty queue Q
Q.enqueue(s)
$i = 1$
while Q is not empty and $i \leq b$ **do**
  $current = Q$.dequeue()
  S.append($current$)
  $i = i + 1$
  for each node n that is adjacent to $current$ (in case of Twitter, follows or is followed by $current$) **do**
    if n not in Q **then**
      Q.enqueue(n)

RW. *Random walk* is a graph traversal technique that consists of sequentially jumping to a random adjacent neighbor (see Algorithm 2).

Algorithm 2 Random Walk

Input: Graph $G(V, E)$, source node s, budget b
Output: Sequence of nodes with repetitions $S \subset V$
$i = 1$
$current = s$
while $i \leq b$ **do**
  S.append($current$)
  $current$ = random node that is adjacent to $current$
  if $current$ not in S **then**
    $i = i + 1$

RWRW. *Re-weighted random walk* is essentially the same way to traverse the graph as described in the Algorithm 2, but with additional re-weighting applied after it is finished. RW are biased toward high-degree nodes, thus, if one wants to estimate a fraction of nodes that have a metric $m(v)$ equal to A:

$$F_A = \frac{|v \in V : m(v) = A|}{|V|}$$

one needs to apply the following re-weighting. Let S be a sequence of nodes from G visited during RW, to estimate F_A we can use the following estimator [Ribeiro 2010]:

$$F_A = \frac{1}{\sum_{v \in S} \frac{1}{deg(v)}} \sum_{\forall v \in S : m(v) = A} \frac{1}{deg(v)}$$

In the following we use this estimator when we refer to RWRW. We treat the graph as undirected, hence $deg(v)$ is the sum of the number of followers and the number of followings of a user.

USDSG. *Unbiased sampling for directed social graphs* [Wang 2010b] is a modification of RW that performs unbiasing in an online fashion by discarding a jump to a high-degree node with some probability. This algorithm is an adaptation of Metropolis-Hasting RW [Gjoka 2010] for directed graphs. Algorithm 3 presents the scheme we use in our implementation.

Algorithm 3 Metropolis-Hasting Random Walk

Input: Graph $G(V, E)$, source node s, budget b
Output: Sequence of nodes with repetitions $S \subset G$

```
i = 1
current = s
while i ≤ b do
    S.append(current)
    current_candidate = random node that is adjacent to current
    α = random number from [0, 1]
    if α ≤ deg(current) / deg(current_candidate) then
        current = current_candidate
        if current not in S then
            i = i + 1
    else
        continue
```

UNI. *Uniform sampling by ID* can be applied to OSNs that have IDs assigned to users. The algorithm consists of probing the *ID space* for existing accounts (see Algorithm 4).

Algorithm 4 Uniform sampling by numeric ID

Input: Graph $G(V, E)$, budget b, minimal ID id_{min}, maximal ID id_{max}
Output: Sample of nodes $S \subset V$

```
i = 0
while i ≤ b do
    current = random number from [id_min, id_max]
    if current exists and current not in S then
        i = i + 1
        S.append(current)
```

5.3 Practical Cost of Crawling the Graph

In this section, we discuss the practical cost of crawling the graph in terms of the time and number of requests that need to be made to the OSN server. We consider the case when we use the API to crawl the OSN. OSN APIs are often subject to strict rate limits to prevent the crawls of the OSN data. In the literature, sampling methods are often compared by the

OSN	Crawl			
	profile	friends	followers	followings
Twitter (user auth)	$\frac{1}{12}$ or $\frac{1}{1200}$ in batch	N/A	$\frac{n}{5000}$	$\frac{n}{5000}$
Twitter (app auth)	$\frac{1}{12}$ or $\frac{1}{400}$ in batch	N/A	$\frac{n}{5000}$	$\frac{n}{2500}$
Facebook (Graph API)	$\frac{10}{3}$	$\frac{n}{1500}$	N/A	N/A

Table 5.1 – **The cost of OSN crawling.** We show the number of minutes it takes to crawl given information from the OSNs provided that we respect the rate limits of the corresponding APIs and that we have a single account. When n is present in the formula, it represents the number of friends, followers, or followings a user has. Starting from the v2.0, the Facebook Graph API does not support methods to retrieve the list of followers and followings.

number of nodes or edges they can crawl, however in reality the structure of the API may impart the real performance of crawling algorithms. APIs of modern OSNs have pagination because the amount of data can be overwhelming. Also requesting the user profile takes an additional request.

Let's first consider uniform sampling by ID. Many OSNs assign a numeric identifier to its users because it is used internally by the software supporting the OSN. These IDs are often invisible for the user, but can be seen in the OSN API. The simplest way to allocate IDs is to make them sequential (*e.g.*, on Twitter) or assign them sequentially in ranges (*e.g.*, Facebook in the past). This open an opportunity of an exhaustive crawl. One can probe the ID space (all the possible IDs in the OSN) to discover all accounts of the OSN. As described above, this technique yields an unbiased sample regardless of the ID allocation scheme. The effectiveness of the crawl will depend on the density of the ID allocation. For example, the density of the ID allocation in Twitter d_{tw} in 2012 can be estimated as follows $d_{tw} = \frac{\#\text{accounts}}{\text{maximal ID}} = \frac{5.37\times10^8}{7.34\times10^8} = 0.73$. It means that if we create a random integer in the range from 1 to 733,804,076 (largest ID determined by probing Twitter), we have 73% probability to discover an existing user account. However it is not always efficient to use this method. The density of the ID allocation for Facebook in 2009 was $d_{fb} = \frac{3.5\times10^8}{2^{32}-1} = 0.08$. Later, Facebook moved from 32bit IDs towards 64bit, and now $d_{fb} = \frac{1.59\times10^9}{2^{64}-1} = 0.86 \times 10^{-10}$ that makes it impractical to apply uniform sampling by ID in Facebook.

Often OSNs have numerical IDs for historical reasons. We may see in the future that these IDs will be discontinued to protect the data of the OSN from being crawled. For example, the OSN may decide to keep only a string identifier making the ID space (all possible permutations of symbols) too large. In this case traversals may be the only way to discover users.

Table 5.1 presents the cost of crawling particular information from the API of some OSNs. Depending on the target of the crawl different crawling techniques may have drastically different cost for the same size of the sample. For example, if one wants to get a uniform sample of user profiles on Twitter of size N, the uniform sample by ID would require $\frac{N}{d_{tw}} \approx 1.34N$, the classical RW with offline bias correction will need $2N$ requests (1 request for the profile and 1 request to get the next hop), and USDSG will need $(2 + t_{st})N$ requests, where t_{st} is the time spent in the node due to a rejected jump (in case of Twitter we observed $t_{st} \approx 50$).

Figure 5.1 summarizes the cost of running the sampling techniques on the Twitter social

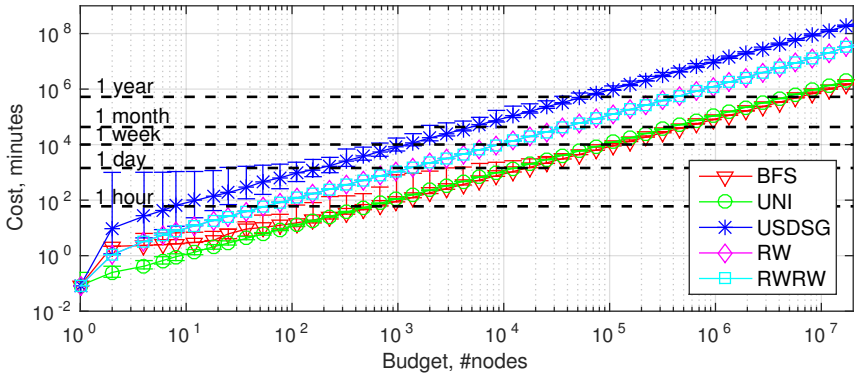


Figure 5.1 – **Crawling cost using different sampling techniques.** Dashed black lines are used to mark the timescale.

graph. *We show that the cost of performing RW and RWRW (resp. USDSG) is one (resp. two) order(s) of magnitude higher than the cost of performing BFS or UNI.*

5.4 Estimation of User Activity

In this section, we use the five sampling techniques to estimate global user activity metrics on Twitter such as number of active users and the volume of tweets. These metrics are rarely published in the reports of Twitter or analytic companies. Moreover, these reports do not give a clear definition of active accounts and do not provide accurate time-frames. We argue that the data on user activity on Twitter is essential to understand the way Twitter works, its use and evolution. We propose to use the following metrics.

Number of active users. We define an *active* account as an account that tweeted during last 24 hours, we also tried other duration such as 2 days or 1 week and didn't observe any qualitative difference.

Number of tweets ever sent. We compute the total number of tweets ever sent on Twitter as the sum of the number of tweets of all Twitter users.

Number of tweets sent daily. Such information as the number of tweets sent daily is not directly available from the Twitter API, but it can be estimated. One can compute the *tweet frequency* for users by diving the number of tweets they have by the lifetime of their account in days. The sum of the tweet frequency of all Twitter users yields the estimate of the number of tweets sent daily. This method can be improved by considering the tweet frequency of active users only[1].

Note that to properly apply these metrics on a sample one needs to know the size of the graph or the relative size of the sample to scale it up to the whole Twitter. There are two ways to get around this limitation. First, we can estimate the current size of the Twitter

[1]Another way to estimate the daily volume of tweets is to monitor the 1% sample of tweets available from the Twitter streaming API. Yet, we will need to rely on the fact that this sample is 1% and is not manipulated. Actually, as of today, the claim about the sample having exactly 1% is removed from the Twitter documentation.

graph by looking at the ID of a newly created account (to get the maximal ID) and polling a set of random user IDs for existing accounts (to estimate the density of ID allocation). Second, the measured values can still be useful for comparative studies, *e.g.*, to follow the evolution of metrics.

Figure 5.2 present the results, we show how good are the five sampling techniques in the estimation of the metrics defined above. Given the cost of the sampling techniques (see Figure 5.1), *the current best option is to use uniform sampling by ID (UNI)* that can provide accurate estimates in reasonable time (*e.g.*, 1 day). In case there is no access to numeric IDs and UNI cannot be used, the best option is to use re-weighted random walk (RWRW). Also, USDSG provides the same estimate as RWRW, but has higher variance and much higher cost. Note that we observe an overestimate of approximately 25% when using RWRW and USDSG due to the fact that graph traversals cannot reach disconnected components that contain approximately 22% of the users with very low activity. BFS and RW give overestimate by orders of magnitude.

5.5 Estimation of the Distribution

In this section, we use sampling techniques to estimate the distribution of the following user characteristics:

- Number of followers, followings, or tweets.
- Number of days since last tweet.

We define the distance between two distributions as follows. Let the distributions have the empirical probability density function $A(x)$ and $B(x)$, then the distance between the two distributions is

$$D(A, B) = \sum_{\forall x} |A(x) - B(x)|.$$

Figure 5.3 present the results. The only sampling technique that is close to the original distribution is UNI. The other techniques are relying on graph traversals; as we discovered in Chapter 3, the largest strongly connected (LSC) component is only 50% of the size of the graph. In our samples that go up to 10% of the graph in size, we observed that 70% to 80% of nodes belong to LSC, as opposed to 50% in the full graph. This leads to overestimate of all the metrics; indeed, the LSC is the most active component (see Section 3.4) and the only one that is strongly connected (see Section 3.3), other components are loosely connected, hence, are less likely to be discovered during traversals. Yet again, *the most accurate way to estimate the distribution of user characteristics is to use UNI to sample the graph*; if UNI cannot be applied, the results of the estimation obtained with graph traversal techniques should be corrected with regard to the structure of the Twitter social graph.

5.6 Discussion

We have applied classical sampling techniques to the largest Twitter dataset ever collected. On the one hand, we showed that all classical sampling techniques introduce bias toward high degree nodes. This bias can completely change the results of the studies that rely on the

partial crawl of the social graph. This motivates the need for a deeper study of the internal structure of social graphs to design an unbiased technique to sample directed OSNs.

We show that it is important to properly address the sampling bias, because classical sampling techniques, such as RW and BFS, are biased towards high degree nodes. Also, we argue that one needs to carefully account for the practical cost of sampling when designing sampling algorithms (*e.g.*, the cost of Metropolis-Hasting RW is up to two orders of magnitude higher than the cost of the BFS with the same number of nodes sampled). We show that we can easily estimate such metrics as number of active accounts or number of tweets sent within the time-frame of one day.

5.7 Future Work

In this section we discuss the future directions of this study of the Twitter social graph sampling.

5.7.1 What's the bias of my sample?

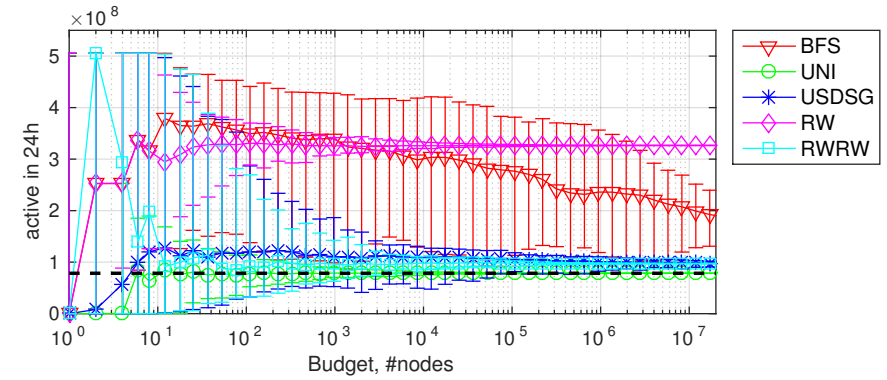
Nowadays, analytic companies heavily use social networks to collect data, and to analyze it to make business decisions. The most common use of social network analytics is to mine the sentiments and the properties of the users using a product. But what is more important is to understand how exactly these users are different from the rest of the Twitter population; this information can be useful to understand the targeting audience of the product or to show how to expand it. This question can be answered by using graph sampling with the proper correction of its biases.

5.7.2 Is Twitter dying?

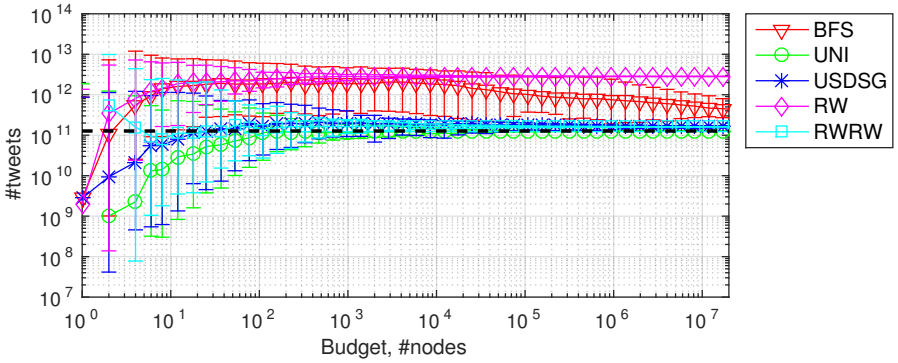
Lately, there has been numerous news articles[2] speculating that Twitter is dying and its audience is decreasing. Such articles often rely on the reports of analytic companies or leaked data of questionable credibility. Indeed, there is very little information available about user activity on Twitter, official reports publish by Twitter often operate with blurred terms and present the data in a way that is profitable to Twitter. Moreover, there is no easy way to verify the data in these reports; especially since Twitter went public and started offering shares on the New York Stock Exchange in November 2013[3], any information leak can cause the shares to drop.

We believe that the only way to validate of invalidate the claim that Twitter is dying is to perform periodic samples of the graph to access the activity of Twitter users.

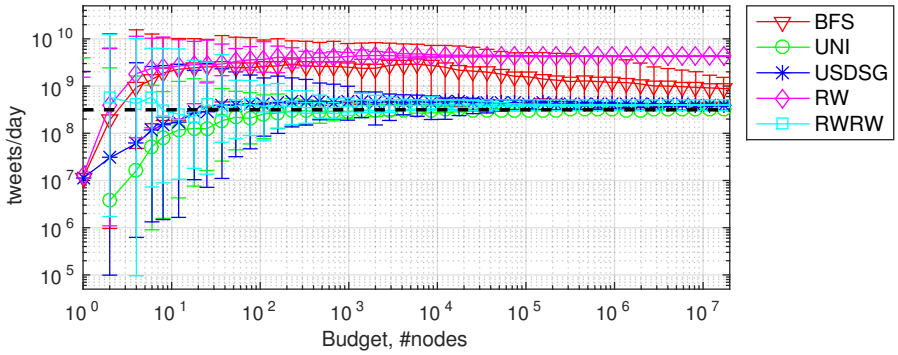
[2] https://www.quora.com/Is-Twitter-dying
[3] http://business.time.com/2013/11/07/live-updates-twitter-goes-public/



(a) Number of users that tweeted in last 24 hours.

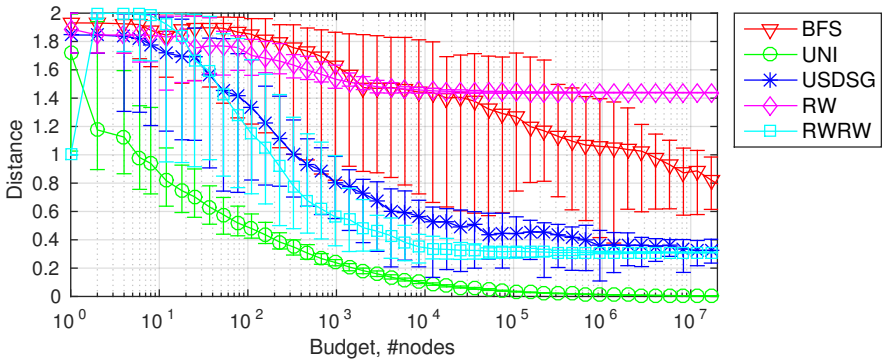


(b) Total number of tweets ever sent.

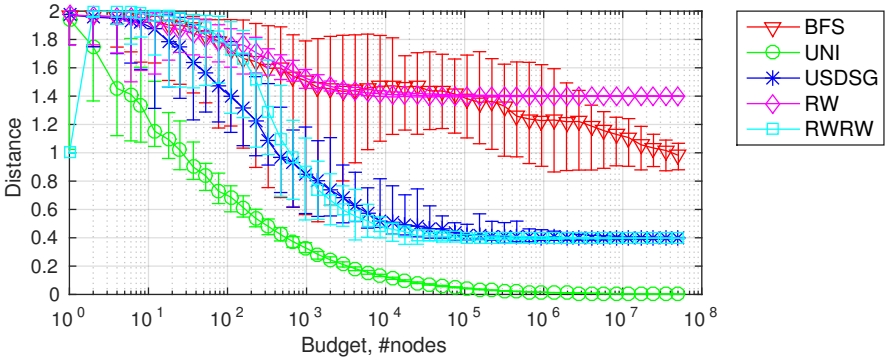


(c) Daily volume of tweets.

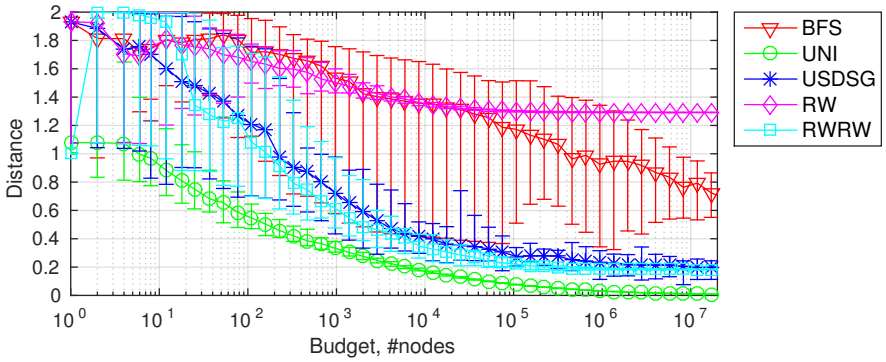
Figure 5.2 – **Estimation of global activity on Twitter.** The black dashed line shows the estimation on the full Twitter social graph. Each of the five sampling techniques were repeated 100 times, the error bars show the 5th and 95th percentile.



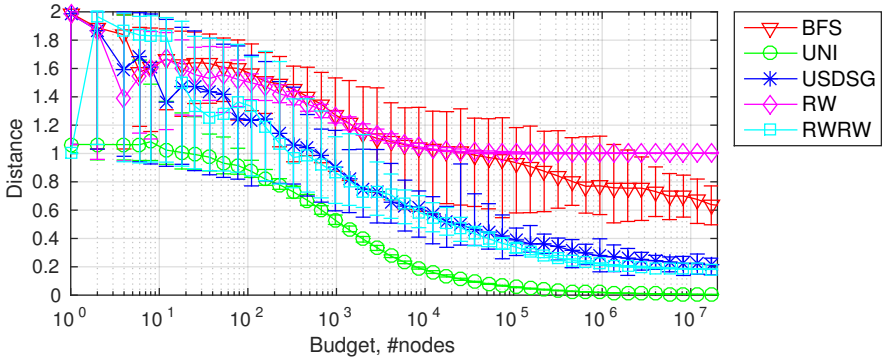
(a) Distribution of the number of followers.



(b) Distribution of the number of followings.



(c) Distribution of the number of tweets.



(d) Distribution of the number of days since last tweet.

Figure 5.3 – **Estimation of the distribution of local metrics.** Each of the five sampling techniques were repeated 100 times, the error bars show the 5th and 95th percentile.

Chapter 6

Conclusion

In this thesis, we make the following contributions.

First, we present the largest, most complete dataset of the Twitter social graph that may be the latest of its size. This graph contains 505 million accounts connected with 23 billion arcs, and its anonymized version is publicly available on the website of our project *soTweet*[1]. In addition, we present a methodology to practically compute the macrostructure of any directed social graph and to exhaustively classify each account to one of the identified components. We applied this methodology to the Twitter social graph and found that only 50.71% of the accounts belong to the LSC component, that 21.35% belong to IN component, and that 21.60% of the accounts (in the DISCONNECTED component) have no path to the other accounts. We show that the main components of the macrostructure of the Twitter social graph correspond to specific usages. For instance, the LSC component hold most of the regular Twitter activity, the IN component holds passive readers and DISCONNECTED component holds abandoned accounts and spammers. Also, we present a simple methodology to explore the evolution of the macrostructure of Twitter with time, we validate this methodology using the public datasets crawled in 2009. We applied this methodology to the Twitter social graph and found that its evolution can be divided into two phases. The first phase (2006 – 2009) is the early time when many users joint Twitter and did not know how to use it. This phase ends in 2009 when celebrities started to join Twitter; this is the year when the first user reached 1 million followers. The second phase (2009 – 2012) is marked by near-exponential growth of the number of Twitter users and an increasing amount of passive readers; it signifies that Twitter is getting more traits of information network.

Second, we performed an analysis of information dissemination focusing on the case of news media URLs shared on Twitter. Sharing behaviors alone has been studied for years; we are the first ones to combine the sharing data from social media with the click data. As we have demonstrated, multiple aspects of the analysis of social media are transformed by the dynamics of clicks. (i) News articles that are not promoted through headlines are responsible for the long tail of content popularity; they generate more clicks both in absolute and relative terms. (ii) Social media attention is more long-lived when clicks are taken into account, it goes in contrast with temporal evolution estimated from shares or receptions. (iii) We show that number of shares, that is nowadays used as a measure of impact and influence, poorly correlates with the number of actual clicks.

We provide the first publicly available dataset to jointly analyze sharing and reading behavior online[1]. We examined the multiple ways in which this information affects previous hypotheses and inform future research. Our analysis of social clicks showed the ability of social media to cater to the myriad taste of a large audience. Our research also highlights future area that require immediate attention. Chiefly among those, predictive models that leverage

[1] http://j.mp/soTweet

temporal property and user influence to predict clicks have been shown to be particularly promising. We hope that our methodology, the data collection effort that we provide, and those new observations will help foster a better understanding of how to best address future users' information needs.

Third, we presented a practical study of sampling methods. We have applied classical sampling techniques to the largest Twitter dataset ever collected. We showed that classical sampling techniques introduce bias toward high degree nodes. This bias can completely change the results of the studies that rely on the partial crawl of the social graph. This motivates the need for a deeper study of the internal structure of social graphs to design an unbiased technique to sample directed OSNs. However, the bias of these techniques towards high degree nodes gives a simple instrument to crawl high degree nodes, which correspond to popular users in the OSN.

We showed the drastic difference in the cost of sampling techniques when practical aspects of the API structure are taking into account. Our study highlights the need to carefully account for the practical cost when designing a sampling technique. Also, we evaluated the efficiency of five sampling techniques to estimate various metrics of Twitter that are vital for understanding its activity.

Appendix A

Résumé des travaux de thèse

A.1 Introduction

Un réseau social peut être défini comme un ensemble d'entités sociales (par exemple, des individus, des groupes, des organisations, etc.) reliées entre elles avec différents types de relations. L'analyse des réseaux sociaux est un domaine académique interdisciplinaire à l'intersection de la sociologie, de la psychologie, des mathématiques et de l'informatique.

L'idée des réseaux sociaux prend ses racines dans les travaux de deux sociologues, Émile Durkheim et Ferdinand Tönnies, publiés au début des années 1890. Ces travaux sur les groupes sociaux annonçaient l'idée des réseaux sociaux. Tout au long du 20e siècle, il y a eu des développements majeurs par plusieurs groupes de chercheurs dans les différents sites cités précédemment et de façon indépendante. À cette époque, l'enregistrement et l'analyse des interactions sociales systématiques ont été effectuées sur des petits groupes, par exemple des groupes de travail ou des classes d'élèves, en raison de la difficulté naturelle de faire de grandes études scientifiques avec de vraies personnes. Dans les années 1970, différentes analyses des réseaux sociaux ont été combinés. En 1969, Travers et Milgram [Travers 1969] créent leur expérience bien connue dans laquelle ils ont demandé à 196 personnes choisies arbitrairement dans le Nebraska et à Boston de remettre une lettre à une personne cible dans le Massachusetts via une chaîne de connaissances. Cette expérience a été révolutionnaire car elle a suggéré que la société humaine est un réseau « petit monde » avec un chemin de longueur courte, le nombre moyen d'intermédiaires entre les personnes initiales et la personne cible était de 5,2. Elle a joué un rôle important dans le développement du concept de « six degrés de séparation » qui suggère que chaque personne dans le monde éloignée d'une autre par six degrés.

En raison du développement rapide d'Internet à la fin des années 1990, les réseaux sociaux en ligne (RSL) ont émergé. Les RSL, également connu comme sites de réseautage social, sont des services basés sur Internet qui permettent aux individus de (a) créer un profil public ou semi-public dans le service, (b) d'établir des relations avec d'autres utilisateurs du service et (c) d'afficher tout ou partie de la relation entre d'autres utilisateurs. La fonctionnalité du service peut également inclure la possibilité d'échanger des messages, du contenu multimédia ou d'exprimer des réactions, mais ces caractéristiques peuvent varier d'un RSL à l'autre.

En 1997, le premier réseau social `SixDegrees.com` a été lancé. SixDegrees était en avance sur son temps, les utilisateurs ne savaient pas quoi faire après s'être enregistrés et connectés à leurs amis (le service a été fermé en 2013). A partir de 2003, de nombreux RSL ont été lancés, et les réseaux sociaux dans Internet se sont déployés (voir l'enquête sur les sites de réseautage social [Boyd 2007]). Aujourd'hui les RSL peuvent être divisés en deux catégories, *généraux* et *spécialisés*[1]. Des exemples notables de RSL à usage général sont Facebook,

[1] `https://en.wikipedia.org/wiki/List_of_social_networking_websites`

Twitter et Google+. Les RSL spécialisés populaires incluent Instagram.com (qui met l'accent sur le partage de photos et de vidéos) et LinkedIn.com (pour le réseautage professionnel). La popularité des sites de réseautage social varie avec le temps et la région géographique. Par exemple, en raison de la réglementation gouvernementale, plusieurs sites Web tels que Facebook, Twitter et Google+ ne sont pas disponibles en Chine, ce qui a conduit à la création de sites de réseaux sociaux locaux tels que Qzone, RenRen et Sina Weibo. Un autre exemple de RSL à usage général est VK.com qui est populaire dans les républiques post-soviétiques.

Les RSL à usage général sont normalement utilisés pour rester en contact avec les amis « hors ligne » mois servent peu à établir de nouvelles relations. Ça implique que nous pouvons voir RSL comme un modèle de relations sociales d'un individu. Nous pouvons conclure que les RSL ne facilitent pas seulement la communication entre leurs utilisateurs, mais créent aussi une source d'information précieuse pour les chercheurs et les entreprises. Avec l'aide des RSL les chercheurs peuvent faire des études à l'échelle planétaire.

A.1.1 Pourquoi étudier les RSL?

Avec le nombre d'utilisateurs enregistrés dans les RSL qui atteint des milliards, il est difficile de nier que les RSL jouent un rôle important dans la vie des gens. Dans ce qui suit, nous exposons les raisons pour lesquelles nous pensons que les RSL devraient être étudiés.

L'analyse du trafic. Selon diverses sources, le trafic des réseaux sociaux est responsable de 30% de tout le trafic vers les contenus publiés en ligne (voir Section 4.1). Cependant, on sait peu de choses sur la façon dont ce trafic est formé.

La publicité sur Internet est le revenu majeur pour de nombreuses sociétés telles que Google, Facebook ou Twitter[2]. En outre, elle est l'une des façons de garder le contenu librement accessible pour des ressources en ligne comme les mass-médias. L'étude du trafic en provenance des RSL est cruciale pour l'entreprise reposant sur les revenus de la publicité.

Mesure de l'opinion publique. Internet est de plus en plus répandu, selon des estimations en 2014, 40% de la population mondiale utilise internet, ce qui en fait l'un des meilleurs moyens de connaître l'opinion du peuple sur des événements tels que des élections politiques ou des lancements de produits. Les RSL sont une source précieuse d'information pour les analystes d'affaires et les chercheurs, mais ils peuvent aussi impacter le marché financier. Par exemple, un tweet unique peut provoquer des fluctuations majeures sur le marché boursier[3].

Les systèmes de recommandation sont devenus extrêmement fréquents au cours des dernières années, ils sont utilisés pour proposer des produits pour le consommateur en fonction de leur intérêt. Des études antérieures montrent que les utilisateurs qui ont des relations dans les RSL sont plus susceptibles de partager des intérêts [Zhang 2014], par exemple, les genres de films ou de musiques. Ainsi, l'exploitation des données des RSL peut améliorer les résultats de la recherche ou la qualité des recommandations.

[2] http://www.statista.com/statistics/460687/digital-ad-revenue-select-companies/
[3] http://j.mp/20P5vUc

En outre, elle peut être utilisée pour améliorer la connectivité dans les systèmes peer-to-peer en créant une connexion supplémentaire entre les utilisateurs ayant des liens sociaux sur un RSL [Zhang 2014].

La confiance sociale et le spam. Dans le monde moderne, la quantité d'information que les gens reçoivent chaque jour est énorme et ne cesse d'augmenter, mais il est difficile de savoir quelle information est digne de confiance et quelle est celle qui ne l'est pas. Un autre sujet étroitement lié est la détection de spams. Bien qu'il soit difficile de comprendre à partir du contenu d'un message particulier s'il est spontané, indésirable ou apocryphe, nous pouvons bénéficier de la connaissance des liens sociaux de l'utilisateur, par exemple, en considérant le contenu provenant des « amis » sur les RSL plus dignes de confiance.

Modélisation des systèmes. De nombreux paramètres, tels que les distributions de popularité, émergent des comportement des gens. Ces paramètres peuvent être révélés par la mesure des RSL et peuvent être utilisées pour la modélisation. Par exemple, les données sur le nombre de *followers* sur Twitter peuvent être utilisés pour modéliser la distribution de popularité pour les réseaux orientés information.

A.1.2 Pourquoi Twitter?

Dans cette thèse, nous avons choisi Twitter comme étude de cas. Twitter est un service de microblogging qui permet à ses utilisateurs d'envoyer des messages courts (tweets) jusqu'à 140 caractères, les utilisateurs peuvent aussi souscrire aux messages des autres. La page principale d'un utilisateur enregistré sur Twitter est appelée un *timeline*, il montre la liste des tweets des personnes auxquelles l'utilisateur est abonné dans un ordre chronologique inverse.

Twitter a été lancé en 2006 et a maintenant plus de 332 millions d'utilisateurs actifs (en Janvier 2016), ce qui en fait le troisième plus grand RSL dans le monde. Bien que Twitter n'est pas le RSL le plus populaire, il a quelques caractéristiques qui le rendent différent des autres RSL et plus intéressant à étudier du point de vue de la recherche; ces caractéristiques sont décrites dans ce qui suit.

Initialement les RSL permettaient à leurs utilisateurs de créer des relations d'amitié entre leurs utilisateurs. Cette relation nécessite une confirmation des deux parties de la relation et est mis en œuvre comme un type majeur de relation dans les RSL tels que SixDegrees, Facebook ou LinkedIn. Cependant l'amitié n'est pas le seul type de relation sociale que nous rencontrons dans le monde. Dans un cadre traditionnel, quand les gens regardent la télévision ou lisent les journaux, ils sont normalement intéressés par le contenu, mais l'éditeur du journal ou le producteur de l'émission de télévision n'a aucun moyen direct pour discuter avec les gens qui consomment le produit. Cette relation éditeur-abonné peut être modélisée par l'introduction d'une relation *follow* quand un utilisateur peut suivre un autre utilisateur et recevoir ses mises à jour sans la confirmation de ce dernier. Les relations de suivi dans les RSL ont été introduites sur LiveJournal en 1999, mais les termes *follow* et *follower* sont plus connus dans le contexte de Twitter. Le concept de cette relation unidirectionnelle est beaucoup plus flexible, car il peut aussi bien modéliser une relation éditeur-abonné, par exemple lorsqu'un utilisateur suit une célébrité qu'une relation d'amitié, par exemple lorsque

deux utilisateurs se suivent. Voilà pourquoi nous croyons que Twitter est l'une des meilleures sources d'information sur les relations sociales entre les gens.

Une autre caractéristique intéressante de Twitter est la limitation de 140 caractères sur la longueur du poste. Initialement mise en place pour sa compatibilité avec la messagerie SMS, la limite de 140 caractères a joué un grand rôle dans la formation de la culture de Twitter. En raison de la nature courte des tweets, Twitter est utilisé pour envoyer des mises à jour à des gens. Voilà pourquoi Twitter est largement utilisé lors des catastrophes naturelles ou des rassemblements publics pour relier les gens et servir de canal de communication supplémentaire. Nous pouvons dire que Twitter est devenu le bouche-à-oreille d'Internet. Nous croyons qu'un message court est une meilleure représentation de ce qui est dans l'esprit des gens qu'une longue publication sur Facebook, les tweets courts nous donnent l'occasion d'observer la réaction instantanée de personnes aux événements.

En outre, cette limite a augmenté l'utilisation des URL shorteners comme `goo.gl` et `bit.ly` de sorte que les URL ne consomment pas la majorité de la limite des 140 précieux caractères . La vaste utilisation de shorteners sur Twitter ouvre la possibilité sans précédent d'étudier la diffusion du contenu de Twitter vers le Web en utilisant les URL shorteners pour fournir des statistiques détaillées sur le nombre réel de clics effectués sur les URL (voir Chapter 4).

Compte tenu de la longueur contrainte du tweet, les utilisateurs de Twitter ont développé une série de conventions qui permettent aux utilisateurs d'ajouter de la structure à leurs tweets [Boyd 2010]. Ces conventions ont émergé des utilisateurs et sont devenues si populaires qu'elles ont été inclus officiellement dans Twitter.

Mentions. Les utilisateurs ont commencé à utiliser la syntaxe `@username` pour mentionner ou s'adresser à des utilisateurs particuliers dans leurs tweets. Désormais, les utilisateurs sont avertis s'ils sont mentionnés dans une conversation.

Re-tweets. Pour partager le message d'un autre utilisateur de Twitter avec leurs *followers*, un utilisateur peut copier le contenu du message et l'afficher en le faisant précéder de la mention `RT @username` ou `via @username`. Maintenant, il existe un bouton retweet dans Twitter.

Hashtags. Les utilisateurs peuvent regrouper les messages par sujet ou type en utilisant un hashtag, qui est un mot préfixé avec un signe « # ». Un clic sur un hashtag donne les messages contenant ce hashtag dans un ordre chronologique inverse.

Toutes ces conventions n'aident pas seulement les utilisateurs à mieux naviguer à travers Twitter, mais aussi à fournir aux chercheurs et aux analystes un moyen plus facile d'interpréter l'information sans mettre en œuvre des outils lourds de traitement du langage naturel. En outre, à la fin de 2009, des *Twitter lists* ont été introduites, ce qui permet aux utilisateurs de créer et de suivre des listes *ad hoc* d'auteurs au lieu des auteurs individuels.

En Juillet 2015, Twitter a étendu la limite pour les messages privés à 10 000 caractères. Plus tard, en Janvier 2016, Jack Dorsey (PDG actuel de Twitter) a révélé que Twitter prévoit d'étendre la limite de caractères pour les tweets egalement; cette limite serait aussi de 10.000 caractères, mais les utilisateurs devront cliquer pour voir quoi au-delà des 140 caractères.

Twitter n'impose pas l'utilisation de noms réels aux utilisateurs. D'une part, cela implique une plus grande liberté d'expression. Certains RSL, par exemple SixDegrees.com, ont perdu

beaucoup d'utilisateurs après l'introduction de l'obligation d'utiliser des noms réels et en suspendant les comptes qui semblaient faux. En outre, il est important de garder l'anonymat lorsque les conversations touchent des sujets tels que la politique, ou lorsque la divulgation de l'information peut causer des dommages à l'expéditeur. Cependant, l'anonymat crée plus de possibilités pour les spammeurs et pour manipuler l'opinion, ce qui est un sujet de recherche intéressant en soi.

A.1.3 Défis et contributions

Dans cette section, nous décrivons l'approche que nous avons pris dans l'analyse de Twitter, les défis auxquels nous nous sommes confrontés, et les contributions qui ont été faites. Notre analyse consiste en trois étapes. (i) Comprendre le médium où l'information se propage, c'est-à-dire est le graphe social Twitter. (ii) Étudier la propagation de l'information sur Twitter en se concentrant sur les articles des mass-médias. (iii) Étudier le problème de l'utilisation de l'échantillonnage de graphe pour l'estimation des divers paramètres lorsque l'accès aux données est limité.

A.1.3.1 Structure de Graphe Social

Pour comprendre la propagation de l'information sur Twitter, nous devons d'abord comprendre la structure du médium où elle se propage. Dans Twitter, ce médium est le *graphe social* qui a les utilisateurs comme sommets et les relations *follow* comme arêtes (arcs). L'information peut se propager dans ce graphe d'un utilisateur à ses *followers*, ensuite elle peut être retweetée et atteindre les *followers* des *followers* d'un utilisateur. Alors que la volonté des utilisateurs de retweeter quelque chose dépend fortement du contenu et des utilisateurs eux-mêmes, il ne fait aucun doute qu'il n'y a aucun moyen pour l'information de circuler entre les utilisateurs qui sont déconnectés dans le graphe social. Nous pouvons dire que la structure de graphe social limite la propagation de l'information. Une première étape dans la compréhension de la façon dont l'information se propage sur Twitter est d'étudier la structure de son graphe social. Nous avons fait face aux défis suivants au cours de cette étude.

Défis.

Collecte de données. Pour analyser la structure du graphe social de Twitter, nous devons d'abord crawler le graphe. Twitter ne donne pas accès à son graphe social, en outre, ils utilisent une infrastructure distribuée pour soutenir le fonctionnement de Twitter.

Le moyen le plus fiable pour obtenir les données est l'interface de programmation d'application Twitter (API). Cependant, l'accès à cette API est soumise à de strictes limites de taux de collecte, au moments où nous avons recueilli les données (2012) ces taux-limites ont été appliqués par adresse IP. Nous avons dû mettre en œuvre un crawler distribué qui utilisait 550 machines réparties dans le monde entier. Il nous a fallu quatre mois pour recueillir le graphe social complet de Twitter, les détails sont présentés dans le Section 3.2.

Traitement des données. Un autre défi consiste à traiter les données de grande taille, le graphe que nous avons recueilli se compose de 537 millions d'utilisateurs et de 24 milliards d'arcs et il nécessite 74GB de RAM pour être stocké dans le format d'une liste d'adjacence.

Nous avons testé plusieurs outils de l'état de l'art, y compris Hadoop[4], NetworkX[5], SNAP[6], et GraphChi [Kyrola 2012]. Map-reduce (Hadoop) et les approches vertex-centric (GraphChi) se sont révélés inefficaces dans l'accomplissement de l'algorithme de parcours en largeur (BFS) qui est essentiel pour le calcul des composants fortement connectés (CFC). NetworkX et SNAP semblent être inutilisables pour les grands graphes ou les graphes avec des nœuds de haut degré. Dans un premier temps, nous avons développé notre propre solution qui a fait les calculs sur le graphe en utilisant un algorithme « divide and conquer », par exemple, nous avons divisé le graphe social de Twitter en blocs qui entrent dans la RAM, calculé la CFC dans chaque bloc, puis fusionné les résultats. Ensuite, lorsque nous avons eu accès à des machines avec plus de RAM, nous avons utilisé une combinaison d'un code écrit Python avec un module écrit en C destiné à coder les structures de données. L'utilisation du C nous aide à éviter la surcharge de la mémoire tout en conservant l'avantage de Python en termes de mise en œuvre rapide et la manipulation de données efficace[7].

Interprétation des résultats Nous avons calculé différentes statistiques sur le graphe social de Twitter, par exemple, la distribution des degrés, les composants faiblement connectés, les CFC et nous avons conçu et appliqué la version généralisée de la décomposition graphique de Broder *et al.* [Broder 2000]. Nous avons obtenu la décomposition du graphe en 8 composants et nous avons cherché la signification physique de ces composants. Cette étape demande le calcul de nombreux paramètres par composant conjointement avec l'inspection manuelle des comptes d'utilisateurs appartenant à ces différents composants, ce qui prend du temps. Les résultats de ces activités sont présentés dans Section 3.4.

Contributions Nous avons recueilli le graphe social complet de Twitter en 2012 qui est probablement le dernier jeu des données d'une telle ampleur. Nous sommes passés par un processus éthique et nous avons partagé une version anonyme de notre jeu des données[8] qui a été approuvée par le COERLE (le comité d'étique d'Inria). Nous avons eu 32 demandes d'accès de chercheurs du monde entier, dont 18 ont signé l'accord de licence et obtenu le jeu des données à des fins diverses comme l'étude de l'échantillonnage non biaisée, la génération de graphe, la propagation d'influence, la crédibilité de nœuds, les tests de logiciel de traitement de graphe (par exemple, problème de partitionnement de graphe pour Apache Spark [9]) ou des algorithmes (par exemple PageRank personnalisé, la détection de communautés, ou l'estimation du diamètre de graphe).

Nous avons amélioré et appliqué la méthode de décomposition de Broder *et*

[4] http://hadoop.apache.org/
[5] https://networkx.github.io/
[6] http://snap.stanford.edu/
[7] Python est un outil excellent pour le développement rapide, mais il a un surcoût important en termes d'utilisation du CPU et en termes de consommation de mémoire (dans Python tout est représenté sous forme d'objets qui ont des en-têtes, par exemple un entier nécessite au moins 12 octets en mémoire, alors que dans C il aurait besoin de 4 octets seulement).
[8] http://j.mp/soTweet
[9] http://spark.apache.org/

al. [Broder 2000] pour le graphe social de Twitter, ce qui nous a permis d'associer les différentes composantes de la décomposition à différents types d'utilisateurs, donc, de voir comment Twitter est utilisé aujourd'hui (voir Section 3.4) et d'observer l'évolution de l'utilisation de Twitter avec le temps (voir Section 3.5).

A.1.3.2 La propagation de l'information

Nous avons étudié la propagation de l'information axée sur la diffusion des URL des mass-médias sur Twitter. Selon Mitchell *et al.* [Mitchell 2014], plus de la moitié des adultes américains utilisent les RSL comme leur principale source d'actualité politique. En outre, il y a deux raisons techniques pour lesquelles nous avons choisi les URL des mass-médias. Premièrement, les URL sont faciles à suivre, car elles ont une syntaxe particulière, de plus les URL des mass-médias pointent vers un contenu de bonne réputation qui a été préparé par les éditeurs et classés. Deuxièmement, nous pouvons utiliser le nombre de clics sur ces URL pour évaluer extérieurement leur popularité. Au cours de cette analyse, nous avons fait face aux défis suivants.

Défis.

Collecte de données. Le volume quotidien des tweets est mesuré en centaines de millions. Il est pratiquement impossible pour une personne en dehors de Twitter, Inc. d'obtenir des données complètes sur ces tweets. Cependant, nous pouvons construire une méthodologie combinant différents paramètres de l'API de Twitter pour obtenir un échantillon de données cohérent. Par exemple, nous avons utilisé l'échantillon de 1% des tweets fournis par Twitter sans frais d'abonnement, ainsi que l'API de recherche de Twitter.

Twitter peut nous donner des informations sur qui partage quoi et quand, mais il ne peut pas nous dire exactement qui a vu cette information sur Twitter. Nous pouvons estimer le nombre de personnes qui l'ont vu en estimant le nombre de réceptions, par exemple, en additionnant le nombre de *followers* de personnes qui ont partagé les informations (nous les appelons posters). Cette approximation est assez naïve, car elle ne prend pas en compte le chevauchement entre les *followers* des posters, et nous n'avons aucun moyen de valider que les gens ont réellement vu l'information dans leur *timeline* Twitter; ils peuvent ne pas vérifier leur *timeline* régulièrement. Cependant, il existe un moyen de savoir si un article particulier des mass-médias partagé sur Twitter a été regardé par des utilisateurs. Nous avons suivi l'échantillon de 1% des tweets et découvert que de 70% à 90% des URL des médias d'information sont raccourcis en utilisant des services tels que `bit.ly`, qui fournit une API pour récupérer les statistiques sur le nombre de clics réalisés sur ses URL raccourcies. De plus, nous pouvons distinguer entre les clics provenant de Twitter et des clics provenant d'autres sites en regardant le referrer. Plus de détails sur le processus de collecte de données peuvent être trouvées dans Section 4.2.1.

Nous sommes les premiers à combiner des données sur le comportement de partage des utilisateurs et leur comportement de clic; ces données dévoilent de multiples aspects de la propagation de l'information, tous auparavant inconnus (voir Chapter 4).

Correction de biais. En raison du fait que nos données sont incomplètes, nous sommes confrontés au risque d'introduire de multiples biais dans nos données. Par exemple, nous utilisons l'échantillon aléatoire de 1% des tweets pour découvrir les URL des médias. Cet échantillon donne un sous-ensemble aléatoire de tweets, mais en termes d'URL il est fortement biaisé vers les URL populaires. La même URL peut être contenue dans plusieurs tweets; si nous échantillonnons les tweets à un taux constant de 1%, une URL partagée sur Twitter une fois a 1% de chance d'apparaître dans notre jeu des données, alors qu'une URL partagée 100 fois a $1-(1-0.1)^{100} \approx 63\%$ de chance d'apparaître dans notre jeu des données. Nous ne pouvons pas récupérer les URL que nous avons manquées, mais nous pouvons corriger les statistiques de notre étude en donnant plus de poids aux URL impopulaires. Notez que sans la correction de ce biais, nous observerions des résultats étonnamment différents. Il y a aussi un biais dans l'estimation du nombre de réceptions que nous avons mentionné dans le paragraphe précédent. Plus de détails sont présentés dans le Section 4.2.3.

Analyse de l'influence des utilisateurs. Une autre question que nous voulons étudier est l'influence des utilisateurs qui partagent le contenu sur Twitter. Nous pouvons le faire en regardant leur degré de réussite en publiant un contenu populaire. Cependant, nous ne connaissons pas de de lien entre les utilisateurs qui partagent un certain contenu et les clics effectués sur ce contenu; nous avons seulement le nombre de clics par contenu, ce qui est bien du point de vue de la vie privée, mais gênant pour nos recherches. Nous analysons l'influence de l'utilisateur en regardant le succès du contenu auquel ils ont participé en le partageant. Nous avons pu valider notre approche sur un sous-ensemble d'utilisateurs dans notre jeu des données, plus de détails dans Section 4.5.2.

Contributions Nous présentons une étude non biaisée de grande échelle des articles de presse partagés sur Twitter. Nous avons recueilli un mois de visites Web à des ressources en ligne qui sont situées dans 5 grands sites d'actualité et qui sont mentionnés dans Twitter. Ce sont les premières données qui combinent les activités de partage avec des clics. Notre jeu des données équivaut à 2,8 millions de posts, ainsi responsables de 75 milliards de vues possibles sur ce média social et 9,6 millions de clics réels à 59,088 ressources uniques. Nous concevons une méthodologie reproductible et corrigeons soigneusement ses biais. Nous envisageons de partager nos données avec la communauté (les détails seront disponibles sur `http://j.mp/soTweet`).

L'analyse des activités de partage avec des clics a révélé de multiples aspects de la diffusion de l'information, tous auparavant inconnus. Tout d'abord, les articles des médias qui ne sont pas promus dans des gros titres et sont responsables de la longue queue de popularité du contenu, génèrent plus de clics à la fois en termes absolus et relatifs. Deuxièmement, l'attention des utilisateurs sur les médias sociaux est en fait à long terme, en contraste avec l'évolution temporelle estimée des actions ou des réceptions. Troisièmement, l'influence réelle d'un utilisateur ou d'un article est mal prédit par leur nombre de posts, qui est OLOLO [pourtant lorgnent] utilisé de nos jours comme une métrique d'impact sur les ressources en ligne.

A.1.3.3 Échantillonnage du graphe

Nous avons étudié le problème de l'échantillonnage d'un graphe social. Nous avons recueilli le graphe social complet de Twitter en 2012; peu de temps après que notre crawl ait été terminé, Twitter a introduit une nouvelle version de l'API Twitter avec une authentification obligatoire pour chaque requête. Le taux limite de requêtes dans la nouvelle API est appliqué par utilisateur, alors qu'auparavant il était appliqué par adresse IP. La nouvelle API a rendu l'exploration du graphe social de Twitter beaucoup plus difficile, car il est beaucoup plus difficile de créer des comptes utilisateur que d'utiliser plusieurs adresses IP (par exemple en utilisant PlanetLab, ou toute autre plate-forme distribuée). Par ailleurs, la création automatisée de plusieurs comptes utilisateur viole les termes d'utilisation de Twitter et tous les comptes créés peuvent être suspendus.

Nous croyons que notre jeu des données collecté en 2012 est le dernier crawl de Twitter d'une telle ampleur. Les entreprises qui gèrent les RSL introduisent des mesures pour empêcher les crawles à grande échelle de leurs données, car de nos jours, l'information est une ressource précieuse pour les entreprises. C'est pourquoi nous croyons que la seule façon de faire des mesures de réseaux sociaux, excepté en cas de collaboration directe avec un fournisseur de RSL, est d'utiliser *l'échantillonnage*.

Cependant, l'utilisation mal maitrisée de l'échantillonnage peut conduire à des résultats faux. Par exemple, lorsque l'on décide d'utiliser BFS pour faire un crawl complet du graphe, on suppose que le graphe possède un composant géant connecté (qui peut être atteint avec BFS) et que les autres composants sont de taille négligeable. Selon l'objectif de l'étude, cet échantillonnage peut conduire à des résultats incorrects, par exemple, nous avons découvert que 20% des comptes de Twitter sont déconnectés, et si l'on utilise une seule direction des liens sur Twitter pour faire le crawl, il manquerait de 25% à 50% du graphe. En outre, le résultat d'une telle analyse serait hautement tributaire de la source pour démarrer le BFS.

Un autre problème apparaît lorsque l'on utilise la marche aléatoire (RW) pour effectuer une mesure du graphe. La théorie derrière le RW suppose que le graphe est connexe et possède une propriété de mélange rapide. Cependant, cette propriété n'a jamais été validée sur les graphes sociaux réels.

Nous souhaitons étudier les propriétés des techniques d'échantillonnage bien connues (par exemple, BFS, RW) sur le graphe social complet de Twitter collectés en 2012. En particulier, nous voulons voir comment ces techniques sont efficaces dans l'estimation des paramètres liés aux activités de Twitter, par exemple le nombre d'utilisateurs actifs, ou le nombre de tweets envoyés. Ces paramètres sont rarement publiés par les fournisseurs des RSL et souvent ne peuvent pas être vérifiés.

Défis.

Traitement des données. Afin d'effectuer des tests de techniques d'échantillonnage sur notre jeu des données de Twitter, nous avons besoin de construire un outil qui émulerait l'API de Twitter. Cette étape n'est pas simple pour deux raisons. Tout d'abord, pour faire un calcul fiable sur le graphe, nous avons besoin de répéter les expériences plusieurs fois. Par exemple, nous ne pouvons rien conclure sur un RW fait à partir d'un nœud particulier dans le graphe. Deuxièmement, en raison de la taille de l'ensemble de données, il est important

d'avoir un accès rapide aux données. Par exemple, si nous décidons de stocker les données sur un disque dur, pour un RW, il faudra nombreuses lectures à partir d'endroits aléatoires du fichier, ce qui est connu pour être extrêmement lent; alors, puisque nous avons besoin de répéter plusieurs fois le calcul (par exemple, pour obtenir des intervalles de confiance), le calcul peut prendre des semaines ou des mois. En fait, la plupart du temps, nous avons besoin de garder deux copies du graphe social de Twitter dans la mémoire: la liste d'adjacence des *followers* et la liste d'adjacence des *followings*. Nous remédions à ce problème en plaçant le graphe dans la mémoire en tant que structure en lecture seule et en faisant des multiples threads de calcul pour bénéficier de l'architecture multi-cœur du serveur.

En outre, certaines des techniques d'échantillonnage (par exemple celles dérivées de RW) exigent des calculs à effectuer sur chaque étape de l'échantillonnage. Par exemple, les RW sont naturellement biaisés en faveur des nœuds de haut degré, par conséquent, il faut de-biaiser les résultats à la volé (Metropolis-Hasting RW) ou hors ligne (RW repondérées). Ces calculs prennent généralement peu de temps, mais puisque nous cherchons à voir l'évolution des différents paramètres de la taille de l'échantillon, nous devons répéter ces calculs plusieurs fois, ce qui entraîne alors une augmentation significative du temps de calcul. L'optimisation de ce calcul se fait par le profilage et le refactoring du code.

Coût du crawl. Un autre problème intéressant est le coût de crawl. Ici, nous proposons de considérer le coût en termes de temps requis pour récupérer les informations des RSL ou en termes de nombre de requêtes que nous devons faire aux serveurs des RSL. En théorie, nous représentons souvent des relations sociales entre les utilisateurs comme un graphe $G(V, E)$ où V est un ensemble d'utilisateurs et E un ensemble de relations reliant les utilisateurs. Une approche naïve serait de considérer un échantillon $S \subset V$ pour avoir coût de $|S|$ ou 1 unité (par exemple requête) par utilisateur. Cependant, dans la pratique, ce n'est pas ce que nous observons. En effet, la façon dont l'information est présentée dans les RSL peut affecter le coût de crawl. Par exemple, en raison de l'énorme quantité d'informations, l'API Twitter pagine les résultats de certaines demandes, considérons que nous devons récupérer la liste des *followers* de Lady Gaga, elle avait 22 millions de *followers* en 2012, comme l'API retourne au plus 5.000 *followers* par page, ça fait $\frac{22\times10^6}{5000} = 4400$ requêtes $= 73,3$ heures pour obtenir l'information; alors que pour un utilisateur ordinaire, nous aurions besoin d'une seule minute. En outre, certaines techniques (par exemple Metropolis-Hasting RW) repose sur une connaissance supplémentaire (par exemple, le degré d'un nœud adjacent) au cours du crawl pour décider si nous sautons vers un nouveau nœud ou si nous restons dans l'actuel; cela peut augmenter le coût de l'exploration par un facteur énorme (50x dans le cas de Twitter).

En résumé, nous cherchons à évaluer la performance des techniques d'échantillonnage pour estimer divers paramètres du graphe social en tenant compte du coût réel de l'échantillonnage.

Contributions Nous montrons qu'il est important de traiter correctement le biais d'échantillonnage, parce que les techniques d'échantillonnage classiques, tels que RW et BFS, sont biaisées en faveur des nœuds de degré élevé. En outre, nous pensons que l'on doit tenir compte attentivement du coût pratique de l'échantillonnage lors de la conception des algorithmes d'échantillonnage (par exemple, le coût de Metropolis-Hasting RW est de deux ordres de grandeur plus élevé que le coût de BFS avec le même nombre de nœuds échantillonnés).

Nous montrons que nous pouvons facilement estimer ces paramètres comme le nombre de comptes actifs ou le nombre de tweets envoyés en un seul jour.

A.1.4 Structure de la thèse

Cette thèse a la structure suivante. Le chapitre 2 contient la description de l'état de l'art. Dans le chapitre 3, nous présentons l'étude de la structure du graphe social Twitter, nous identifions huit composants sur la base de la connectivité du graphe et associons ces composants à un usage particulier de Twitter. Dans le chapitre 4, nous étudions la diffusion des articles des mass-médias sur Twitter en surveillant les URL de cinq médias populaires. Dans le chapitre 5, nous reconnaissons que compte tenu de la croissance rapide des RSL et des contraintes posées par les RSL sur l'accès à leurs données, il sera difficile, voire impossible, de recueillir des grands jeux des données, nous discutons du problème de l'échantillonnage des RSL et adressons la biais de ces échantillons. Le chapitre 6 conclut la thèse.

A.2 Graphe social de Twitter

Twitter est l'un des plus grands réseaux sociaux qui utilisent des liens exclusivement dirigés entre comptes. Cela rend le graphe social Twitter beaucoup plus proche du graphe social qui permet les communications de la vie réelle que, par exemple, Facebook. Cependant, on sait peu de choses sur la façon dont la propagation de l'information sur Twitter est limitée par sa structure interne.

Dans ce chapitre, nous présentons une étude approfondie de la structure macroscopique du graphe social de Twitter en dévoilant les routes sur lesquelles les tweets se propagent, l'activité des utilisateurs associée à chaque composante de cette structure macroscopique, et l'évolution de cette structure macroscopique avec le temps sur les 6 dernières années. Pour cette étude, nous avons crawlé Twitter pour récupérer tous les comptes et toutes les relations sociales (liens de *follow*) entre les comptes; le crawl s'est achevé en juillet 2012 avec 505 millions de comptes reliés entre eux par 23 milliards de liens[10]. Ensuite, nous présentons une méthodologie pour dévoiler la structure macroscopique du graphe social de Twitter. Cette structure macroscopique se compose de 8 composants définis par leurs caractéristiques de connectivité. Chaque composant regroupe les utilisateurs avec une utilisation spécifique de Twitter. Par exemple, nous avons identifié des composants qui rassemblent des spammeurs ou des célébrités. Enfin, nous présentons une méthode pour approximer la structure macroscopique du graphe social de Twitter dans le passé, nous validons cette méthode en utilisant des anciens jeux des données, et nous discutons de l'évolution de la structure macroscopique du graphe social de Twitter au cours des 6 dernières années.

Ce travail a été accepté et présenté à ACM SIGMETRICS 2014 à Austin, TX, USA [Gabielkov 2014b].

Twitter est l'un des plus grands réseaux sociaux avec plus de 500 millions de comptes enregistrés. Cependant, il se distingue des autres grands réseaux sociaux, tels que Facebook et Google+, car il utilise exclusivement des arcs entre les comptes[11]. Par conséquent, la façon

[10] http://j.mp/soTweet

[11] Les arcs — qui sont des liens dirigés — représentent la relation de *follow* (ou suivi) sur Twitter. Si A suit B, A reçoit tweets de B, mais B ne recevra pas des tweets de A, à moins que B suive A.

dont l'information se propage sur Twitter est proche de la façon dont l'information se propage dans la vie réelle. En effet, les communications de la vie réelle se caractérisent par une forte asymétrie entre les producteurs d'information (tels que les médias, les célébrités, etc.) et les consommateurs de contenu. Par conséquent, la compréhension de la façon dont l'information se propage sur Twitter a des implications au-delà de l'informatique.

Cependant, l'étude de la propagation de l'information sur un grand réseau social est une tâche complexe. En effet, la propagation de l'information est une combinaison de deux phénomènes. Tout d'abord, le contenu des messages envoyés sur le réseau social déterminera sa chance d'être relayé. D'autre part, la structure du graphe social limitera la propagation des messages. Dans ce chapitre, nous nous concentrons particulièrement sur la façon dont la structure du graph social de Twitter contraint la propagation de l'information. Ce problème est important parce qu'il permet d'identifier les routes utilisées par les flux d'information. Pour atteindre cet objectif, nous devons surmonter deux défis. Tout d'abord, nous avons besoin d'un graphe social à jour et complet. Les plus récents jeux des données de Twitter accessibles au public datent de 2009 [Kwak 2010, Cha 2010], à l'époque Twitter était 10 fois plus petit qu'en juillet 2012. De plus, ces jeux des données ne sont pas exhaustifs, donc certaines propriétés subtiles peuvent ne pas être visibles. Deuxièmement, nous avons besoin d'une méthodologie révélant les relations sociales sous-jacentes entre les utilisateurs, une méthodologie qui peut passer à l'échelle de centaines de millions de comptes et de dizaines de milliards d'arcs. Les métriques standards globales telles que la distribution du degré ne sont d'aucune aide parce que nous avons besoin d'identifier les routes du graphe suivi par les messages. Par conséquent, nous avons besoin d'une méthodologie pour à la fois réduire la taille du graphe social et garder sa structure principale.

Dans ce chapitre, nous surmontons ces défis avec les contributions suivantes.

1. Nous avons recueilli l'ensemble du graphe social de Twitter, qui représente 505 millions de comptes connectés grâce à 23 milliards d'arcs. C'est le plus grand graphe social *complet* jamais collecté.

2. Nous dévoilons une structure macroscopique dans le graphe social de Twitter qui préserve les routes de propagation de l'information. Notre méthode étend celle de Broder *et al.* [Broder 2000] et peut être appliquée à tous types de graphes dirigés.

3. Nous montrons que non seulement la structure macroscopique du graphe social de Twitter limite la propagation de l'information, mais que chaque composant de la macrostructure correspond à un groupe d'utilisateurs avec une utilisation spécifique de Twitter. En particulier, nous montrons que les comptes réguliers, abandonnés et malveillants ne sont pas uniformément répartis sur les composants de la structure macroscopique du graphe social de Twitter. Ce résultat est important pour comprendre comment Twitter est utilisé, où sont les utilisateurs avec un usage spécifique, et comment échantillonner Twitter sans biais significatif.

4. Nous présentons une méthodologie simple pour explorer l'évolution de la structure macroscopique de Twitter avec le temps, nous validons cette méthodologie, et montrons que les anciens jeux des données à partir de 2009 ne représentent pas la structure actuelle du graphe social de Twitter. Nous explorons cette évolution dans le temps pour comprendre les changements dans l'utilisation de Twitter depuis sa création.

A.3 Clics sociaux

Les sites Web d'actualité comptent de plus en plus sur les médias sociaux pour générer du trafic vers leurs sites. Pourtant, nous savons étonnamment peu de choses sur la façon dont une conversation sur les médias sociaux mentionnant un article en ligne génère en fait des clics. Les comportements de partage, en revanche, ont été entièrement ou partiellement examinés au cours des années. Alors que cela a conduit à de multiples hypothèses sur la diffusion de l'information, chaque hypothèse a été faite en ignorant les clics réels.

Nous présentons une étude non biaisée, de grand échelle, des clics sociaux qui est la première en son genre. Nous avons collecté un mois de visites vers des ressources en ligne qui sont situées dans 5 grands sites d'actualité et qui sont mentionnées dans Twitter. Notre jeu des données équivaut à 2,8 millions de posts responsables de 75 milliards de vues possibles et 9,6 millions de clics réels vers 59,088 ressources uniques. Nous concevons une méthodologie reproductible et corrigeons soigneusement ses biais. Les propriétés des clics impactent plusieurs aspects de la diffusion de l'information, tous inconnus auparavant. (i) Les ressources secondaires, qui ne sont pas promues dans les gros titres et sont responsables de la longue queue de la popularité du contenu, génèrent plus de clics à la fois en termes absolus et relatifs. (ii) L'attention des lecteurs des médias sociaux est en fait à long terme, en contraste avec l'évolution temporelle estimée des partages ou des réceptions. (iii) L'influence réelle d'un intermédiaire ou d'une ressource est mal prédite par leur nombre des posts, et nous montrons comment cette prédiction peut être rendue plus précise.

Ce travail a été accepté et présenté à ACM Sigmetrics / IFIP Performance 2016 à Antibes Juan-les-Pins, France [Gabielkov 2016].

Les médias sociaux ont une popularité croissant et sont en train de changer radicalement la façon dont nous accédons aux ressources Web. En effet, il a été estimé pour la première fois en 2014 que la façon la plus commune pour atteindre un site Web est de cliquer sur une URL citées dans une media social[12]. Les médias sociaux représentent 30% des visites globales des sites Web, ce qui est plus élevé que les visites en raison de résultats de recherche organiques a partir de moteurs de recherche. Cependant, le contexte et la dynamique des références sur ces médias sociaux restent étonnamment inexplorés.

Pour voir une URL sur un moteur de recherche, un utilisateur a besoin de faire une demande explicite, et la réponse sera adaptée au profil de l'utilisateur en utilisant une analyse comportementale, une analyse lexicale, ou des algorithmes de personnalisation. Au contraire, sur les médias sociaux, un utilisateur a juste besoin de créer une relation sociale avec d'autres utilisateurs, puis il recevra automatiquement le contenu produit par ces utilisateurs. En première approximation, un service de recherche Web fournit des informations filtrées par des algorithmes et les médias sociaux fournissent une information filtrée par des humains. En fait, nos résultats confirment que la dynamique temporelle des clics dans ce cas est très différente.

Pendant longtemps, les études de clics dans les médias sociaux ont été entravées par l'absence de données, mais ce chapitre prouve que, aujourd'hui, cela peut être surmonté[13].

[12] http://j.mp/1qHkuzi

[13] Comme toutes les études sur les médias sociaux, nous nous appuyons sur des API ouvertes et des fonctions de recherche qui peuvent être rendues obsolètes après des changements de politique, mais toute la collection que nous présentons dans ce chapitre suit les conditions d'utilisation au moment de la collecté, et nous

Cependant, aucune information individuelle sensible n'est divulguée dans les données que nous présentons. En utilisant plusieurs techniques de collecte de données, nous sommes en mesure d'étudier conjointement les conversations sur Twitter et les clics pour les URL de cinq sites au cours d'un mois de l'été 2015. Notez que nous ne disposons pas de données complètes sur les clics, mais nous analysons soigneusement ce biais de sélection et nous avons constaté que nous avons recueilli environ 6% de toutes les URL, et observé 33% de tous les clics. Nous avons choisi d'étudier des sites d'actualité pour de multiples raisons décrites ci-dessous.

Tout d'abord, les actualités sont largement discutée dans les médias sociaux[14]. En particulier, Twitter est classé troisième derrière Facebook et Pinterest sur le volume total du trafic web de référence, et apparaît souvent deuxième après Facebook lorsque les internautes discutent de leur exposition à des actualités[12]. Savoir que les médias sociaux génèrent du trafic vers site d'actualités est important pour comprendre comment les utilisateurs sont exposés au les actualités.

Deuxièmement, la diffusion d'actualités est généralement considérée comme très influente. L'opinion politique est façonnée par les actualités, mais aussi par les intermédiaires qui réussissent à susciter l'intérêt pour ces actualités. Cela est également vrai pour toutes sortes d'opinions publiques sur des sujets allant d'une catastrophe naturelle aux prochains blockbusters de cinéma. Savoir qui relaie l'information et qui génère du trafic vers un site d'actualités est important pour identifier les véritables influenceurs.

Enfin, les actualités présentent de multiples formes de diffusion qui peuvent varier d'une promotion traditionnelle de haut en bas dans les gros titres à un effet de bouche-à-oreille du contenu partagé à l'origine par des utilisateurs ordinaires. Sachant comment les actualités sont relayées dans les médias sociaux est important pour comprendre les mécanismes derrière influence.

Nous présentons maintenant les contributions suivantes.

- Nous validons une méthode pour étudier comment les médias sociaux dirigent du trafic vers les sites d'information. Nous partirons des URL shorteners, de leurs API associées, et de plusieurs API de recherche des médias sociaux. Nous identifions quatre sources de biais mois nous validons que l'on peut soigneusement minimiser ou corriger leur impact pour obtenir un échantillon représentatif des posts/réceptions/clics pour toutes les URL des sites d'actualités étudiés. A titre d'exemple, un biais de sélection conduit à collecter seulement 6,41% des URL, mais on valide expérimentalement que ce biais peut être entièrement corrigé. (Section 4.2)

- Nous montrons les limites des études précédentes qui ont ignoré les clics. Tout d'abord, nous analysons les contenus peu populaire principalement a cause d'URL mentionnées qui ne passent pas dans les gros titres et nous montrons que leur effet est largement sous-estimé.

  Ces contenus génèrent la *majorité* des clics, mettant en évidence un processus de sélection social efficace. En fait, nous avons constaté que le nombre de retweet, affiché aujourd'hui sur les pages Web des médias pour indiquer la popularité, est une mesure inexacte du lectorat réel. (Section 4.3)

rendons nos donnés disponibles. http://j.mp/soTweet

[14]Par exemple, plus d'un adultes américains sur deux indiquent qu'ils utilisent les médias sociaux comme la principale source d'actualités politiques, ce qui est plus qu'un réseau traditionnel tel que CNN.

- Nous montrons que les utilisateurs peuvent cliquer sur les liens reçus par des médias sociaux des jours après leur réception. Cela contraste fortement avec *le comportement de partage* des médias sociaux, précédemment montré comme presque entièrement concentré dans les premières heures, comme nous le confirmons avec de nouveaux résultats. Une analyse approfondie révèle que le contenu populaire a tendance à attirer beaucoup de clics pour une longue période de temps, cet effet n'a pas d'équivalent dans le comportement de partage des utilisateurs. (Section 4.4)

- Nous proposons une analyse de l'influence basée sur les clics et les Clics-Per-Follower. L'influence des URL et des utilisateurs peut être mis à profit pour prédire une fraction significative des futurs clics sociaux, avec de multiples applications sur la performance des services Web en ligne. (Section 4.5)

A.4 Échantillonnage de Twitter

Les réseaux sociaux en ligne (RSL) sont une source importante d'information pour les scientifiques dans différents sites tels que l'informatique, la sociologie ou l'économie. Cependant, il est difficile d'étudier les RSL car ils sont très grands. Par exemple, Facebook a 1,28 milliard d'utilisateurs actifs en Mars 2014 et Twitter revendique 255 millions d'utilisateurs actifs en Avril 2014. En outre, les entreprises prennent des mesures pour prévenir les analyses de leurs RSL et s'abstiennent de partager leurs données avec la communauté des chercheurs. Pour ces raisons, nous soutenons que les techniques d'échantillonnage seront la meilleure alternative pour étudier les RSL à l'avenir.

Dans ce travail, nous prenons une approche expérimentale pour étudier les caractéristiques des techniques d'échantillonnage bien connues sur un graphe social complet de Twitter crawlé en 2012 (voir Chapter 3). Notre contribution est d'évaluer le comportement de ces techniques sur un graphique dirigé réel.

Les réseaux sociaux en ligne (RSL)ont fondamentalement changé la façon dont les gens communiquent. La principale raison de ce succès est que les RLS permettent la création de liens sociaux entre les utilisateurs, imitant dans un monde virtuel les relations sociales existantes dans la vie réelle, et brisant la distance entre les utilisateurs, en fournissant des communications transparentes.

En conséquence, les RSL sont un sujet d'étude intéressant pour de nombreux domaines scientifiques tels que la sociologie, ou l'informatique; la pierre angulaire de ces études est l'accès à des données telles que les profils utilisateur et leur liens sociaux. Cependant, l'accès à ces données est souvent un défi. Le nombre d'utilisateurs des RSL est en constante augmentation. Les utilisateurs peuvent définir des paramètres de confidentialité qui empêchent l'accès à des informations spécifiques (comme les liens sociaux), et même si les RSL fournissent les API publiques pour accéder aux données, ils limitent la vitesse de cet accès. En effet, les entreprises prennent des mesures pour prévenir les analyses de leurs réseaux sociaux, par exemple Twitter a abandonné l'API 1.0; la nouvelle API 1.1 requiert une authentification par utilisateur ce qui rend les crawls de plus en plus durs à effectuer.

Une solution pour contourner ces limitations est d'échantillonner les RSL, mais il est difficile d'évaluer le biais d'un échantillon et ses implications sur la propriété que nous voulons observer. En 2012, nous avons recueilli le graphe complet de Twitter (voir Chapter 3). Nous

utilisons ce graphe pour voir comment les techniques d'échantillonnage classiques fonctionnent avec un budget d'échantillonnage limité.

A.5 Conclusion

Dans cette thèse, nous faisons les contributions suivantes.

Tout d'abord, nous présentons le plus grand et le plus complet jeu des données du graphe social de Twitter. Ce graphe contient 505 millions de comptes et 23 milliard d'arcs, et sa version anonymisée est accessible au public sur le site web de notre projet *soTweet*[15]. En outre, nous présentons une méthodologie pour calculer la macrostructure d'un graphe social dirigé et classer chaque compte à l'un des composants identifiés. Nous avons appliqué cette méthodologie au graphe social de Twitter et constaté que seulement 50,71% des comptes appartiennent à la composante LSC, que 21,35% appartiennent au composant IN, et que 21,60% des comptes (dans le composant DISCONNECTED) n'a pas de chemin vers les autres comptes. Nous montrons que les principales composantes de la macrostructure du graphe social de Twitter correspondent à des usages spécifiques. Par exemple, le composant LSC regroupe la majeure partie de l'activité régulière de Twitter, le composant IN regroupe les lecteurs passifs et le composant DISCONNECTED regroupe les comptes abandonnés et les spammeurs . En outre, nous présentons une méthodologie simple pour explorer l'évolution de la macrostructure de Twitter dans le temps, nous validons cette méthodologie en utilisant les jeux des données publics crawlés en 2009. Nous avons appliqué cette méthodologie au graphe social de Twitter et nous avons constaté que son évolution peut être divisée en deux phases. La première phase (2006 – 2009) est le début de l'époque où de nombreux utilisateurs ont rejoint Twitter et ne savaient pas comment l'utiliser. Cette phase se termine en 2009 quand des célébrités ont commencé à joindre Twitter; 2009 est l'année où le premier utilisateur a atteint 1 million de *followers*. La deuxième phase (2009 - 2012) est marquée par une croissance quasi-exponentielle du nombre d'utilisateurs de Twitter et par un nombre croissant de lecteurs passifs; cela signifie que Twitter evolue vers à un réseau d'information classique.

Deuxièmement, nous avons effectué une analyse de la diffusion de l'information en mettant l'accent sur le cas des URL de sites d'actualités partagés sur Twitter. Les comportements de partage seul ont été étudiés pendant des années; nous sommes les premiers à combiner les données de partage des médias sociaux avec les données sur les clics. Comme nous l'avons démontré, plusieurs aspects de l'analyse des médias sociaux sont transformés par la dynamique des clics. (i) Les articles qui ne sont pas promus dans les gros titres des journaux sont responsables de la queue longue de la popularité du contenu; ils génèrent plus de clics à la fois en termes absolus et relatifs. (ii) De plus, alors que l'attention des utilisateurs des médias sociaux est courte en ce qui concerne les posts, elle est étonnamment longue lorsque l'on regarde les clics. Nous montrons que le nombre de posts, qui est aujourd'hui utilisé comme une mesure de l'impact et de l'influence, est mal corrélée avec le nombre de clics réels.

Nous fournissons le premier jeu des données accessible au public pour analyser conjointement le comportement de partage et de lecture en ligne[15]. Notre analyse des clics sociaux a montré la capacité des médias sociaux à répondre aux goûts différentes d'un large public. Nos résultats monteront que des modèles prédictifs qui permettent de prédire les clics se sont

[15] http://j.mp/soTweet

révélés particulièrement prometteurs. Nous espérons que notre méthodologie, l'effort de collecte des données que nous fournissons, et ces nouvelles observations aideront à promouvoir une meilleure compréhension de la popularité sur Internet.

Troisièmement, nous avons présenté une étude pratique des méthodes d'échantillonnage. Nous avons appliqué les techniques d'échantillonnage classiques au plus grand jeu des données de Twitter. Nous avons montré que les techniques d'échantillonnage classiques introduisent un biais vers les nœuds de haut degré. Ce biais peut changer complètement les résultats des études qui reposent sur l'analyse partielle du graphe social. Cela motive la nécessité d'une étude plus approfondie de la structure interne des graphes sociaux concevoir une technique non biaisée pour échantillonner les RSL dirigés. Cependant, le biais de ces techniques vers des nœuds de degré élevé donne un instrument simple pour analyser les nœuds de haut degré, qui correspondent aux utilisateurs populaires dans le RSL.

Nous avons montré la différence drastique de coût des techniques d'échantillonnage lorsque les aspects pratiques de la structure de l'API sont pris en compte. Notre étude met en évidence la nécessité de tenir compte attentivement du coût pratique lors de la conception d'une technique d'échantillonnage. En outre, nous avons évalué l'efficacité des cinq techniques d'échantillonnage pour estimer divers paramètres de Twitter qui sont essentiels à la compréhension de son activité.

Bibliography

[Abisheva 2014] Adiya Abisheva, Venkata Rama Kiran Garimella, David Garcia and Ingmar Weber. *Who watches (and shares) what on YouTube? And when?: using Twitter to understand YouTube viewership.* In Proc. of ACM WSDM'14, New York, NY, USA, February 2014.

[Adamic 2005] Lada A. Adamic and Natalie Glance. *The political blogosphere and the 2004 U.S. election: divided they blog.* In Proc. of ACM SIGKDD LinkKDD'05, Chicago, IL, USA, August 2005.

[An 2011] Jisun An, Meeyoung Cha, Krishna Phani Gummadi and Jon Crowcroft. *Media Landscape in Twitter: A World of New Conventions and Political Diversity.* In Proc. of the Fifth AAAI International Conference on Weblogs and Social Media, July 2011.

[Bakshy 2011] Eytan Bakshy, Jake M. Hofman, Winter A. Mason and Duncan J. Watts. *Everyone's an influencer: quantifying influence on Twitter.* In Proc. of ACM WSDM'11, Hong Kong, PRC, February 2011.

[Benevenuto 2010] F. Benevenuto, G. Magno, T. Rodrigues and V. Almeida. *Detecting Spammers on Twitter.* Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS), July 2010.

[Boanjak 2012] Matko Boanjak, Eduardo Oliveira, José Martins, Eduarda Mendes Rodrigues and Luís Sarmento. *TwitterEcho: a distributed focused crawler to support open research with twitter data.* In Proc. of WWW '12 Companion, 2012.

[Böhringer 2009] M. Böhringer. *Really Social Syndication: A Conceptual View on Microblogging.* Sprouts: Working Papers on Information Systems, vol. 9, no. 31, 2009.

[Boyd 2007] Danah N. Boyd and Nicole B. Ellison. *Social Network Sites: Definition, History, and Scholarship.* Journal of Computer-Mediated Communication, vol. 13, no. 1, pages 210–230, 2007.

[Boyd 2010] Danah Boyd, Scott Golder and Gilad Lotan. *Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter.* In System Sciences (HICSS), 2010 43rd Hawaii International Conference, volume 0, pages 1–10, Los Alamitos, CA, USA, January 2010. IEEE.

[Broder 2000] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins and Janet Wiener. *Graph structure in the Web.* In Proc. of WWW'00, Amsterdam, The Netherlands, April 2000.

[Cha 2009] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn and Sue Moon. *Analyzing the video popularity characteristics of large-scale user generated content systems.* IEEE/ACM Transactions on Networking (TON, vol. 17, no. 5, pages 1357–1370, 2009.

[Cha 2010] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto and Krishna Phani Gummadi. *Measuring User Influence in Twitter: The Million Follower Fallacy.* In Proc. of AAAI ICWSM'10, Washington, DC, USA, May 2010.

[Cha 2012] M. Cha, F. Benevenuto, H. Haddadi and K. P. Gummadi. *The World of Connections and Information Flow in Twitter.* Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions, vol. 42, no. 4, pages 991–998, 2012.

[Chu 2012] Zi Chu, S. Gianvecchio, Haining Wang and S. Jajodia. *Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?* Dependable and Secure Computing, IEEE Transactions on, vol. 9, no. 6, pages 811–824, December 2012.

[Crane 2008] Riley Crane and Didier Sornette. *Robust Dynamic Classes Revealed by Measuring the Response Function of a Social System.* In PNAS, volume 105, pages 15649–15653, October 2008.

[Degroot 1974] Morris H. Degroot. *Reaching a Consensus.* Journal of the American Statistical Association, vol. 69, no. 345, pages 118–121, March 1974.

[Easley 2010] David Easley and Jon Kleinberg. Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge University Press, New York, NY, USA, 2010.

[Ediger 2010] David Ediger, Karl Jiang, Jason Riedy, David A. Bader and Courtney Corley. *Massive Social Network Analysis: Mining Twitter for Social Good.* In Proc. of IEEE ICPP'2010, September 2010.

[Gabielkov 2012] Maksym Gabielkov and Arnaud Legout. *The Complete Picture of the Twitter Social Graph.* In ACM CoNEXT 2012 Student Workshop, Nice, France, December 2012.

[Gabielkov 2014a] Maksym Gabielkov, Ashwin Rao and Arnaud Legout. *Sampling Online Social Networks: An Experimental Study of Twitter.* ACM SIGCOMM 2014, December 2014. Poster.

[Gabielkov 2014b] Maksym Gabielkov, Ashwin Rao and Arnaud Legout. *Studying Social Networks at Scale: Macroscopic Anatomy of the Twitter Social Graph.* In ACM Sigmetrics 2014, Austin, United States, April 2014.

[Gabielkov 2016] Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau and Arnaud Legout. *Social Clicks: What and Who Gets Read on Twitter?* In ACM SIGMETRICS / IFIP Performance 2016, Antibes Juan-les-Pins, France, June 2016.

[Ghosh 2011] Saptarshi Ghosh, Gautam Korlam and Niloy Ganguly. *Spammers' networks within online social networks: a case-study on Twitter.* In Proc. of WWW '11, 2011.

[Ghosh 2012] Saptarshi Ghosh, Bimal Viswanath, Farshad Kooti, Naveen Kumar Sharma, Gautam Korlam, Fabricio Benevenuto, Niloy Ganguly and Krishna Phani Gummadi. *Understanding and combating link farming in the Twitter social network.* In Proc. of WWW'12, Lyon, France, April 2012.

[Gjoka 2010] M. Gjoka, M. Kurant, C. T. Butts and A. Markopoulou. *Walking in Facebook: A Case Study of Unbiased Sampling of OSNs.* In INFOCOM, 2010 Proceedings IEEE, pages 1–9, March 2010.

[Goel 2010] Sharad Goel, Andrei Broder, Evgeniy Gabrilovich and Bo Pang. *Anatomy of the long tail: ordinary people with extraordinary tastes.* In Proc. of ACM WSDM'10, New York, NY, USA, February 2010.

[Gomez-Rodriguez 2012] Manuel Gomez-Rodriguez, Jure Leskovec and Andreas Krause. *Inferring Networks of Diffusion and Influence.* ACM TKDD, vol. 5, no. 4, February 2012.

[Hegde 2013] Nidhi Hegde, Laurent Massoulié and Laurent Viennot. *Self-Organizing Flows in Social Networks.* In Proc. of SIROCCO'13, pages 116–128, Ischia, Italy, July 2013.

[Hu 2013] Pili Hu and Wing Cheong Lau. *A survey and taxonomy of graph sampling.* arXiv preprint arXiv:1308.5865, 2013.

[Huberman 2008] Bernardo Huberman, Daniel Romero and Fang Wu. *Social networks that matter: Twitter under the microscope.* First Monday, vol. 14, no. 1, 2008.

[Humphreys 2010] Lee Humphreys, Phillipa Gill and Balachander Krishnamurthy. *How much is too much? Privacy issues on Twitter.* In Proc. of the Conference of International Communication Association, Singapore, June 2010.

[Java 2007] Akshay Java, Xiaodan Song, Tim Finin and Belle Tseng. *Why we Twitter: understanding microblogging usage and communities.* In Proc. of WebKDD/SNA-KDD'07, San Jose, California, August 2007.

[Java 2008] Akshay Java. *Mining Social Media Communities and Content.* PhD thesis, University of Maryland, Baltimore County, December 2008.

[Katz 1957] Elihu Katz. *The Two-Step Flow of Communication: An Up-To-Date Report on an Hypothesis.* Public Opinion Quarterly, vol. 21, no. 1, page 61, 1957.

[Kempe 2003] David Kempe, Jon M. Kleinberg and Éva Tardos. *Maximizing the spread of influence through a social network.* Proc. of ACM SIGKDD KDD'03, August 2003.

[Kleinberg 2007] Jon M. Kleinberg. *Cascading behavior in networks: Algorithmic and economic issues.* Algorithmic game theory, 2007.

[Krishnamurthy 2008] Balachander Krishnamurthy, Phillipa Gill and Martin Arlitt. *A few chirps about Twitter.* In Proc. of WOSN'08, Seattle, WA, USA, August 2008.

[Kwak 2010] Haewoon Kwak, Changhyun Lee, Hosung Park and Sue Moon. *What is Twitter, a social network or a news media?* In Proc. of WWW'10, Raleigh, NC, USA, May 2010.

[Kwak 2011a] Haewoon Kwak, Hyunwoo Chun and Sue Moon. *Fragile online relationship: a first look at unfollow dynamics in Twitter.* In Proc. of ACM CHI'11, Vancouver, BC, Canada, 2011.

[Kwak 2011b] Haewoon Kwak, Changhyun Lee, Hosung Park, Hyunwoo Chun and Sue Moon. *Novel aspects coming from the directionality of online relationships: a case study of Twitter*. SIGWEB Newsl., pages 5:1–5:4, July 2011.

[Kyrola 2012] Aapo Kyrola, Guy Blelloch and Carlos Guestrin. *GraphChi: Large-Scale Graph computation on Just a PC*. In Proc. of USENIX OSDI'12, Hollywood, CA, USA, October 2012.

[Lee 2008] Jong Gun Lee and Kavé Salamatian. *Understanding the characteristics of online commenting*. In Proceedings of the 2008 ACM CoNEXT Conference, CoNEXT'08, pages 59:1–59:2, New York, NY, USA, 2008. ACM.

[Lee 2010] Jong Gun Lee, Panayotis Antoniadis and Kavé Salamatian. *Faving Reciprocity in Content Sharing Communities: A Comparative Analysis of Flickr and Twitter*. In Proc. of ASONAM'10, Odense, Denmark, August 2010.

[Lelarge 2009] Marc Lelarge. *Efficient control of epidemics over random networks*. In Proc. of ACM SIGMETRICS'09, Seattle, WA, USA, June 2009.

[Lelarge 2012] Marc Lelarge. *Diffusion and cascading behavior in random networks*. Games and Economic Behavior, vol. 75, no. 2, pages 752–775, July 2012.

[Leon-Garcia 2008] A. Leon-Garcia. Probability, statistics, and random processes for electrical engineering. Pearson/Prentice Hall, 2008.

[Lerman 2012] Kristina Lerman, Rumi Ghosh and Tawan Surachawala. *Social Contagion: An Empirical Study of Information Spread on Digg and Twitter Follower Graphs*. February 2012.

[Liben-Nowell 2008] D. Liben-Nowell and Jon M. Kleinberg. *Tracing information flow on a global scale using Internet chain-letter data*. In PNAS, volume 105, page 4633, 2008.

[Lord 1979] Charles G. Lord, Lee Ross and Mark R. Lepper. *Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence*. Journal of Personality and Social Psychology, vol. 37, no. 11, page 2098, November 1979.

[Lussier 2011] Jake T. Lussier and Nitesh V. Chawla. *Network Effects on Tweeting Discovery Science*. volume 6926 of *Lecture Notes in Computer Science*, chapter 18, pages 209–220. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2011.

[Mao 2011] Huina Mao, Xin Shuai and Apu Kapadia. *Loose tweets: an analysis of privacy leaks on Twitter*. In Proc. of WPES'11, Chicago, IL, USA, October 2011.

[Massoulié 2015] Laurent Massoulié, Mesrob I. Ohannessian and Alexandre Proutiere. *Greedy-Bayes for Targeted News Dissemination*. In Proc. of ACM SIGMETRICS'15, Portland, OR, USA, June 2015.

[May 2014] Avner May, Augustin Chaintreau, Nitish Korula and Silvio Lattanzi. *Filter & Follow: How Social Media Foster Content Curation*. In Proc. of ACM SIGMETRICS'14, Austin, TX, USA, June 2014.

[McMahan 2013] H. B. McMahan, G. Holt, D. Sculley, M. Young and D. Ebner. *Ad Click Prediction: a View from the Trenches*. In Proc. of ACM SIGKDD KDD'13, Chicago, IL, USA, August 2013.

[Meeder 2011] Brendan Meeder, Brian Karrer, Amin Sayedi, R. Ravi, Christian Borgs and Jennifer Chayes. *We know who you followed last summer: inferring social link creation times in twitter*. In Proc. of WWW'11, Hyderabad, India, March 2011.

[Mitchell 2014] A. Mitchell, J. Gottfried, J. Kiley and K. E. Matsa. *Political Polarization & Media Habits*. Technical report, Pew Research Center, October 2014.

[Moore 2009] Robert J. Moore. *Twitter Data Analysis: An Investor's Perspective*. http://bit.ly/KwOsm, October 2009.

[Neglia 2014] Giovanni Neglia, Xiuhui Ye, Maksym Gabielkov and Arnaud Legout. *How to Network in Online Social Networks*. In 6th IEEE INFOCOM International Workshop on Network Science for Communication Networks (NetSciCom 2014), pages 819–824, Toronto, Canada, May 2014.

[Ok 2014] Jungseul Ok, Youngmi Jin, Jinwoo Shin and Yung Yi. *On maximizing diffusion speed in social networks*. In Proc. of ACM SIGMETRICS'14, Austin, TX, USA, June 2014.

[Peddinti 2014] Sai Teja Peddinti, Keith W. Ross and Justin Cappos. *"On the Internet, Nobody Knows You'Re a Dog": A Twitter Case Study of Anonymity in Social Networks*. In Proceedings of the Second ACM Conference on Online Social Networks, COSN '14, pages 83–94, New York, NY, USA, 2014. ACM.

[Pinto 2012] Pedro Pinto, Patrick Thiran and Martin Vetterli. *Locating the Source of Diffusion in Large-Scale Networks*. Physical review letters, vol. 109, no. 6, page 068702, August 2012.

[Ramachandran 2014] Arthi Ramachandran, Yunsung Kim and Augustin Chaintreau. *"I Knew They Clicked when I Saw Them with Their Friends": Identifying Your Silent Web Visitors on Social Media*. In Proceedings of the Second ACM Conference on Online Social Networks, COSN '14, pages 239–246, New York, NY, USA, 2014. ACM.

[Ribeiro 2010] Bruno Ribeiro and Don Towsley. *Estimating and Sampling Graphs with Multidimensional Random Walks*. In Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, IMC '10, pages 390–403, New York, NY, USA, 2010. ACM.

[Russell 2011a] Matthew Russell. 21 recipes for mining Twitter. Real Time Bks. O'Reilly Media, Inc., 2011.

[Russell 2011b] Matthew Russell. Mining the social web: Analyzing data from Facebook, Twitter, LinkedIn, and other social media sites. Head First Series. O'Reilly Media, Inc., 2011.

[Sadikov 2009] Eldar Sadikov and Maria Montserrat Medina Martinez. *Information Propagation on Twitter*. CS322 project report, Stanford University, 2009.

[Saroop 2011] A. Saroop and A. Karnik. *Crawlers for social networks amp; structural analysis of Twitter.* In Proc. of IMSAA'2011, December 2011.

[Sharma 2012] Naveen Kumar Sharma, Saptarshi Ghosh, Fabricio Benevenuto, Niloy Ganguly and Krishna Phani Gummadi. *Inferring who-is-who in the Twitter social network.* In Proc. of ACM WOSN'12, Helsinki, Finland, August 2012.

[Szabo 2010] Gabor Szabo and Bernardo A. Huberman. *Predicting the popularity of online content.* Communications of the ACM, vol. 53, no. 8, August 2010.

[Tarjan 1971] Robert Tarjan. *Depth-first search and linear graph algorithms.* In Proc. of 12th Annual Symposium on Switching and Automata Theory, 1971.

[Thomas 2011] Kurt Thomas, Chris Grier, Dawn Song and Vern Paxson. *Suspended accounts in retrospect: an analysis of Twitter spam.* In Proc. of ACM SIGCOMM IMC'11, Berlin, Germany, November 2011.

[Travers 1969] Jeffrey Travers and Stanley Milgram. *An Experimental Study of the Small World Problem.* Sociometry, vol. 32, pages 425–443, 1969.

[Wang 2010a] Alex Hai Wang. *Don't follow me: Spam detection in Twitter.* In Proc. of SECRYPT'2012, July 2010.

[Wang 2010b] Tianyi Wang, Yang Chen, Zengbin Zhang, Peng Sun, Beixing Deng and Xing Li. *Unbiased Sampling in Directed Social Graph.* In Proceedings of the ACM SIGCOMM 2010 Conference, SIGCOMM '10, pages 401–402, New York, NY, USA, 2010. ACM.

[Wang 2016] L. Wang, A. Ramachandran and A. Chaintreau. *Measuring click and share dynamics on social media: a reproducible and validated approach.* In Proc. of AAAI ICWSM NECO'16, Cologne, Germany, May 2016.

[Webberley 2011] Will Webberley, Stuart Allen and Roger Whitaker. *Retweeting: A study of message-forwarding in Twitter.* In Proc. of the First NSS Workshop on Mobile and Online Social Networks, September 2011.

[Wong 2015] Felix Ming Fai Wong, Zhenming Liu and Mung Chiang. *On the Efficiency of Social Recommender Networks.* In Proc. of IEEE INFOCOM'15, pages 1–13, Hong Kong, PRC, April 2015.

[Wu 2007] Fang Wu and Bernardo A. Huberman. *Novelty and collective attention.* In PNAS, volume 104, pages 17599–17601, November 2007.

[Wu 2011] Shaomei Wu, Jake M. Hofman, Winter A. Mason and Duncan J. Watts. *Who says what to whom on Twitter.* In Proc. of WWW'11, Hyderabad, India, March 2011.

[Xu 2014] Jiaming Xu, Rui Wu, Kai Zhu, Bruce Hajek, R. Srikant and Lei Ying. *Jointly clustering rows and columns of binary matrices.* In Proc. of ACM SIGMETRICS'14, Austin, TX, USA, June 2014.

[Yang 2011] Jaewon Yang and Jure Leskovec. *Patterns of temporal variation in online media.* In Proc. of ACM WSDM'11, Hong Kong, PRC, February 2011.

[Yang 2012] Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin and Guofei Gu. *Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter.* In Proc. of WWW'12, Lyon, France, 2012.

[Ye 2010] Shaozhi Ye and S. Felix Wu. *Measuring message propagation and social influence on Twitter.com.* In Proc. of SocInfo'10, Laxenburg, Austria, October 2010.

[Zadeh 2013] Reza Bosagh Zadeh, Ashish Goel, Kamesh Munagala and Aneesh Sharma. *On the precision of social and information networks.* In Proc. of ACM COSN'13, Boston, MA, USA, October 2013.

[Zhang 2011] Chao Michael Zhang and Vern Paxson. *Detecting and analyzing automated activity on Twitter.* In Proc. of PAM'11, Atlanta, GA, USA, March 2011.

[Zhang 2014] Honggang Zhang, Benyuan Liu, Bin Nie, Zhiyong Xu, Xiayin Weng and Chao Yu. *Leveraging online social friendship to improve data swarming performance.* Computer Networks, vol. 71, pages 130 – 143, 2014.