Social Networks 38 (2014) 16–27

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Assessing the bias in samples of large online networks

Sandra González-Bailón[a,*], Ning Wang[b], Alejandro Rivero[c], Javier Borge-Holthoefer[d], Yamir Moreno[c,e,f]

[a] Annenberg School for Communication, University of Pennsylvania, United States
[b] Oxford Internet Institute, University of Oxford, United Kingdom
[c] Institute for Biocomputation and Physics of Complex Systems, University of Zaragoza, Spain
[d] Qatar Computing Research Institute, Qatar Foundation, Qatar
[e] Department of Theoretical Physics, Faculty of Sciences, University of Zaragoza, Zaragoza 50009, Spain
[f] Complex Networks and Systems Lagrange Lab, Institute for Scientific Interchange, Turin, Italy

ARTICLE INFO

Keywords:
Social media
Twitter
Political communication
Social protests
Measurement error
Graph comparison

ABSTRACT

We consider the sampling bias introduced in the study of online networks when collecting data through publicly available APIs (application programming interfaces). We assess differences between three samples of Twitter activity; the empirical context is given by political protests taking place in May 2012. We track online communication around these protests for the period of one month, and reconstruct the network of mentions and re-tweets according to the search and the streaming APIs, and to different filtering parameters. We find that smaller samples do not offer an accurate picture of peripheral activity; we also find that the bias is greater for the network of mentions, partly because of the higher influence of snowballing in identifying relevant nodes. We discuss the implications of this bias for the study of diffusion dynamics and political communication through social media, and advocate the need for more uniform sampling procedures to study online communication.

1. Introduction

An increasing number of studies use Twitter data to investigate a wide range of phenomena, including information diffusion and credibility, user mobility patterns, spikes of collective attention, and trends in public sentiment (Bakshy et al., 2011; Bollen et al., 2011; Castillo et al., 2011; Cha et al., 2012; Dodds et al., 2011; Lehmann et al., 2012; Paltoglou and Thelwall, 2012; Quercia et al., 2012; Romero et al., 2011; Wu et al., 2011). This boost of attention to Twitter activity responds to the prominence of the platform as a means of public communication, and to its salience in policy discussions on issues like privacy regulation, freedom of speech or law enforcement. However, the rising attention that Twitter has received from researchers is also explained by the relatively easy access to the data facilitated by the platform: unlike other prominent social networking sites (like Facebook), Twitter is public by default and the messages exchanged through the network can be downloaded at scale through the application programming interface (API) that the platform makes available to developers and, by extension, researchers.

The type of access the API offers to the underlying database of Twitter activity has changed over the years, becoming increasingly more restrictive. Currently, there are two main channels to collect messages from Twitter: the search API, which can collect messages published during the previous week but applies a rate limit to the number of queries that can be run[1]; and the streaming API, which allows requests to remain open and pushes data as it becomes available but, depending on volume, still captures just a portion of all activity taking place in Twitter (about 1% of the 'firehose' access, or complete stream of all tweets, which currently requires a commercial partnership with the platform). The questions this paper considers are: How do the data collected through the two APIs compare to each other? Do they allow a similar estimation of the underlying (unobserved) network of communication? If not, what is the nature of the bias and what are the theoretical implications for the interpretation of the data?

* Corresponding author at: Annenberg School for Communication, University of Pennsylvania, 3620 Walnut Street, Philadelphia, PA 19104, United States. Tel.: +1 215-898-4775.
E-mail address: sgonzalezbailon@asc.upenn.edu (S. González-Bailón).

[1] According to the Twitter developers' page, "search will be rate limited at 180 queries per 15 min window for the time being, but we may adjust that over time" (see https://dev.twitter.com/docs/rate-limiting, accessed in January 2014).

http://dx.doi.org/10.1016/j.socnet.2014.01.004
0378-8733/

These questions respond to a motivation that falls in line with previous work on network-based sampling and the reliability of estimates drawn from incomplete network data. Network analysts have long considered the effects of sampling on network statistics when the population under study has no clear frame, either because it is hard to reach or because of the boundary specification problem and the empirical difficulty of defining rules for inclusion (Erickson and Nosanchuk, 1983; Erickson et al., 1981; Frank, 1978; Frank and Snijders, 1994; Granovetter, 1976). Much of this previous work relies on survey-elicited network data and it pays special attention to the effects of snow-balling (Burt and Ronchi, 1994; Butts, 2003; Costenbader and Valente, 2003; Frank, 1977; Illenberger and Flötteröd, 2012; Newman, 2003). More recently, the increasing availability of online observational data has facilitated further work on sampling from large networks (Kossinets, 2006; Leskovec and Faloutsos, 2006; Manos et al., 2013; Morstatter et al., 2013; Wang et al., 2012; Yan and Gregory, 2013). Here the concern is often not data limitation but data abundance, and the question of how to build good representations of large networks that help reduce the processing costs of working with large data. Another concern is how to evaluate the effects of noise in the form of missing or misrepresented data, as when it is difficult to disambiguate records in a database or when the sample is censored by the observation window.

In the context of social media – and Twitter in particular – sampling can introduce two types of measurement error: one affects the coverage and representativeness of the messages returned by the APIs; a second error affects the networks of communication that can be reconstructed from the messages sampled. Social media allow users to engage in direct communication. In Twitter, this takes the form of mentions or replies to other users (which are tagged with the @handle convention), or the broadcasting of messages previously published by someone else (via re-tweets or RTs). Messages that are missed by the sample can dent the reconstruction of communication networks because they might prevent the identification of users, or under-represent the bandwidth (or intensity) of communication between the users identified. Collecting data through the publicly available APIs introduces, in other words, two potential sources of bias: one affecting the list of messages retrieved (first-order bias); and one affecting the networks of communication estimated from those messages (second-order bias). This paper pays special attention to the second form of bias, although it also considers the first as a specific form of boundary specification.

In addition to the API restrictions, the parameters used in the queries also affect sampling accuracy and reliability. This is particularly the case when filters are applied to the collection of messages in order to capture communication on a particular topic or in a given geographical location. Filters exclude users and content by further delimiting the boundaries of data collection. They are an important research design choice because they effectively define the empirical focus of study. To the extent that the API rate limits depend on total volume of communication, filtering information on the basis of content or location might yield better estimations because it reduces the scope of interest and maximises the information retrieved; but it can also exclude users and relevant content if the filtering parameters are misspecified. How sensitive Twitter communication networks are to these parameters remains an empirical question.

We consider this question by comparing the networks that result from two independent samples, collected using the search and the streaming APIs, and applying different parameters in the form of a more or less inclusive list of hashtags (i.e. the labels contributed by users themselves to categorise their messages under specific topics). Our findings suggest that the structure of the sampled networks is significantly affected by both the API and the number of hashtags used to retrieve messages. Using the same list of keywords results in smaller networks when queries are run through the search API (compared to the streaming API), which underestimates centrality scores; the bias, however, is greater when different parameters are used to retrieve messages: a less extensive list of keywords used with the same API results also in smaller networks but centralization is, in this case, overestimated. The biases are especially noticeable for the communication networks formed by mentions, where a higher proportion of users are added in the second wave of data collection, that is, after snowballing from seed messages. Our findings also suggest that on the aggregate level some network features are more robust than others to the biases introduced during data collection.

We think these findings are important for two reasons: first, because they contribute novel evidence on the effects of sampling on network estimation, borrowing some of the insights of previous work to address the challenges created by social media data; and second, because they help address the theoretical implications of the bias, which is likely to affect the answers to questions of increasing interest for social scientists – for instance, how online networks co-evolve with offline political events and behaviour, including mass mobilisations that emerge with the support of social media (Farell, 2012). The aim of this paper is, ultimately, to provide evidence that can help correct measurement errors introduced by research choices in the form of search parameters, or by filters that operate outside the control of researchers (i.e. APIs).

2. Previous research and sampling strategies

Since the launch of Twitter in 2006, an increasing body of research has tried to identify the topological properties of this communication network (Huberman et al., 2009; Java et al., 2007; Kwak et al., 2010), the position and characteristics of influential users (Bakshy et al., 2011; Cha et al., 2010), the dynamics of information exchange (boyd et al., 2010; Cha et al., 2012; Gaffney, 2010; Gonçalves et al., 2011; Honey and Herring, 2009), the existence of polarisation (boyd et al., 2010; Conover et al., 2011), and how information propagates from user to user (Borge-Holthoefer et al., 2011; Harrigan et al., 2012; Jansen et al., 2009; Romero et al., 2011; Wu et al., 2011). A search to the Web of Knowledge database for articles with Twitter as main topic returns, at the time of writing, more than 850 entries, spanning research published in conference proceedings for computer science and engineering, but also in journals of communication, media, sociology, and behavioural sciences. Although all these studies are concerned with how communication takes place through the online network, the diversity of sampling frames and procedures (not to mention the theoretical aims) prevent a direct comparison of their findings. Table 1 summarises the characteristics of the samples used in this previous research, giving a sense of the diversity of approaches that have been employed in the past.

The references in Table 1 are not an exhaustive list of all research done with Twitter data, but they are representative of the different sampling frames that have been applied so far to analyse Twitter communication. There are two main things to highlight from this table: one is the overlapping of observation windows across studies that used different data collection strategies; this results in redundancies in the acquisition and management of data resources, and limits the comparability of findings: although some studies have the same observation window, they do not necessarily apply the same parameters to filter the data analysed. The second message is that the samples analysed were submitted to very different manipulations: in some cases, the focus is on the properties of the underlying following-follower structure, measured as a global network (Kwak et al., 2010) or at the level of dyads (Takhteyev et al., 2012); in other cases, it is on the more direct