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

Table 1
Sample characteristics of previous research on Twitter. List of representative studies using Twitter data. This list reveals the diversity of data collection strategies and sampling frames applied, which undermines comparability of findings.

Reference	Number of messages	Observation window (DD/MM/YY)	Source[a]	Networks?		
				Following	@mentions	RTs
Java et al. (2007)	1.5 million	01/04/07 to 30/05/07	Public timeline	Y	N	N
Huberman et al. (2009)	79 million	not specified	Search API	Y	Y	N
Jansen et al. (2009)	150 thousand	04/04/08 to 03/07/08	Search API	N	N	N
boyd et al. (2010)	720 thousand	26/01/09 to 13/06/09	Public timeline	N	N	Y
	203 thousand	20/04/09 to 13/06/09	Search API	N	N	Y
Cha et al. (2010)	1.8 billion	01/05/2006 to 30/08/2009[b]	Search API (white-listed)	Y	Y	Y
Gaffney (2010)	770 thousand	16/06/2009 to 23/10/2009	Search API (white-listed)	N	N	Y
Kwak et al. (2010)	106 million	06/06/2009 to 31/06/2009	Search API (white-listed)	Y	N	Y
Yardi and boyd (2010)	30 thousand	31/05/2009 to 01/06/2009	Search API (white-listed)	N	N	N
Bakshy et al. (2011)	1.03 billion	13/09/2009 to 15/11/2009	Streaming API (firehose)	Y	N	Y
Borge-Holthoefer et al. (2011)	189 thousand	25/04/2011 to 26/05/2011	Streaming API	N	Y	N
Conover et al. (2011)	250 thousand	14/09/2010 to 01/11/2010	Streaming API (gardenhose)	N	Y	Y
Gonçalves et al. (2011)	382 million	01/05/2006 to 30/05/09[b]	Streaming API (firehose)	Y	Y	N
Romero et al. (2011)	3 billion	01/08/2009 to 01/01/2010	Search API	N	Y	N
Wu et al. (2011)	5 billion	28/07/2009 to 08/03/2010	Streaming API (firehose)	Y	N	Y
Takhteyev et al. (2012)	481 thousand	01/08/2009 to 07/08/2009[b]	Public timeline	Y	N	N
Grabowicz et al. (2012)	12 million	20/11/2008 to 11/12/2008	search API (white-listed)	Y	Y	Y
Morstatter et al., 2013	1.2 million	14/12/2011 to 10/01/2012	Streaming API (firehose)	N	N	Y

[a] 'firehose' gives access to the full stream of messages (it currently requires commercial agreement); the free access stream API (a.k.a. 'spritz') returns about 1% of the full firehose; 'gardenhose' access, now discontinued, returned about 10% of the stream. White-listed accounts could run up to 20k queries per hour using the search API (white-listing is no longer provided).
[b] Actual dates of data collection not specified (only months or years).

channels of communication created by @mentions and RTs, which are employed to reconstruct ties between users (Conover et al., 2011; Grabowicz et al., 2012) or cascades of information diffusion (Bakshy et al., 2011), but also to illuminate communicative practices from a more qualitative perspective (boyd et al., 2010; Honey and Herring, 2009).

In addition, these studies focus on different layers of their data depending on the information domain that is relevant to their research question; for instance, some focus on the subset of messages about trending topics (Cha et al., 2010), others analyse content related to commercial products (Jansen et al., 2009), and others analyse messages labelled with specific hashtags (Borge-Holthoefer et al., 2011; Morstatter et al., 2013; Romero et al., 2011). These studies reveal that the dynamics of communication change across different information domains; however, it is difficult to determine how many of those differences result from actual use as opposed to research design choices, especially those related to data acquisition. Crucially, some of these studies cannot be replicated because they had a level of data access that is not available to most researchers.

This absence of a unified approach to sampling and data management has parallels in other online means of communication, like blogs, for which the effects of applying different sampling procedures have already been discussed (Butts and Cross, 2009). Blog networks also change significantly depending on the sampling strategy and, in particular, on the selection of seed blogs from where snow-balling starts. The difference with social media is that in these platforms the researcher does not have full control on how communication is sampled due to the added constraints imposed by API query limits. The API policies have become more restrictive as the number of users joining the platform increased; as a result, the volume of activity that can be downloaded for analysis is an increasingly smaller percentage of the full stream of activity. In the analysis of blogs, the most important choice is the original set of seeds from where crawling starts: all channels for information exchange are captured as long as the crawl is set up to identify and follow all links. In Twitter, by contrast, sampling usually starts with messages, which are the main entry point to the identification of relevant content and communication patterns between users. Researchers choose how to filter messages according to some substantive criteria (i.e. topic, location); what they do not choose is the filter imposed by the APIs.

Most researchers acknowledge that the APIs return a semi-random sample of messages, but there is no systematic account of the bias. A recent exception is Morstatter et al. (2013), where the authors analyse the full stream of messages containing hashtags, geo-location, and user handles related to the Syrian Revolution for the period of one month; they compare these data to the 1% sample yielded by the publicly available streaming API, using the same filter parameters. The paper shows that accuracy increases when the coverage of the streaming API is greater: because of activity fluctuations (i.e. the volume of messages containing the search keywords changes over time), the 1% sample is more or less comprehensive depending on the time of observation; the reliability of the estimations drawn from this sample goes up with a more comprehensive coverage. Larger samples, in other words, are also better in the context of Twitter data. An implication of this is that aggregating days of data can increase accuracy and improve the estimations of the underlying, unobserved network.

Access to the full stream of activity (or firehose) provides the natural benchmark to assess the bias that derives from using the more restrictive APIs. This access, however, is out of reach for most research organisations, which makes the assessment of the bias even more relevant: it is likely to affect most researchers. In this paper we compare the samples that result from the search and the streaming APIs, both publicly available. The streaming API requires more infrastructure and technical management; the search API needs less resources but yields lower volumes of data (more, however, if data collection needs to go back in time). As Table 1 shows, these two sources of data have been consistently used in previous research, but they have never been directly compared; the analyses that follow aim to fill that gap.

The comparison across APIs is relevant from a methodological point of view (it tests the accuracy and reliability of different data collection strategies) but also for more substantive or theoretical reasons. Recent events have triggered much discussion, within and outside academia, about how and why social media facilitate collective action and political uprisings (Andersen, 2011; Farell, 2012). This interest feeds back into a longer discussion on how online technologies are changing the logic of collective action (Bimber

Table 2
Size of datasets collected. Summary of three samples of Twitter activity collected using different filters (hashtags) and APIs (search and streaming).

	Sample A	Sample B1 (short # list)	Sample B2 (long # list)
Number of Tweets	272,944	434,905	1,026,291
Number of unique authors	71,118	100,368	216,716
Number of unique hashtags	14,324	18,691	46,546

et al., 2005; Earl and Kimport, 2011; Lupia and Sin, 2003). Social networks have traditionally been analysed as the main channels through which participants are recruited and a critical mass can be attained (Diani, 2003; Marwell and Oliver, 1993). The dynamics of information diffusion and political mobilisation – previous research suggests – are very sensitive to the structure of the underlying communication network. With online data, communication networks are often reconstructed from messages. In Twitter, ties between users are created after parsing the messages such that if user *i* mentions or re-tweets user *j*, an arc is created from *i* to *j*. The key issue is: if the sample of messages returned by the two APIs differ, so will the networks reconstructed on the basis of those messages. This bias has theoretical implications because it affects the assessment of how digital technologies shape information diffusion and political mobilisation.

Digital media have their own peculiarities, and the meaning of connections in online networks is not necessarily equivalent to ties measured with surveys (McAdam, 1986; McAdam and Paulsen, 1993). However, online activity can potentially improve measurement in two important respects: first, it captures the actual intensity of communication that is directly relevant to political protests; and second, it offers a richer picture of longitudinal dynamics. In the context of political communication (especially when it is oriented to activate and coordinate protests) a random sample of all Twitter activity would not be very useful; what is illuminating is the subset of all messages related to the specific event, and how the volume of those messages, and the networks of communication between users, evolve over time. Hashtags are keywords that help identify relevant messages in Twitter: they are labels contributed by users that allow them to classify the topic of their messages, and signal to the public that they are part of a community of users with similar interests.

The selection of the keywords that are used in data collection creates the first sampling filter: queries run through the APIs usually request messages that contain certain hashtags (Borge-Holthoefer et al., 2011; Gaffney, 2010; Yardi and boyd, 2010); or, once a data set has been collected, keywords are used to identify relevant streams and filter down larger data sets (Cha et al., 2012; Conover et al., 2011; Romero et al., 2011). Choosing these hashtags is equivalent to specifying the boundaries for data collection: working with the wrong list of keywords might cause relevant data to be missed. This limitation has two sides: it can either be that the list of hashtags used for data collection is incomplete; or that users publish relevant messages without using any hashtags (so these messages are missed as well). While it is difficult to estimate how many messages dispense with hashtags, it makes sense to assume that they will be a minority, especially since using hashtags was a convention developed by Twitter users themselves.

Having a misspecified list of hashtags is, therefore, a potential source of missing information. APIs, and their restrictions to data acquisition, create a second source of bias, which might affect results even when researchers are working with the appropriate list of keywords. The following sections assess the impact that these two sources of bias (i.e. the search parameters chosen by the researcher, and the APIs) have on the estimated networks.

3. Data

We sampled Twitter activity around the same political protests for the period 30 April to 30 May 2012. These protests were organised to celebrate the first anniversary of the Spanish 'indignados' or outraged movement, which erupted in 2011 to protest against spending cuts and the management of the economic crisis. The data were collected as a follow up of previous work analysing the emergence of the movement in May 2011 (González-Bailón et al., 2013, 2011). We collected two independent samples of messages related to the protests: the first sample (A) was collected from the UK using the search API and a list of six hashtags: these include the top five hashtags (in frequency of use) according to the sample collected in 2011 plus a new hashtag created to identify content about the 2012 mobilisations (#12M15M). The second sample (B) was collected from Spain using the streaming API and the more extensive list of 70 hashtags used in the 2011 data collection. For direct comparison with sample A, and to test for the effects of filtering parameters, we also retrieved a shorter version of sample B based on the reduced list of six hashtags. A complete list of all keywords used during data collection, ranked by frequency of use, can be found in Appendix.

The search API used in sample A was easier to implement than the streaming API used in sample B, which required more system resources and more programming work to maintain the long-run connection with the Twitter servers. While the search API is more suitable for precise keyword data collection, and benefits from being able to retrieve messages published up to one week in the past, it is, as explained above, also more restricted in the number of calls it can make to the servers. All else equal, samples collected using the search API will be smaller than samples collected using the streaming API, which is more comprehensive if the researcher has a clear idea of the messages that are of interest, via a selection of hashtags or keywords; however, since rates apply to the full stream, fluctuations in the relative volume of messages of interest will also affect the sample.

Table 2 summarises the datasets that resulted from using these alternative methods. Most of the activity captured in sample A is contained in the larger sample B2, but the overlapping is not complete: 2.5% of the Tweets, 1% of the authors, and 1.3% of the hashtags that appear in the small sample do not appear in the large sample. The higher volume of messages captured in the large sample, on the other hand, accounts for many interactions that go unnoticed in the smaller sample. The differences between the two streaming samples (B1 and B2) are also substantive: more than twice unique authors are identified when a longer list of hashtags is used to retrieve messages. To the extent that the top 5 hashtags appear in most of the messages (see list in Appendix) this difference is surprising: it points at the importance of a large periphery of low-activity users that can be easily missed from the network.

The numbers in Table 2 summarise the first-order bias affecting the messages retrieved, which begs the question of how the missed information impacts on the reconstruction of communication networks. This second-order bias can adopt two forms: false negative nodes (i.e. missed users) and false negative edges (i.e. underrepresented connections); other sources of error are fake user accounts, which lead to false positive nodes, or problems with disambiguation, which can treat different authors as the same and

vice versa (Wang et al., 2012), but these are likely to affect all samples in the same way. Because the smaller samples A and B1 miss a significant number of the users captured by the large sample, they might draw a wrong picture of the centrality of the users observed. By missing messages, they might underestimate the actual bandwidth of communication between users, that is, the intensity of their interactions, often used as a proxy to the strength of connections or the diversity of information exchanged (cf. Aral and Van Alstyne, 2011).

Given the practice of using more than one hashtag in the same message, the list of unique hashtags contained in the samples grows considerably out of the original lists used to collect the data. The overlap for hashtags across samples is high, but the relatively smaller subset of hashtags found in samples A and B1 suggests that there is also less diversity of topics; this matters if the content of the messages is to be used as an approximation to opinions exchanged or to the framing of (in this case) political protests.

On the basis of these messages, we reconstructed protest-relevant communication networks using two Twitter conventions: @mentions and RTs. The first, @mentions, target other users by explicitly mentioning their username; RTs, on the other hand, are employed to broadcast messages previously published by other users (boyd et al., 2010; Honey and Herring, 2009). The networks reconstructed using these two features offer a more accurate approximation to actual communication exchange than the underlying following/follower structure: this network offers the potential for information flow but that potential is realised differently depending on the information domain, as previous research suggests (Cha et al., 2010; Grabowicz et al., 2012; Romero et al., 2011). To reconstruct the @mentions and RTs networks, we parsed the text of the messages captured in the two samples and created a directed link between users i and j every time i mentioned or re-tweeted j.

Because of the way these conventions work, all RTs include a mention (with the @handle of the user being re-tweeted) but not all mentions are RTs. For this reason, and because the literature reviewed in the previous section (and summarised in Table 1) does not always specify how communication networks are reconstructed, we build three different structures: the network of all mentions (including RTs); the network of RTs; and the network of mentions only (i.e. excluding RTs). These networks are directed and weighted, and connections are not necessarily reciprocated. Table 3 summarises the descriptive statistics for the three networks as reconstructed from the three samples. On this level of aggregation, the three networks look very similar: they all contain about 90% of the users in a single giant connected component; the degree distribution is (in line with most online networks) very skewed, with a tail that is especially long for indegree; and the mean path length is relatively short, compared to the size of the networks; they are also characterised by a negative degree correlation, with a tendency of central users to interact with peripheral users.

The vertices in these networks are subsets of the authors captured in the original samples of messages: many users sending protest tweets (about 15–25%, depending on the sample) did not engage in direct interaction with other users (via mentions or RTs), so they are not part of the networks. However, additional users were identified and added to the networks during the snowball that resulted from parsing the messages searching for mentions or RTs. As Fig. 1 illustrates, the percentage of users added to the @mention networks during this second wave is nearly double for the larger sample (B2). This ameliorates the problem of false negative nodes that results from an incomplete sample of messages: reconstructing networks snowballing from seed messages off-sets to a certain extent the first-order bias; however, users added during the second wave are less central in the network than users captured in the first

Table 3
Descriptive network statistics. Summary of the communication networks reconstructed using mentions and re-tweets. On this level of aggregation, networks show similar structures: skewed centrality distribution, global connectivity, and a tendency of central users to interact with peripheral users.

	Sample A (search API, short hashtag list)			Sample B1 (streaming API, short hashtag list)			Sample B2 (streaming API, long hashtag list)		
	@mentions (including RTs)	RTs	@mentions (no RTs)	@mentions (including RTs)	RTs	@mentions (no RTs)	@mentions (including RTs)	RTs	@mentions (no RTs)
N (num of vertices)	68,196	56,196	35,852	95,886	80,912	46,195	222,017	163,873	122,725
M (num of edges)	200,039	121,492	85,870	306,496	199,980	119,235	706,176	450,857	284,441
<k> (mean degree)	6	4	5	6	5	5	6	6	5
max(k_{in}) (max indegree)	5204	3930	1584	7333	5893	2597	13,420	11,460	4634
max(k_{out}) (max outdegree)	619	259	600	798	334	768	2817	450	2802
l (mean path length)	5.18	5.17	5.72	5.15	5.18	5.66	5.32	5.53	5.84
C (clustering)	0.014	0.008	0.017	0.013	0.008	0.016	0.012	0.009	0.010
r (degree correlation)	−0.098	−0.097	−0.110	−0.107	−0.107	−0.102	−0.095	−0.089	−0.154
# strong components	60,857	52,816	32,690	85,375	75,613	41,838	198,223	151,566	113,649
N giant component	62,403 (92%)	51,037 (91%)	31,747 (89%)	89,112 (93%)	75,011 (93%)	41,226 (89%)	198,587 (90%)	149,070 (91%)	103,446 (84%)
N 2nd component	2	2	2	2	2	2	5	2	5

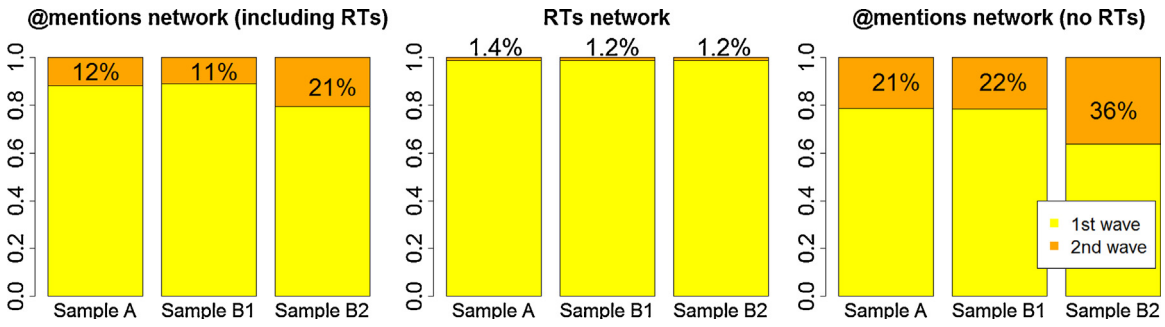

Fig. 1. First and second wave of data collection. Percentage of users added to the networks during the second wave of data collection. Snowballing from mentions leads to the addition of more users to the network, which suggests that mentions are used to target users that are not active in the exchange of messages related to the protests (but probably visible in other domains).

wave, which is an artefact of data collection. Because in the RTs network there are fewer snowballed nodes, the network bias might be less severe.

Fig. 1 also suggests that there are differences in the way users employ the @mention and RT features in the context of political protest: while RTs involve the most active users (who, by sending more messages, are more likely to be captured in the first wave), mentions are used to target users that are more peripheral in this stream of information; the fact that they are targeted nonetheless suggests that these users might be visible in other domains (or that they are more central in the underlying following-follower network), which makes them be perceived as potential diffusers of information and hence common targets. This interpretation is consistent with qualitative insights on how these conventions are generally used in Twitter (boyd et al., 2010; Honey and Herring, 2009); on the basis of observational data, however, it is difficult to separate the confounding motivations behind the use of mentions and RTs.

The upper panel of Fig. 2 shows that most hashtags captured by the three samples are used just once, but there are a few outliers that appear in most messages. The two most often used hashtags, '12M15M' and '15M' (which refer to the dates of the demonstrations for 2012 and 2011) appear in close to three quarters of all messages captured by the samples. Most of the other hashtags co-appeared with these two main keywords; the larger the samples, the more low-frequency hashtags are collected. The scatterplots in the lower panel of Fig. 2 show that the three samples agree most closely in the upper tail of the frequency distribution, and that there is more disagreement for hashtags that are used infrequently: the smaller samples tend to underestimate their use; however, compared to the search API, the streaming API also underestimates the actual frequency counts of some of these keywords. In general, though, the top 2 hashtags (labelled in the scatterplots) would have yielded the vast majority of the messages. This builds a case for restricting queries to a smaller subset of keywords, which can be identified using, for instance, trending topics; this approach will return a similar sample of messages than using a more exhaustive list of hashtags. However, excluding lower-ranked hashtags can lead to different maps of the communication networks created with those messages, especially if author diversity (i.e. the number of nodes in the network) depends on low-frequency hashtags. The following section explores this second-order bias in more detail.

4. Differences in network structure and change

The previous section showed that the three samples allow reconstructing networks that, on the aggregate, share structural features and contain similar information. Larger samples lead to a more inclusive list of hashtags (a proxy to content diversity) but a small fraction of all hashtags appear in the vast majority of messages; these hashtags are the same across samples, differences lying in the periphery of low-frequency content. This section is concerned with the estimation of individual network positions and changes over time: Do users have comparable network positions in the three samples or do these samples give us a different picture of what happens at the local level? Does the similarity of networks increase or decrease over time? In order to address longitudinal changes, we reconstructed the daily structure for the three networks, and calculated the Jaccard coefficient comparing the composition of those networks. This coefficient is a similarity statistic that calculates the ratio of the size of the intersection between sets (in this case, the nodes in the networks) over their union:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

A higher coefficient indicates that networks are more similar. Fig. 3 tracks changes in this coefficient over time, comparing the networks that result from daily aggregations during our observation window. The upper panel compares the networks retrieved from the search and streaming APIs; the second panel compares the networks that result from the streaming API, using the long and short list of hashtags as filtering parameters; and the lower panel displays changes in network size (based on the @mentions networks that include RTs) for reference. The figure tells us two things: first, that using the two different APIs with the same list of hashtags returns more similar networks (coefficients are above 0.5 for most days) than using the same API with different hashtags (coefficients are mostly below 0.5); and second, when the volume of activity goes up (i.e. around protest days), the two streaming samples agree more than the search and streaming samples; in other words, the consistency of the networks inferred from messages is greater when there is a higher volume of messages, for which the streaming API can extract more observations. This finding is consistent with previous work comparing the full stream of messages with those returned by the streaming API (Morstatter et al., 2013): more activity translates into more accurate samples.

These trends suggest that the picture the three samples give of online communication can be quite different; they also indicate that the resolution of the picture changes over time, becoming less accurate when there is a lower volume of information exchange. One conclusion we can draw from this is that the search API, which has higher limitations in data collection, might be a less appropriate tool for periods of low activity or for information streams that are not very abundant. If larger is better, the streaming API will yield more accurate results in high activity days; this is also a good reason to aggregate data for wider time periods: during days of low activity, the bias will affect inferences more drastically.

It is one thing, however, to measure similarity at the level of network composition (i.e. ratio of overlapping nodes); it is another to

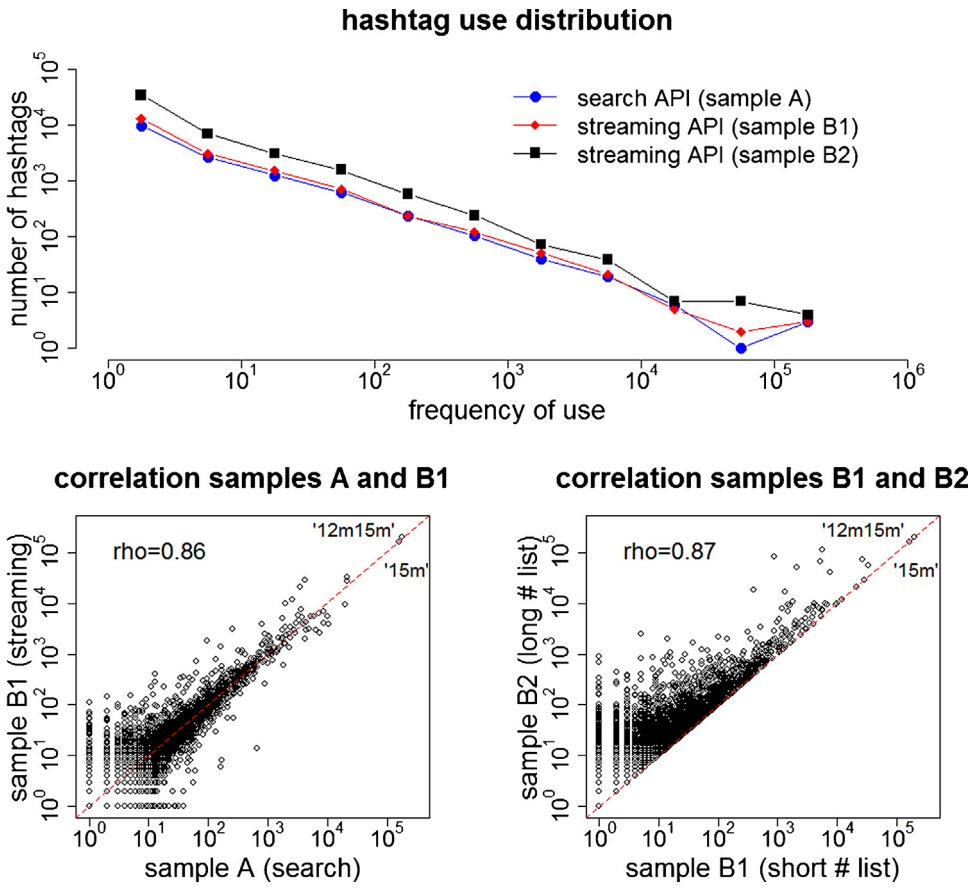

Fig. 2. Distribution of Hashtags by frequency of use. The upper panel summarises the popularity of hashtags: a small number of them are used in the vast majority of messages. The lower scatterplots show that the three samples agree most closely in the upper tail of the frequency distribution, with most disagreement affecting infrequently used hasthags. The top 2 hashtags are labelled.

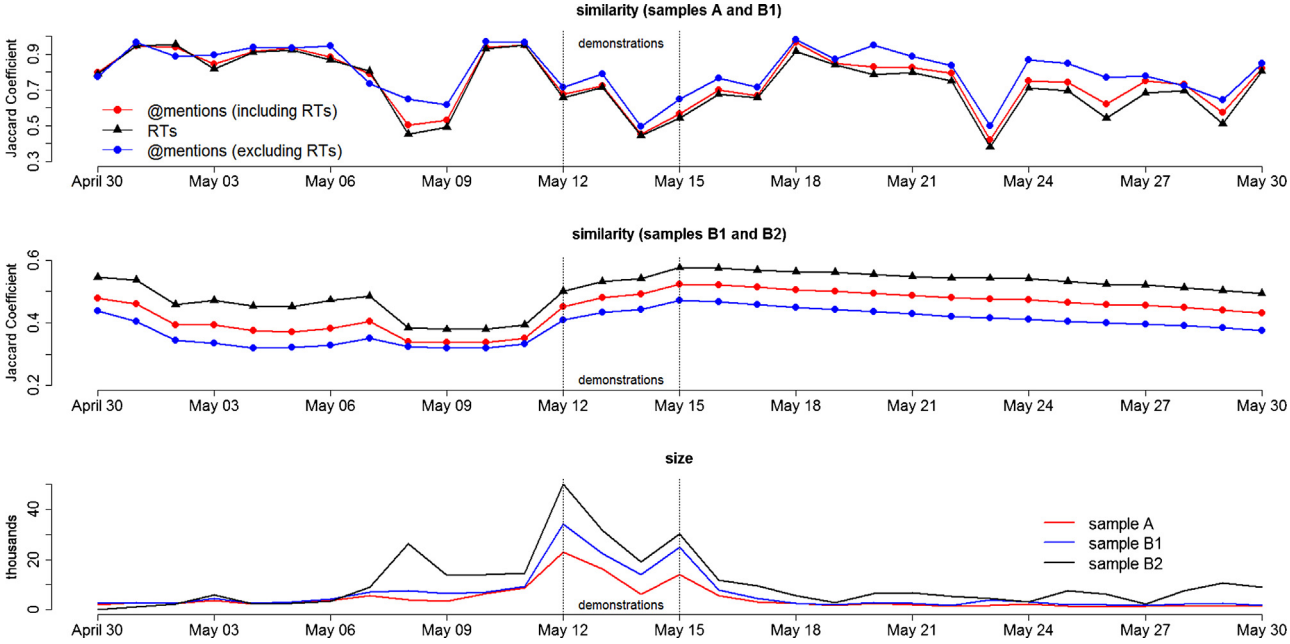

Fig. 3. Similarity of networks over time. Changes in the Jaccard coefficient measuring the similarity of the sampled networks in terms of composition. A higher coefficient indicates that networks are more similar. For reference, the lower panel displays changes in network size. Using two different APIs with the same list of hashtags returns more similar networks (coefficients are overall above 0.5) than using the same API with different hashtags (coefficients are mostly below 0.5). In general, more activity translates into more accurate samples.

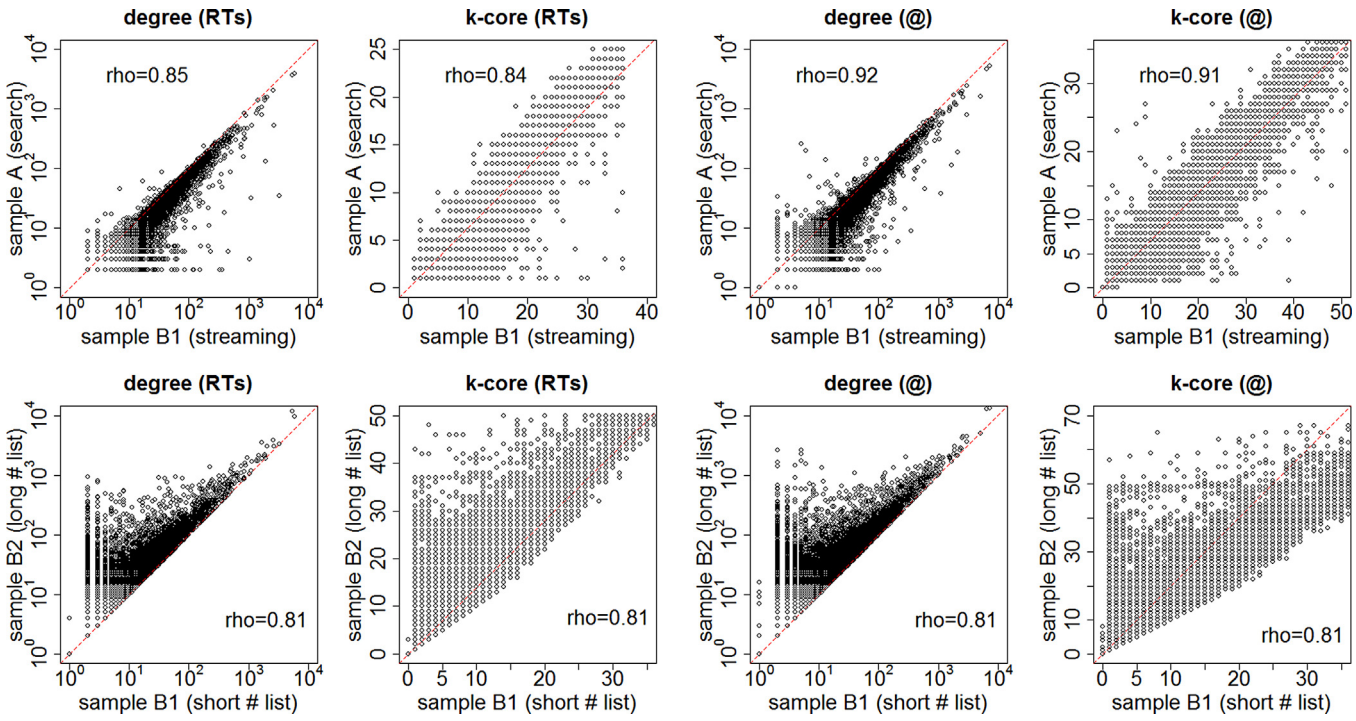

Fig. 4. Correlation of centrality measures. The upper panels summarise the association of individual centrality measures between the search and streaming samples (same list of hashtags); the lower panels summarise the association between the streaming samples (short and long list of hashtags). For users that are less well connected, the smaller samples underestimate centrality more clearly. The extra information captured by the streaming API lies mostly at the fringes of the communication network.

assess similarity in terms of network positions. This is what Fig. 4 assesses. The upper panels illustrate the association of measures between samples A (search) and B1 (streaming, short hashtags list); the lower panels illustrate the association between samples B1 and B2 (both streaming, but with a short and long list of hashtags). The plots consider two measures of user centrality, degree and k-core (Bonacich, 1987; Seidman, 1983); they were calculated using the networks aggregated for the full observation period. The figure shows that there is a clear association of individual centrality scores, especially for the higher ranks of the distributions: in the case of degree, these are users with roughly a hundred or more connections. The k-core shows also a monotonic association, which means that the closer a user is to the core of the network according to the larger samples, the closer the same user is to the core of the network according to the smaller samples. However, for users that are less well connected, the small sample underestimates centrality more clearly. This is because larger samples are better at getting peripheral messages; these messages are absent from the smaller samples and, as a result, undervalue the centrality of many users.

Centrality in this context means that someone is mentioned more often or re-tweeted more times in the flow of protest-related communication. It is a direct measure of visibility in this information domain, but also of outreach: it helps identify who is being more active in promoting protest-related talk. The disagreement between samples on the centrality scores of many users has different implications depending on the research question at hand. If the topic of interest involves identifying the leaders of the mobilisation (in terms of who is seeding the network with relevant information) then both samples would give a similar picture of the users that populate the core of activity. But if the question is how that information spreads, given that most users are peripheral and that this is the subset of nodes where most disagreement takes place, conclusions should be more cautious, especially for the smaller samples. Here the notions of centre and periphery relate to the density of connections, and to the extent to which the network has a class of actors that are loosely connected to a more cohesive subgraph; the existence of a core/periphery structure in networks does not necessarily imply a hierarchical organisation (Borgatti and Everett, 1999). What the findings above suggest is that the search API is worse at tapping those peripheral, loosely connected users – who happen to be the majority in online networks. The extra information captured by the streaming API lies mostly at the fringes of the communication network.

5. Measuring the bias

The analyses above suggest that the smaller samples (A compared to B1 or B1 compared to B2) are not a random subset of the larger samples. Previous work has shown evidence that the streaming API does not offer a random sample of the full stream of communication either, but that larger samples are more accurate (Morstatter et al., 2013). In this section we measure the bias in the smaller samples by comparing the networks that result from them with the networks that would result from a random selection of nodes in the larger samples. The process consists of iterating a random selection process so that in every round a subgraph is extracted from the larger network by randomly selecting N vertices (and the adjacent edges), where N equals the size of the smaller RTs and @mentions (excluding RTs) networks. This is repeated 1000 times and network statistics are then calculated for the thousand random networks, which help build a theoretical probability distribution with which to assess the magnitude of the observed statistics – and the extent to which they depart from what would be expected by random chance. This comparison makes sense to the extent that the overlap between samples is close to complete.

Fig. 5 shows the results of these randomizations, comparing the search and streaming samples (panel A), and the two streaming

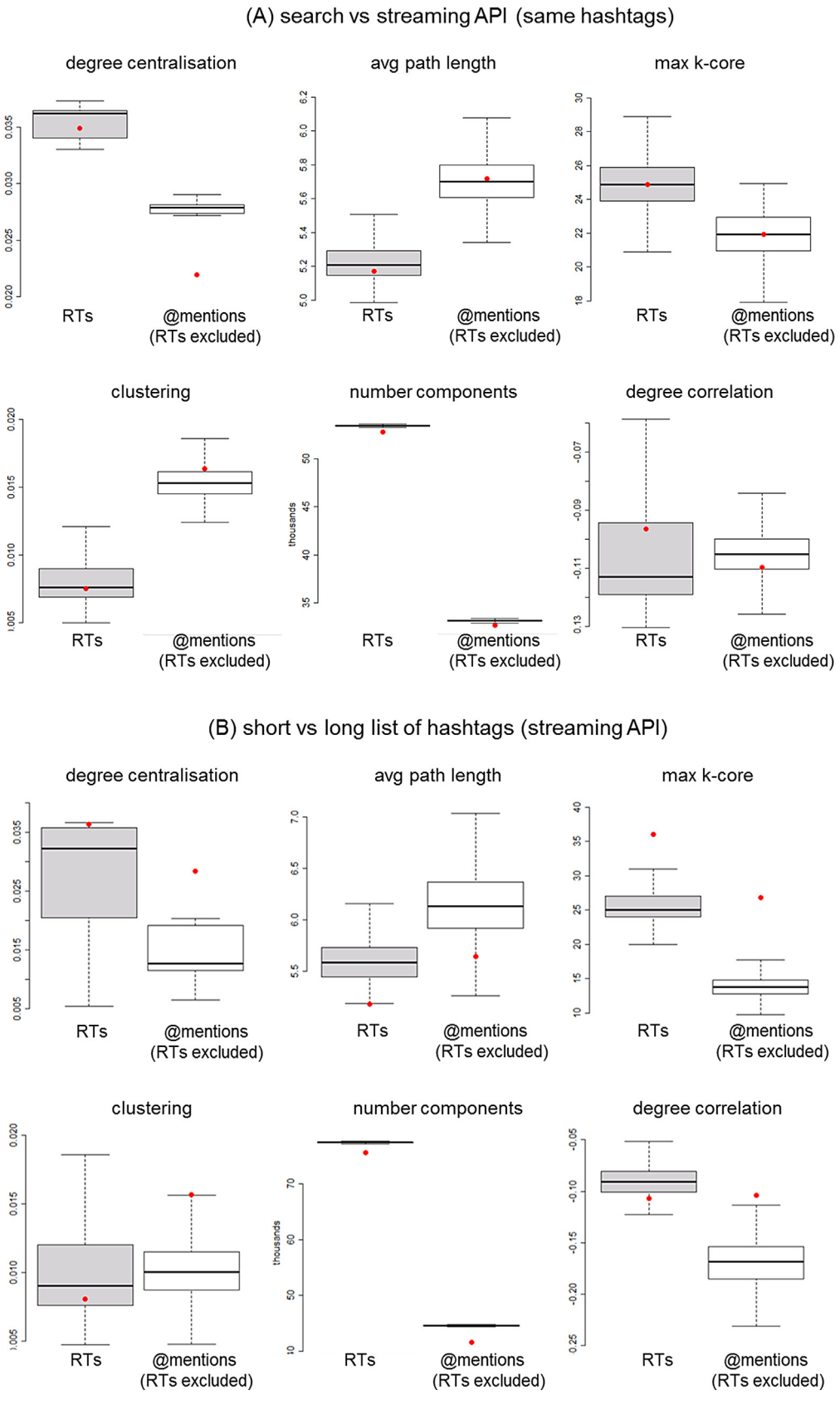

Fig. 5. Bias in Network Statistics Estimated using Small Samples. Red dots correspond to observed statistics (search API sample for panel A; small streaming API sample for panel B). Boxplots summarise the range of network statistics calculated using random draws from the larger samples (1000 subgraphs). The comparison of observed values with random draws suggests that bias is less noticeable for RT networks, especially for degree centralization; also that the influence of filtering parameters is more important than the API used. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

samples based on the short and long list of hashtags (panel B). The red dots correspond to the statistics found in the observed networks; the boxplots give the distribution of network statistics calculated using the random draws from the larger samples (B1 for panel A and B2 for panel B). The results suggest that size is less of a problem for RT networks, which overall offer more accurate estimations of what would be expected under random sampling: the bias is greater for @mention networks, probably due to the higher number of users in these networks collected during the second wave, a percentage that depends on sample size. A second finding is that the influence of filtering parameters can be more important than the actual API used: as panel B shows, a shorter list of hashtags results in networks that overestimate centralization, clustering, and degree correlation; this is due to the fact that more peripheral activity is included in the sample with a more exhaustive list of parameters; using top-ranked hashtags in data collection taps well into the centre of the network, but underestimates the effects of the larger, unobserved periphery.

On the basis of these findings we can conclude two things: first, that networks formed by mentions are more biased because of the way in which mentions, as a convention, are used: users who are mentioned very often are not necessarily proactive in sending messages, whereas users who are re-tweeted tend to be very active themselves. This finding relates to previous work on the effects of snowballing in network sampling (for instance, Granovetter, 1976; Frank, 1977, 1978; Costenbader and Valente, 2003) but it is not directly comparable because of the way in which data is retrieved from online sources: instead of asking people to point to other network members, connections are inferred from messages; to the extent that some users send many more messages than others, they have more chances of being captured in the sample (along with the users they mention). A second conclusion is that this bias can be exacerbated if the parameters used to retrieve data (in this case, hashtags) are misspecified. Although a minority of hashtags concentrate most of the activity, many peripheral nodes in the network can only be captured if less prominent hashtags are included in the list. Although the frequency of use of hashtags (which can be assessed by means of aggregated time series like trending topics) can be used as a rule to identify the boundaries of data collection, it is important to remember that this will artificially crop a periphery of activity – in a similar way as name generators often impose an artificial limit to the size of ego-networks (Klofstad et al., 2009; Wang et al., 2012).

The implications of these biases will vary with the research question, but it is particularly relevant for the study of core-periphery dynamics and, in the context of online collective action, the emergence of a critical mass. Peripheral users do not have many connections, and they do not offer the important voices in terms of impact or reach; but their superiority lies in the numbers: most users qualify as peripheral (this is what the skewed degree distributions tell us) and they are the mass that is activated when political protests are successful. The bias in the data means that it is difficult to assess how large that periphery is and, by extension, how long it takes to activate those users; it also undermines the study of diffusion dynamics because it artificially shrinks the actual size of the population of interest.

6. Discussion

The rising interest in digital media and social interactions mediated by online technologies is boosting the research outputs in an emerging field that is multidisciplinary by nature: it brings together computer scientists, sociologists, physicists, and researchers from a wide range of other disciplines. The collaborative nature of this research is making the number of studies increase fast, often to catch up with the rise of new technologies; but the speed at which the field moves means that standards often do not have time to mature and consolidate. Procedures for data collection and sampling offer a particularly good example of this lack of consolidation: they vary widely across disciplines, and often adapt differently to the peculiarities of the platforms from where the data are gathered. As a result, it is difficult to integrate research outputs – especially when proprietary data are involved in the analyses, which hampers the ability to replicate findings and engage in cumulative research and theory building. A lack of attention to previous work investigating the effects of sampling in networks is another limitation undermining the consolidation of robust methods for the analysis of online activity.

This paper has tackled this lack of common standards by focusing on one particular platform, Twitter, currently one of the most popular SNSs (at least in Western societies), and at the centre of much public and media attention for its alleged role in protests and mass mobilisations. In particular, the paper has compared the networks of communication that can be reconstructed when three different sampling strategies are applied to the same underlying population of messages. What the comparison reveals is that there is a bias in the reconstructed networks that goes unnoticed in most research but that might have important theoretical implications for some of the questions that have been posed to the data in the past.

The virtues of online data are so great (in terms of their scale and resolution) that the problems and deficiencies are often dismissed without a systematic account of how they might interfere with the analyses. The conclusions of this paper are limited to the particular information context analysed (i.e. politically relevant communication in the context of mass mobilisations), and are subject to the limitations of not being able to access unfiltered data. But they provide enough evidence to defend the claim that a more careful account of data quality and bias, and the creation of standards that can facilitate the comparability of findings, would benefit an emerging area of research, especially if it is to yield insights that can survive specific technologies and short-sighted research aims. The Library of Congress announced in 2010 its plans to archive every tweet since 2006; however, more than three years later these plans have yet to fructify. In the meantime, researchers interested in reconstructing the trails of online communication will have to factor in their analysis the bias inherent to their data; in so doing, they can also contribute to the broader goal of understanding the effects of missing information on the structure and dynamics of networks.

Acknowledgements

Thanks to the audience of the Nuffield-OII networks seminar for suggestions, especially Scott Hale, Bernie Hogan, Wybo Wiersma, and Taha Yasseri. We are also grateful to the two anonymous reviewers for their insightful comments. SGB acknowledges funding received from the Fell Fund and Google while still at the Oxford Internet Institute; NW also thanks Google for funding support. This work has been partially supported by MINECO through Grant FIS2011-25167; Comunidad de Aragón (Spain) through a grant to the group FENOL, and by the EC FET-Proactive Project MULTIPLEX (grant 317532).

Appendix.

Table A1.

Table A1
List of 70 hashtags used in API queries ranked by frequency of use.

Rank	Hashtag
1	#acampadasol*
2	#spanishrevolution*
3	#nolesvotes*
4	#15m*
5	#nonosvamos*
6	#democraciarealya
7	#notenemosmiedo
8	#yeswecamp
9	#15mani
10	#acampadasevilla
11	#globalcamp
12	#acampadavalencia
13	#acampadagranada
14	#acampadamalaga
15	#acampadazgz
16	#consensodeminimos
17	#italianrevolution
18	#estonosepara
19	#acampadaalicante
20	#tomalacalle
21	#europeanrevolution
22	#acampadapamplona
23	#worldrevolution
24	#acampadapalma
25	#tomalaplaza
26	#acampadas
27	#15mpasalo
28	#cabemostodas
29	#nonosmovemos
30	#3puntosbasicos
31	#frenchrevolution
32	#estonoseacaba
33	#acampadatoledo
34	#nonosrepresentan
35	#acampadalondres
36	#globalrevolution
37	#acampadazaragoza
38	#acampadaparis
39	#takethesquare
40	#periodismoeticoya
41	#hastalasgenerales
42	#irishrevolution
43	#democraziarealeora
44	#democraciaparticipativa
45	#15mpamplona
46	#barcelonarealya
47	#dry_jaen
48	#usarevolution
49	#dry_caceres
50	#dryasturies
51	#democraziareale
52	#democratiereelle
53	#dry_cadiz
54	#dry_toledo
55	#acampadasvlla
56	#drybizkaia
57	#dry_santander
58	#15mayovalencia
59	#dry_pisa
60	#dryginebra
61	#DRY_Algeciras
62	#demorealyaib
63	#DRYGipuzkoa
64	#DryValladolid
65	#ItalRevolution
66	#BolognaDRY
67	#DRY_Pavia
68	#DRY_Almeria
69	#15mayoCordoba
70	#ciudades-dry

Note: Hashtags marked by * were used to collect sample A (search API) and the shorter version of sample B (streaming API).

References

Andersen, K., 2011. The Protester, Time.
Aral, S., Van Alstyne, M., 2011. The Diversity-Bandwidth Trade-off. American Journal of Sociology 117, 90–171.
Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J., 2011. Everyone's an influencer: quantifying influence on Twitter. In: Proceeding of the Fourth International Conference on Web Search and Data Mining (WSDM 2011).
Bimber, B., Flanagin, A., Sthol, C., 2005. Reconceptualizing collective action in the contemporary media environment. Communication Theory 15, 365–388.
Bollen, J., Mao, H., Zeng, X.-J., 2011. Twitter mood predicts the stock market. Journal of Computational Science 2, 1–8.
Bonacich, P., 1987. Power and centrality: a family of measures. American Journal of Sociology 92, 1170–1182.
Borgatti, S.P., Everett, M.G., 1999. Models of core/periphery structures. Social Networks 21, 375–395.
Borge-Holthoefer, J., Rivero, A., García, I., Cauhé, E., Ferrer, A., Ferrer, D., Francos, D., Iñiguez, D., Pérez, M.P., Ruiz, G., Sanz, F., Serrano, F., Viñas, C., Tarancón, A., Moreno, Y., 2011. Structural and dynamical patterns on online social networks: the Spanish May 15th Movement as a Case Study. PLoS One 6, 6e23883, http://dx.doi.org/10.1371/journal.pone.0023883.
boyd, d., Golder, S.A., Lotan, G., 2010. Tweet, Tweet, Retweet: conversational aspects of retweeting on Twitter. In: HICSS-43 IEEE, Kauai, HI.
Burt, R.S., Ronchi, D., 1994. Measuring a large network quickly. Social Networks 16, 91–135.
Butts, C.T., 2003. Network inference, error, and informant (in)accuracy: a Bayesian approach. Social Networks 25, 103–140.
Butts, C.T., Cross, R.B., 2009. Change and external events in computer-mediated citation networks: English Language Weblogs and the 2004 U.S. Electoral Cycle. Journal of Social Structure, 10.
Castillo, C., Mendoza, M., Poblete, B., 2011. Information credibility on Twitter. In: Proceedings of the 20th International World Wide Web Conference (WWW 2011).
Cha, M., Benevenuto, F., Haddadi, H., Gummadi, K.P., 2012. The world of connections and information flow in Twitter. IEEE Transactions on Systems, Man and Cybernetics 42, 991–998.
Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P., 2010. Measuring user influence in Twitter: the million follower fallacy. In: Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM 2010).
Conover, M.D., Ratkiewicz, J., Francisco, M., Goncalves, B., Flammini, A., Menczer, F., 2011. Political polarization on Twitter. In: International Conference on Weblogs and Social Media (ICWSM'11).
Costenbader, E., Valente, T.W., 2003. The stability of centrality measures when networks are sampled. Social Networks 25, 283–307.
Diani, M., 2003. Networks and social movements: a research programme. In: Diani, M., McAdam, D. (Eds.), Social Movements and Networks. Relational Approaches to Collective Action. Oxford University Press, New York.
Dodds, P.S., Harris, K.D., Kloumann, I.M., Bliss, C.A., Danforth, C.M., 2011. Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. PLoS One 6, e26752.
Earl, J., Kimport, K., 2011. Digitally Enabled Social Change: Activism in the Internet Age. MIT, Cambridge, MA.
Erickson, B.H., Nosanchuk, T.A., 1983. Applied network sampling. Social Networks 5, 367–382.
Erickson, B.H., Nosanchuk, T.A., Lee, E., 1981. Network sampling in practice: some second steps. Social Networks 3, 127–136.
Farell, H., 2012. The consequences of the internet for politics. Annual Review of Political Science 15, 35–52.
Frank, O., 1977. Survey sampling in graphs. Journal of Statistical Planning and Inference 1, 235–264.
Frank, O., 1978. Sampling and estimation in large social networks. Social Networks 1, 91–101.
Frank, O., Snijders, T.A., 1994. Estimating the size of hidden populations using snowball sampling. Journal of Official Statistics 10, 53–67.
Gaffney, D., 2010. #iranElection: quantifying online activism. In: Proceedings of the WebSci10: Extending the Frontiers of Society On-Line.
Gonçalves, B., Perra, N., Vespignani, A., 2011. Modeling users' activity on Twitter networks: validation of Dunbar's number. PLoS One 6, e22656.
González-Bailón, S., Borge-Holthoefer, J., Moreno, Y., 2013. Broadcasters and hidden influentials in online protest diffusion. American Behavioral Scientist 57, 943–965.
González-Bailón, S., Borge-Holthoefer, J., Rivero, A., Moreno, Y., 2011. The dynamics of protest recruitment through an online network. Scientific Reports 1, 197.
Grabowicz, P.A., Ramasco, J.J., Moro, E., Pujol, J.M., Eguiluz, V.M., 2012. Social features of online networks: the strength of intermediary ties in online social media. PLoS One 7, e29358.
Granovetter, M., 1976. Network sampling: some first steps. American Journal of Sociology 81, 1287–1303.
Harrigan, N., Achananuparp, P., Lim, E.-P., 2012. Influentials, novelty, and social contagion: the viral power of average friends, close communities, and old news. Social Networks 34, 470–480.
Honey, C., Herring, S.C., 2009. Beyond microblogging: conversation and collaboration via Twitter, system sciences, 2009. In: 42nd Hawaii International Conference on HICSS '09, pp. 1–10.

Huberman, B.A., Romero, D.M., Wu, F., 2009. Social networks that matter: Twitter under the microscope, First Monday 14, http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2317/2063
Illenberger, J., Flötteröd, G., 2012. Estimating network properties from snowball sampled data. Social Networks 34, 701–711.
Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A., 2009. Twitter power: Tweets as electronic word of mouth. Journal of the Association for Information Science and Technology 60, 2169–2188.
Java, A., Song, X., Finin, T., Tseng, B.,2007. Why we Twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. ACM, San Jose, CA, pp. 56–65.
Klofstad, C.A., McClurg, S.D., Rolfe, M., 2009. Measurement of political discussion networks. A comparison of two name generator procedures. Public Opinion Quarterly 73, 462–483.
Kossinets, G., 2006. Effects of missing data in social networks. Social Networks 28, 247–268.
Kwak, H., Lee, C., Park, H., Moon, S., 2010. What is Twitter, a social network or a news media? In: Proceedings of the 19th International World Wide Web Conference (WWW 2010).
Lehmann, J., Goncalves, B., Ramasco, J.J., Catuto, C., 2012. Dynamical Classes of Collective Attention in Twitter. World Wide Web, Lyon, France.
Leskovec, J., Faloutsos, C.,2006. Sampling from large graphs. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Philadelphia, PA, USA, pp. 631–636.
Lupia, A., Sin, G., 2003. Which public goods are endangered? How evolving communication technologies affect. The logic of collective action. Public Choice 117, 315–331.
Manos, P., Gautam, D., Nick, K., 2013. Sampling online social networks. IEEE Transactions on Knowledge and Data Engineering 25, 662–676.
Marwell, G., Oliver, P., 1993. The Critical Mass in Collective Action. Cambridge University Press, Cambridge.
McAdam, D., 1986. Recruitment to high-risk activism: the case of freedom summer. American Journal of Sociology 92, 64–90.
McAdam, D., Paulsen, R., 1993. Specifying the relationship between social ties and activism. American Journal of Sociology 99, 640–667.
Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M., 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. ICWSM, Boston, MA.
Newman, M.E.J., 2003. Ego-centered networks and the ripple effect. Social Networks 25, 83–95.
Paltoglou, G., Thelwall, M., 2012. Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media. ACM Transactions on Intelligent Systems and Technology (TIST) 3, 66, 61–19.
Quercia, D., Capraz, L., Crowcroft, J., 2012. The SocialWorld of Twitter: Topics, Geography, and Emotions. AAAI, ICWSM, Dublin.
Romero, D.M., Meeder, B., Kleinberg, J., 2011. Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter. International World Wide Web Conference, Hyderabad, India.
Seidman, S.B., 1983. Network structure and minimum degree. Social Networks 5, 269–287.
Takhteyev, Y., Gruzd, A., Wellman, B., 2012. Geography of Twitter Networks. Social Networks 34, 73–81.
Wang, D.J., Shi, X., McFarland, D.A., Leskovec, J., 2012. Measurement error in network data: a re-classification. Social Networks 34, 396–409.
Wu, S., Hofman, J.M., Mason, W.A., Watts, D.J., 2011. Who says what to whom on Twitter. In: Proceedings of the 20th International World Wide Web Conference (WWW 2011).
Yan, B., Gregory, S., 2013. Identifying communities and key vertices by reconstructing networks from samples. PLoS One 8, e61006.
Yardi, S., boyd, d., 2010. Dynamic debates: an analysis of group polarization over time on Twitter. Bulletin of Science, Technology & Society 30, 316–327.