UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Mining one percent of Twitter: collections, baselines, sampling

Gerlitz, C.; Rieder, B.

Publication date
2013
Document Version
Final published version
Published in
M/C Journal

Link to publication

Citation for published version (APA):
Gerlitz, C., & Rieder, B. (2013). Mining one percent of Twitter: collections, baselines, sampling. *M/C Journal*, *16*(2), [620]. http://journal.media-culture.org.au/index.php/mcjournal/article/viewArticle/620

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:25 Sep 2021

an mc publication MEDIA-CULTURE.ORG.AU
Other Publications...
mc journal
A JOURNAL OF MEDIA AND CULTURE

M/C JOURNAL
M/C HOME
CURRENT ISSUE
UPCOMING ISSUES
ARCHIVES
CONTRIBUTORS
ABOUT M/C JOURNAL
LOG IN / REGISTER

SUBSCRIPTIONS
ATOM 1.0
RSS 2.0
RSS 1.0

USER
USERNAME
PASSWORD
REMEMBER ME
Log In

JOURNAL CONTENT
SEARCH
All
Search
BROWSE
BY ISSUE
BY AUTHOR
BY TITLE

INFORMATION
FOR READERS
FOR AUTHORS
FOR LIBRARIANS

FONT SIZE
A A A

JOURNAL HELP
JOURNAL HELP

LANGUAGE

M/C Journal, Vol. 16, No. 2 (2013) - 'mining'

Mining One Percent of Twitter: Collections, Baselines, Sampling

Carolin Gerlitz, Bernhard Rieder Volume 16 | Issue 2 | May 2013 | 'mining'

Reading Tools
- Review policy
- About the author
- How to cite this
- Indexing metadata
- Print version
- Notify colleague*
- Email the author*
- Add comment*
- Finding References

This work is licensed under a Creative Commons Attribution - Noncommercial - No Derivatives 3.0 License.

* Requires registration

Introduction

Social media platforms present numerous challenges to empirical research, making it different from researching cases in offline environments, but also different from studying the "open" Web. Because of the limited access possibilities and the sheer size of platforms like Facebook or Twitter, the question of *delimitation*, i.e. the selection of subsets to analyse, is particularly relevant. Whilst sampling techniques have been thoroughly discussed in the context of social science research (Uprichard; Noy; Bryman; Gilbert; Gorard), sampling procedures in the context of social media analysis are far from being fully understood. Even for Twitter, a platform having received considerable attention from empirical researchers due to its relative openness to data collection, methodology is largely emergent. In particular the question of how smaller collections relate to the entirety of activities of the platform is quite unclear. Recent work comparing case based studies to gain a broader picture (Bruns and Stieglitz) and the development of graph theoretical methods for sampling (Papagelis, Das, and Koudas) are certainly steps in the right direction, but it seems that truly large-scale Twitter studies are limited to computer science departments (e.g. Cha *et al.*; Hong, Convertino, and Chi), where epistemic orientation can differ considerably from work done in the humanities and social sciences.

The objective of the paper is to reflect on the affordances of different techniques for making Twitter collections and to suggest the use of a random sampling technique, made possible by Twitter's Streaming API (Application Programming Interface), for baselining, scoping, and contextualising practices and issues. We discuss this technique by analysing a one percent sample of all tweets posted during a 24-hour period and introduce a number of analytical directions that we consider useful for qualifying some of the core elements of the platform, in particular hashtags. To situate our proposal, we first discuss how platforms propose particular affordances but leave considerable margins for the emergence of a wide variety of practices. This argument is then related to the question of how medium and sampling technique

English
OPEN JOURNAL SYSTEMS
OPEN JOURNAL SYSTEMS

are intrinsically connected.

Indeterminacy of Platforms

A variety of new media research has started to explore the material-technical conditions of platforms (Rogers`; Gillespie; Hayles), drawing attention to the performative capacities of platform protocols to enable and structure specific activities; in the case of Twitter that refers to elements such as tweets, retweets, @replies, favourites, follows, and lists. Such features and conventions have been both a subject and a starting point for researching platforms, for instance by using hashtags to demarcate topical conversations (Bruns and Stieglitz), @replies to trace interactions, or following relations to establish social networks (Paßmann, Boeschoten, and Schäfer). The emergence of platform studies (Gillespie; Montfort and Bogost; Langlois *et al*.) has drawn attention to platforms as interfacing infrastructures that offer blueprints for user activities through technical and interface affordances that are pre-defined yet underdetermined, fostering sociality in the front end whilst mining for data in the back end (Stalder). Doing so, they cater to a variety of actors, including users, developers, advertisers, and third-party services, and allow for a variety of distinct use practices to emerge. The use practices of platform features on Twitter are, however, not solely produced by users themselves, but crystallise in relation to wider ecologies of platforms, users, other media, and third party services (Burgess and Bruns), allowing for sometimes unanticipated vectors of development. This becomes apparent in the case of the retweet function, which was initially introduced by users as verbatim operation, adding “retweet” and later “RT” in front of copied content, before Twitter officially offered a retweet button in 2009 (boyd, Golder, and Lotan). Now, retweeting is deployed for a series of objectives, including information dissemination, promotion of opinions, but also ironic commentary.

Gillespie argues that the capacity to interface and create relevance for a variety of actors and use practices is, in fact, the central characteristic of platforms (Gillespie). Previous research for instance addresses Twitter as medium for public participation in specific societal issues (Burgess and Bruns; boyd, Golder, and Lotan), for personal conversations (Marwick and boyd; boyd, Golder, and Lotan), and as facilitator of platform-specific communities (Paßmann, Boeschoten, and Schäfer). These case-based studies approach and demarcate their objects of study by focussing on particular hashtags or use practices such as favoriting and retweeting.

But using these elements as basis for building a collection of tweets, users, etc. to be analysed has significant epistemic weight: these sampling methods come with specific notions of use scenarios built into them or, as Uprichard suggests, there are certain “a priori philosophical assumptions intrinsic to any sample design and the subsequent validity of the sample criteria themselves” (Uprichard 2). Building collections by gathering tweets containing specific hashtags, for example, assumes that a) the conversation is held together by hashtags and b) the chosen hashtags are indeed the most relevant ones. Such assumptions go

beyond the statistical question of *sampling bias* and concern the fundamental problem of how to go fishing in a pond that is big, opaque, and full of quickly evolving populations of fish. The classic information retrieval concepts of *recall* (How many of the relevant fish did I get?) and *precision* (How many fish caught are relevant?) fully apply in this context. In a next step, we turn more directly to the question of sampling Twitter, outlining which methods allow for accessing which practices – or not – and what the role of medium-specific features is.

Sampling Twitter

Sampling, the selection of subsets from a larger set of elements (the population), has received wide attention especially in the context of empirical sociology (Uprichard; Noy; Bryman; Gilbert; Gorard; Krishnaiah and Rao). Whilst there is considerable overlap in sampling practices between quantitative sociology and social media research, some key differences have to be outlined: first, social media data, such as tweets, generally pre-exist their collection rather than having to be produced through surveys; secondly, they come in formats specific to platforms, with analytical features, such as counts, already built into them (Marres and Weltevrede); and third, social media assemble very large populations, yet selections are rarely related to full datasets or grounded in baseline data as most approaches follow a case study design (Rieder).

There is a long history to sampling in the social sciences (Krishnaiah and Rao), dating back to at least the 19th century. Put briefly, modern sampling approaches can be distinguished into probability techniques, emphasising the representative relation between the entire population and the selected sample, and non-probability techniques, where inference on the full population is problematic (Gilbert). In the first group, samples can either be based on a fully random selection of cases or be stratified or cluster-based, where units are randomly selected from a proportional grid of known subgroups of a population. Non-probability samples, on the contrary, can be representative of the larger population, but rarely are. Techniques include accidental or convenience sampling (Gorard), based on ease of access to certain cases. Purposive non-probability sampling however, draws on expert sample demarcation, on quota, case-based or snowball sampling techniques – determining the sample via *a priori* knowledge of the population rather than strict representational relations. Whilst the relation between sample and population, as well as access to such populations (Gorard) is central to all social research, social media platforms bring to the reflection of how samples can function as "knowable objects of knowledge" (Uprichard 2) the role of medium-specific features, such as built-in markers or particular forms of data access.

Ideally, when researching Twitter, we would have access to a *full sample*, the subject and phantasy of many *big data* debates (boyd and Crawford; Savage and Burrows), which in practice is often limited to platform owners. Also, growing amounts of daily tweets, currently figuring around 450 million (Farber), require specific logistic efforts, as a project by Cha *et al*. indicates: to access the

tweets of 55 million user accounts, 58 servers to collect a total amount of 1.7 billion tweets (Cha *et al*.). Full samples are particularly interesting in the case of *exploratory data analysis* (Tukey) where research questions are not set before sampling occurs, but emerge in engagement with the data.

The majority of sampling approaches on Twitter, however, follow a non-probabilistic, non-representative route, delineating their samples based on features specific to the platform.

The most common Twitter sampling technique is *topic-based sampling* that selects tweets via hashtags or search queries, collected through API calls (Bruns and Stieglitz, Burgees and Bruns; Huang, Thornton, and Efthimiadis) Such sampling techniques rest on the idea that content will group around the shared use of hashtags or topical words. Here, hashtags are studied with an interest in the emergence and evolution of topical concerns (Burgees and Bruns), to explore brand communication (Stieglitz and Krüger), during public unrest and events (Vis), but also to account for the multiplicity of hashtag use practices (Bruns and Stieglitz). The approach lends itself to address issue emergence and composition, but also draws attention to medium-specific use practices of hashtags.

Snowball sampling, an extension of topic-based sampling, builds on predefined lists of user accounts as starting points (Rieder), often defined by experts, manual collections or existing lists, which are then extended through “snowballing” or triangulation, often via medium-specific relations such as following. Snowball sampling is used to explore national spheres (Rieder), topic- or activity-based user groups (Paßmann, Boeschoten, and Schäfer), cultural specificity (Garcia-Gavilanes, Quercia, and Jaimes) or dissemination of content (Krishnamurthy, Gill, and Arlitt). Recent attempts to combine random sampling and graph techniques (Papagelis, Das, and Koudas) to throw wider nets while containing technical requirements are promising, but conceptually daunting.

Marker-based sampling uses medium-specific metadata to create collections based on shared language, location, Twitter client, nationality or other elements provided in user profiles (Rieder). This sampling method can be deployed to study the language or location specific use of Twitter. However, an increasing amount of studies develop their own techniques to detect languages (Hong, Convertino, and Chi).

Non-probability selection techniques, topic-, marker-, and basic graph-based sampling struggle with representativeness (Are my results generalisable?), exhaustiveness (Did I capture all the relevant units?), cleanness (How many irrelevant units did I capture?), and scoping (How “big” is my set compared to others?), which does – of course – not invalidate results. It does, however, raise questions about the generality of derived claims, as case-based approaches only allow for sense-making from inside the sample and not in relation to the entire population of tweets. Each of these techniques also implies commitments to *a priori* conceptualisations of Twitter practices: snowball sampling presupposes coherent network topologies, marker-based sampling has to place a lot of faith in Twitter’s capacity to identify language

or location, and topic-based samples consider words or hashtags to be *sufficient* identifiers for issues. Further, specific sampling techniques allow for studying issue *or* medium dynamics, and provide insights to the negotiation of topical concerns versus the specific use practices and medium operations on the platform.

Following our interest in relations between sample, population and medium-specificity, we therefore turn to *random sampling*, and ask whether it allows to engage Twitter without commitments – or maybe *different* commitments? – to particular *a priori* conceptualisations of practices. Rather than framing the relation between this and other sampling techniques in oppositional terms, we explore in what way it might serve as baseline foil, investigating the possibilities for relating non-probability samples to the entire population, thereby embedding them in a "big picture" view that provides context and a potential for inductive reasoning and exploration. As we ground our arguments in the analysis of a concrete random sample, our approach can be considered *experimental*.

Random Sampling with the Streaming API

While much of the developer API features Twitter provides are "standard fare", enabling third party applications to offer different interfaces to the platform, the so-called Streaming API is unconventional in at least two ways. First, instead of using the common query-response logic that characterises most REST-type implementations, the Streaming API requires a persistent connection with Twitter's server, where tweets are then pushed in near real-time to the connecting client. Second, in addition to being able to "listen" to specific keywords or usernames, the logic of the *stream* allows Twitter to offer a form of data access that is circumscribed in quantitative terms rather than focussed on particular entities. The so called *statuses/firehose* endpoint provides the full stream of tweets to selected clients; the *statuses/sample* endpoint, however, "returns a small random sample of all public statuses" with a size of one percent of the full stream. (In a forum post, Twitter's senior partner engineer, Taylor Singletary, states: "The sample stream is a random sample of 1% of the tweets being issues [*sic*] publicly.") If we estimate a daily tweet volume of 450 million tweets (Farber), this would mean that, in terms of standard sampling theory, the 1% endpoint would provide a representative and high resolution sample with a maximum margin of error of 0.06 at a confidence level of 99%, making the study of even relatively small subpopulations within that sample a realistic option.

While we share the general prudence of boyd and Crawford when it comes to the validity of this sample stream, a technical analysis of the Streaming API indicates that some of their caveats are unfounded: because tweets appear in near real-time in the queue (our tests show that tweets are delivered via the API approx. 2 seconds after they are sent), it is clear that the system does not pull only "the first few thousand tweets per hour" (boyd and Crawford 669); because the sample is most likely a simple filter on the *statuses/firehose* endpoint, it would be technically impractical to include only "tweets from a particular segment of

the network graph" (ibid.). Yet, without access to the complete stream, it is difficult to fully assess the selection bias of the different APIs (González-Bailón, Wang, and Rivero). A series of tests in which we compared the sample to the full output of high volume bot accounts can serve as an indicator: in particular, we looked into the activity of *SportsAB*, *Favstar_Bot*, and *TwBirthday*, the three most active accounts in our sample (respectively 38, 28, and 27 tweets captured). Although Twitter communicates a limit of 1000 tweets per day and account, we found that these bots consistently post over 2500 messages in a 24 hour period. *SportsAB* attempts to post 757 tweets every three hours, but runs into *some* limit every now and then. For every successful peak, we captured between five and eight messages, which indicates a pattern consistent with a random selection procedure. While more testing is needed, various elements indicate that the *statuses/sample* endpoint provides data that are indeed representative of all public tweets.

Using the soon to be open-sourced *Digital Methods Initiative Twitter Capture and Analysis Toolset* (DMI-TCAT) we set out to test the method and the insights that could be derived from it by capturing 24 hours of Twitter activity, starting on 23 Jan. 2013 at 7 p.m. (GMT). We captured 4,376,230 tweets, sent from 3,370,796 accounts, at an average rate of 50.65 tweets per second, leading to about 1.3GB of uncompressed and unindexed MySQL tables. While a truly robust approach would require a longer period of data capture, our main goal – to investigate how the Streaming API can function as a "big picture" view of Twitter and as baseline for other sampling methods – led us to limit ourselves to a manageable corpus. We do not propose our 24-hour dataset to function as a baseline in itself, but to open up reflections about representative metrics and the possibilities of baseline sampling in general. By making our scripts public, we hope to facilitate the creation of (background) samples for other research projects. (DMI-TCAT is developed by Erik Borra and Bernhard Rieder. The stream capture scripts are already available at https://github.com/bernorieder/twitterstreamcapture.)

A Day of Twitter

Exploring how the Twitter one percent sample can provide us with a contrast foil against other collection techniques, we suggest that it might allow to create relations between entire populations, samples and medium-specific features in different ways; as illustration, we explore four of them.

a) Tweet Practices Baseline:

Figure 1 shows the temporal baseline, giving indications for the pace and intensity of activity during the day. The temporal pattern features a substantial dip in activity, which corresponds with the fact that around 60% of all tweets have English language settings, which might indicate sleeping time for English-speaking users.

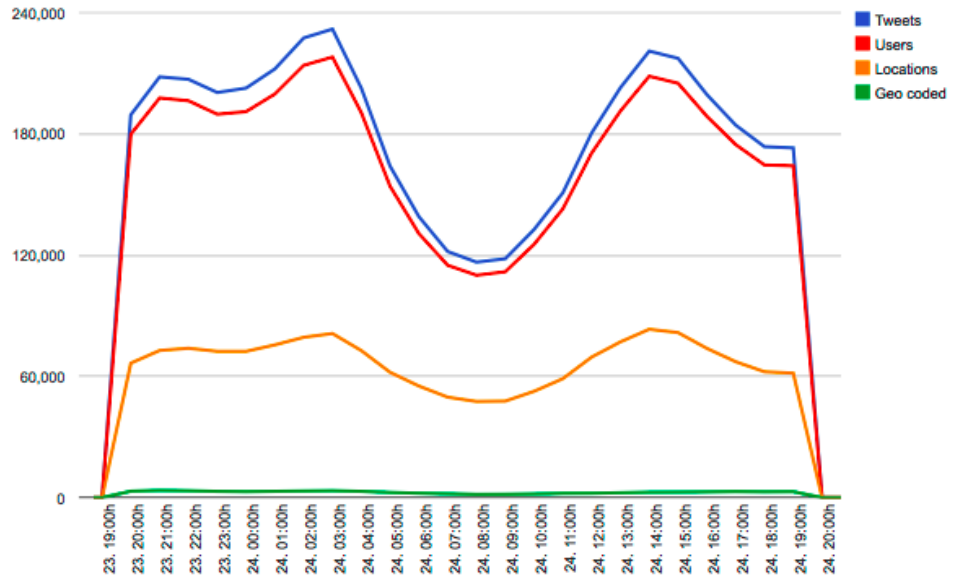

Figure 1: temporal patterns

Exploring the composition of users, the sample shows how "communicative" Twitter is; the 3,370,796 unique users we captured mentioned (all "@username" variants) 2,034,688 user accounts. Compared to the random sample of tweets retrieved by boyd *et al*. in 2009, our sample shows differences in use practices (boyd, Golder, and Lotan): while the number of tweets with hashtags is significantly higher (yet small in relation to all tweets), the frequency of URL use is lower. While these averages gloss over significant variations in use patterns between subgroups and languages (Poblete *et al*.), they do provide a baseline to relate to when working with a case-based collection.

Tweets containing	boyd *et al*. 2010	our findings
a hashtag	5%	13.18%
a URL	22%	11.7%
an @user mention	36%	57.2%
tweets beginning with @user	86%	46.8%

Table 1: Comparison between boyd *et al*. and our findings

b) Hashtag Qualification:

Hashtags have been a focus of Twitter research, but reports on their use vary. In our sample, 576,628 tweets (13.18%) contained 844,602 occurrences of 227,029 unique hashtags. Following the typical power law distribution, only 25.8% appeared more than once and only 0.7% (1,684) more than 50 times. These numbers

are interesting for characterising Twitter as a platform, but can also be useful for situating individual cases against a quantitative baseline. In their hashtag metrics, Bruns and Stieglitz suggest a categorisation derived from *a priori* discussions of specific use cases and case comparison in literature (Bruns and Stieglitz). The random sample, however, allows for alternative, *a posteriori* qualifying metrics, based on emergent topic clusters, co-appearance and proximity measures.

Beyond purely statistical approaches, co-word analysis (Callon *et al*.) opens up a series of perspectives for characterising hashtags in terms of how they appear together with others. Based on the basic principle that hashtags mentioned in the same tweet can be considered *connected*, networks of hashtags can be established via graph analysis and visualisation techniques – in our case with the help of *Gephi*.

Our sample shows a high level of connectivity between hashtags: 33.8% of all unique hashtags are connected in a giant component with an average degree (number of connections) of 6.9, a diameter (longest distance between nodes) of 15, and an average path length between nodes of 12.7. When considering the 10,197 hashtags that are connected to at least 10 others, the network becomes much denser, though: the diameter shrinks to 9 and the average path length of 3.2 indicates a “small world” of closely related topic spaces.

Looking at how hashtags relate to this connected component, we detect that out of the 1,684 hashtags with a frequency higher than 50, 96.6% are part of it, while the remaining 3.4% are spam hashtags that are deployed by a single account only. In what follows, we focus on the 1,627 hashtags that are part of the giant component.

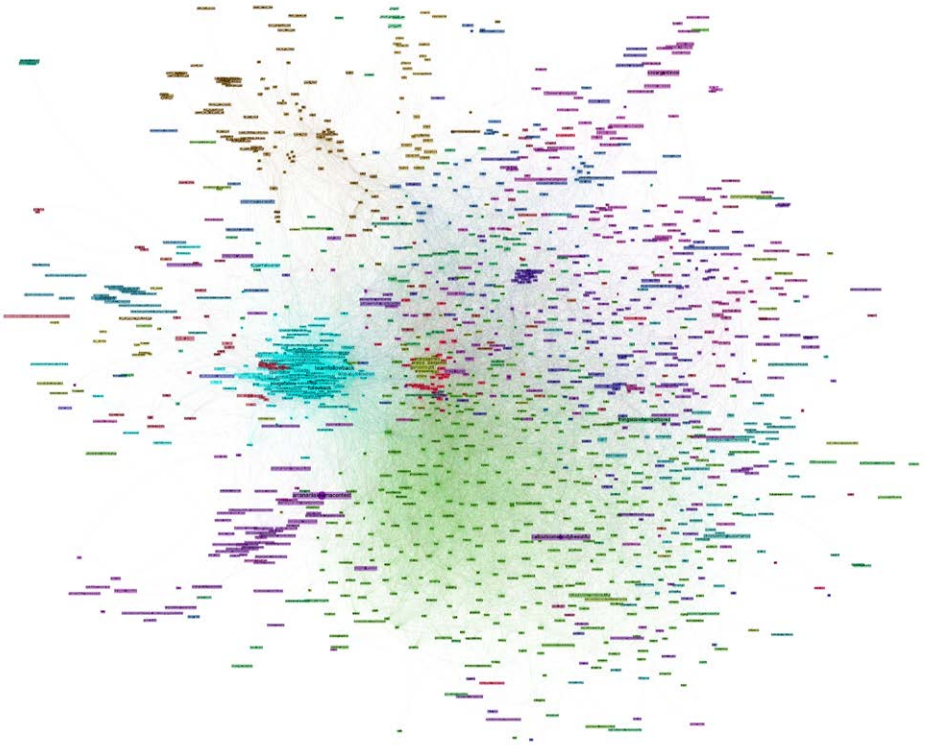

Figure 2: Co-occurrence map of hashtags
(spatialisation: Force Atlas 2; size: frequency of occurrence; colour: communities detected by modularity)

As shown in Figure 2, the resulting network allows us to identify topic clusters with the help of "community" detection techniques such as the *Gephi modularity* algorithm. While there are clearly identifiable topic clusters, such as a dense, high frequency cluster dedicated to following in turquoise (#teamfollowback, #rt, #followback and #sougofollow), a cluster concerning Arab countries in brown or a pornography cluster in bright red, there is a large, diffuse zone in green that one could perhaps most fittingly describe as "everyday life" on Twitter, where food, birthdays, funny images, rants, and passion can coexist. This *zone* – the term cluster suggesting too much coherence – is pierced by celebrity excitement (#arianarikkumacontest) or moments of social banter (#thingsidowhenigetbored, #calloutsomeonebeautiful) leading to high tweet volumes.

Figures 3 and 4 attempt to show how one can use network metrics to qualify – or even classify – hashtags based on how they connect to others. A simple metric such as a node's *degree*, i.e. its number of connections, allows us to distinguish between "combination" hashtags that are not topic-bound (#love, #me, #lol, #instagram, the various "follow" hashtags) and more specific topic markers (#arianarikkumacontest, #thingsidowhenigetbored, #calloutsomeonebeautiful, #sosargentinosi).

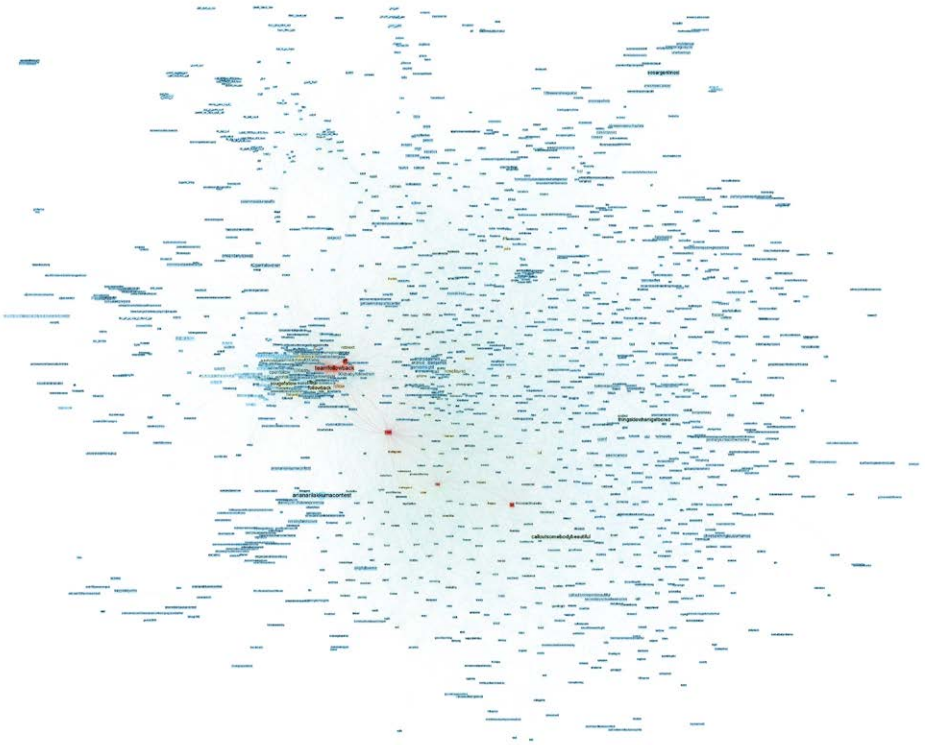

Figure 3: Co-occurrence map of hashtags
(spatialisation: Force Atlas 2; size: frequency of occurrence; colour (from blue to yellow to red): degree)

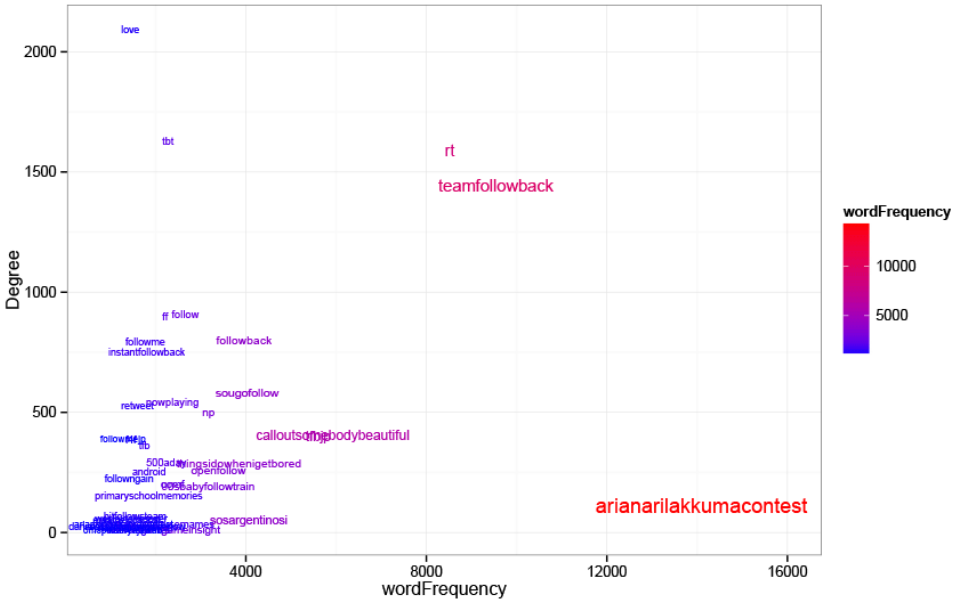

Figure 4: Hashtag co-occurrence in relation to frequency

Another metric, which we call “user diversity”, can be derived by dividing the number of unique users of a hashtag by the number of tweets it appears in, normalised to a percentage value. A score

of 100 means that no user has used the hashtag twice, while a score of 1 indicates that the hashtag in question has been used by a single account. As Figures 5 and 6 show, this allows us to distinguish hashtags that have a "shoutout" character (#thingsidowhenigetbored, #calloutsomeonebeautiful, #love) from terms that become more "insisting", moving closer to becoming spam.

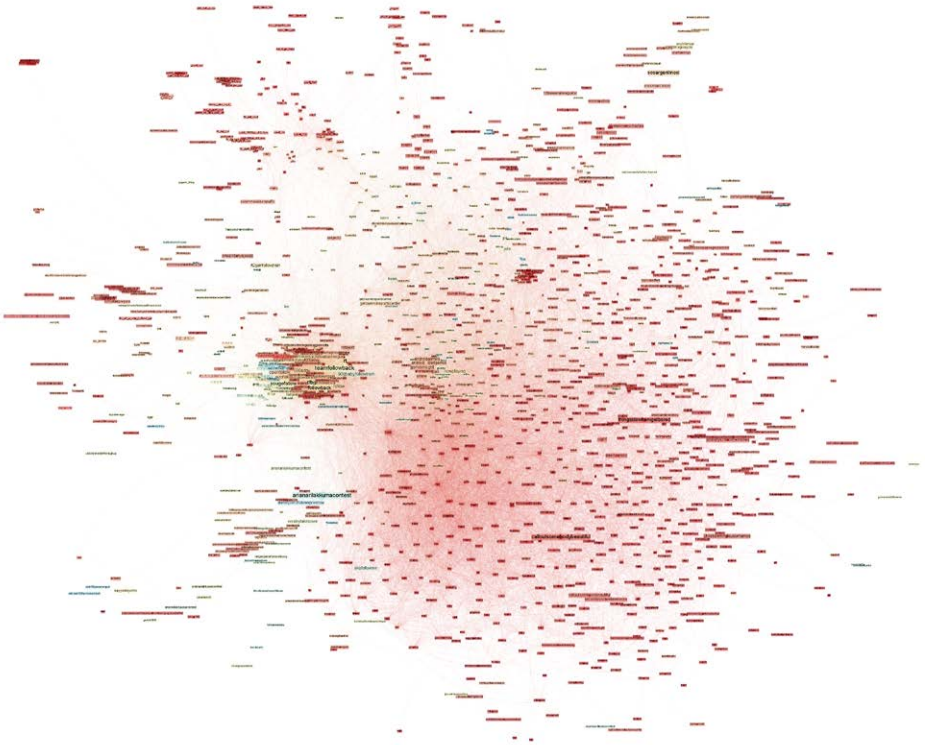

Figure 5: Co-occurrence map of hashtags
(spatialisation: Force Atlas 2; size: frequency of occurrence; colour (from blue to yellow to red): user diversity)

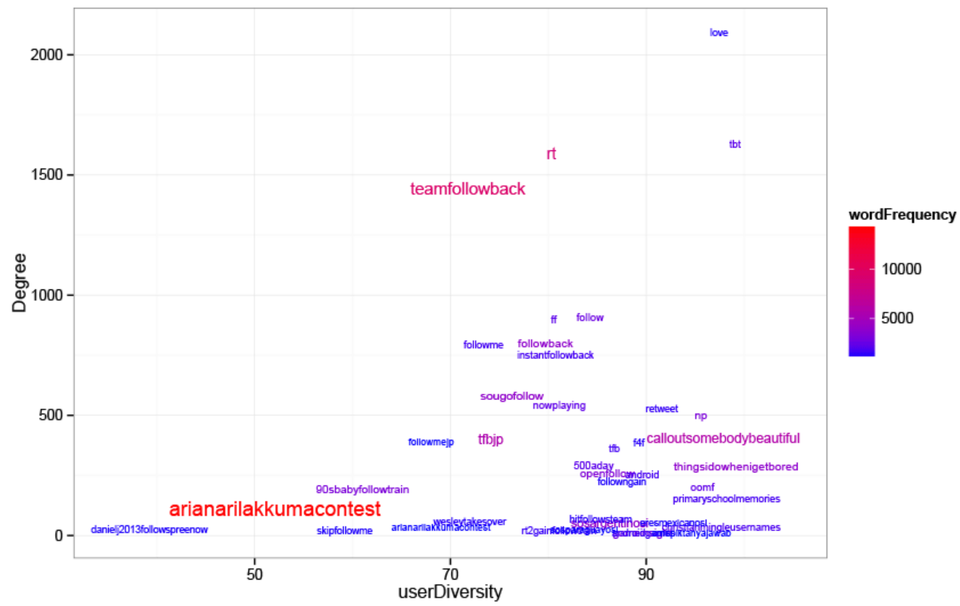

Figure 6: Hashtag user diversity in relation to frequency

All of these techniques, beyond leading to findings in themselves, can be considered as a useful backdrop for other sampling methods. Keyword- or hashtag-based sampling is often marred by the question of whether the “right” queries have been chosen; here, co-hashtag analysis can easily find further related terms – the same analysis is possible for keywords also, albeit with a much higher cost in computational resources.

c) Linked Sources:

Only 11% of all tweets contained URLs, and our findings show a power-law distribution of linked sources. The highly shared domains indicate that Twitter is indeed a predominantly “social” space, with a high presence of major social media, photo-sharing (Instagram and Twitpic) and Q&A platforms (ask.fm). News sources, indicated in red in figure 7, come with little presence – although we acknowledge that this might be subject to daily variation.

Figure 7: Most mentioned URLs by domain, news organisations in red

d) Access Points:

Previously, the increase of daily tweets has been linked to the growing importance of mobile devices (Farber), and relatedly, the

sample shows a proliferation of access points. They follow a long-tail distribution: while there are 18,248 unique sources (including tweet buttons), 85.7% of all tweets are sent by the 15 dominant applications. Figure 8 shows that the Web is still the most common access point, closely followed by the iPhone. About 51.7% of all tweets were sent from four mobile platforms (iPhone, Android, Blackberry, and Twitter's mobile Web page), confirming the importance of mobile devices. This finding also highlights the variety and complexity of the contexts that Twitter practices are embedded in.

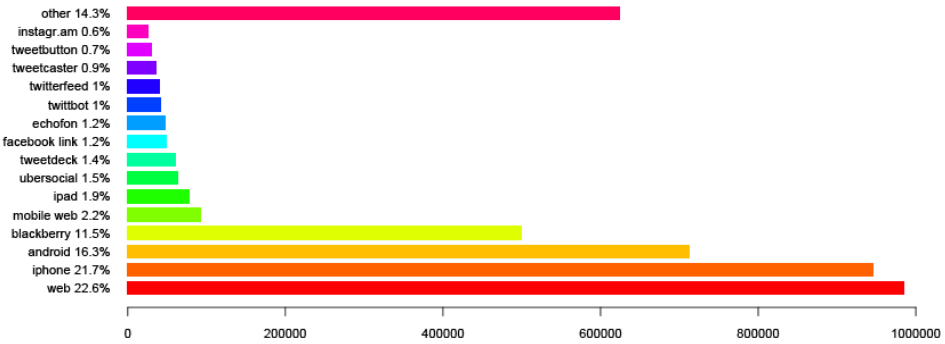

Figure 8: Twitter access points

Conclusion

Engaging with the one percent Twitter sample allows us to draw three conclusions for social media mining. First, thinking of sampling as the making of "knowable objects of knowledge" (Uprichard 2), it entails bringing data points into different relations with each other. Just as Mackenzie contends in relation to databases that it is not the individual data points that matter but the relations that can be created between them (Mackenzie), sampling involves such bringing into relation of medium-specific objects and activities. Small data collection techniques based on queries, hashtags, users or markers, however, do not relate to the whole population, but are defined by internal and comparative relations, whilst random samples are based on the relation between the sample and the full dataset.

Second, thinking sampling as assembly, as relation-making between parts, wholes and the medium thus allows research to adjust its focus on either issue or medium dynamics. Small sample research, we suggested, comes with an investment into specific use scenarios and the subsequent validity of how the collection criteria themselves are grounded in medium specificity. The properties of a "relevant" collection strategy can be found in the extent to which use practices align with and can be utilised to create the collection. Conversely, a mismatch between medium-specific use practices and sample purposes may result in skewed findings. We thus suggest that sampling should not only attend to the internal relations between data points within collections, but also to the relation between the collection and a baseline.

Third, in the absence of access to a full sample, we propose that the random sample provided through the Streaming API can serve

as baseline for case approaches in principle. The experimental study discussed in our paper enabled the establishment of a starting point for future long-term data collection from which such baselines can be developed. It would allow to ground *a priori* assumptions intrinsic to small data collection design in medium-specificity and user practices, determining the relative importance of hashtags, URLs, @user mentions. Although requiring more detailed specification, such accounts of internal composition, co-occurrence or proximity of hashtags and keywords may provide foundations to situate case-samples, to adjust and specify queries or to approach hashtags as parts of wider issue ecologies. To facilitate this process logistically, we have made our scripts freely available.

We thus suggest that sampling should not only attend to the internal or comparative relations, but, if possible, to the entire population – captured in the baseline – so that medium-specificity is reflected both in specific sampling techniques and the relative relevance of practices within the platform itself.

Acknowledgements

This project has been initiated in a Digital Methods Winter School project called “One Percent of Twitter” and we would like to thank our project members Esther Weltevrede, Julian Ausserhofer, Liliana Bounegru, Guilio Fagolini, Nicholas Makhortykh, and Lonneke van der Velden. Further gratitude goes to Erik Borra for his useful feedback and work on the DMI-TCAT. Finally, we would like to thank our reviewers for their constructive comments.

References

boyd, danah, and Kate Crawford. “Critical Questions for Big Data.” *Information, Communication & Society* 15.5 (2012): 662–679.

———, Scott Golder, and Gilad Lotan. “Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter.” *2010 43rd Hawaii International Conference on System Sciences*. IEEE, (2010). 1–10.

Bruns, Axel, and Stefan Stieglitz. “Quantitative Approaches to Comparing Communication Patterns on Twitter.” *Journal of Technology in Human Services* 30.3-4 (2012): 160–185.

Bryman, Alan. *Social Research Methods*. Oxford University Press, (2012).

Burgess, Jean, and Axel Bruns. “Twitter Archives and the Challenges of ‘Big Social Data’ for Media and Communication Research.” *M/C Journal* 15.5 (2012). 21 Apr. 2013 ‹http://journal.media-culture.org.au/index.php/mcjournal/article/viewArticle/561›.

Callon, Michel, *et al*. “From Translations to Problematic Networks: An Introduction to Co-word Analysis.” *Social Science Information* 22.2 (1983): 191–235.

Cha, Meeyoung, *et al*. “Measuring User Influence in Twitter: The

Million Follower Fallacy." *ICWSM '10: Proceedings of the International AAAI Conference on Weblogs and Social Media*. (2010).

Farber, Dan. "Twitter Hits 400 Million Tweets per Day, Mostly Mobile." *cnet*. (2012). 25 Feb. 2013 ‹http://news.cnet.com/8301-1023_3-57448388-93/twitter-hits-400-million-tweets-per-day-mostly-mobile/›.

Garcia-Gavilanes, Ruth, Daniele Quercia, and Alejandro Jaimes. "Cultural Dimensions in Twitter: Time, Individualism and Power." (2006). 25 Feb. 2013 ‹http://www.ruthygarcia.com/papers/cikm2011.pdf›.

Gilbert, Nigel. *Researching Social Life*. Sage, 2008.

Gillespie, Tarleton. "The Politics of 'Platforms'." *New Media & Society* 12.3 (2010): 347–364.

González-Bailón, Sandra, Ning Wang, and Alejandro Rivero. "Assessing the Bias in Communication Networks Sampled from Twitter." 2012. 3 Mar. 2013 ‹http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2185134›.

Gorard, Stephan. *Quantitative Methods in Social Science*. London: Continuum, 2003.

Hayles, N. Katherine. *My Mother Was a Computer: Digital Subjects and Literary Texts*. Chicago: University of Chicago Press, 2005.

Hong, Lichan, Gregorio Convertino, and Ed H Chi. "Language Matters in Twitter : A Large Scale Study Characterizing the Top Languages in Twitter Characterizing Differences Across Languages Including URLs and Hashtags." Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (2011): 518–521.

Huang, Jeff, Katherine M. Thornton, and Efthimis N. Efthimiadis. "Conversational Tagging in Twitter." *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia – HT '10* (2010): 173.

Krishnamurthy, Balachander, Phillipa Gill, and Martin Arlitt. "A Few Chirps about Twitter." *Proceedings of the First Workshop on Online Social Networks – WOSP '08*. New York: ACM Press, 2008. 19.

Krishnaiah, P R, and C.R. Rao. *Handbook of Statistics*. Amsterdam: Elsevier Science Publishers, 1987.

Langlois, Ganaele, *et al*. "Mapping Commercial Web 2 . 0 Worlds: Towards a New Critical Ontogenesis." Fibreculture 14 (2009): 1–14.

Mackenzie, Adrian. "More Parts than Elements: How Databases Multiply." *Environment and Planning D: Society and Space* 30.2 (2012): 335 – 350.

Marres, Noortje, and Esther Weltevrede. "Scraping the Social? Issues in Real-time Social Research." *Journal of Cultural Economy* (2012): 1–52.

Marwick, Alice, and danah boyd. “To See and Be Seen: Celebrity Practice on Twitter.” *Convergence: The International Journal of Research into New Media Technologies* 17.2 (2011): 139–158.

Montfort, Nick, and Ian Bogost. *Racing the Beam: The Atari Video Computer System*. MIT Press, 2009.

Noy, Chaim. “Sampling Knowledge: The Hermeneutics of Snowball Sampling in Qualitative Research.” *International Journal of Social Research Methodology* 11.4 (2008): 327–344.

Papagelis, Manos, Gautam Das, and Nick Koudas. “Sampling Online Social Networks.” *IEEE Transactions on Knowledge and Data Engineering* 25.3 (2013): 662–676.

Paßmann, Johannes, Thomas Boeschoten, and Mirko Tobias Schäfer. “The Gift of the Gab. Retweet Cartels and Gift Economies on Twitter.” *Twitter and Society*. Eds. Katrin Weller *et al*. New York: Peter Lang, 2013.

Poblete, Barbara, *et al*. “Do All Birds Tweet the Same? Characterizing Twitter around the World Categories and Subject Descriptors.” *20th ACM Conference on Information and Knowledge Management, CIKM 2011, ACM, Glasgow, United Kingdom*. 2011. 1025–1030.

Rieder, Bernhard. “The Refraction Chamber: Twitter as Sphere and Network.” *First Monday* 11 (5 Nov. 2012).

Rogers, Richard. *The End of the Virtual – Digital Methods.* Amsterdam: Amsterdam University Press, 2009.

Savage, Mike, and Roger Burrows. “The Coming Crisis of Empirical Sociology.” *Sociology* 41.5 (2007): 885–899.

Stalder, Felix. “Between Democracy and Spectacle: The Front-End and Back-End of the Social Web.” *The Social Media Reader*. Ed. Michael Mandiberg. New York: New York University Press, 2012. 242–256.

Stieglitz, Stefan, and Nina Krüger. “Analysis of Sentiments in Corporate Twitter Communication – A Case Study on an Issue of Toyota.” *ACIS 2011 Proceedings*. (2011). Paper 29.

Tumasjan, A., *et al*. “Election Forecasts with Twitter: How 140 Characters Reflect the Political Landscape.” *Social Science Computer Review* 29.4 (2010): 402–418.

Tukey, John Wilder. Exploratory Data Analysis. New York: Addison-Wesley, 1977.

Uprichard, Emma. “Sampling: Bridging Probability and Non-Probability Designs.” *International Journal of Social Research Methodology* 16.1 (2011): 1–11.

This work is licensed under a Creative Commons Attribution - Noncommercial - No Derivatives 3.0 License.

an mcpublication | Supported by QUT creative industries |

Copyright © M/C, 1998-2008 | ISSN 1441-2616 |