EXPLORING THE GLOBAL DEMOGRAPHICS OF TWITTER

Axel Bruns
Queensland University of Technology

Daryl Woodford
Queensland University of Technology

Troy Sadkowsky
Queensland University of Technology

Abstract

In spite of the substantial international success of Twitter as a social media platform, reliable information about its userbase is surprisingly difficult to come by. Other than the 232 million “monthly active users” reported in the company’s disclosures to the U.S. Securities and Exchange Commission ahead of its listing on the stock exchange, and some high-level breakdowns of account numbers across a number of key markets, most other assumptions about the Twitter userbase remain guesswork or are based on surveys with comparatively limited sample sizes. This paper takes a different approach to exploring the demographics of the platform: by undertaking a long-term crawl process across the entire Twitter user ID numberspace, we have gathered the publicly available details on every Twitter user account created between the platform’s emergence in 2006 and the conclusion of our crawl in 2013. By identifying the key patterns within this database of some 872 million accounts existing during our collection period, we are able to provide a much more comprehensive overview of Twitter’s footprint across the globe, its patterns of growth, and of typical user careers as listeners, followers, hubs and communicators than has been possible in any previous study.

Introduction: Known Unknowns

Both in terms of overall user numbers and in terms of its recognition in mainstream media coverage, Twitter is one of the most successful general-purpose transnational social media platforms in the world; amongst comparable platforms, its total userbase appears to be second only to market leader Facebook and on par with that of Weibo, which remains centered largely on mainland and diaspora Chinese users. Detailed information on the shape and make-up of the global Twitter userbase remains difficult to come by, however: Twitter’s own publicly announced userbase information is limited to headline figures released to the Securities and Exchange Commission (Twitter, Inc., Suggested Citation (APA): Bruns, A., Sadkowsky, T., & Woodford, D. (2014, October 22-24). Exploring the global demographics of Twitter. Paper presented at Internet Research 15: The 15th Annual Meeting of the Association of Internet Researchers. Daegu, Korea: AoIR. Retrieved from http://spir.aoir.org.
2013), and to similar figures quoted in other public releases designed to prove the company’s success and to attract further investment. From such statements, it appears that there are 232 million “monthly active users”, in Twitter’s own definition, but it is unclear what proportion of the total number of registered accounts this number represents, or whether there is a greater number of users who are active less often than monthly. Further, there is a general absence of large-scale longitudinal studies which are able to chart the growth of the Twitter userbase over time and to pinpoint specific increases or decreases in such growth that may be correlated with media coverage or world events.

The data which would offer answers to such questions are largely available from the Twitter Application Programming Interface (API) in the form of the publicly accessible profile information for user accounts (other than for the small minority of “protected”, private accounts, for which only less detailed information is available). However, the significant effort involved in gathering such data for the entire Twitter userbase has meant that few researchers have thus far attempted to do so.

A Comprehensive Crawl of the Twitter User ID Numberspace

Our project addresses this identified gap in existing knowledge about the Twitter userbase. During 2013, we conducted a simple, ‘brute force’ crawl of the entire Twitter user ID numberspace, beginning from user ID 0 and continuing through (at the time of writing) to user ID 1,967,860,448. This approach builds on the fact that – at least at present – Twitter does not re-use user IDs even if an existing account is deleted and thus since the launch of the Twitter platform in 2006, user IDs have steadily counted upwards.

In crawling the numberspace in this way, we gathered public user information provided for each account by the Twitter API’s users/lookup function, including datapoints such as the account’s registration date, its number of followers, followees, tweets posted, timezone, interface language, free-text description and location information, and a range of other details. At the time of writing, we have identified a total of 872 million accounts, and using this dataset (of some 1.97TB of profile information in JSON format), we are now able to identify a number of global and local patterns within the overall Twitter userbase.

Some Preliminary Patterns in the Global Twitter Userbase

Our conference paper will present a variety of key patterns within the global Twitter userbase, including an overall growth curve for the global userbase and a breakdown into languages, timezones, and other demographic information; we present these in aggregate in order to avoid the re-identification of specific individual users. For the purposes of this short paper, we outline a number of notable preliminary findings, and indicate a range of further questions that our conference paper will address.
Growth over Time

By plotting the creation dates of existing Twitter accounts over time, it becomes possible to retrace the growth curve of the platform and to pinpoint significant moments in the history of Twitter (and their relationship with external events). Such trends in registrations may also be further investigated in correlation with other relevant user profile details. By way of example, Fig. 1 shows the daily growth rate in Twitter accounts which selected a specific timezone in their profile settings, and points to a sharp increase above the long-term average for accounts selecting a Tokyo or (less markedly so) Hawai‘i timezone in the hours and days following first reports of the 11 March 2011 earthquake & tsunami.

![Graph showing daily account creation rates for Twitter accounts with Tokyo, Hawai‘i, and Alaska timezones during the first four months of 2011.](image)

Fig. 1: Daily account creation rates for Twitter accounts with Tokyo, Hawai‘i, and Alaska timezones during the first four months of 2011

These patterns point to the fact that major acute events (Burgess and Crawford, 2011) not only create significant activity amongst existing Twitter users – enabling them, for example, to gather in ad hoc publics around key hashtags (Bruns & Burgess, 2011) – but also lead to the creation of a substantial number of new Twitter accounts whose owners are likely to have signed up to Twitter expressly for the purpose of following the platform’s up-to-date news coverage and joining in with the conversation. In this context, it should also be noted that our dataset, gathered in 2013, only contains those accounts from the 2006/7 and 2011 timeframes depicted in fig. 1 that were still in existence in 2013.

Geographic and Language Distribution

In the conference paper, we also intend to pursue the question of the geographic distribution of Twitter users. While timezone, GPS, and free-form location information each provide substantial challenges for an accurate pinpointing of nationality or geographic positioning, a triangulation between these various datapoints nonetheless
provides an opportunity to identify at least general global activity patterns. Fig. 2 illustrates this by using data for Twitter accounts with IDs in the 0-1,000,000 numberspace: using only the timezone information (as set when we gathered the data in 2013).

![Map of Twitter accounts by timezone](image)

**Fig. 2: Twitter accounts with user IDs between 0 and 1,000,000 by timezone**

**Typical Account Careers**

A third major area of investigation concerns patterns in the ‘account careers’ of the profiles we have identified. Here, we intend to explore the relative prominence of accounts with specific ratios between their key activity indicators, to examine whether a range of common user types could be defined. These categories will emerge as we explore more comprehensively the patterns in these and other user metrics within our dataset. An additional metric of interest, for example, is also the use of the more recently introduced favouriting functionality by Twitter users; our working hypothesis is that more recent users, for whom Twitter favouriting may appear as analogous to Facebook liking, use this functionality more readily than older users who came to know Twitter before favouriting was introduced.

**Conclusion**

As the first scholarly investigation of the Twitter userbase which is able to draw on a comprehensive (as of late 2013) set of public profile data, our study will be able to identify key fundamental features of the Twitter userbase. This also provides an important baseline for many other studies of Twitter users and their activities, by making...
available far more accurate data on the size and growth of the global Twitter population than have been available from reliable sources to date.

References


