4-2015

Rich People Don't Have More Followers! Overcoming Social Inequality With Social Media

Hemank Lamba
Carnegie Mellon University

Momin M. Malik
Carnegie Mellon University

Constantine Nakos
Carnegie Mellon University

Jurgen Pfeffer
Carnegie Mellon University

Published In
Proceedings of the International Social Computing, Behavioral-Cultural Modeling and Prediction Conference (SBP) Grand Data Challenge, 2015.

Rich People Don't Have More Followers! Overcoming Social Inequality With Social Media

Hemank Lamba, Momin M. Malik, Constantine Nakos, Jürgen Pfeffer
(Alphabetical Order)

Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
contact: `jpfeffer@cs.cmu.edu`

Abstract. Previous work on personal networks has shown that higher socioeconomic status results in larger and more powerful networks. With the Internet, in particular with social media, it has become easier to establish and maintain relationships, suggesting an equalizing effect. However, people of different socioeconomic status use these new resources in different ways creating a *digital divide*. In this article we study popularity on Twitter based on estimated socioeconomic status in real life. We collect 1 billion geo-coded Tweets from the United States and connect the geographic position of the sender with socioeconomic data at the level of Census block groups. We show that people tweeting from higher income areas do not have more followers. Rather, there is a small negative correlation between income and number of followers.

1 Introduction

Research into social networks has shown that socioeconomic status (SES) is positively related to network range and composition [Campbell et al., 1986]. That is, people of higher socioeconomic status have better networks. With the rise of the Internet, early optimism about its democratizing potential (inevitability, even) gave way to worries about access—that those of lower socioeconomic status were unable to access the Internet and that this was subverting that democratizing potential. Research has consistently shown, however, that access is not sufficient [Hargittai and Litt, 2013] [Hargittai and Hsieh, 2012]; there have been numerous studies that even with equal amounts of access, people of different socioeconomic backgrounds use digital resources in very different ways. In particular, people of higher socioeconomic status engage in far higher amounts of information seeking for news, finance, and health.

We propose to revisit the work on SES and network resources in light of research on digital inequalities. Already, research has shown that the Internet use is associated with creating and maintaing more weak ties [Chen, 2013], and socioeconomic status plays a role in activating online network ties [Boase et al., 2006] [Smith et al., 2012]. However, the differential nature of resources specifically in online networking along lines of SES has not yet been assessed.

In this article, we use utilize Twitter data to analyze whether socioeconomic factors influence the number of social connections, i.e. followers. Over the course of 1.5 years we have collected approximately 1 billion geo-coded tweets from the United States of America. These tweets were mapped to the geographic area of 200,000 US Census *block groups*. Attributes of the tweets (e.g. number of followers of sender) were correlated with socioeconomic data from the US Census Bureau's 2008-2012 American Community Surveys.

Research question: Is popularity on Twitter related to socio-economic status of the users? In particular, do people that send tweets from higher status areas of the United States have more followers?

2 Related Work

In terms of sizes of real-world networks for different socio-economic groups, much work has been done to describe *core* networks. These *ego networks* are often collected with Burt's *name generator* survey instrument [Burt, 1984]. This asks people for the names of those with whom they discuss "very important topics". We know from this work that women have larger core networks than men [Moore, 1990], that married and cohabited people have smaller networks than singles [Johnson and Leslie, 1982], and that age is not a significant factor for differences in network size [Fung et al., 2001].

Twitter has about 300 Million active monthly users who sent 500 Million tweets every day. The tweets of a particular user are visible to their followers. Every user on Twitter has on average 208 followers. This is in line with *Dunbar's number*, i.e. every human is limited to about 150 contacts due to cognitive constraints. Especially from a network perspective, Twitter followers include very weak ties [Granovetter, 1973] and ties that hardly can be classify as network *ties*, e.g. following celebrities or companies. In opposite to Facebook [**?**] (avg. 140), Twitter follower relationships are not reciprocal, which results in a very skewed follower distribution of 80+% of users having 50 or less followers and some celebrities with millions of followers[1].

In order to link Twitter to socioeconomic status, we rely on geo-coded tweets. These have been found [Liu et al., 2014] to be 1.23% of the total volume of tweets, raising concerns about representativeness [Ruths and Pfeffer, 2014]. However, a nationally representative sample survey by Pew [Zickuhr, 2013] on the use of "geosocial services" shows that the users most likely to use such services are not necessarily young, white, male, urban users with a large household income; in fact, in 2011, Pew found that "men, African Americans and Hispanics, adults in households earning less than $30,000 per year, and those who have not gone to college were significantly more likely than other social media users to use automatic location-tagging." In 2013, it was users under age 50, and those living in suburban areas who most used location tagging, with "no significant differences by gender, education level, or household income." Consequently, we

[1] More Twitter statistics: `http://www.beevolve.com/twitter-statistics`

can assume that our collection of geo-coded Tweets is representative enough to draw conclusions about Twitter as a whole.

3 Data and Analysis

3.1 Geo-Coded Twitter Data

By utilizing the Twitter Streaming API we have collected approximately 1 billion tweets with a geographic boundary box around the continental US plus Hawaii (124,7625W 24.5210N - 66.9326W 49.3845N) from April 2013 to December 2014. Consequently, all our tweets are geo-coded with lat/long GPS coordinates.

3.2 US Census Block Groups

The area of the United States (including Puerto Rico) is subdivided by the US Census Bureau into 11,155,486 *census blocks*. Blocks are geographic areas bounded by streets, rivers, etc. For instance, in cities, blocks normally align with street blocks. Census blocks are grouped into *block groups*, which are grouped into *census tracts*, which aggregate into counties that are the administrative units that sub-divide States.

Our analysis has been accomplished on *block group* level as these are the smallest geographic entities for which Census data are available. The 2010 US Census (including Puerto Rico and other US territories) included 211,267 block groups. Census block groups consist of 600 to 3,000 inhabitants and have a unique 12 digit identifier, the so-called *FIPS Codes*. For these block group level we have collected two sets of data:

Socioeconomic Data. Socioeconomic data of census block groups published by the Census Bureau is based on data collection from a fraction of the households in all geographic areas of the United States. The American Community Survey (ACS) is an ongoing survey with a broad array of social and economic topics. We selected the dataset collected from 2008 to 2012 and downloaded socioeconomic variables (age, gender, race, education, household income) from the Census Bureau website[2] for every census block group.

Geospatial Data. The US Census Bureau also provides geographic boundary files (shapefiles) for census data. These shapefiles are available per state[3] and are a collection of the GPS coordinates of the borders of every Census block (and therefore also of every block group). We have combined the shapefiles of 49 states (all US states excluding Alaska) for our purpose.

[2] http://www.census.gov/acs/www/
[3] http://www2.census.gov/geo/tiger/TIGER_DP/2010ACS/

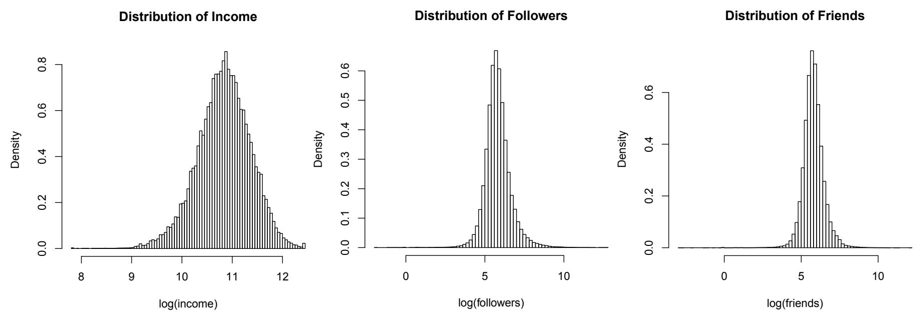

Fig. 1: Distribution of average followers, average friends, and average median income of 1 billion Tweets mapped to socioeconomic data of US Census block groups.

3.3 Putting it All Together

With Python code (utilizing the 'shapely' package) we identified which Census block group each Tweet fell into. The block group FIPS codes of the geospatial data were used to connect the Tweets via geographic ares to the socioeconomic data of the US Census.

3.4 Results

To increase the robustness of our analysis we remove block groups with less than 10 users that have sent Tweets from these regions. The logs of the resulting data vectors (median income, followers, friends) have normal distribution characteristics and can are therefore suitable for correlation as well as for regression analysis (see Figure 1).

The primary research question of this paper is whether people tweeting from higher income areas have more followers. Therefore, we correlate the followers and friends data vector with the median household income. We use Pearson correlation as well as Spearman's rank correlation to incorporate nonparametric aspects.

The results in Table 1 show the weak negative correlation between followers and median household income. The correlation is even stronger for the number of friends, but this was not in the focus of our research question.

	Pearson			Spearman		
	r2	t	p	rho	S	p
Followers	-0.0335	-13.0911	$<2.2e-16$	-0.0559	6.2e+14	$<2.2e-16$
Friends	-0.0717	-28.0613	$<2.2e-16$	-0.0844	6.4e+14	$<2.2e-16$

Table 1: Correlations of followers and friends with median household income of Census block groups.

3.5 Data Release

After publication of this article we will publish the Tweet IDs of approximately 1 Billion geo-coded Tweets that have been used for this study. We will organize the published Tweets by county and time so that other researchers can easier access the Tweets of their geographic area and/or time frame of interest.

4 Conclusions

We examined whether higher socioeconomic status is related to having more followers on Twitter, in order to investigate whether higher SES is related to greater network resources in online social networking. We operationalized this question by connecting geo-located tweets to US Census data of the particular areas in which the tweets were sent. Our results indicate there is no positive correlation between median income of geographic areas and the number of followers of Twitter users sending from these areas but we found a slightly negative correlation.

In this article we showed evidence that Twitter is an instrument to overcome social inequality indicating that social media has the potential of overcoming social inequality—at least in terms of social networks, information exchange, and online prestige. The nature of the exchanged information, the communities of prestige, and the nature of network resources is the topic for future work, in which we will go into greater detail analyzing socioeconomic variables.

References

Boase et al., 2006. Boase, J., Horrigan, J. B., Wellman, B., and Rainie, L. (2006). The strength of internet ties: The internet and email aid users in maintaining their social networks and provide pathways to help when people face big decisions. Technical Report Pew Internet & American Life Project, Pew Research Center, 615 L ST., NW — Suite 700, Washington, D.C. 20036.

Burt, 1984. Burt, R. S. (1984). Network items and the general social survey. *Social Networks*, 6:293–340.

Campbell et al., 1986. Campbell, K. E., Marsden, P. V., and Hurlbert, J. S. (1986). Social resources and socioeconomic status. *Social Networks*, 8(1):97–117.

Chen, 2013. Chen, W. (2013). Internet use, online communication, and ties in americans networks. *Social Science Computer Review*, 31(4):404–423.

Fung et al., 2001. Fung, H. H., Carstensen, L. L., and Lang, F. R. (2001). Age-related patterns in social networks among european americans and african americans: implications for socioemotional selectivity across the life span. *International journal of aging & human development*, 52(3):185–206. PMID: 11407486.

Granovetter, 1973. Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1680.

Hargittai and Hsieh, 2012. Hargittai, E. and Hsieh, Y. P. (2012). Digital inequality. In Dutton, W. H., editor, *The Oxford Handbook of Internet Studies*, pages 129–150. Oxford University Press, Oxford.

Hargittai and Litt, 2013. Hargittai, E. and Litt, E. (2013). New strategies for employment? internet skills and online privacy practices during people's job search. *IEEE Security and Privacy*, 11(3):38–45.

Johnson and Leslie, 1982. Johnson, M. P. and Leslie, L. (1982). Couple involvement and network structure: A test of the dyadic withdrawal hypothesis. *Social Psychology Quarterly*, 45(1):34.

Liu et al., 2014. Liu, Y., Kliman-Silver, C., and Mislove, A. (2014). The tweets they are a-changin': Evolution of twitter users and behavior. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, ICWSM '14.

Moore, 1990. Moore, G. (1990). Structural determinants of men's and women's personal networks. *American Sociological Review*, 55(5):726.

Ruths and Pfeffer, 2014. Ruths, D. and Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, 346(6213):1063–1064.

Smith et al., 2012. Smith, E. B., Menon, T., and Thompson, L. (2012). Status differences in the cognitive activation of social networks. *Organization Science*, 23(1):67–82.

Zickuhr, 2013. Zickuhr, K. (2013). Location-base services. Technical Report Pew Internet and American Life Project, Pew Research Center.