The Digital Divide among Twitter Users and its Implications for Social Research:

Grant Blank

University of Oxford

Please cite as:

Blank, Grant. (2016). The digital divide among Twitter users and its implications for social research. *Social Science Computer Review*. DOI: 10.1177/0894439316671698.

Abstract

Hundreds of papers have been published using Twitter data but few previous papers report the digital divide among Twitter users. British Twitter users are younger, wealthier and better educated than other Internet users, who in turn are younger, wealthier and better educated than the off-line British population. American Twitter users are also younger and wealthier than the rest of the population but they are not better educated. Twitter users are disproportionately members of elites in both countries. Twitter users also differ from other groups in their online activities and their attitudes. These biases and differences have important implications for research based on Twitter data. The unrepresentative characteristics of Twitter users suggest that Twitter data are not suitable for research where representativeness is important such as forecasting elections or gaining insight into attitudes, sentiments or activities of large populations. In general, Twitter data seem to be more suitable for corporate use than for social science research.

Keywords: Social media; Twitter; representativeness; selection bias; elites; Oxford Internet Survey; OxIS; Pew Internet and American Life

The Digital Divide among Twitter Users and its Implications for Social Research:

The digital divide has now been intensively investigated for over a decade, but it also remains an area of active and innovative research. In this article, I provide a multivariate empirical analysis of the digital divide among Twitter users. I extend this by comparing Twitter users and non-users with respect to their characteristic patterns of Internet activity and to certain key attitudes. This is important for two reasons. First, this fills a gap in our knowledge about an important social media platform, and it joins a surprisingly small number of studies that describe the population that uses social media (Wells & Link, 2014).

Second, the implications of who uses Twitter stretch beyond digital divide questions. The simplicity of 140-character tweets and the extreme ease with which they can be collected have proven attractive to hundreds of researchers (Bruns & Weller, 2014). Researchers have attempted to predict the behavior of populations by looking solely at the content of tweets. Examples include attempts to predict disease outbreaks, election results, film box office gross, and stock market movements; see Kalampokis et al. (2013) for a review. Who uses Twitter has major implications for research attempts to use the content of tweets for inference about population behavior. For what populations are Twitter data appropriate? Questions about the representativeness of Twitter users have been raised before (e.g. boyd & Crawford, 2013; Mislove, Lehmann, Ahn, Onnela, Rosenquist, 2011; Tufekci, 2014) but prior answers have been limited by lack of data, or lack of multivariate models or both. I explore this issue in the discussion section.

The plan of this article is as follows. I begin by summarizing the digital divide literature with special attention to Twitter. This gives us a basis for predictions of the shape of the digital divide on Twitter. I then analyze Twitter users, comparing United Kingdom data with United

States data to show where they are similar and where they differ on demographics, attitudes and Internet use. Finally, I discuss research attempts to use Twitter content to predict behavior. I explore the implications of these results for attempts to use Twitter content as a research tool.

Digital divides

The digital divide has been a major focus of online research; for representative examples see Bonfadelli (2002), van Dijk (2005) or van Deursen & van Dijk (2013). Digital inequality can take many forms that have been explored in the United Kingdom and the United States, as well as elsewhere. In Britain the Oxford Internet Survey (OxIS) has charted 10 years of trends in the online population (Dutton & Blank, 2011, 2013). These studies document that British Internet users have been younger, better educated and wealthier than the off-line population since the earliest wave in 2003. Some differences between the online and off-line populations have disappeared, such as the gender gap which was important in the early 2000s but disappeared by 2011. Students have been the most likely to use the Internet, although employed people have been closing the gap. Retired people are least likely to be Internet users. Disabled people are about half as likely to use the Internet as non-disabled, although this gap has been declining. Black and Asian minorities are more likely to use the Internet than whites. Urban-rural differences are not significant. In the United States the Pew Internet and American life project has collected similar data (Pew Research Center, 2014). The characteristics of the online British population broadly parallel American Internet users: Users are younger, wealthier, better educated and less likely to be disabled. Students are most likely to be Internet users and retired people least likely. Blacks are least likely to use the Internet. A difference is that American users are more likely to be urban or suburban. Similar studies in other countries have also documented the characteristics of the online population. The World Internet Project (www.worldinternetproject.org) has data from over 30 nations. These examples suggest considerable interest in the characteristics of the online

population, and it is striking that few comparable studies exist of the characteristics of the population of Twitter users.

Prior attempts to estimate demographics

Previous studies of Twitter demographics have been attempts to estimate the demographic characteristics of Twitter users based on tweets and other public data like profiles completed by Twitter users (e.g. Mislove et al 2011; Pennacchiotti & Popescu, 2011; Rao, Yarowsky, Shreevats & Gupta, 2010; Sloan, Morgan, Housley, Burnap, & Williams, 2015; Sloan et al., 2013). These studies attempt to infer characteristics like gender, age, region, race, political orientation or other attributes using machine learning or other computational techniques. They compare their computational classifications to reference datasets that have been hand-coded. This tells them how well their algorithms match hand-coded results; they can obtain a number that measures this kind of accuracy, but it is a weak measure of accuracy. They have few variables on which to compare their results with actual, real-world demographic data on Twitter users, Twitter non-users, or non-users of the Internet. So they can't tell us much about how Twitter users compare to anyone else. Mislov et al. (2011) and Sloan et al. (2013) make the most methodologically sophisticated attempts to find Twitter demographics.[1] Both are remarkable projects. Using first and last names,

[1] Mislov et al. (2011) infer geographic location using the self-reported location field in Twitter user profiles: 75.3% of publicly visible users enter something in that field and Mislov et al. can attach reported locations to latitude and longitude on a US map in 8.8% of the 75.3%. They compare this data to US Census 2000 data at the county level. They infer gender based on first names and race/ethnicity based on last names. They are able to match gender to 64.2% of users and race/ethnicity to 18.5% of users (p. 557). The relatively low percentages are not encouraging.

Mislov et al. (2011, p. 557) infers gender in 64.2% of cases and race/ethnicity in 18.5%. Also using names, Sloan et al. (2013) match gender for 48%. This is less than the 50% correct matches one would expect from random guessing. With such a low percentage of matches it is hard to be confident of the results.

Three papers attempted to infer age of Twitter users. Rao, Yarowsky, Shreevats, and Gupta (2010) categorize age as a binary variable, over or under age 30 (They do not explain why they chose this division). Their best prediction using language independent variables is 74.11%, which is lower accuracy than the random prediction described above. Sloan, Morgan, Housley, Burnap and Williams (2015) infer numerical age using information in the Twitter profile. They are successful in 0.37% of the cases: 1,470 out of 398,452 Twitter profiles. Other papers have attempted to infer liberal-conservative political orientation (Pennacchiotti & Popescu, 2011; Rao et al., 2010), race (Chang, Rosenn, Backstrom & Marlow, 2010), occupation (Sloan et al., 2015) and ethnicity (Chang et al., 2010). I am not aware of any attempt to infer education, marital status, income or socio-economic status. Part of the problem of these studies is that the variables they predict are relatively simple. Gender is often used as a binary variable, but other variables often are not. White versus non-white is a very important distinction but it is a pale shadow of the racial and ethnic complexity of the US or the UK in the 21st century. This problem is even more pronounced with age. The first law of the Internet is that everything is related to age. Age has effects all along its range from the very youngest to the oldest. A binary age variable is theoretically and empirically inadequate. When researchers infer actual numerical age (Sloan et al., 2015) their success rate is very low (0.37%). In short it is not clear that these efforts are useful.

Sloan et al. (2013) also infer gender from first names. They match 48% using their name database, which is 40,000 Namen (Michael, 2007).

The Pew Internet and American Life project has been the major source of high quality information about social media users in the United States, including Twitter. Although Pew supplies descriptive tables, it does not draw out the implications of its data for the digital divide or for social science. The first demographic breakdowns of Twitter users were published by Duggan and Brenner (2013), and Duggan and Smith (2014) based on 2012 and 2013 Pew surveys of the United States. They report the gender, age, race, education, income and urban-rural status of American Twitter users. This is a considerable step forward, but these reports lack any comparative frame since they include only Twitter users. To understand Twitter users we need comparative data that shows us how Twitter users differ (or not) from other online and off-line populations. Only when we have comparative data can we understand the potential biases of conclusions based solely on Twitter users. As far as I have been able to determine, no one has published the actual comparative demographic breakdown of Twitter users. An exception to this statement is Hargittai and Litt (2012) who presenet comparative demographics for a sample of young Twitter users, age 18-24. Hargittai (2015) presents multivariate results, discussed below, but not comparative demographics. The data needed for comparison are available in the OxIS and the Pew Research Center's Internet and American Life Project.

Methods

OxIS collects data on British Internet users and non-users. Conducted biennially since 2003, the surveys are random samples of more than 2,000 individuals aged 14 and older in England, Scotland, and Wales. Interviews are conducted face-to-face by an independent survey research company. See Dutton and Blank (2013) for details of the data collection and sample. The analyses below are based on the 480 Twitter users, 1,270 social network site (SNS) users or the 1,610 Internet users out of the full 2013 sample of 2,053 respondents.

The Pew Research Center's Internet and American Life project conducts regular telephone surveys examining the impact of the Internet. In May 2013 Pew (2013) asked a random sample of the American population age 18 or older about SNS use including Twitter. Analyses of the Pew dataset are based on 341 Twitter users, 1,377 SNS users or 1,912 Internet users (using Pew's definition of Internet users) out of the full sample of 2,252 respondents (see Table 1).

Table 1. British Twitter users compared to other groups

	Twitter users	Non-Twitter SNS users	Non-SNS Internet users	Off-line population	Full sample
Number	480	789	341	442	2,052
Percentage of all Respondents	23.4	38.5	16.6	21.5	100.0
Percentage of all Internet users	29.8	49.0	22.4	—	—

Note: Total SNS users (including Twitter users): 1,270, total Internet users (including Twitter users): 1,610, and the total adds to 2,052 not 2,053 because two SNS users who Don't Know if they use Twitter are omitted, but rounding the weighted frequencies adds one case. SNS = social network site.

We also use nine standard, self-reported demographic variables. We use four education categories: no degree, secondary education degree, further education, and university undergraduate or post-graduate degree. Race is coded as white versus non-white. Place is coded as urban versus rural. Lifestage is a four-category variable: students, employed, unemployed and retired. Marital status has five categories: single, married, living with partner, divorced, widowed. We also include age and binary measures of disability and gender. All of these variables are available in the OxIS dataset; the Pew dataset contains age, education, income, race, lifestage, marital status and gender. To enhance comparability I have attempted to make the demographic categories as similar as possible in the two datasets; I note any differences below.

OxIS contains measures of participation in 43 activities that people do on the Internet (The Pew data set does not contain activity measures).The activities cover an extremely wide range, from buying online, to blogging, to making travel plans, to listening to music, to finding out health information, to reading celebrity news or gossip. Each activity was measured using an identical 6-category Likert scale ranging from 0 = *never participate* to 5 = *do several times per*

day. We use these variables to measure amount, variety and types of Internet use. We create variables following definitions and procedures used by Blank and Groselj (2014). Like Blank and Groselj (2014) we did a principal components analysis to reduce the 43 variables to a more manageable set of types of Internet use. After Varimax rotation and Kaiser normalization the PCA yielded 10 components with eigenvalues greater than 1.0. We constructed binary variables measuring participation or non-participation in each of the 10 activity types. To be counted as participating in a type, a respondent had to report doing the activities that load strongly on a component an average of more than never. Since Twitter use is *not* included among these 43 variables we can compare Twitter user's vs non-users participation in 10 types.

We also compare Twitter users on how varied is their use of the Internet. Variety is measured by the count of the number of different activities that a respondent does more than never, with a theoretical range from zero to 43. Finally, we compare users on total amount of use. Amount of use measures how much any respondent does on the Internet. Since the Likert scales ask how often people do each activity, we can simply sum all of the activities to measure total amount of Internet use. This scale is continuous, with a theoretical range from zero to 215.

Results

We classify respondents into four mutually exclusive groups based on their relation to Twitter and the Internet. The groups are Twitter users, users of other SNSs who do not use Twitter, Internet users who do not use SNSs, and non-users of the Internet who we will call the "off-line population". Notice that each respondent in the survey will be a member of one and only one group. This allows us systematically to compare how Twitter users differ from the other three groups: SNS users, Internet users, and off-line non-users. Table 1 reports the Ns and percents comparing these four groups. In 2013 in Britain, Twitter users are about 23% of all respondents and 30% of all Internet users.

Table 2 contains the same data from the Pew dataset. Notice that American Twitter use is about 8 percentage points less than British Twitter use (15.2% of all respondents compared to 23.4%), although the proportion of Americans who use the Internet is higher.

Table 2. American Twitter users compared to other groups

Characteristics	Twitter users	Non-Twitter SNS users	Non-SNS Internet users	Off-line population	Full sample
Number	341	1,054	512	340	2,252
Percentage of all Respondents	15.2	46.9	22.8	15.1	100.0
Percentage of all Internet users	17.9	55.3	26.8	—	—

Note: Total SNS users (including Twitter users) = 1,377, total Internet users (including Twitter users) = 1,912, and four respondents who refused to answer the SNS question have been omitted. SNS = social network site.

Demographic comparisons of Twitter users to other groups are in Table 3 for both Great Britain and the United States. I first talk about the British data, followed by comparisons to the American data. The columns in both of these tables correspond to the columns in Tables 1 and 2. The nine demographic variables are ordered approximately from the strongest effects to the weakest.

I discuss each demographic variable in the British part of Table 3 in succession, beginning with age. Age differences are large. Twitter users are more likely to be young by about 26 percentage points: 30% of Twitter users are between age 18-24 compared to 4.3% of the off-line population. Comparatively few Twitter users are over age 55, whereas over 70% of the off-line population is over age 55. Looking at education, Twitter users are less likely to have no educational qualifications by 60 percentage points compared to off-line respondents. Twitter users are more likely to have graduated from college. Thirty-five percent of Twitter users have at least one higher education degree, about seven percentage points more than non-Twitter users and 29 percentage points more than people who are off-line. For marital status, Twitter users are about 17 percentage points more likely to be single compared to SNS users who don't use Twitter and about 15 percentage points less likely to be married. Among the other marital status categories,

Twitter users are more likely to be living with a partner but less likely to be divorced, separated or widowed than the other categories. In terms of lifestage, Twitter users are eight percentage points more likely to be students than non-Twitter SNS users and 17 percentage points more likely than the off-line population. Twitter users are also more likely to be employed, and they tend not to be unemployed or retired. Income also shows large differences. Twitter users are about five percentage points more likely to have incomes of £50,000/year or more compared to non-Twitter SNS users and about seven percentage points more likely to have incomes of between £40,000/year and £50,000/year. Compared to the off-line population they are over 50 percentage points less likely to earn less than £12,500 per year. Twitter users are about 10 percentage points more likely to be white than those who are not online. They are slightly more likely to be male than off-line people. Twitter users are somewhat less likely to be disabled than non-Twitter SNS users by about six percentage points. They are about 11 percentage points *less* likely to be disabled than off-line respondents.

Table 3: Demographic Comparison of Twitter Users to Other Groups (%)

British Data

	Twitter Users	Non-Twitter SNS Users	Non-SNS Internet Users	Off-line population
Age				
Age 18-24	30.2	13.6	3.8	4.3
Age 25-34	29.9	21.2	5.3	4.9
Age 35-44	19.0	22.0	17.4	4.4
Age 45-54	12.9	20.8	23.0	12.0
Age 55-64	6.5	14.6	22.2	16.6
Age 65-74	1.5	5.5	19.0	25.2
Age 75+	0.1	2.4	10.4	32.6
Education				
No qualifications	5.6	14.9	18.5	65.7
Secondary degree	39.5	38.5	39.5	24.9
Further education	20.3	19.2	9.5	3.8
Higher education	34.6	27.4	32.6	5.6
Marital status				
Single	46.1	29.2	12.6	14.0
Married	31.4	46.3	62.1	41.9
Living with partner	18.0	16.2	9.5	5.4
Divorced/separated	4.1	6.6	8.3	11.2
Widowed	0.5	1.7	7.5	27.5
Lifestage				
Students	17.7	9.2	4.1	0.2
Employed	64.0	57.7	45.8	17.3
Retired	3.7	10.5	36.8	63.2
Unemployed	14.6	22.6	13.3	19.3

American Data

	Twitter Users	Non-Twitter SNS Users	Non-SNS Internet Users	Off-line population
Age				
Age 18-24	33.3	16.1	4.7	1.0
Age 25-34	18.0	22.1	9.6	5.0
Age 35-44	18.5	21.9	14.3	8.1
Age 45-54	19.0	19.2	27.2	11.6
Age 55-64	8.7	13.0	21.9	25.1
Age 65+	2.8	7.8	22.3	49.2
Education				
Less than HS	5.8	6.2	7.9	26.4
HS diploma	24.8	30.9	30.2	48.6
Some college	37.1	34.0	31.9	17.8
Higher education	32.3	29.9	30.0	7.2
Marital status				
Single	42.1	27.1	16.5	9.1
Married	37.7	49.0	56.0	42.9
Living with partner	6.4	7.5	5.8	4.1
Divorced/separated	11.8	13.4	13.8	18.9
Widowed	2.0	3.2	7.9	25.0
Lifestage				
Students	2.7	0.70	0.2	0.0
Employed	74.5	68.8	58.0	24.9
Retired	5.2	12.1	25.9	45.9
Unemployed	17.5	18.4	16.0	29.2

Yearly Household Income (UK £				
Less than £12,500	23.7	29.9	34.6	75.1
>£12,500-£20,000	21.9	25.1	31.9	14.2
>£20,000-£30,000	17.1	22.7	18.2	6.1
>£30,000-£40,000	16.4	13.5	7.5	3.2
>£40,000-£50,000	10.3	3.5	3.7	0.7
More than £50,000	10.6	5.2	4.1	0.8
Race				
White	83.7	87.8	95.3	93.2
Non-white	16.3	12.3	4.7	6.8
Gender				
Male	52.2	46.6	54.1	47.1
Female	47.8	53.4	45.9	52.9
Disability				
Not disabled	93.2	87.8	83.4	82.2
Disabled	6.8	12.2	16.6	17.9
Urban-rural				
Rural	8.8	11.8	15.3	17.0
Urban	91.2	88.2	84.7	83.0

Yearly Household Income (US $)				
Less than $20,000	17.4	17.2	14.5	48.9
>$20,000-$30,000	6.4	14.8	12.7	15.8
>$30,000-$40,000	10.4.	11.2	8.6	13.5
>$40,000-$50,000	7.8	9.8	12.2	6.0
>$50,000-$75,000	19.1	16.7	16.7	7.6
More than $75,000	38.9	30.3	35.3	8.3
Race				
White	64.4	76.4	77.8	77.2
Non-white	35.6	23.6	22.2	22.8
Gender				
Male	49.7	45.8	52.4	46.7
Female	50.5	54.3	47.6	53.3

Note: Disability and Urban-Rural are not available in the Pew dataset.
SNS = social network site; HS = high school.

The American data in Table 3 show patterns that are sometimes similar and sometimes different than the British. The age portion of the American table looks very similar to the British table. The modal users are the youngest respondents and Twitter use declines monotonically with age. In the education categories, in both countries highly educated people are more likely to be Twitter users. The modal Twitter user has some college in the US but is a secondary school graduate in Britain. The best explanation is that this reflects differences in the meaning of "further education" and "some college" in the two countries. In other words, it is the difference in the definition of the category. For marital status, the modal category in both countries is single, followed by married. The other differences reflect differences in the American and British populations more than differences in the way people of different marital status's use Twitter. In lifestage we see a large difference: Only about 3% of American students use Twitter compared to about 18% of British students. The lack of American student use is striking. Almost three-quarters of American employed respondents use Twitter compared to less than two-thirds of British employed persons. Income is another area where the two countries differ. Among the lowest income category, Americans show 30 percentage points difference between Twitter users and Internet non-users; low income British show over 50 percentage points difference. Among high income Americans, 30 percentage points separate Twitter users from Internet non-users, compared to about 10 percentage points in Britain. For race, nonwhites are 13 percentage points more likely to use Twitter in the US compared to nine percentage points in Britain. Finally, gender doesn't seem to matter in the U.S. data.

In summary, the big demographic differences between Twitter users and other groups are that, in both countries, Twitter users are more likely to be younger, better educated, students or employed, single, and wealthier. Specifically they are younger than other SNS users, who are in turn younger than other Internet users, who are younger than non-users. This monotonic relationship shows up in most of the strong relationships in the table. It shows up in all education

categories, in marital status, and in lifestage. But Twitter use is not identical in the countries: There are three notable differences. In America, students are much less likely to use Twitter. American non-whites are more likely to use Twitter and differences across income categories are smaller in the United States than in Britain.

Multivariate analyses

Many of the demographic variables in Tables 2 and 3 seem to be associated with Twitter use, at least at a zero-order level, but are they really all associated? Youth, being single and being a student are often related. Could we be seeing indirect effects of only one or two variables (like age), and not such a broad range of causal factors? We can answer this question by using a multivariate model to predict Twitter use. Table 4 shows the odds ratios from logistic regressions that use the nine British demographic variables and the seven American variables to predict whether or not a respondent uses Twitter. In the British data, only four demographic variables are statistically significant: age, income, education and lifestage. Age, entered as a continuous variable, shows that each year of age decreases the odds of being a Twitter user by about 6%. People with incomes of over £40,000 per year are 3-4 times more likely to be Twitter users. Secondary school graduates and university graduates are about twice as likely to be Twitter users. After controlling for age, education and income, retired people are about three times more likely to be Twitter users than students. None of the other five variables are statistically significant.

The American data are similar with respect to age and income. The odds of being a Twitter user decline by about 3% per year. People with high incomes are about twice as likely to be Twitter users compared to people with incomes under $20,000 per year. Americans differ in four variables. Education doesn’t seem to influence American Twitter use at all. Unemployed are about one-quarter as likely to be Twitter users as students. Married people are about half as likely to use Twitter as singles. Finally, non-whites are about 56% more likely to use Twitter

than whites. Hargittai's (2015, Table 1) multivariate analysis of Twitter using the same Pew dataset produces similar results, although difference in model specification make detailed comparisons difficult.

Table 4: Logistic Regressions Predicting Twitter Use
British and American Data

	British Odds Ratios		**American Odds Ratios**
Age	0.94***	Age	0.97***
Household income			
£12.5-£20,000	0.83	$20-$30,000	0.52
£20-£30,000	0.80	$30-$40,000	1.15
£30-£40,000	1.47	$40-$50,000	1.08
£40-£50,000	4.12***	$50-$75,000	1.78
£50,000 or more	2.81**	$75,000 or more	2.11**
Education			
Secondary graduate	2.33*	High school graduate	0.88
Further education	1.80	Some college	0.92
Higher ed. degree	2.27*	College degree	1.02
Lifestage			
Employed	1.87	Employed	0.35
Retired	3.22*	Retired	0.29
Unemployed	1.29	Unemployed	0.28*
Marital status			
Married	0.83	Married	0.57*
Living with partner	0.91	Living with partner	0.69
Divorced/Separated	1.14	Divorced/Separated	0.90
Widowed	0.61	Widow	0.89
Ethnicity	1.27	Race	1.56*
Gender	0.88	Gender	1.29
Disability	1.07	(not available)	
Urban-rural	1.24	(not available)	
Constant	1.26	Constant	2.02
N	1,298		1,518
McFadden's R^2	0.16		0.09

Note: Omitted categories. British: < £12,500, no qualifications, student, single, white, men, not disabled, rural. American: < $20,000, less than high school, student, single, white, men. Two variables, disability and urban-rural, are not available in the Pew dataset.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

The bottom line is that, while there are certainly similarities in the effects of age and income, there are considerable differences in terms of education, lifestage, marital status, and

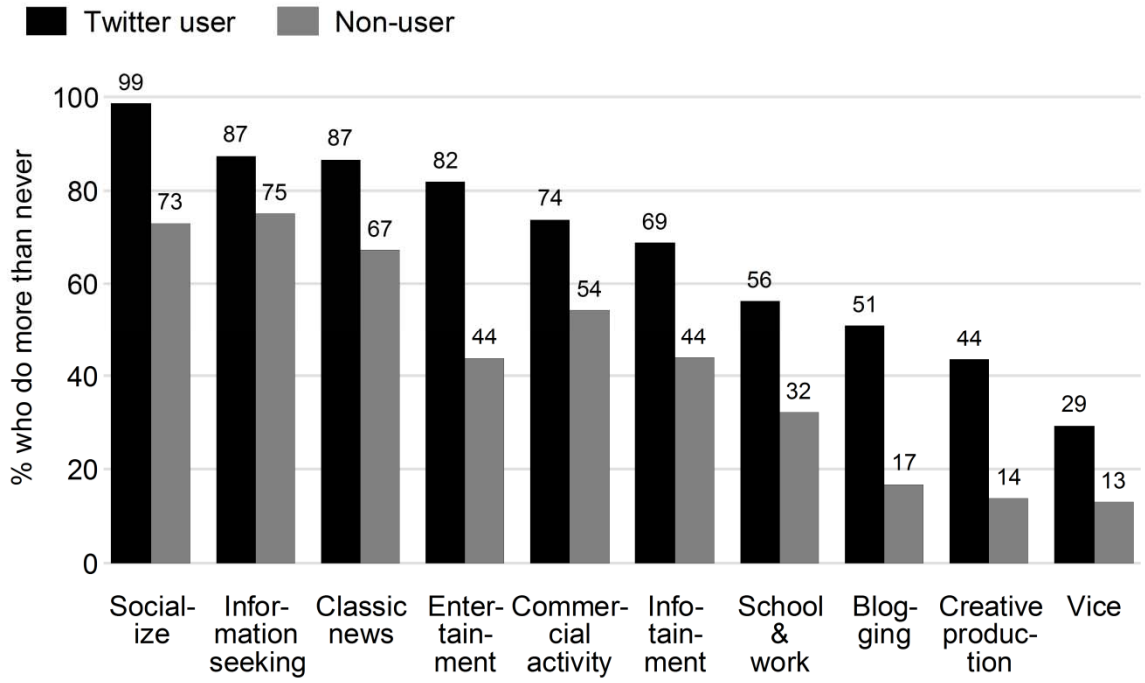
race. The differences in education, lifestage and race seem important and worth thinking about. In terms of who uses Twitter, these countries have different patterns of use.

The preceding pages describe differences between the Twitter users and others. What is the impact of these differences? The next two sections explore this question, comparing activities and attitudes of Twitter users to others.

Activities of Twitter users

Once people are online, they act in a variety of ways. Do Twitter users act differently from other users? Figure 1 compares Twitter users to non-users on 10 types of activities. From this point onward, we can no longer compare British and American Twitter use because this Pew dataset doesn't have any of the activity variables; these tables use only OxIS data. The striking result is how much more Twitter users do. In every activity, Twitter users are 12-38 percentage points more likely to participate than non-Twitter users. This gap extends across the entire range of the 10 activities, from the most frequent activity—socializing—to the least frequent—vice. The smallest differences are information seeking (12 percentage points) and vice (16 percentage points). The largest differences are in entertainment use of the Internet (38 percentage points), blogging (34 percentage points) and creative production (30 percentage points). It is notable that the biggest differences tend to be in the most challenging online activities; the blogging category includes not only blogging but maintaining a personal website, and creative production includes posting videos and posting anything the respondent considers 'creative'. The entertainment category is an exception. It is important to reiterate that Twitter use is not included in any of the 10 activity scales. Considering all the errors inherent in self-reported survey variables, this figure shows surprisingly large differences between Twitter users and non-users.

Figure 1: Activities of Twitter Users & Non-users

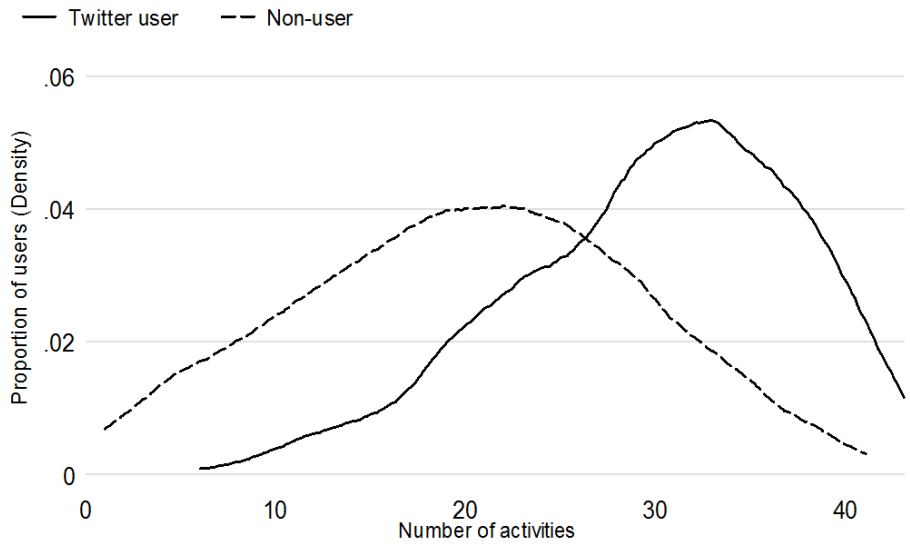


OxIS current users: 2013 N=1,613

Figure 2 shows the variety of Internet use by Twitter users and non-users. The figure displays a kernel density using an Epanechnikov kernel, based on the 43 activities included in the activities scales. It displays how many out of the 43 activities each user does more than never. You can readily see that Twitter users do a wider range of activities on the Internet than non-users. Twitter users do an average of 30 activities (median: 31) compared to 20 activities (median: 20) for non-users.

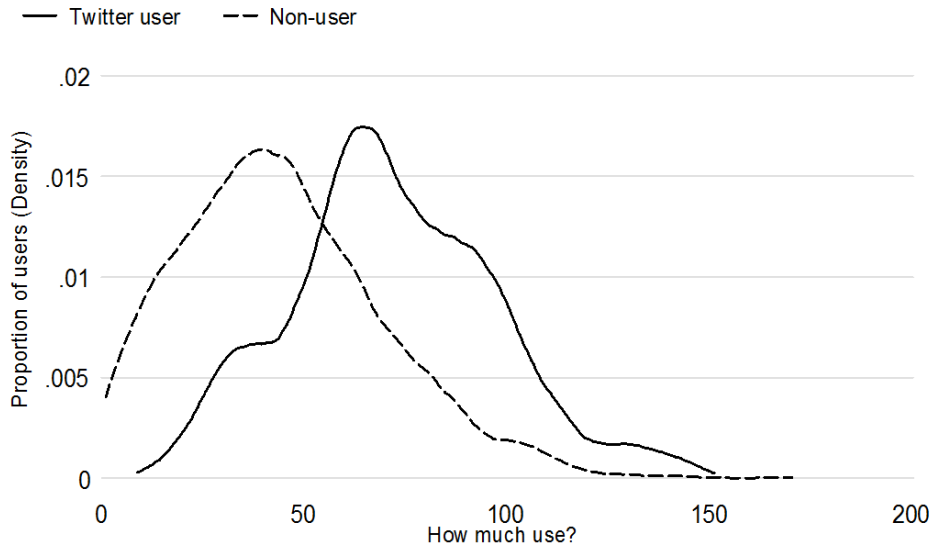
Figure 3 shows a kernel density graph of the amount of use by Twitter users and non-users. Amount of use is a continuous variable measuring the frequency of Internet use in day-to-day life, not the length of time someone has been using the Internet. Amount of use (often called "frequency of use") is a relevant property since Internet users can vary extensively in how much time they spend online – some people use it for many hours each day, others only once a week. By now the story is familiar: on a day-to-day basis Twitter users make much more intensive use of the Internet than non-users.

Figure 2: Variety of Internet use



OxIS current users: 2013 N=1,613

Figure 3: Amount of Internet Use



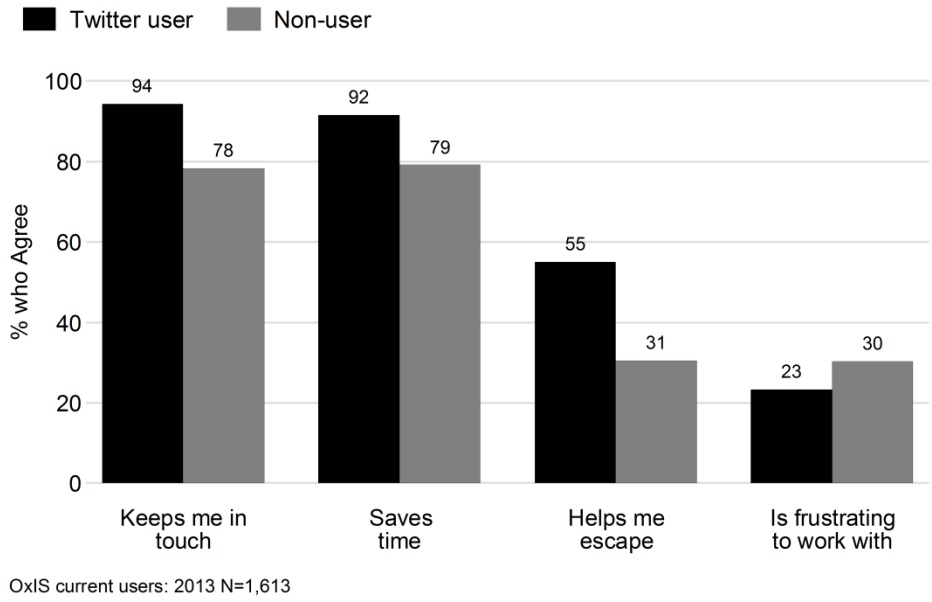
OxIS current users: 2013 N=1,613

Attitudes of Twitter users

Finally we can look at attitudes of Twitter users. There are some 48 attitude variables in the 2013 wave of OxIS, most of which show statistically significant differences between Twitter users and non-users. Figure 4 contains four variables selected to illustrate the range of variables on which there are differences. Twitter users are between 13 and 26 percentage points more

likely to agree that the Internet keeps them in touch with people, saves time and helps them escape. They are seven percentage points *less* likely to agree that the Internet is frustrating to work with. The overall pattern is that Twitter users have a much more positive attitude toward all aspects of the Internet than non-users.

Figure 4: Attitudes of Twitter Users & Non-users



OxIS current users: 2013 N=1,613

The digital divide is strong among Twitter users: Twitter users are unlike other groups in many ways. Furthermore, Twitter users in the UK and the US have different characteristics. What does this mean? We explore the implications of these differences for research using Twitter.

Discussion

The implications of the digital divide in Twitter use are different than for most digital divide research. Digital divide research has usually been concerned that large elements of the population are missing the benefits of being online. Since off-line people are more likely to be poor, uneducated, or elderly, the lack of access to Internet benefits reinforces their marginal

status. The digital divide in Twitter use is important because of what it implies for research based on Twitter data. Since Twitter has been so attractive for researchers it is important to ask, are the people being studied representative of any important population? Can researchers generalize from Twitter users to any important population?

Twitter data have many qualities that appeal to researchers. They are extraordinarily easy to collect. Furthermore, they are available in very large quantities; millions of tweets are not unusual. With a simple 140-character text limit and few options (hashtags, retweets, links) they are easy to analyze. As a result of these attractive qualities, over 1,400 papers have been published using Twitter data (Bruns & Weller, 2014). Easy availability of Twitter data links nicely to a key goal of computational social science. If researchers can find ways to impute user characteristics from social media, then the capabilities of computational social science would be greatly extended. Sloan et al. (2013) express the hope of these studies: "Twitter can be conceptualised as a 'digital agora' (Housley, Edwards, Williams, & Williams, 2013) that provides an insight into mass user generated opinions, sentiments and reactions to social events."

But, who are the mass users? Do Twitter users share identical characteristics with some population of interest? Twitter users in both the United Kingdom and the United States are a subset of Internet users, who are in turn a subset of the whole population. The Twitter subset is unrepresentative in terms of demographics, attitudes and engagement with the Internet. To the extent that Twitter users do not share the characteristics of a population, then inferences drawn from analysis of Twitter data are not evidence of population characteristics. Under these circumstances, any collection of tweets will be biased and inferences based on analysis of such Tweets will not match the population characteristics. This particular bias cannot be corrected by collecting more data. A biased sample remains biased regardless of how many millions of tweets are in the sample. Mislov et al. (2011, p. 557) summarize the problem "Most existing work does

not address the sampling bias, simply applying machine learning and data mining algorithms without an understanding of the Twitter user population."

The preceding analysis suggests both the limits and value of using Twitter as a data source for research on social and political events. Twitter users fit the profile of young, well-educated, wealthy elites. This means that research using Twitter data is not a proxy for research on the population as a whole or even the subset of the population that is online. Instead Twitter users reflect the interests, values, skills, priorities, and biases of elites.

This may not be a bad thing. Twitter use appears to be an emerging channel for transmission of elite influence. If the goal of a research project is to study how elites influence politics, culture, or society, then Twitter may be an excellent research site. Similarly, studies of democratisation or the public sphere could benefit from more work on how elites use their dominance of Twitter to enhance their influence.

There are additional reasons to question the representativeness of Twitter data. Prior research has raised questions about the effect of research access on representativeness. Boyd and Crawford (2013) point out that most researchers do not have access to all tweets (the "firehose") because it is too expensive. Instead they use one of the free subsets (the "garden hose" or the "spritzer"). The criteria for selection of tweets in either subset have never been made public.[2] Investigations into these subsets indicates that they are not random samples (Morstatter, Pfeffer, Liu & Carley, 2013), so researchers using these subsets cannot assume that they are getting a representative sample of the population of tweets. In any event, data available to researchers does not include private or "protected" tweets, and the proportion of protected tweets has never been made public. The problem, then, is that even if the characteristics of Twitter users are

[2] Twitter also makes a "sample hose" available, but whether it is an actual random sample it has never been independently verified.

known, most researchers do not have access to a representative collection of their tweets. This introduces a further bias of unknown size and unknown direction. Boyd and Crawford (2013, p. 669) summarize their discussion of difficulties of research access to a representative collection of tweets by saying "it is difficult for researchers to make claims about the quality of the data that they are analyzing."

Furthermore, among Twitter users rates of tweeting are extremely skewed. Forty-four percent of Twitter users are lurkers: they have never sent a tweet (Koh, 2014). Among those who have sent tweets, the top 1% of users account for 20% of tweets and the top 15% of users generate 85% of all tweets (Leetaru, Wang, Cao, Padmanabhan, & Shook, 2013). Leetaru et al. conclude that "a very small number of core users thus drive the majority of Twitter's traffic." So, not only are Twitter users unrepresentative but so are Tweets. Tweets are mostly representative of a small group of active users, and not representative of the Twitter user population (Mustafaraj, Finn, Whitlock, & Metaxas, 2011).

In addition to the tweets themselves, there are questions of how people use Twitter compared to other social media (cf. Tufekci, 2014). To the extent that people do not use Twitter like other social media—say, they tweet on different topics or in different situations—then Twitter data do not represent the same forms of use. Affordances, the characteristics of the platform that make some actions easy, some things difficult, and other things impossible, are the key point. Twitter has lightweight programming model, limited message length, and a simple interface. This tends to make it easy to use on devices with relatively restricted input capabilities and limited screens; that is, mobile phones. It is particularly suited for rapid, short messages. Twitter can be used for many purposes, on many devices, but it is unusually well suited for political demonstrations, sports events, crowds, disasters, emergencies and other live events where being an on-the-spot observer is valuable. In general, it is often a prominent communication method in mobile or low-bandwidth situations. Other social media, like

Facebook or Google+, have richer, more complex interfaces, much longer text, integration of photos, music, and graphics. These characteristics make them less suitable for smaller mobile devices, like phones. They also change the way people use them: Longer texts extend response times; complex interfaces make them more difficult to use on limited bandwidth devices, small devices or in mobile situations, photos require bigger displays to see detail. We can summarize this by saying that not only are Twitter *users* unrepresentative of any online or off-line population (other than themselves), but also the characteristics of the Twitter platform structure behaviour so that *usage* is skewed and unrepresentative of social media in general.

Is post-stratification weighting an answer?

One possible response to this research is the thought that, even though Twitter users are unrepresentative of the British or American population, it may be possible to construct post-stratification weights that would re-weight Twitter data to be representative of the population. This idea perishes when we consider the implications of comparing the British and American data.

To some extent American Twitter use is similar to British. American Twitter users are also young and wealthy. So far, they are the same. But American Twitter users are also non-white, unmarried and unemployed, none of which have any statistically significant counterpart in Britain. Furthermore, education, which is important in Britain, seems to have no influence in the United States. This suggests that American and British Twitter users are different enough that they cannot simply be thought of as interchangeable. There are national differences in Twitter use.

National differences are broader than the United States and the United Kingdom. A six-country study by Wilkinson and Thelwall (2012, p. 1638) found that “types of topics vary by country… both in terms of differences in the types of topics discussed and in terms of the specific topics discussed.” This means that any re-weighting would have to take into account the

nationality of the person tweeting. Since few tweets are geocoded, it is impossible for researchers to take into account the source of tweets. Only in special situations, where the subject of the tweets is mostly of local interest can will tweets be known. There are studies, like Dubois' (2013) study of a local issue in Canadian politics, where this is possible. In most cases, the country where the user is located is unknown so country-based reweighting is impossible.

Inability to determine the geographic source of tweets introduces further noise into the signal given by Twitter. Tufekci (2014) describes the problem using Egypt. She points out that in cases where the topic is of global interest (think: Arab Spring) Twitter could generate more tweets from outside Egypt than from inside the country. Under those circumstances tweets will be some combination of local and global commentary; they will not be a good measure of local activity or local interest. To the extent that tweets from around the globe focus on topics that are different than the local Twitter focus, analysis could lead to misleading and incorrect conclusions. Since there is no reliable way to know the source of tweets, researchers using Twitter will be unaware that their conclusions are wrong.

Predicting Elections: polling versus Twitter

The unrepresentative characteristics of Twitter users and tweets explain some of the results of Twitter research. For example, forecasting elections using Twitter data has well-known problems. To understand why Twitter cannot forecast elections, it is helpful to compare a successful method of predicting an election, public opinion polling, with use of Twitter. Polling benefits from decades of research into sampling, language and questionnaire construction. Pollsters have control over how the sampling is conducted, although they are often constrained

by costs and a perceived need for speed.[3] Because of government-collected census microdata, they know a great deal about who they are sampling. Well known theories explain wording and question order effects (Zaller 1992). A set of techniques, including cognitive interviews, focus groups and pretests, are available to help clarify meanings of different wording among different subgroups in the population. Such tools give pollsters the ability to write items where the language has the same meaning for everyone from teenagers to 80-year-olds and where meanings are stable across gender, ethnic, religious, and racial boundaries. By controlling both sampling and questions, polls have control over much of the *process* that produces their data.

Now contrast pollsters' control with the situation of researchers using Twitter. Twitter researchers have no control over wording or language. Since the location of almost all tweets is unknown, there is no equivalent to the use of census microdata and researchers do not know the age, gender, ethnicity or race of the writer. Since there is no way to contact the Twitter user, researchers cannot request important information that is not spontaneously supplied. The data generating process for Twitter is out of the control of researchers. Unlike a sample survey, Twitter data is created spontaneously by users; it is not designed from the ground up to produce valid, reliable data for scientific purposes. Twitter researchers must make largely untestable assumptions about their ability to map the relationship between off-line behavior and online texts, while using convenience samples. As the analyses above show, Twitter users have different attitudes than non-users. Election predictions, for example, have to confront the volatility of issues, party competition, and changes in the legal rules from one election to another. Furthermore, elections provide incentives for partisans to engage in strategic behavior

[3] The effort to produce cheap, fast polls has sometimes led pollsters to cut corners on samples, producing unrepresentative samples and inaccurate predictions. This appears to be the source of the erroneous predictions of the 2015 British General Election (Sturgis et al. 2016).

to attempt to manipulate online sentiment, which means what they write may contain false or misleading information. This mix of instabilities and perverse incentives in the political system suggest that elections may never be predictable using social media data. Huberly (2015) comprehensively reviews attempts to forecast elections using Twitter (and other social media). His conclusion is blunt: "All known forecasting methods based on social media have failed when subjected to the demands of true forward-looking electoral forecasting" (p. 992).

Successful Twitter use

There are examples of successful uses of Twitter to forecast off-line behavior; for example, film box office gross (Asur & Huberman, 2010; Hennig-Thurau, Wiertz & Feldhaus, 2015), book sales (Gruhl, Guha, Kuman, Novak, & Tomkins, 2005), and music sales (Frick, Tsekouras & Li, 2014). It is notable that these are all commercial products. Commercial products face a relatively standardized, stable environment in comparison to elections. Furthermore, they are often consumed by the age cohort where Twitter use is high: young people. The goal of finding entertainment is often less complex than decisions about who to vote for so making a decision based on the sentiment in a tweet makes more sense. The lack of representativeness is less important. If people who see the tweet are influenced, the extent of the influence can be measured. The fact that tweets are most likely to be seen by elites would be important only for products consumed by people in non-elite categories. There are few or no incentives for users to create strategic tweets that don't reflect their true personal opinions. Commercial predictions are typically very short term, for example opening weekend gross sales (Hennig-Thurau et al., 2015) or next day album sales (Frick et al., 2014). Unlike elections, shifting sentiments over weeks or months need not figure into the methodology. What this means is that researchers need to make fewer untestable assumptions and the link between online and off-line behavior is stronger. Finally, consider the goal of the prediction. The prediction errors for individual products may not matter as long as the aggregate prediction is accurate.

It is important to note that these issues apply not just to tweets, but also to attempts to infer content. Sentiment analysis can be done, just as long as the sentiments do not need to be representative of any larger population. Even if researchers learn how to impute user characteristics from the content of tweets—something that no one can now do—the value of this achievement would be severely hamstrung by the fact that Twitter users are not representative of any population other than themselves. The characteristics of successful predictions suggest a conclusion: Twitter may be much more valuable for corporate use than for social research.

Not all scholarly research requires representative samples. For example, Twitter can be used as a coordination medium or as a means of exerting influence on events, even if other media are also used. Studying Twitter influence compared to Whatsapp influence or Facebook Messenger influence or television influence, etc could be very interesting. In situations like this, Twitter data could be a valuable research site.

The dominant reason to focus on Twitter has often been that the data are so easy to collect. Using other social media like Whatsapp or Facebook as a research site is hampered by the difficulty of collecting the data (in the case of Whatsapp) or the complexity of the data (in the case of Facebook). This seems to be a case where researchers, enamored with making their research easier, have often prioritized easy data collection and only later looked for substantive problems to which the data could be attached. For substantive issues requiring data representative of a population, Twitter data unlikely to be appropriate.

This paper points toward several new lines of research. The simplicity of Twitter will continue to appeal to researchers, consequently we can expect to continue to see a large stream of research. There are many opportunities in this research. Twitter users tend to be elites and they can be a valuable window into elite activities. The affordances of Twitter foster its use during live, breaking events. It is one of the few ways of gathering research data about those events as they happen. In these and other areas Twitter offers unusual opportunities. The payoff

from this research will be stronger if the empirical and methodological basis for research using Twitter is firmer. The differences between the United States and the United Kingdom raise many questions. America and Britain are similar in many ways, including the percentage of their populations that are online. It is not at all clear why Twitter use should be different. Why do so few American students use Twitter? Do they not use short messages or have they found a functional replacement in instant messaging (IM) or text messages? If this is true, then why don't British students use IM or text messages in the same way? Why do married Americans use Twitter so much less than singles? Are they again using some functional replacement or are they not communicating via micro-blogs? If they are not, then why not? Answers to questions like these are crucial if we are to understand who tweets, and to reach reliable conclusions based on Twitter data.

Author Information

Grant Blank (Ph.D. University of Chicago) is the Survey Research Fellow at the Oxford Internet Institute, University of Oxford, United Kingdom. He is a sociologist specializing in the political and social impact of computers and the Internet, the digital divide, statistical and qualitative methods, and cultural sociology. He is currently working on a project asking how are cultural hierarchies understood in online reviews of cultural attractions. His other project links sample survey data with census data to generate small area estimates of Internet use in Great Britain. He can be reached at grant.blank@oii.ox.ac.uk; see also http://www.oii.ox.ac.uk/people/blank/.

References

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. In Proceedings of the IEE/WIC/ *ACM International Conference on Web Intelligence and Intelligent Agent Technology* (pp. 492-499). doi:10.1109/WI-IAT.2010.

Blank, G. & Groselj, D. (2014). Dimensions of Internet use: Amount, variety and types. *Information, Communication & Society*, 17, 417-435. doi: 10.1080/1369118X.2014.889189

Bonfadelli, H. (2002). The Internet and knowledge gaps: A theoretical and empirical investigation. *European Journal of Communication*, 17, 65–84. doi: 10.1177/0267323102017001607

boyd, d. & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication, & Society*, 5(15), 662-679. doi: 10.1080/1369118X.2012.678878.

Bruns, A., & Weller, K. (2014) Twitter data analytics or the pleasures and perils of studying Twitter. *Aslib Journal of Information Management*, 66. doi:10.1108/AJIM-02-2014-0027

Chang, J., Rosenn, I., Backstrom, L., Marlow, C. (2010). EPluribus: Ethnicity on social networks. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. pages 18-25. Washington, DC. May 23–26, 2010.

Dubois, E. (2013). *Telling Vic Everything: Digital Contention and the Traditional Media*. Paper presented at the Annual meeting of the American Sociological Association. New York, USA.

Duggan, M. & Brenner, J. (2013). *The Demographics of Social Media Users — 2012*. Pew Research Center. Retrieved from http://pewinternet.org/Reports/2013/Social-media-users.aspx.

Duggan, M. & Smith, A., (2014, January). *Social Media Update 2013*. Pew Research Center. Retrieved from http://pewinternet.org/Reports/2013/Social-Media-Update.aspx

Dutton, W. H., & Blank, G. (2011). *Next Generation Users: The Internet in Britain, 2011*. Oxford: Oxford Internet Institute. Retrieved from http://oxis.oii.ox.ac.uk/reports

Dutton, W. H., & Blank, G. with Groselj, D. (2013). *Cultures of the Internet: The Internet in Britain, 2013*. Oxford: Oxford Internet Institute. Retrieved from http://oxis.oii.ox.ac.uk/reports

Frick, T., Tsekauras, D., & Li, T. (2014). The times they are a-changin: Examining the impact of social media on music album sales and piracy. Paper presented at the Annual Meeting of the Academy of Management, Philadelphia, PA.

Gruhl, D., Guha, R., Kumar, R., Novak, J. and Tomkins, A. (2005), The predictive power of online chatter, *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 78-87. doi: 10.1145/1081870.1081883

Hargittai, E. (2008). Whose space? Differences among users and non-users of social network sites. *Journal of Computer-Mediated Communication*, 13 (1), 276–297. doi:10.1111/j.1083-6101.2007.00396.x.

Hargittai, E., & Litt, E. (2012). Becoming a tweep. *Information, Communication & Society*, 15, 680–702.

Hargittai, E. (2015). Is bigger always better? Potential biases of big data derived from social network sites. *Annals of the American Academy of Political and Social Science*, 659, 63-76. DOI: 10.1177/0002716215570866.

Hennig-Thurau, T., Wiertz, C., & Feldhaus, F. (2015). Does Twitter matter? The impact of microblogging word of mouth on consumer's adoption of new movies. Journal of the Academy of Marketing Science, 43, 375–394. doi:10.1007/s11747-014-0388-3

Housley, W. Edwards, A. Williams, M. & Williams, M. (Eds.) (2013). Special Issue, Computational Social Science: Research, Design and Methods, *International Journal of Social Research Methods*, 16(3).

Huberly, M. (2015). Can we vote with our tweet? On the perennial difficulty of election forecasting with social media. *International Journal of Forecasting*, 31(3), 992-1007.

Kalampokis, E., Tamouris, E., & Tarabanis, K. (2013) Understanding the predictive power of social media. *Internet Research,* 23 (5), 544-559.

Koh, Y. (2014). Report: 44% of Twitter accounts have never sent a tweet. *Wall Street Journal*. 11 April. Retrieved from http://blogs.wsj.com/digits/2014/04/11/new-data-quantifies-dearth-of-tweeters-on-twitter.

Leetaru, K.H., Wang, S., Cao, G., Padmanabhan, A. & Shook, E. (2013) Mapping the global heartbeat: The geography of Twitter. *First Monday*. 18 (5). 6 May.

Michael, J. (2007). *40000 Namen, Anredebestimmung anhand des Vornamens*. Retrieved from http://www.heise.de/ct/ftp/07/17/182/ (in German).

Mislove A., Lehmann S., Ahn Y., Onnela J., Rosenquist J. (2011). Understanding the demographics of Twitter users. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. July 17-21, 2011; Barcelona. (pp. 554 – 557). Palo Alto, CA: AAAI.

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 '13 Proceedings of the 7th International AAAI Conference on Weblogs and Social Media.* (pp. 400-408).

Mustafaraj, E., Finn, S., Whitlock, C. and Metaxas, P. (2011), "Vocal minority versus silent majority: discovering the opinions of the long tail", *Proceedings of IEEE PASSAT/ SocialCom*, pp. 103-110.

Pennacchiotti, M., & Popescu, A.-M. (2011, August 21-24,). Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (p. 430-438). San Diego, CA. doi:10.1145/2020408.2020477

Pew Research Center. (2014). *The Web at 25*. Retrieved from http://www.pewinternet.org/2014/02/25/the-web-at-25-in-the-u-s

Pew Research Center. (2013, May). Online Dating (Prelim). Retrieved from http://www.pewinternet.org/datasets/may-2013-online-dating-prelim/.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010, October 30). Classifying Latent User Attributes in Twitter. *Proceedings of SMUC2010: 2nd International Workshop on Search and Mining User-generated Contents*. Toronto, Ontario. Pp 37-44. Doi:10.1145/1871985.1871993

Sloan, L., Morgan, J., Housley, W., Burnap, P. & Williams, M. (2015) Who tweets? Deriving the demographic characteristics of age, occupation & social class from Twitter user meta-data. *PLOS One*, 10(3), e0115545. DOI:10.1371/journal.pone.0115545

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P. and Rana, O. (2013). Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter. *Sociological Research Online*, 18(3), 7. Retrieved from http://www.socresonline.org.uk/18/3/7.html. Doi: 10.5153/sro.3001

Sturgis, P. Baker, N. Callegaro, M. Fisher, S. Green, J. Jennings, W. Kuha, J. Lauderdale, B. and Smith, P. (2016) *Report of the Inquiry into the 2015 British general election opinion polls,* London: Market Research Society and British Polling Council.

Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. Forthcoming in *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*.

Van Deursen, A.J.A.M & van Dijk, J.A.G.M. (2013) Digital divide shifts to differences in usage. *New Media & Society*. 16: 507-526. DOI: 10.1177/1461444813487959

van Dijk, J. A. G. M. (2005). *The deepening divide: Inequality in the information society*. Thousand Oaks, CA: Sage.

Van Dijk, J.A.G.M. (2005). *The deepening divide: Inequality in the information society*. London: Sage.

Wells, T. & Link, M. (2014) Facebook user research using a probability-based sample and behavioral data. *Journal of Computer-Mediated Communication*. 19: 1042-1052. doi:10.1111/jcc4.12058.

Wilkinson, D., & Thelwall, M. (2012) Trending Twitter topics in English: An international comparison. *Journal of the American Society for Information Science and Technology*, 63(8),1631–1646. DOI: 10.1002/asi.22713

Williams, S. A., Terras, M. M., & Warwick, C. (2013a) What do people study when they study Twitter? Classifying Twitter related academic papers. *Journal of Documentation*, 69(3), 384-410.

Zaller, John. (1992) *Nature and origins of mass opinion*. New York: Cambridge University Press.