Check for updates

Article

Potential Biases in Big Data: Omitted Voices on Social Media

Social Science Computer Review
1-15
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0894439318788322
journals.sagepub.com/home/ssc

Eszter Hargittai[1]

Abstract

While big data offer exciting opportunities to address questions about social behavior, studies must not abandon traditionally important considerations of social science research such as data representativeness and sampling biases. Many big data studies rely on traces of people's behavior on social media platforms such as opinions expressed through Twitter posts. How representative are such data? Whose voices are most likely to show up on such sites? Analyzing survey data about a national sample of American adults' social network site usage, this article examines what user characteristics are associated with the adoption of such sites. Findings suggest that several sociodemographic factors relate to who adopts such sites. Those of higher socioeconomic status are more likely to be on several platforms suggesting that big data derived from social media tend to oversample the views of more privileged people. Additionally, Internet skills are related to using such sites, again showing that opinions visible on these sites do not represent all types of people equally. The article cautions against relying on content from such sites as the sole basis of data to avoid disproportionately ignoring the perspectives of the less privileged. Whether business interests or policy considerations, it is important that decisions that concern the whole population are not based on the results of analyses that favor the opinions of those who are already better off.

Keywords

big data, data bias, sampling, sampling bias, survey, social media, Facebook, Twitter

Much enthusiasm has accompanied the massive amounts of data readily available about people's opinions and behavior, that is, "big data," with some going so far as to claim "the end of theory" (Anderson, 2008), although most focusing on the opportunities without rejecting the scientific enterprise (Bail, 2014; Goldberg, 2015). The many promising avenues for new studies of the social world not notwithstanding, the enthusiasm has been accompanied by literature raising critical questions about big data ranging from ethical to privacy considerations and beyond (e.g., Baym, 2013; boyd & Crawford, 2012; Gitelman, 2013; Lazer et al., 2009; Neff, 2013). One limitation of much such work is that most of it is based on hypothetical cases rather than being grounded in empirical research (Hidalgo, 2014, elaborates on this point). This article empirically tackles a

[1] University of Zurich, Zurich, Switzerland

Corresponding Author:
Eszter Hargittai, University of Zurich, Andreasstrasse 15 IKMZ, Zurich 8050, Switzerland.
Email: pubs@webuse.org

significant question concerning the ethics of big data: Who is most likely to be excluded from data sets often used as the basis of big data studies? In other words, whose voices, opinions, and behavior are the least likely to be reflected in certain types of big data that often make up the basis of big data studies?

Many big data studies rely on content collected from social network sites such as Facebook and Twitter even when the questions they are asking are not about the use of such sites (e.g., Asur & Huberman, 2010; Bakshy, Messing, & Adamic, 2015; O'Connor, Balasubramanyan, Routledge, & Smith, 2010; Schwartz et al., 2013). While some research has questioned the viability of such approaches and has pointed out their shortcomings (Gayo-Avello, 2013; Hargittai, 2015), generally speaking, little critique focuses on the core question of sampling bias. When studies ask questions about the larger population rather than focusing on platform-based behavior, knowing whether site users are representative of the population is important. Relying on social media data to generalize opinions and behaviors to the larger population assumes that people select into the use of such sites randomly. Yet scholarship has shown this not to be the case both generally for social network sites (Haight, Quan-Haase, & Corbett, 2014) and regarding specific ones (Blank, 2016; boyd, 2011; Hargittai, 2007, 2015; Hargittai & Litt, 2012). Missing from prior work is analysis on national data that include not just multiple platforms but also measures of Internet skills, a variable that digital inequality research has identified as crucial in how people are incorporating the Internet into their lives (Litt, 2013). It is this gap in the literature that this article fills.

What Do We Know About Who Uses Social Network Sites?

There have been a few attempts to establish the representativeness of specific social network site users, although the studies are often not framed as addressing that question. The earliest such work was by boyd (2007) and Hargittai (2007) looking at the differences in which youth adopted MySpace versus Facebook. These researchers found racial, ethnic, and socioeconomic differences in adoption, boyd through qualitative methods, Hargittai through survey data analysis. Their subsequent work confirmed those initial findings (boyd, 2011; Hargittai, 2011) showing that these differences persisted over time. Once Twitter started diffusing to the population, analyzing panel data of a group of diverse young adults, Hargittai and Litt (2011) showed that its adoption varied by race and ethnicity as well as Internet skills.

In the early years of social media, few studies were able to report on such findings as most data collected about social media users either did not specify platforms (e.g., the Pew Research Center did not start disaggregating by site until 2012), focused on certain undergraduate student populations where most were users of the platform making comparison with nonusers difficult (Ellison, Steinfield, & Lampe, 2007), or only collected data about users of a particular platform without data about nonusers (Ahn, 2013). With the rise of different social network sites and work showing that they function quite differently (Papacharissi, 2009; van Dijck, 2013), fortunately more and more data-collection efforts started disaggregating between social network sites rather than assuming that they were interchangeable. With such data, it was then possible to conduct more analyses of how different population groups were represented on each platform. For example, drawing on U.S. national data from the Pew Research Center, Hargittai (2015) showed that those with more education and higher income were more likely to be on Facebook, LinkedIn as well as Twitter than the less privileged. Analyzing panel data about a group of young adults, she also found that higher Internet skills from 3 years prior were related to Twitter, LinkedIn, and Tumblr use. She was not able to consider the role of Internet skills for social media adoption in a national sample, however, lacking an appropriate data set.

In a welcome extension of previous studies that had all focused on U.S. users, Blank and Lutz (2017) examined social media use by user background in the UK also finding unequal adoption.

Although there were no differences by educational level, higher income was associated with a higher likelihood of using LinkedIn and Twitter. Because that paper does not report on results from models without the Internet experience variables included, it is hard to know what the sociodemographic relationships are like without taking those variables into consideration. That study included both self-efficacy and skill measures finding that one or the other mattered for all sites' adoption. Koiranen and Räsänen (2017) examined social media platform adoption among Finnish Internet users and also found socioeconomic differences. These findings from the UK and Finland suggest that variation in adoption by user characteristics is not restricted to the U.S. raising concerns about big data derived from social media in varying national contexts.

Other types of research have also shed light on potential biases that stem from basing data collection on particular platforms. Stern, Bilgen, McClain, and Hunscher (2017) recruited survey participants through placing ads on both Facebook and Google. They found that those who sign up through Facebook are demographically different from those who sign up through a more inclusive mode (in that study's case, Google). Those who took the survey through the Facebook recruitment mode were more likely to be White, higher educated, and with a higher income than those who were recruited through Google (and also compared to the general population). This approach is a helpful alternative to surveying people about their social media usage in establishing potential biases stemming from different platform users and suggests that survey researchers must also be careful about what platforms they use to recruit respondents even in the case of platforms that are used by a significant portion of the population.

An additional challenge of studies that rely on big data derived from social network sites is that often they are not even representative of the particular site's users and content. For example, numerous studies based on Twitter data do not include all potentially relevant tweets, rather, they select on ones that include hashtags, which are a specific feature of the platform. Yet, the few studies that have reported on what percentage of tweets include hashtags have found this to be a feature of a minority of posts ranging from 5% to 18% depending on language (14% in English and 11% on average for all languages examined; Hong, Convertino, & Chi, 2011). Another study found it to be 10% of 74 million tweets, but a much larger percent (21%) of retweets, which is worth mentioning as that is yet again a type of Twitter post that is disproportionately represented in big data studies derived from Twitter (Suh, Hong, Pirolli, & Chi, 2010). In all such cases, a considerable portion of tweets would be ignored were a researcher to use hashtags to sample on tweets. In a more recent study, Rafail (2017) examined the prevalence of specific hashtags among tweets addressing a particular topic. Even in such a focused corpus of Twitter data, he found that less than a third of posts contained a topic-specific hashtag, in many cases considerably fewer. Since hashtag use is likely a reflection of skilled Twitter use (i.e., knowing what it is and using it is a type of skill), authors of tweets that include a hashtag may well not be representative of the average Twitter user adding another layer of potential sampling bias to such data sets.

Another way that researchers may end up with a nonrepresentative sample of social network site users is to sample on a subgroup of users of the site. Such is the case with studies that have come out of the myPersonality app connected to Facebook, which advertises itself as an app that reveals details to users about their personality (Schwartz et al., 2013). It is reasonable to assume that people who are inclined to download and partake in the services of such an app are not representative on personality measures of the average Facebook user. Accordingly, it may well be problematic to use it as a sampling frame for studies that concern personality traits.

While some of the authors of the piece drawing on the myPersonality app have gone on to claim that the personality measures of those who use their app are not different from other samples (Rife, Cate, Kosinski, & Stillwell, 2016), their comparison samples have serious limitations making those claims questionable. They compared the traits of the myPersonality app users with undergraduate students and data from Web users in 1999 and 2000, that is, from over a decade earlier than when

they collected their own data. While the authors are right to note that much research in the field of psychology is based on undergraduate students, it is important to recognize that such a group is quite WEIRD (i.e., "Western, Educated, Industrialized, Rich, and Democratic" as per Henrich, Heine, & Norenzayan, 2010) and thus not at all representative of the larger population. Taking Web users from 1999 and 2000 as their baseline comparison is also problematic since those online at that time and taking surveys through the Web are hardly representative of today's Web users or the American population as a whole (Horrigan, 2000; Lenhart, 2000). It turns out that one need not go so far as comparisons to the general population to detect potential pitfalls of relying on specific sites when studying personality. A study comparing related measures of people recruited through Facebook and Twitter found personality differences by site questioning the validity of data based on myPersonality app users all drawn from Facebook when it comes to generalizations of their findings to the population (Hughes, Rowe, Batey, & Lee, 2012).

Given that existing work has already established variations by user background in who shows up on social network sites, what can this piece add to the conversation? First, as various social media continue to gain popularity, there is value in revisiting the relationships between socio-demographics and social media use to see whether the relationships persist over time. Also, while Blank and Lutz (2017) asked about skills, they did so using a suboptimal measure so revisiting the relationship of Internet skills to social network site adoption is worthwhile. Blank and Lutz (2017) relied on a question that asked respondents: "How would you rate your ability to use the Internet," which is a type of measure of skills that research has shown to be a bias-prone way of asking about skills (Hargittai & Shafer, 2006) especially when it comes to gender differences in responses. The present study sidesteps this bias by using a less direct and more detailed skills measure (Hargittai, 2005; Wasserman & Richmond-Abbott, 2005). Additionally, the data set here differs as it includes information about Americans' race and ethnicity, variables that prior work has shown to matter in the U.S. context. Also, none of the cited papers have considered variations in Reddit adoption, a site that has also served as the basis of both public conversations about social behavior (regarding political topics and harassment; Massanari, 2017) as well as big data studies (e.g., Mills, 2017).

Data and Method

To examine whether people select into the use of social network sites randomly, I analyze survey data from a national sample of U.S. adults (18 years old or over) collected in summer 2016. The survey instrument incorporates detailed measures of individuals' background attributes and Internet experiences and skills, with information about their use of various social media platforms.

Data Collection

The survey was administered through the independent research organization National Opinions Research Center (NORC) at the University of Chicago (NORC subsequently) using its AmeriSpeak panel online. The panel is representative of the U.S. population using "area probability sampling and includes additional coverage of hard-to-survey population segments such as rural and low-income households that are underrepresented in surveys relying on address-based sampling" (National Opinion Research Center, 2017). After pretesting the survey with 23 respondents and updating it based on the results in early May 2016, the survey ran on May 25 to July 5, 2016. The instrument included an attention-check question and only responses from participants who passed this question are included in the data set. In total, valid responses exist for 1,512 American adults 18 and over, which constitutes a 37.8% survey response rate.

Measures: Independent Variables

Demographic and socioeconomic factors. Background variables about respondents such as their age, gender, education, income, and race or ethnicity were supplied by NORC based on their earlier data collection about the AmeriSpeak panel. Here, I describe what coding I used for these measures. I report age as a continuous variable. I created three education categories: high school or less, some college, and college degree or more. Income was reported in 18 categories, which I recoded to their midpoint values to make it a continuous variable. In the regression analyses, I use the square root of income as this transformation produces a distribution much closer to normal. Race and ethnicity are dummy variables for White, Hispanic, African American, Asian American, Native American, and Other. There is a dummy variable for those employed either full time or part time. There is also a dummy variable signaling rural residence.

Internet experiences and skills. Following prior literature, I include measures for how long people have been Internet users, how much autonomy they have in accessing the Internet when and where they want, how much time they spend online, and their Internet skills. The instrument asked respondents when they first started using the Internet offering the following answer options with their recoded values in parentheses: *within the past year* (1), *1 to 5 years ago* (2.5), *more than 5, but less than 10 years ago* (7.5), and *10 or more years ago* (12.5). To measure autonomy of use, the survey asked "At which of these locations do you have access to the Internet, that is, if you wanted to you could use the Internet at which of these locations?" followed by nine options including home, workplace, and friend's home. The following question assessed frequency of use: "On an average weekday, not counting time spent on e-mail, chat, and phone calls, about how many hours do you spend visiting Web sites?", which was also asked for the "average Saturday or Sunday." The answer options ranged from *none* to *6 hr or more* with six options in between. I calculated weekly hours spent on the Web by multiplying the answer to the first question by five, the second question by two, and summing the two figures.

For measuring Internet skills, I use a validated, established index (Hargittai & Hsieh, 2012; Wasserman & Richmond-Abbott, 2005). Respondents were presented with 13 Internet-related terms (such as tagging, PDF, and spyware) and were asked to rank their level of understanding of these items on a five-point scale ranging from *no understanding* to *full understanding*. I then calculate the mean for all items as the Internet skills measure (Cronbach's $\alpha = .94$).

Measures: Dependent Variables

Early in the survey, we asked respondents: "Have you ever heard of the following sites and services?" with various social network sites on the list. The majority of respondents reported having heard of the sites. In this article, I focus on five social network sites that all include considerable text: Facebook, LinkedIn, Twitter, Tumblr, and Reddit. I exclude social media platforms such as Instagram, Pinterest, and Snapchat despite their popularity as they are mainly image-based and perhaps for this reason have not tended to be the basis of nearly as many studies as the five platforms listed above.

The least known site of the ones explored here was Reddit at 54% followed by Tumblr at 71% then LinkedIn at 86%. Almost everyone had heard of Twitter (99%) and Facebook (a few respondents short of 100%). We then asked the people who had heard of the respective sites:

> Have you ever visited the following sites and services? For each site, indicate if no, you have never visited it; yes, you have visited it in the past, but do not visit it nowadays; yes, you currently visit it sometimes; yes, you currently visit it often.

Table 1. Sample Descriptives.

	Percent	*M*	*SD*	*N*
Background				
Age (18–94)		48.74	16.87	1,512
Income in US$ 1,000 s (2.5–225)		71.48	54.40	1,512
Female	51			1,512
Employed	62			1,512
Rural resident	13			1,512
Education				
High school or less	26			1,512
Some college	32			1,512
Bachelor's or higher	43			1,512
Race and ethnicity				
White	71			1,511
Hispanic	12			1,511
Black	11			1,511
Asian	3			1,511
Native American	2			1,511
Other	1			1,511
Internet experiences				
Internet autonomy (0–9)		4.80	2.28	1,512
Internet use frequency (0–42)		14.75	10.75	1,491
Years of Internet use (1–12.5)		11.11	2.78	1,512
Internet skills (1–5)		3.37	1.08	1,511

Current use of a social network site is the recode of the answer to this question so that people who responded "yes, you currently visit it sometimes" or "yes, you currently visit it often" are included in the user category.

The Sample

Table 1 reports summary statistics for all of the measures. There is close to equal representation of women (51%) and men. The average age is 48.7 years. The majority are White (71%) followed by Hispanics (12%), African Americans (11%), Asian Americans (3%), Native Americans (2%), and people who reported "Other" races (1%). The median income is US$55,000, the mean income is US$71,478. Just over a quarter of the group (26%) has no more than a high school education, just under a third (32%) has some college education, and 43% has a college degree or more. Sixty-two percent are employed either full time or part time, 13% live in a rural area. In sum, while a diverse sample, it is more educated and has a higher income than the average American, as is the case with Internet users generally (Pew Research Center, 2017).

Regarding online experiences, very few of the respondents are new to the Internet, not surprisingly given that Internet use statistics have plateaued in the United States in recent years (Pew Research Center, 2017). The average participant has been using the Internet for over 10 years. The median number of access locations is 5, the mean is 4.8. The median number of hours participants spend online weekly is 12 hr, the mean is 14.8. The skill measure ranges from 1 to 5, the median is 3.4 as is the mean (standard deviation: 1.1), showing that respondents vary considerably in their online know-how.

Table 2. Use of Social Network Sites by User Background (Percentage of Full Sample).

	Facebook	LinkedIn	Twitter	Tumblr	Reddit
Age (18–34)	90***	35	39***	24***	26***
Age (35–50)	82	39	36***	10	14
Age (51–62)	79	36	23*	7*	5***
Age (63–94)	67***	23***	11***	3***	2***
Men	72***	35	27	13	16***
Women	87***	32	27	10	8***
White	79	33	26	10	12
Hispanic	84	32	29	14	12
Black	84	35	36*	12	12
High school or less	77	15***	20***	8	6***
Some college	83	28*	28	13	15
College or more	79	48***	31*	12	14
Income LQ	85***	22***	24	12	11
Income HQ	78	49***	32*	10	14
Currently working	82**	41***	33***	14***	15***
Currently not working	76**	21***	19***	8***	8***
Urban or suburban resident	84	36***	28	12	13**
Rural resident	79	17***	24	8	6**
Internet experiences and skills					
Autonomy of use LQ	73***	23***	19***	6***	5***
Autonomy of use HQ	89***	50***	43***	21***	25***
Frequency of use LQ	72***	25***	15***	5***	5***
Frequency of use HQ	89***	35	43***	21***	21***
Number of use years (LQ)	74*	16***	19***	8	5***
Number of use years (HQ)	81*	38***	30***	12	14***
Internet skills LQ	69***	11***	9***	3***	2***
Internet skills HQ	88***	53***	46***	24***	31***

Note: Asterisks signify statistically significant differences between the subgroups of the category: *$p < .01$. **$p < .005$. ***$p < .001$. LQ = lowest quartile; HQ = highest quartile.

Analyses

First, I present bivariate statistics to show how background characteristics and online experiences relate to the use of various social network sites. Then, I show results from logistic regression analyses to establish what factors explain the use of each individual social network site while holding all other variables constant. There are no issues of multicollinearity in the data set (the correlation matrix is available from the author).

Who Uses Various Social Media?

First, I discuss the binary relationship of social network site usage with user characteristics and then report on the results of logistical regression analyses. Table 2 shows the relationship of all socio-demographic and Internet experience variables for the five social media platforms. The asterisks denote significance at the *$p < .01$, **$p < .005$, and ***$p < .001$ levels. A quick look at the table shows that unlike findings from the past, there is little variation by race and ethnicity in site adoption. (The table excludes figures for Asian Americans, Native Americans, and those reporting Other races as these groups make up very small percentages of the data set.) The only finding for race and ethnicity is the higher likelihood of African Americans on Twitter, something that has been the case for many years (Hargittai & Litt, 2011).

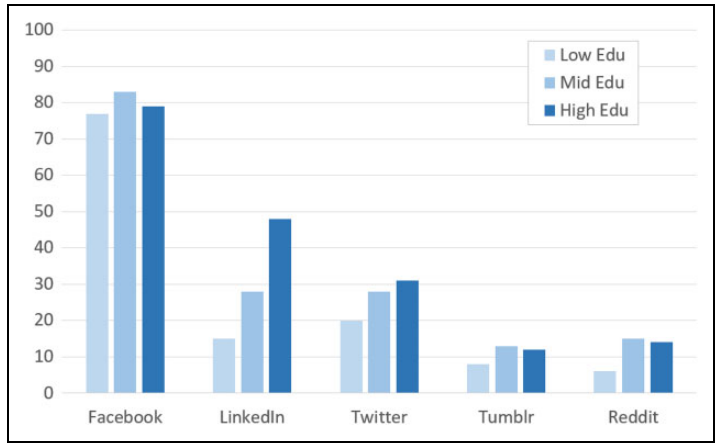

Figure 1. Use of social network sites by level of education.

Regarding age, those 18–34 (the lowest quartile of the age distribution) are more likely to use most platforms other than LinkedIn. The middle age ranges do not show significant differences for Facebook and LinkedIn, but the oldest age category (63–94) is the least likely to be on all such sites. Gender differences exist on two platforms: Facebook and Reddit. Women are considerably more likely to be on the former while men are considerably more likely to use the latter. Differences by level of education are evident for LinkedIn, Twitter, and Reddit, but not for Facebook and Tumblr. There are also no income differences for either Tumblr or Reddit, but there are for Facebook, where those making less are more likely to be on the site. The relationship of income to LinkedIn and Twitter is the reverse; those who make more are significantly more likely to use these sites. Those who are employed are more likely to be on all five platforms. Those in rural areas are significantly less likely to use LinkedIn and Reddit.

The bottom part of Table 2 shows the relationship of Internet experiences and skills with social media platform adoption. In most cases, autonomy of use, frequency of use, number of years of use all show differences for site adoption. The most significant relationship of site adoption to an independent variable is Internet skills. The differences between those who scored in the lowest quartile versus the highest quartile of the skills distribution are considerable in all cases.

Figures 1 and 2 show the relationship of education and social network site use and the relationship of Internet skills and social network site use, respectively. This visual representation is especially helpful for conveying the magnitude of difference by these two variables in who is and who is not present on the various social media platforms.

Next, I report the results of logistic regression analyses to see what relationships persist when controlling for other factors. First, I look at sociodemographic variables only and then I add the Internet experiences and skills variables. Tables 3 and 4 present these results (split solely due to space layout issues). When it comes to being on Facebook, both age and gender remain significant even when controlling for other factors. Namely, younger people and women are considerably more likely to be on this platform. No other sociodemographic factors matter. When controlling for Internet experiences and skills, both age and gender remain significant, but we also see variation by autonomy of use, use frequency, and skills, all of which are positively related to the likelihood of using Facebook.

Other than LinkedIn, older people are less likely to be on the social media platforms examined here than younger people, although it is important to keep in mind the figures presented in Table 2 to recognize that these findings are likely driven by the oldest age category rather than those in middle

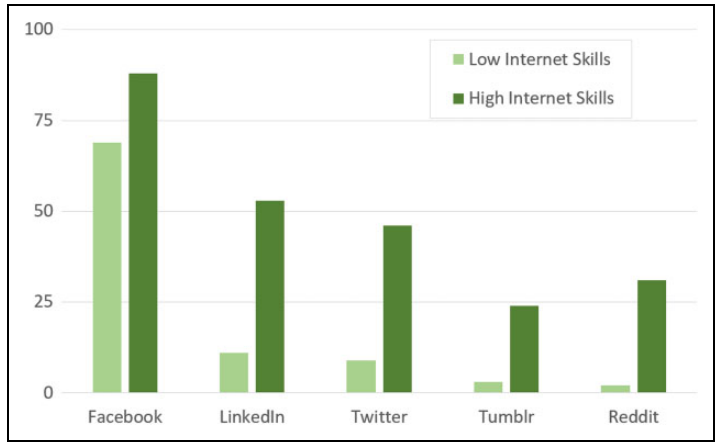

Figure 2. Use of social network sites by Internet skills (lowest quartile vs. highest quartile).

Table 3. Logistic Regression on Using Social Network Sites (Odds Ratios with Standard Errors in Parentheses).[c]

	Facebook	Facebook	LinkedIn	LinkedIn	Twitter	Twitter
Age	0.97 (0.00)***	0.98 (0.01)*	0.99 (0.00)	1.01 (0.00)	0.96 (0.00)***	0.98 (0.00)***
Female	2.49 (0.35)***	2.72 (0.40)***	0.95 (0.11)	1.07 (0.14)	1.00 (0.12)	1.14 (0.15)
Hispanic[a]	0.87 (0.20)	1.03 (0.25)	0.85 (0.16)	1.05 (0.22)	0.79 (0.15)	0.86 (0.18)
Black[a]	1.07 (0.25)	1.04 (0.25)	1.18 (0.23)	1.20 (0.25)	1.44 (0.27)	1.32 (0.26)
Asian American[a]	0.47 (0.16)	0.46 (0.17)	0.73 (0.24)	0.81 (0.28)	0.87 (0.29)	0.87 (0.30)
Native American[a]	1.06 (0.61)	1.11 (0.65)	0.71 (0.35)	0.91 (0.47)	0.18 (0.14)	0.17 (0.13)
Other race[a]	0.47 (0.29)	0.42 (0.27)	3.26 (2.04)	3.68 (2.53)	1.27 (0.81)	0.98 (0.68)
High school or less[b]	0.65 (0.11)	0.89 (0.17)	0.24 (0.04)***	0.35 (0.07)***	0.60 (0.10)**	0.83 (0.15)
Some college[b]	1.17 (0.20)	1.26 (0.22)	0.53 (0.07)***	0.56 (0.08)***	0.94 (0.14)	1.00 (0.15)
Income (square root)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)***	1.00 (0.00)***	1.00 (0.00)**	1.00 (0.00)**
Employed	0.96 (0.15)	0.91 (0.15)	2.02 (0.29)***	1.84 (0.28)***	1.26 (0.18)	1.24 (0.19)
Rural resident	1.50 (0.32)	1.66 (0.37)	0.48 (0.10)***	0.53 (0.11)**	0.97 (0.18)	1.09 (0.22)
Autonomy of use		1.10 (0.04)**		1.08 (0.03)*		1.05 (0.03)
Frequency of use		1.03 (0.01)***		1.00 (0.01)		1.04 (0.01)***
Number of use years		1.03 (0.03)		1.06 (0.03)		1.00 (0.03)
Internet skills		1.27 (0.10)**		1.92 (0.15)***		1.65 (0.13)***
Intercept	15.16 (5.89)	0.95 (0.55)	0.45 (0.14)	0.01 (0.00)	1.36 (0.43)	0.04 (0.02)
N	1,505	1,484	1,497	1,476	1,583	1,482
Pseudo R^2	.083	.118	.122	.188	.083	.142

[a]The omitted category is White. [b]The omitted category is college or more. [c]*$p < .01$. **$p < .005$. ***$p < .001$.

Table 4. Logistic Regression on Using Social Network Sites (Odds Ratios with Standard Errors Are in Parentheses).[c]

	Tumblr	Tumblr	Reddit	Reddit
Age	0.95 (0.01)***	0.97 (0.01)***	0.92 (0.01)***	0.94 (0.01)***
Female	0.65 (0.11)	0.76 (0.14)	0.38 (0.07)***	0.48 (0.09)***
Hispanic[a]	0.94 (0.24)	0.98 (0.27)	0.59 (0.16)	0.58 (0.18)
Black[a]	0.89 (0.24)	0.81 (0.23)	0.80 (0.22)	0.72 (0.22)
Asian American[a]	0.82 (0.38)	0.94 (0.45)	1.02 (0.42)	1.19 (0.51)
Native American[a]	1.50 (0.88)	1.59 (1.00)	1.40 (0.86)	1.69 (1.14)
Other race[a]	7.08 (4.45)**	6.41 (4.45)**	0.66 (0.73)	0.50 (0.55)
High school or less[b]	0.58 (0.14)	0.85 (0.23)	0.37 (0.10)***	0.61 (0.18)
Some college[b]	0.98 (0.20)	1.10 (0.23)	1.06 (0.21)	1.25 (0.27)
Income (square root)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
Employed	1.09 (0.23)	1.05 (0.23)	0.80 (0.17)	0.74 (0.17)
Rural resident	0.76 (0.22)	0.87 (0.26)	0.46 (0.15)	0.47 (0.17)
Autonomy of use		1.08 (0.05)		1.07 (0.05)
Frequency of use		1.03 (0.01)***		1.02 (0.01)**
Number of use years		0.98 (0.04)		1.04 (0.05)
Internet skills		2.12 (0.26)***		3.00 (0.42)***
Intercept	1.71 (0.71)	0.02 (0.02)	6.09 (2.56)	0.01 (0.01)
N	1,498	1,478	1,489	1,469
Pseudo R^2	.112	.181	.211	.305

[a]The omitted category is White. [b]The omitted category is college or more. [c]*$p < .01$. **$p < .005$. ***$p < .001$.

ages. While there is no gender variation in LinkedIn, Twitter, and Tumblr adoption, the findings about women being much more likely to use Facebook and much less likely to use Reddit hold even when controlling for other variables. As with the binary analyses presented in Table 2, the regression results confirm that there are no racial or ethnic differences in the adoption of these sites. Worth noting that once we control for other factors, African Americans are no longer more likely to be on Twitter than others so the binary finding regarding that relationship does not hold up.

Those with no more than a high school education are less likely to use several sites compared to those with a college degree or more, namely, LinkedIn, Twitter, and Reddit. However, only in the case of LinkedIn does this relationship hold once we include Internet experiences and skills in the model. It may well be that differentiated skills that are very much related to education are the reason fewer less educated people are on certain social media platforms. Income matters similarly to education, there is a difference for LinkedIn and Twitter, but not the other sites. Employment status and rural residence only matter for LinkedIn use. It is worth noting that less than 5% of the sample is "not working, looking for work," and comparing this group to the rest of the sample shows that such people are more likely to use LinkedIn than others who are in the "not working" category, although less likely than those who are employed.

Looking at Internet experiences, number of user years does not make a difference for adoption of any of the sites, not surprisingly given that most users had been online for numerous years at the time of the survey. Autonomy of use matters for Facebook and LinkedIn use. Frequency of use is related to all, but not LinkedIn use.

The binary relationships observed with Internet skills hold in the regression analyses even once we control for sociodemographics and online experiences. For all sites, those who are more skilled about using the Internet are more likely to use them. It is the inclusion of Internet experiences and skills in the models that removes the statistical significance of education for two of the three sites

where education matters: Twitter and Reddit use. Only with LinkedIn does education persist as an important variable in explaining differential rates of usage once we consider those factors. Overall, there are several variables that matter to who adopts what social media platforms. Their uses are not random across the population and they attract different types of people.

Discussion and Conclusion

Scholarship has seen a proliferation of articles based on data about social behavior derived from social network sites (in lieu of offering an endless list of citations, I invite the reader to run a search on Google Scholar for examples using search phrases like >Twitter data<). Many of these studies do not limit their research questions to social media use, rather, they ask questions about social behavior more generally. By doing so without acknowledging the biases that go into who selects into the use of such sites in the first place, they simply assume that the various sites' users are representative of the larger population (whether, that is the larger Internet-user population or beyond). But they do so erroneously given the findings presented in this article and elsewhere (e.g., Blank & Lutz, 2017; Hargittai, 2015; Stern et al., 2017), showing that the users of these sites bias toward those who are more educated and more skilled at using the Internet.

Analyzing national survey data of American adults' Internet uses collected in summer 2016, this article shows that there is large variation in people's experiences with social media. First, it is important to note that other than Facebook, most such sites are used by only a minority of Internet users. Second, selection into their uses is far from random. People from lower socioeconomic status are less likely to be on several such sites and Internet skills are an important correlate of use with higher skilled users much more likely to show up on all such platforms. Given that the data are cross-sectional, it may be that using social media improves people's general Internet skills and that is why we see that correlation. It is worth noting, however, that prior work on panel data has shown that general Internet skills from earlier years explain differential rates of social network site adoption years later (Hargittai, 2015). Additionally, the Internet skills measure is not based on social media skills, so the two measures are not too closely tied.

The findings about nonrandom selection into the use of social media platforms suggest that the opinions and behavioral traces of the more privileged are more likely to be represented in data sets that use social media as their sampling frames than the views and actions of the less privileged. Consequently, studies that rely on such research design must be explicit about the limitations of their findings' generalizability, that is, they must acknowledge that the opinions and behaviors they uncover represent only certain parts of the population.

Why does it matter that big data derived from social media bias against certain types of people? Increasingly, businesses and governments alike may use data derived from social network site users to make decisions about how to communicate information about their products and services, that is, how to reach constituents. If they are basing their decisions on data that underrepresent the less privileged then their resulting actions may not meet the needs of people from different backgrounds equally. During the winter holidays of 2010, then-mayor of Newark, New Jersey (USA), Cory Booker received considerable positive press coverage for using Twitter to reach people in the city whose streets had not been plowed of snow (e.g., Gregory, 2010). While this approach likely helped some stranded behind walls of snow, it was precisely the people who likely needed help the most who were the least likely to be on that medium in the first place. Older adults with less ability to dig themselves out and those with fewer resources to hire snow plows were then and still are today among the least likely users of Twitter. Using a medium whose users skew toward the more privileged to make decisions about the allocation of scarce resources is problematic as it is likely to exacerbate social inequalities.

What is to be done, then, given the appeal of using social media–based big data? Too often when asked why a researcher relied on social media data, the response is along the lines of "it was easily available." On its own, this is rarely a sound scientific justification. Yes, social media data are available and have the potential to shed light on issues that may otherwise be difficult to study. Nonetheless, recognizing the limitations of such data is essential and researchers should discuss explicitly what the lack of representativeness of social media users implies for their findings. Additionally, use of different methods in a single study can be helpful and offer a much-needed complementary perspective. The present study itself has the limitation of relying on only one method, but alternatives are possible. For example, Sánchez, Craglia, and Bregt (2017) use both Twitter data and survey data of the general population to look at the prevalence of people contacting political officials. In the article, they discuss the pros and cons of both methods including an explicit mention of the fact that Twitter data are not representative of the population.

The goal of this article is not to dismiss all social media–based data sets. Rather, it is to call for an understanding of and reflection about the biases that such data represent. When answering questions about social behavior more generally, analyses of big data derived from social network sites must discuss explicitly what the biases may mean for their findings. A simple throwaway comment is not enough as the implications may be considerable. Ideally, such studies will increasingly include a triangulation of methodologies (Sloan & Quan-Haase, 2017) so that research can be conscious of whose voices are represented in big data being analyzed and whose voices are left behind.

Author's Note

The author is grateful to the editors and reviewers for their helpful comments on the article. She would like to thank Aaron Shaw for working together on the survey on which this study draws. She is grateful to Merck (Merck is known as MSD outside the United States and Canada) and the Robert and Kaye Hiatt Fund at Northwestern University for supporting the data collection. The data set in this article is available from the author by writing to pubs@webuse.org

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was financially supported by Merck (Merck is known as MSD outside the United States and Canada) and the Robert and Kaye Hiatt Fund at Northwestern University as well as time made available through the April McClain-Delaney and John Delaney Professorship at Northwestern University to conduct some of this work.

References

Ahn, J. (2013). *What can we learn from Facebook activity? Using social learning analytics to observe new media literacy skills*. Presented at the LAK'13—Third International Conference on Learning Analytics and Knowledge (pp. 135–144). Leuven, Belgium: ACM. Retrieved from https://doi.org/10.1145/2460296.2460323

Anderson, C. (2008). *The end of theory: The data deluge makes the scientific method obsolete. Wired*. June 23. Retrieved from https://www.wired.com/2008/06/pb-theory/

Asur, S., & Huberman, B. A. (2010) *Predicting the future with social media* Presented at the Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 1914092 (pp. 492–499). IEEE Computer Society. Retrieved from https://doi.org/10.1109/wi-iat.2010.63

Bail, C. A. (2014). The Cultural Environment: Measuring culture with big data. *Theory and Society*, *43*, 465–482.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*, 1130–1132. Retrieved from https://doi.org/10.1126/science.aaa1160

Baym, N. K. (2013). Data not seen: The uses and shortcomings of social media metrics. *First Monday*, *18*. Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/4873/3752

Blank, G. (2016). The digital divide among Twitter users and its implications for social research. *Social Science Computer Review*. Retrieved from https://doi.org/10.1177/0894439316671698

Blank, G., & Lutz, C. (2017). Representativeness of social media in Great Britain: Investigating Facebook, LinkedIn, Twitter, Pinterest, Google+, and Instagram. *American Behavioral Scientist*, *61*, 741–756. Retrieved from https://doi.org/10.1177/0002764217717559

boyd, d. (2007). Viewing American class divisions through Facebook and MySpace. Retrieved from http://www.danah.org/papers/essays/ClassDivisions.html

boyd, d. (2011). White flight in networked publics? How race and class shaped American teen engagement with MySpace and Facebook. In L. Nakamura & P. Chow-White (Eds.), *Race after the Internet* (pp. 203–222). Abingdon, England: Routledge.

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*, 662–679. Retrieved from https://doi.org/10.1080/1369118X.2012.678878

Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, *12*, 1143–1168. Retrieved from https://doi.org/10.1111/j.1083-6101.2007.00367.x

Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review*, *31*, 649–679. Retrieved from https://doi.org/10.1177/0894439313493979

Gitelman, L. (Ed.) (2013). *"Raw data" is an oxymoron*. Cambridge, MA: MIT Press.

Goldberg, A. (2015). In defense of forensic social science. *Big Data & Society*, *2*, 1–3. Retrieved from https://doi.org/10.1177/2053951715601145

Gregory, S. (2010, December 29). Cory Booker: The mayor of Twitter and Blizzard superhero. *Time*. Retrieved from http://content.time.com/time/nation/article/0,8599,2039945,00.html

Haight, M., Quan-Haase, A., & Corbett, B. A. (2014). Revisiting the digital divide in Canada: The impact of demographic factors on access to the Internet, level of online activity, and social networking site usage. *Information, Communication & Society*, *17*, 503–519. Retrieved from https://doi.org/10.1080/1369118X.2014.891633

Hargittai, E. (2005). Survey measures of web-oriented digital literacy. *Social Science Computer Review*, *23*, 371–379. Retrieved from https://doi.org/10.1177/0894439305275911

Hargittai, E. (2007). Whose space? Differences among users and non-users of social network sites. *Journal of Computer-Mediated Communication*, *13*, 276–297. Retrieved from https://doi.org/10.1111/j.1083-6101.2007.00396.x

Hargittai, E. (2011). Open doors, closed spaces? Differentiated adoption of social network sites by user background. In Lisa Nakamura & P. Chow-White (Eds.), *Race after the Internet* (pp. 223–245). Abingdon, England: Routledge.

Hargittai, E. (2015). Is bigger always better? Potential biases of big data derived from social network sites. *The ANNALS of the American Academy of Political and Social Science*, *659*, 63–76. Retrieved from https://doi.org/10.1177/0002716215570866

Hargittai, E., & Hsieh, Y. P. (2012). Succinct survey measures of web-use skills. *Social Science Computer Review*, *30*, 95–107. Retrieved from https://doi.org/10.1177/0894439310397146

Hargittai, E., & Litt, E. (2011). The tweet smell of celebrity success: Explaining variation in Twitter adoption among a diverse group of young adults. *New Media & Society*, *13*, 824–842. Retrieved from https://doi.org/10.1177/1461444811405805

Hargittai, E., & Litt, E. (2012). Becoming a tweep: How prior online experiences influence Twitter use. *Information, Communication and Society*, *15*, 680–702. Retrieved from https://doi.org/10.1080/1369118x.2012.666256

Hargittai, E., & Shafer, S. (2006). Differences in actual and perceived online skills: The role of gender. *Social Science Quarterly*, *87*, 432–448. Retrieved from https://doi.org/10.1111/j.1540-6237.2006.00389.x

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–135. Retrieved from https://doi.org/10.1017/S0140525X0999152X

Hidalgo, C. A. (2014, April 29). Saving big data from big mouths. *Scientific American*. Retrieved from https://www.scientificamerican.com/article/saving-big-data-from-big-mouths/

Hong, L., Convertino, G., & Chi, E. H. (2011). Language matters in Twitter: A large scale study. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. AAAI. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2856/3250

Horrigan, J. (2000). *New Internet users: What they do online, what they don't, and implications for the "net's" future*. Washington, DC: Pew Internet and American Life Project.

Hughes, D. J., Rowe, M., Batey, M., & Lee, A. (2012). A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior*, *28*, 561–569. Retrieved from https://doi.org/10.1016/j.chb.2011.11.001

Koiranen, I., & Räsänen, P. (2017). *Are the social media use patterns changing? Findings from Finland*. Presented at the European Symposium Series on Societal Challenges in Computational Social Science, London, United Kingdom.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., … Alstyne, M. V. (2009). Computational social science. *Science*, *323*, 721–723. Retrieved from https://doi.org/10.1126/science.1167742

Lenhart, A. (2000). Who's not online: 57% of those without Internet access say they do not plan to log on. Washington, DC: Pew Research Center.

Litt, E. (2013). Measuring users' Internet skills: A review of past assessments and a look toward the future. *New Media & Society*, *15*, 612–630. Retrieved from https://doi.org/10.1177/1461444813475424

Massanari, A. (2017). #Gamergate and the Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, *19*, 329–346. Retrieved from https://doi.org/10.1177/1461444815608807

Mills, R. A. (2017). Pop-up political advocacy communities on reddit.com: SandersForPresident and the Donald. *AI & SOCIETY*, 1–16. Retrieved from https://doi.org/10.1007/s00146-017-0712-9

National Opinion Research Center. (2017). "AmeriSpeak: NORC's Breakthrough Panel-Based Research Platform. Retrieved from http://www.norc.org/Research/Capabilities/pages/amerispeak.aspx

Neff, G. (2013). Why big data won't cure us. *Big Data*, *1*, 117–123. Retrieved from https://doi.org/10.1089/big.2013.0029

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (Vol. *11*, pp. 122–129). Washington, DC: The AAAI Press. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1536/1842/

Papacharissi, Z. (2009). The virtual geographies of social networks: A comparative analysis of Facebook, LinkedIn and ASmallWorld. *New Media & Society*, *11*, 199–220. Retrieved from https://doi.org/10.1177/1461444808099577

Pew Research Center. 2017. *Internet/Broadband Fact Sheet*. Washington, D.C. Retrieved from http://www.pewinternet.org/fact-sheet/internet-broadband/

Rafail, P. (2017). Nonprobability sampling and Twitter: Strategies for semibounded and bounded populations. *Social Science Computer Review*. Retrieved from https://doi.org/10.1177/0894439317709431

Rife, S. C., Cate, K. L., Kosinski, M., & Stillwell, D. (2016). Participant recruitment and data collection through Facebook: The role of personality factors. *International Journal of Social Research Methodology*, *19*, 69–83. Retrieved from https://doi.org/10.1080/13645579.2014.957069

Sánchez, C. R., Craglia, M., & Bregt, A. K. (2017). New data sources for social indicators: The case study of contacting politicians by Twitter. *International Journal of Digital Earth*, *10*, 829–845. Retrieved from https://doi.org/10.1080/17538947.2016.1259361

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., . . . Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS One*, *8*, e73791. Retrieved from https://doi.org/10.1371/journal.pone.0073791

Sloan, L., & Quan-Haase, A. (2017). A retrospective on state of the art social media research methods: Ethical decisions, big-small data rivalries and the spectre of the 6Vs. In L. Sloan & A. Quan (Eds.), *The handbook of social media research methods* (pp. 662–672). London, England: Sage.

Stern, M. J., Bilgen, I., McClain, C., & Hunscher, B. (2017). Effective sampling from social media sites and search engines for web surveys: Demographic and data quality differences in surveys of Google and Facebook users. *Social Science Computer Review*, *35*, 713–732. Retrieved from https://doi.org/10.1177/0894439316683344

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010). *Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network*. Presented at the IEEE International Conference on Social Computing, Minneapolis, Minnesota, USA.

van Dijck, J. (2013). *The culture of connectivity: A critical history of social media*. Oxford, England: Oxford University Press.

Wasserman, I. M., & Richmond-Abbott, M. (2005). Gender and the Internet: Causes of variation in access, level, and scope of use. *Social Science Quarterly*, *86*, 252–270. Retrieved from https://doi.org/10.1111/j.0038-4941.2005.00301.x

Author Biography

Eszter Hargittai (PhD, sociology, Princeton) is a professor and chair of Internet Use and Society at the Institute of Communication and Media Research of the University of Zurich. She has published several articles over the years about biases in social media adoption, starting with differences among the users of MySpace and Facebook. She is working on a database of survey questions concerning the study of people's Internet uses. She is editing the *Handbook of Digital Inequality*. Email: pubs@webuse.org