Study Designs for Quantitative Social Science Research Using Social Media

Leticia Bode[1], Pamela Davis-Kean[2], Lisa Singh[1], Tanya Berger-Wolf[3], Ceren Budak[2], Guangqing Chi[4], Andy Guess[5], Jennifer Hill[6], Adam Hughes[7], Brad Jensen[1], Frauke Kreuter[8], Jonathan Ladd[1], Margaret Little[1], Zeina Mneimneh[2], Kevin Munger[5], Josh Pasek[2], Trivellore Raghunathan[2], Rebecca Ryan[1], Stuart Soroka[2], and Michael Traugott[2]

July 1, 2020

[1] Georgetown University
[2] University of Michigan
[3] The Ohio State University
[4] Pennsylvania State University
[5] Princeton University
[6] New York University
[7] Pew Research Center
[8] University of Maryland

This series of white papers is funded by National Science Foundation awards #1934925 and #1934494 as a collaboration between Georgetown University and University of Michigan. To learn more about The Future of Quantitative Research in Social Science research project, visit www.smrconverge.org.

1. Overview

In 2019, a group of computer scientists and social scientists began a project focused on the convergence of these two areas, hoping to yield insight into using data from social media to improve our understanding of human behavior. Social scientists use multiple methods for collecting data but rely most commonly on either self-report measures or observations that are designed to correspond to specific research questions and require significant intervention on the part of researchers. At the same time, however, individuals have started using social media, search engines, smart devices, and other technologies that record their moment-to-moment behaviors (digital traces). Unlike self-reports, these organic forms of data have no set structure and are not designed to answer specific hypotheses or questions; instead, the data are provided "as is," which often means they are not well-oriented for most statistical software packages and are both complex and highly dense in nature. Moreover, these data have unique privacy and bias aspects not typically taken into account by traditional statistical approaches. Beyond these differences in design and intervention, digital trace data can offer value for social scientific questions because they provide a window into social and behavioral phenomena that may not be accessible with other methods (e.g. because individuals often lack the self-awareness to respond accurately). Due to the magnitude and complexity of the data, methods are needed for managing and making sense of these data, and these methods are most commonly found in the toolkits of computer scientists.

Employing methods from computer science to wrangle digital trace data to answer social scientific questions presents a number of challenges. First, it is often unclear who, exactly, produced the data that appear in social media datasets and how that set of individuals might relate to the larger groups of people that social science researchers would like to understand. For many questions, social media data contain information about the presence of a behavior, but do not have the necessary covariates to understand why that behavior may have occurred--which is a key area of focus for many social scientists. Further, ethical questions around the use of digital trace data in research contexts is a nascent field that may require collaboration across these disciplines. Understanding what we need to know about the data that are gathered, what other data are needed to supplement them to answer central research questions, and how to do so responsibly is critical for trace data to live up to their full potential. Thus, bringing together social scientists who try to understand human behavior and computer scientists who design and deploy algorithms to solve computational problems is a necessary convergence of two research realms.

Currently, the medium where we are obtaining the highest amount of data on human behavior and so see the most "convergence" between these disciplines is social media. Social media provides a rich amount of data on the everyday lives, opinions, thoughts, beliefs and behaviors of individuals and organizations in near real time. Leveraging these data effectively and responsibly should therefore improve our ability to understand political, psychological, economic, and sociological behaviors and opinions across time. For computer scientists, designing algorithms with a new set of constraints and optimizing existing algorithms for this large-scale, real time domain, while addressing privacy, bias, and algorithmic fairness concerns, will not only advance computer science research, but can also help to make their data and models more usable by social science researchers and more applicable to societal challenges.

In order to start the important conversations between social scientists and computer scientists, a set of topical meetings have been planned that will assemble these groups with the goal of creating consensus for how social and computer science could converge to answer important questions about complex human behaviors and dynamics. These topics include study design; data acquisition and sampling; data curation and measurement; model construction; analyses and storytelling; and criteria for the responsible conduct of research with social media data. This is the first in a series of white papers that will provide a summary of the discussions and future directions that are derived from these topical meetings.

2. Convergence in Research Design

There are many ways to begin thinking about research design. We chose to identify four common research designs, and then looked at how they might be applied to social media data. Obviously, there are other ways we could have organized this discussion - based on the goal of the research (causality, generalizability, description, inference, change over time, etc.), or the units of analysis used (individuals, groups, dyads, networks, etc.), or many other dimensions.

We chose to focus on common research designs for two main reasons. First, we thought this would allow everyone to start on the same page - irrespective of discipline, researchers are likely familiar with the general approaches we outline below. Second, social science is a very broad umbrella, with dozens of research designs in common use. We could not include everything, so this was a way to effectively narrow the scope and highlight key points of convergence and challenges in doing so. For these reasons, the designs we consider here are not meant to be comprehensive. Instead this document aims to integrate many disciplinary perspectives about common designs into a single framework. Over time, we hope this framework will be used to also rethink other traditional study designs and invent new ones in the context of studies involving social media, as well as the use of various designs for other types of digital trace data.

It is important to recognize the cultural differences that exist across disciplines. One area where there are significant differences is the way researchers across disciplines design their studies or experiments. In computer science, research designs vary significantly across different branches of computer science. Within theoretical computer science, researchers study the design and analysis of computation itself. They formulate a detailed task/problem description and prove the existence or nonexistence of efficient algorithms under different conditions and constraints. Some of the core areas of theoretical computer science include algorithm design, complexity theory, information theory, coding theory and cryptography. Here, the study design focuses on formulating the computational process and the constraints precisely so that the efficiency of the computation and limits can be proven mathematically. This branch of computer science does not typically work on experiments with data, yet do inform the feasibility of approximate solutions when ideal approaches are computationally intractable.

The other main branch of computer science is applied computer science. This branch focuses on the practical implementations of theoretical principles and designs of computing. Because applied computer science spans a large number of areas –security, operating systems, networking, databases, artificial intelligence, human computer interaction, programming languages, and software engineering to name a few – study design and experimental design vary considerably. In system's design research, the system must be designed and implemented to satisfy an explicit set of requirements before more traditional experiments are run to evaluate the accuracy, performance, and usability of the system. For a scholar studying security, a study design may be centered on developing an algorithm to observe the world, e.g. traffic across a network, or to simulate different events or conditions to understand their impact, e.g. an attack that brings down web servers. In areas like human computer interaction, the study designs are similar to those of social scientists. What differs are the questions computer scientists are trying to answer, and their fundamental approach to answering the question -- this approach typically involves developing an algorithm or a system and testing it to understand its accuracy, efficiency, robustness, and possibly usability under different conditions. Some studies are designed with explicit hypotheses, but many are not. More recently, computational data science and artificial intelligence are the applied computer science fields that are increasingly used in the social sciences and, similar to the social sciences, the goal is to produce a model of the observed world as represented by data.

While the goal of the computational data-driven models are typically descriptive and predictive, the social sciences models are typically explanatory and causal in nature and tend to be hypothesis driven. Ideally, a researcher articulates a question of interest, and subsequently decides on a method and designs a study to answer that question. The different approaches between social science and computer science motivate much of the discussion below, which we divide into three basic research designs that are commonly used in social science and are applicable to research using social media data: *qualitative observation*, *experiments*, and *surveys.* We also discuss a fourth design that is primarily informed by computer science, *non-designed data,* but that can inform social science research. After a brief discussion of the general approach of these designs and their applicability for use with social media data, we discuss the challenges associated in their use with social media data and potential solutions for “convergence” of these methods for future quantitative research in the social sciences.

3. Social Science Designs

3.1. Qualitative Observational Design

When designing studies, social scientists approach the collection of data about beliefs and behaviors both quantitatively and qualitatively. The quantitative side will be addressed in the other study design sections. This section will highlight how studies are designed using more qualitative data collection by watching (either in person or on recorded video) or listening (in person or recorded audio) to individuals in either a laboratory or a naturalistic setting and recording what is seen or heard. These observations of behavior can range from highly controlled laboratory settings where tasks are provided for individuals to very unstructured settings where the researcher has no control over what the individual is doing. Some common research designs included in this category are structured and unstructured in-depth interviews, focus groups, participant observations, and ethnographies.

The information generated by these approaches are analyzed for the presence of certain terms, themes, or concepts. Using this research tool, the data are converted from written notes to numbers by creating themes for the data and then coding various soundbites or video clips into these themes in order to analyze their prevalence, understand how they relate to one-another, and determine their relationship to other variables of interest. In order to reduce the error in theme generation and coding, multiple coders are asked to code the same piece of information and interrater reliability is assessed (typically as a chance-adjusted amount of agreement between these coders). After the coding and reliability assessment, the data can be used to make inferences about these themes, their implications, and the conditions under which they emerge. Typically, qualitative studies have a small number of participants, ranging from 10s to 100s. Larger samples are typically too expensive and time-consuming to collect, code, and analyze.

Application to Social Media.

Given that this is a method for collecting unstructured data in various formats, there are aspects of this study design that closely parallel social media data, which often consists of descriptive data about how an individual is interacting with those around them in an uncontrolled setting (social media platform) with no researcher-initiated interventions. Social media is ideal for collecting large amounts of data, text, pictures or video, and using different forms of content analysis to examine themes and constructs. It is important to consider that social media data are complex data on human behaviors, often reflecting individuals' beliefs and opinions coming in at large volumes which is not the traditional way that data generated by qualitative observation designs are considered. In some sense, social media data are large-scale qualitative data that are generated without a research question or hypothesis. Typically, qualitative observational studies in social science would involve data collection designed around a research question that has been predetermined to test a hypothesis that has been proposed. In the case of social media data, researchers need to look to find data that can be used to answer questions of interest, understand the strengths and limitations of using these data, and then generate theory-driven hypotheses.

The volume of data (millions of observations in some cases), however, makes social media data impractical to annotate manually as we might do with traditional qualitative data. Combining the existing qualitative methods with ways that computer scientists and computational linguists process and classify primarily text data to derive meaning from different representations of text is foundational for understanding large quantities of qualitative data describing human behavior.

Computer scientists and computational social scientists have written many different algorithms to help measure or quantify different individual attributes using text. For example, recommendation systems aggregate user purchasing patterns with behavioral data (e.g. reviews) and network data to determine which products to recommend. Typically, computer scientists are given large volumes of well-labeled data for this task. Social media data present a number of new algorithmic challenges that computer scientists are still trying to understand. First, researchers sometimes only have access to very short posts with typos, new words (e.g. hashtags), and limited context. From very limited information, algorithms need to label themes, tone, stance, emotion, etc. While algorithms for these tasks are well understood for more coherent, longer documents, the noisy, short text associated with social media content requires redesigning algorithms to consider these new conditions. Also, researchers do not want to pay to label or code large numbers of posts. So there is a need to use only a small number of labeled posts, typically 100s or 1000s, to build a model that can be applied to millions of posts. Unfortunately, manual labeling of posts can be hard for certain tasks, e.g. sentiment. If the labeled posts are poor quality, the accuracy of models built using the manually labeled data will be reduced. New active learning or reinforcement learning algorithms need to be considered to mitigate this challenge. Another challenge has to do with post *intent*, humans are good at identifying contextual issues like sarcasm, humor, and posts containing multiple languages. These types of subtleties are harder for computers, and text analysis tools vary considerably for different languages. Finally, data samples obtained opportunistically from social media are often biased towards various subsets associated with specific data collection parameters. This in turn results in biased patterns being extracted from those data computationally. Fairness in machine learning and data science is a growing area that is beginning to explicitly address issues of data bias, fairness in machine learning algorithms, data transparency, and ethical uses of data in computational data mining.

This creates an excellent opportunity for a convergence between computer science and social science as "big data" is used with more qualitative approaches to understanding the themes and constructs that are emerging within and across multiple social media platforms.

3.2 Experimental Designs

Experiments are research designs in which some portion of participants are randomly assigned to a treatment group (that receives some intervention), and some portion are assigned to a control group (that does not receive the intervention, or receives a different intervention). The main goal of experimental research is *causal inference.* Because participants are randomly assigned to control and treatment conditions the two groups should be virtually the same on average in terms of all of their pretreatment characteristics (observed and unobserved). Therefore, we can be confident that any differences in average outcomes between the two groups are due to treatment, rather than to characteristics of the respondents or their environments. This makes experiments a key research design for all social scientists who are interested in answering causal questions.

In the world of computer science, there are a number of sub-disciplines that use experimental designs. One example is in Human Computer Interaction. There the goal is to understand how to improve devices, interfaces, and software to make them more usable, to increase productivity, or to automate a manual task. For example, a researcher may have a new device for a computer, e.g. an electronic pen, and want to see whether or not the electronic pen and a tablet can be used as effectively as pen and paper. In this case, the treatment group can be given the electronic pens and tablet and the control group pen and paper. Similar to social science experiments, a causal relationship between the electronic devices and productivity can be measured.

Application to Social Media.

Social media can serve experimental methods in three main ways. First, it can serve as a way to recruit subjects. Researchers have had some success in using online platforms to recruit participants in studies, including experimental research. These include traditional online platforms intended for this use (Mechanical Turk, Prolific, etc.), as well as platforms intended for other uses, but repurposed for recruitment. Examples of this include recruiting from Facebook groups, online message boards, Twitter ads, and Craigslist ads (Schneider & Harknett, 2019). The benefits of using these platforms include lower cost than traditional recruitment methods (e.g. going through a recruitment firm), and a more varied sample of subjects as compared to more traditional subject pools like undergraduates.

Second, it can serve as a means by which to apply a treatment. That is, social media users can be used as experimental subjects, and assigned to a treatment or control group, accordingly. Some interesting examples of treatments that computer scientists may be able to help with include developing algorithms to identify misinformation and then building a tool or app that alerts users who were exposed to posts with misinformation. Many ethical concerns and challenges with this approach have been discussed (Gallego et al., 2019), such as figuring out how to appropriately obtain informed consent from users, and how to debrief them after an experiment.

Finally, social media data can serve as a way to measure outcomes of interest, after engaging in some intervention. That intervention could happen within social media (Munger, 2016), as outlined above, or could happen externally - the treatment could occur offline, or on another website, or in a survey experiment, while the outcome of interest is measured on social media. The challenge with the latter is successfully linking the people who are in the treatment or control groups with the user profiles of interest. One solution to this is thinking about natural experiments (Zhang et al., 2017a; Budak et al., 2017) and/or quasi-experiments (Oktay et al., 2010; Zhang et al., 2017b), where we have enough information to determine whether a social media user was exposed to the as-if-random treatment or not.

3.3. Survey Designs

A survey is a systematic method for gathering information from a sample of entities to generate quantitative descriptions of certain attributes of the larger population which the entities are a part of. Thus, constructing valid measures and representing the target population are two essential features of any survey. There are two primary distinctions in survey design, probability samples and non-probability samples. Probability samples are designed to ensure that every element in the population has a known non-zero probability of being selected in a sample. This is an essential feature for applying the proper methods of estimating sampling variances and generalizing to the population. In most non-probability samples, however, the probability of selection is unknown, potentially jeopardizing the representation of the sample.

Construct measurement is achieved by carefully crafting survey questions to reflect constructs of interest. Commonly, measures are pre-tested and validated to ensure that they are reliably and accurately measuring what we hope to know. The wording of questions is carefully considered to ensure that it does not lead respondents to answer in a particular way, and to minimize social desirability and acquiescence biases. Questions are crafted to ascertain characteristics, attitudes, and behaviors of participants. Survey questions may be open-ended, where participants offer a text response, or closed-ended, where they choose from predetermined answers provided by the researchers.

Surveys rely on self-reports from participants (whether the mode is interviewer mediated or not), which is both a pro and a con. Sometimes self-reports are the only way to uncover attitudes, since otherwise they exist only in subjects' heads. On the other hand, self-reports are subject to a variety of biases, including recall and social desirability bias as people are inherently wired to try to answer "correctly" and to provide socially appropriate answers.

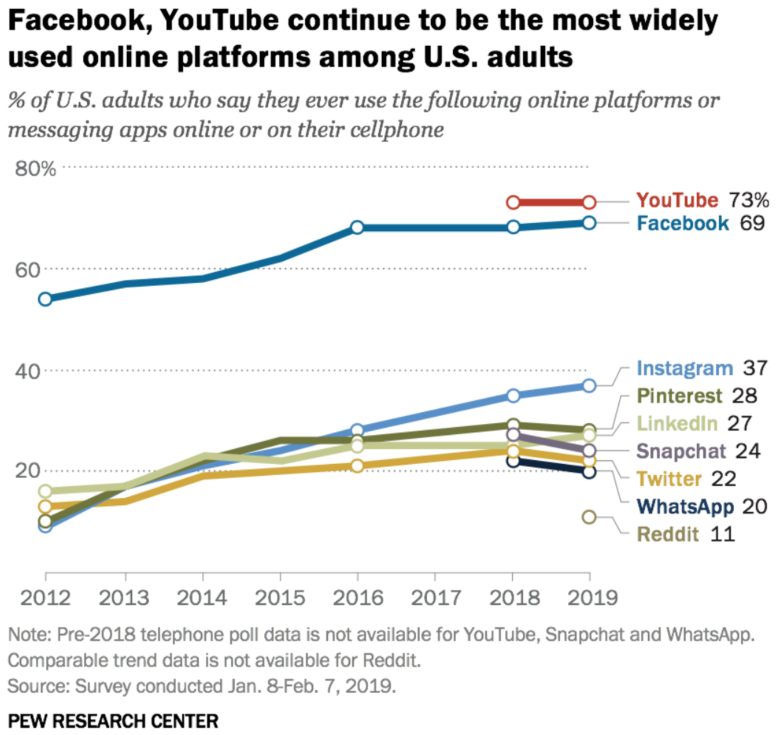
Application to Social Media.

There have been several attempts to use information from one source, such as surveys, to generalize to the other, such as social media users, with limited success (Budak, 2019). One reason is that the active use of social media platforms by the general public is limited and differs by platform. As data from a recent Pew survey illustrates (see Figure 1), the most ubiquitous self-reported use of a platform is on YouTube and Facebook (about 7 in 10 adults) while only one-in-five (22%) use Twitter. The next highest level of use is on Instagram (37%) (Perrin & Anderson, 2019). These rates suggest differential problems of representation of the adult population in the United States on social media platforms. A further complication is that information from Facebook and Instagram is not available at the individual level. Information from Twitter, on the other hand, is available at the individual level, but the coverage rate is much lower, making representativeness more of a concern.

This also creates a problem for researchers who start with a sample of individuals to survey about their social media use. Since the use of social media platforms in the adult population are low, a first issue is the sample size needed to produce a subsample of Twitter users, for example, in order to analyze their personal or political characteristics in relation to use patterns. Another issue is that the use levels are self-reported and subject to a number of potential well-known biases (Henderson et al., 2019; Ernala et al., 2020; Guess et al., 2018). In addition, if a researcher is interested in looking at individual tweets, IRB protocols require that the survey respondents are asked for their permission to link to this content. Research shows that agreement to this condition is about 20% in a general population survey (Mneimneh et al., 2020), but could be as high as 76% in special samples of social media users (Wojcik & Hughes 2019).

Figure 1.

% of U.S. Adults Who Say They Ever Use Online Platforms, 2012-2019.



Facebook, YouTube continue to be the most widely used online platforms among U.S. adults

% of U.S. adults who say they ever use the following online platforms or messaging apps online or on their cellphone

Note: Pre-2018 telephone poll data is not available for YouTube, Snapchat and WhatsApp. Comparable trend data is not available for Reddit.
Source: Survey conducted Jan. 8-Feb. 7, 2019.

PEW RESEARCH CENTER

Note. Reprinted from Pew Research Center, by A. Perrin & M. Anderson, 2019, https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/. Copyright 2019 by Pew Research Center.

Typically, population-based/survey design requires treating individual social media users as the unit of analysis, at least initially. After obtaining consent from survey respondents to access and use their social media data, researchers link survey responses with accounts and begin data collection. This "focus on the user" is a defining attribute of the population-based/survey design and imposes additional constraints on data collection - namely, that all users ought to have an equal chance of selection (regardless of how often they participate on the platform) and that non-human social media users (i.e., bots) will not be selected.

At the user level, this approach allows researchers to generate easily interpretable summaries of social media behavior. For example, we can leverage the representativeness of random samples to obtain unbiased estimates of social media posting behavior, or discussion of particular topics, for defined populations of social media users. In addition, we can leverage self-reported attributes of social media users to explain patterns in behavior: for example, examining whether strong partisan identity relates to political discussion online.

However, social media behavioral data are generated in a dynamic way, and individuals have the ability to both engage in new kinds of behavior and make changes to past behavior (e.g.

editing or deleting posts; making posts private or public, removing geotagging, etc). Survey responses, by contrast, are more of a snapshot of user attributes - fixed to the time of survey administration, changing only when researchers conduct additional surveys of the same respondents. This dynamic nature may be a new dimension for social scientists who typically use survey designs, but it is a fairly standard dimension for many data mining and data science researchers. Understanding how to construct meaningful measures from these dynamic data that can be used in conjunction with the traditional survey responses is an important area for interdisciplinary work.

This design may be less appropriate when researchers intend to use posts, rather than users, as the unit of analysis. Since the recruitment of social media users is not based on post volume or post content (since population weighting targets for those dimensions are often unknown), many survey respondents might not make any posts at all - and a small number likely generates a large share of them on a select number of topics. Thus a sample design centered on all social media users - rather than those who generate the posts of interest - is likely to be inefficient.

Finally, computer scientists have a challenging time finding ground truth data from social media for developing robust machine learning algorithms. When social media data are linked to survey responses, the survey responses can serve as a ground truth dataset for these machine learning algorithms. For example, if respondents specify their political affiliation in the survey, algorithms that use their posts can attempt to infer their political affiliation and the accuracy of these algorithms can be measured using the reported survey value.

3.4. Non-Designed, Complex, “Big” Data

As noted previously, social science data are generally designed to answer hypothesis driven questions. Even when large, population representative datasets are collected as infrastructure data for the community of scholars, the intent is to use the data to test hypotheses of the social scientists. With the advent of large amounts of data being produced and collected on individuals through wearables, smart phones, commercial datasets, genome data, brain scans, and multiple other technology and machine data, the amount of data on behavior in the population has increased exponentially. These are the types of datasets that applied computer scientists are comfortable working with. In contrast, social scientists who are used to “long” or “wide” data formats are now presented with a situation of lots of data about human behavior but without structure or design or in ways that flowed from theories or hypotheses. We consider these types of data non-designed data. Non-designed data are any data that have been collected opportunistically from sources where the data were not intended for answering specific theory or disciplinary questions. Instead, they are data collected either for administrative use (people applying for non-employment benefits), data reflecting behaviors people engage in (voting turnout), data describing locations of people (mobile phone data), or data collected to determine how effective a program or curriculum might be (testing data in schools, for instance).

Given the size and complexity of these data, computer algorithms are needed to extract and manage the data in ways not normally done in statistical software. This type of data and the use of it demands both programming and statistical skills in order to appropriately use the data for understanding the underlying human behavior. The task of extracting information from the data not designed or intended for answering questions of theories or hypotheses is a challenge for the typical way that social science is done, but has the potential to greatly increase our understanding of behavior across contexts and time.

Though these data have attractive qualities, like providing information about the behavior of individuals, they also pose challenges for linking relevant datasets to characteristics of the individual (demographic data, but also beliefs, intents, decision making) that are often important to social scientists. They also lack information about those not participating in the conversation. What tends to make these data attractive is that they are larger-scale and that they are readily available. Furthermore, various natural experiments affect behavior of individuals/organizations generating these found data. As such, these can even be used to ask causal questions, despite its observational nature, through the use of causal inference or quasi-experimental design techniques (Jensen et al., 2008; Zhang et al., 2017a).

Application to Social Media.

The application of non-designed, complex data to social media data is perhaps the most straightforward of the four examples we consider. Social media is just a newer and generally larger source of data than researchers have relied upon in the past. Rather than looking at administrative data regarding monthly unemployment applications, for instance, we might collect mentions of unemployment applications on Twitter. There is greater noise in the latter than in the former, but in general they are both valuable data sources that are not created for researchers to use, but rather for some other purpose entirely. This may be particularly valuable for variables that are not easily accessible from surveys or administrative datasets, e.g. shared stories about harassment or abuse. Many of the same challenges exist in using social media data as exist for using other "found" or "organic" data, although that is not to say those challenges are easier to overcome (Singh et al., 2020).

Social media data by itself is important for understanding instant reactions to events such as political/election events, policy changes, and cultural changes (Bode et al., 2020; Bowman Williams et al., 2020; Anderson et al., 2018; Anderson & Toor, 2018). Finally, for societal scale issues that lack traditional survey data, these data can provide insight and evidence that may be partial, but can still enhance our understanding of a complex situation, e.g. humanitarian crises

(Singh et al., 2019; Earle et al., 2012; Ramakrishnan et al., 2014), public health concerns (De Choudhury et al., 2013; Dredze, 2012), etc.

In 2008, for example, Chris Anderson a writer for WIRED magazine wondered if the scientific method of designing studies from theory and then carefully controlling the collection of data would continue with the influx (or data deluge as he termed it) of data from websites and search engines (Anderson, 2008). More than a decade later, the “data deluge” is even higher with more data on individuals’ behavior being created in seconds than what can be gained in years from traditional qualitative and survey data collection methods. In many ways, this would be an ideal way to obtain data where the data collection on important social behavior is collected unobtrusively and perhaps somewhat free of individuals actively managing their responses in order to give what are perceived as socially desirable responses to surveys. However, non-designed data also comes with less knowledge and control of the structure of the data. Thus, trying to make predictions from these data or trying to generalize findings from the data becomes problematic unless the data can be matched to a representative set of data that would allow for these generalizations.

4. Challenges for Designing Studies Using Social Media Data

Now that we have described the basic designs of social science data and how they have or could incorporate social media data to answer questions on human behavior, broadly construed, we turn to the challenges that are faced when trying to use social media data to answer relevant social science questions. We highlight the particular issues of using the unstructured data from social media platforms and how they present issues for how social scientists approach their research paradigms. Importantly, we note that aspects of the computer science paradigms for gathering these data and analyzing them is important for understanding and overcoming the challenges of converging the social science and computer science methods together.

4.1. Understanding the Population of Social Media Users

Social media data does not come from a scientifically selected sample of an easily identifiable population. This is because not everyone is on social media, not everyone who is on social media posts to the same extent, not everyone on social media posts about all possible topics researchers would want to study, and not everyone makes every post publicly accessible. Thus, there are biases in the data related to who chooses to be on a social media platform, chooses to interact or post on that platform publicly, and chooses to post about a topic of interest to social science. This selection process is unknown to a large extent (perhaps unknowable), and therefore the data generation process is also unknown. Lack of knowledge about the characteristics of the sample in our designs is not a new problem in the social sciences and not all of the disciplines are concerned about well characterized samples. Social media data, however, is perhaps a more extreme example of the lack of clarity about a sample and about the population from which it was derived. In some cases, we do not know anything about the participant except what he or she does on the platform itself. In other cases, a participant will share a large amount of demographic information. This is particularly problematic on platforms without real name policies, where users may be completely anonymous, making basic attributes like gender difficult or impossible to identify without the assistance of computational models. The extreme variation and lack of uniformity in terms of missing information limits the ability of social science researchers to generalize to the broader population from social media data. This is a major challenge to be addressed when using organic as opposed to designed data.

4.2. Non-response – What Does It Mean?

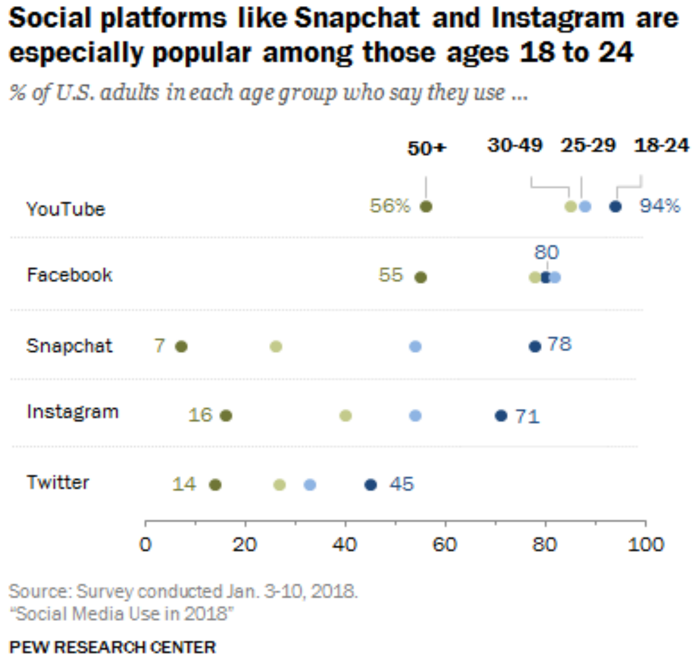
A related challenge is how representative the social media data are to the beliefs and behaviors of the population. There is no standard stimulus to which people are responding on social media. Sometimes the posts are related to an event or topic that is popular at the time (e.g. COVID-19) and other times they are just reflections of normal activities during the day of the individual. Since we do not understand the underlying model for how the data are generated, we do not understand what it means to respond or not respond to something happening around social media users. A non-response or lack of posting on a topical or individual issue can indicate lots of different things - disinterest in the topic, unwillingness to post about it publicly, ambivalence, lack of identification regarding the topic, management of an online persona, or a host of other possibilities. Distributions of user participation on social platforms are often not normal. Such distributions frequently have a long tail - that is, many users engage very little with a small number of users producing most of the content, or most of the communication, etc. (Ortega et al., 2008; Hughes et al., 2019). This may pose a challenge computationally, though we have good offline research exemplars to draw from in that regard. It may also pose a challenge conceptually - thinking about the modal or median user versus thinking about different types of users (high producers, high consumers, lurkers, etc.) may sometimes be appropriate and is an assumption worth stating in this type of research. If user online behavior is the outcome of some intervention, effects may only be interpretable among some subsets of individuals. Thus, the normative control that researchers would have on designing and managing the data generation so that inferences can be made about individuals' behavior is challenged. This means that while many approaches for considering social media data in a more descriptive, qualitative way are available, more care needs to be taken when conducting causal inference or interpreting non-response.

4.3. Increase Potential for Algorithmic Bias

Because every platform is unique, they each have different interfaces for researchers to collect data. This means that inconsistency exists with the way platforms construct samples for researchers (e.g. by users, by posts, by keywords), and each platform also captures different types of data about users (e.g. photos, video, text). All of these issues create sampling, algorithmic and other computational challenges. For example, if samples are based on the fraction of posts shared as opposed to the fraction of individuals sharing a post, the sample will be biased to those who post more frequently. This may bias algorithms that are being developed to understand beliefs and opinions of those on social media. For example, these algorithms may have very high accuracy for individuals who post frequently, but have a much lower accuracy for those who post less frequently. Other inconsistencies may arise based on demographic characteristics. since different user groups may prefer different social media platforms. For example, social media platform choice varies by age (see Figure 2), geography (Yeung, 2019), and topic (Menon et al., 2017). Research needs to be conducted to determine appropriate data sources to account for biases in data with respect to the question of study. Currently, much of this is being done in an ad hoc manner. Given all the diversity in the data formats and populations using different social media platforms, a number of computational questions also need to be addressed, including what database models are sufficiently robust enough to handle large volumes of these diverse data, how these data should be indexed for quick access and use, i.e. which database models and indexing strategies are optimal for querying, preprocessing and analysis of these data in order to be able to evaluate them efficiently and at scale.

Figure 2.

% of U.S. Adults by Age Group Who Say They Use Online Platforms, 2018.



Social platforms like Snapchat and Instagram are especially popular among those ages 18 to 24

% of U.S. adults in each age group who say they use ...

Source: Survey conducted Jan. 3-10, 2018.
"Social Media Use in 2018"

PEW RESEARCH CENTER

Note. Reprinted from Pew Research Center, by A. Perrin & M. Anderson, 2019, https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/. Copyright 2019 by Pew Research Center.

4.4. Inferring Demographic Characteristics

Although qualitative data are rarely representative of a broader population and is not intended as such, there is generally knowledge about the nonresponse rates and exclusionary conditions that are imposed upon the data. Additionally, some effort is made to assess the potential biases due to self-selection into the respondent pool, mostly based on demographic data of participants. Because this information is often unknown for social media data, our ability to understand who is generating the data is limited. A number of algorithms that infer different demographics of individuals on social media have been proposed. Issues arise when using these algorithms because they assume knowledge of either all the posts of the individuals, their network, or other account information. These extra data are not readily accessible on different platforms to researchers. Therefore, there is much room to develop new algorithms that can handle large amounts of variability in the underlying user data, and for identifying the types of user data that are most useful for the machine learning algorithm (active learning). A further concern is the ethics of inferring different demographic characteristics that are not shared. For example, users may not care about their gender being inferred if they share pictures online, but they may be concerned about political affiliation, sexual preferences, or cultural identification being inferred if they choose to not share the information themselves. As we try to learn about the population, we must be mindful of what is reasonable to infer and what is unethical, especially for discovered and repurposed data, where user consent is not explicitly given.

In general, it is important to remember that social media data are composed entirely of people who self-select into the platform. There is the potential to try to structure unstructured data by connecting it to a known research design, such as a survey. The aim of such an approach is to assess biases with reference to the target population of interest and validate aspects of the characteristics of the individuals in the study (e.g. parent/non-parent, voter/nonvoter, US citizen, non-citizen, etc.) and will be discussed more in the next section.

4.5. Data Access Inconsistencies

Social media platforms are not universally available to all users and researchers (Hughes et al., 2019; Sloan, 2017). Twitter and Reddit, for example, have general data availability, but other platforms are more idiosyncratic on what data are made available to researchers (if any). Even when data are available, the data access is still subject to API restrictions. Further, there are restrictions in some countries and even in various institutions on what type of social media can be accessed. Thus, the data that are available from social media are more akin to a sample of the data than representing the full population of data. This sample, like many social science samples, may contain a bias with regard to the type of information available for the researcher to make inferences to the population to which they would like to generalize. Thus, careful thought and perhaps additional data collection needs to be done when using existing social media data in order to address these potential biases in the data and account for the differences in policies and access across platforms.

4.6. Platform Content Biases

A more difficult selection issue to consider is how the algorithm for the particular social media platform is derived and thus what information is presented to the user in this platform that may influence their responses. In the words of Larry Lessig (2000), the platform defines the architecture which constrains the digital realities of its participants. This can be experienced in the form of the platform user interface (e.g., addition of a like button on Facebook, or the order in which information is provided) or the computations done at the backend to identify what types of information will be provided to the user. Past research shows this architecture indeed shapes user behavior (Budak et al., 2017). Some platforms provide certain content to the user based on their user profile or the choices they make about what to click on or otherwise engage with (Gillespie, 2014). For this reason, the content a user posts is likely not only a function of their own interests, thoughts, and behaviors, but also contingent upon what content they are seeing within the platform. For example, Facebook famously showed an emotion contagion effect - when users saw fewer happy posts, they were also less likely to post happy content (Kramer et al., 2014). Even more problematic, many of the components of platform algorithms are entirely unknown and are constantly changing. The analogy in survey research is the presentation order of questions and how questions are asked and the bias they can have on responses. Here the researcher can examine these potential biases and even randomize the questions across respondents. There is no similar control with existing social media data readily available to the researcher. The potential for collaborations with social media companies is a possible solution to help in adjusting for this potential bias. Still, even if researchers are able to determine the possible biases, because platforms continually change their layouts and content, replication of studies becomes very difficult, if not impossible.

4.7. Active Management of Data

With structured data, researchers set the parameters and the characteristics (e.g. multiple choice, short response, Likert scale) of how the questions will be asked and the format of the data that will be received from the individual answering the questions. In the case of social media data, there are no restrictions of how the individual will talk about a topic (if at all) or format of the data. Individuals are also potentially providing different types of data across multiple platforms. Some individuals are actively managing their data and others are just reacting to other posts on the platforms. Thus, another bias is that not all individuals are using the same set of platforms or use them differentially to represent or manage their identity on these platforms in various ways. When using existing social media data to understand the role of the individual, the researcher needs to understand the profile of how the individual is interacting with single or multiple platforms. There may be ways to compare the comments on each platform to construct a "social media identity" but that may not be adequate for comparing to someone who uses only one platform. Understanding the beliefs or behaviors of individuals will vary depending on how open they are in sharing information on the platform and whether they share consistent information across them. Indeed, there may be a bias on who is likely to use multiple platforms versus those who use only single platforms on what we can infer regarding their beliefs or behaviors.

4.8. Data Ethics and Expectations on Consent

Social media data can be accessed in many different ways for experiments. Researchers can collect data from the platform's API. Researchers can scrape all of the data directly from the webpage using automated data scraping robots. Researchers can ask questions and get responses using the platform's advertising infrastructure. Researchers can directly befriend users or request access to private posts. Different approaches are needed for different study designs. Unfortunately, some of these approaches are more invasive than others and may lead to study designs that are ethically questionable (Kramer et al., 2014; Kanich et al., 2008). Social scientists have established procedures for ensuring the safety and privacy of researcher participants, overseen by their university's institutional review boards (IRBs). Computer scientists have less training (if any at all) in this area. Unfortunately, IRBs have not kept pace with the new uses of social media data, leading to a number of ethically questionable studies. It is also unclear when it is reasonable to use these data without consent and when it is not for some platforms. Consistent standards need to be developed and all researchers, including computer scientists, need to be trained using these standards (Singh, 2016).

Similarly, as we design our studies, not only must we think about the adjustments necessary when using social media for our individual study, but we must also consider the ethical challenges associated when combining multiple studies. Can we, for example, quantify the level of privacy lost when information from multiple studies are combined? We do not know what **will be** discoverable from the data we collect now and what privacy and security risks they may pose in the future. Differential privacy and privacy filters do not protect, for example, against longitudinal linking or identifying **an individual** from a group (e.g. LGBTQ). Our new hybrid methods must involve innovation of both social science and computer science research to tackle the challenges identified.

4.9. Reproducibility of Findings

A final concern is the challenge of reproducibility. This manifests in two main ways for social media. First, because platforms often limit the ability to share original content (i.e. Twitter will not let you share tweets obtained through the API with others), there is no meaningful way to create an archive of such content in a traditional sense. Social media data are difficult to archive. This is due to its size, but also to terms of service limitations imposed by the social media platforms themselves. This challenge makes reproducing research using social media data more difficult than with traditional sources of data. That is, an archive may exist, but would be inaccessible due to the amount of computing power needed to use the archive, making the reproduction of the results difficult. Second, because platforms change quickly, sometimes invisibly, and almost always without warning, a study fielded in time 1 may not be able to be reproduced in time 2 because the platform itself has changed in some way (e.g. character limits are longer, algorithmic curation is different, formats change, etc.). Thus, even a perfect reproduction of a study might produce different results (Munger, 2019).

5. Solutions for Designing Studies Using Social Media Data

As with all science, there are many challenges in validating the data that we use to describe or make inferences or causal claims. Indeed, surveys of early social media work show that scholars tended to use the word “influence” while referring to correlated behavior (Agrawal et al., 2011). We detailed a few of these challenges in the previous section but now turn to thinking about how to address these challenges as we bring the social science and computer science disciplines together. The social science data landscape is changing (Ledford, 2020; Singh et al., 2020) and having methods that converge the social science expertise on understanding characteristics of the population with the computer science expertise in managing and analyzing large and complex data is an important step in having robust and valid data on beliefs and behaviors for future research.

5.1. Moving Toward Generalizability

The issue of generalizability can be addressed in various ways. We do need to have representative datasets with information about social media membership and these can be used to bound our generalizations. Thus, the basic demographic characteristics of social media users need to be known and to be regularly updated. Those can be compared to the non-user population to make general but important points about how those who have generated social media data are different from the general population, or from a specified target population. But researchers should not stop there. It would be important to have more fine-grained demographic information about those who post on different topics or use social media for different reasons (e.g. among users, who are using Twitter or Reddit for politics, for work, for parenting, for socializing). The other way to consider bounding or testing the generalizability of social media data is to compare answers to similar questions across social media platforms with those in a survey. Comparing survey responses among social media users and non-users with regard to questions asked with social media data alone is key - although it does make conducting and validating social media research more cumbersome.

Another potential solution is to think of social media as a new kind of non-probability survey. Surveys are designed to ascertain participant attitudes and behaviors, and sometimes we can do that less invasively using what individuals post on social media instead of designing and executing a formal survey. Social media users may share their opinions on a topic, like how they feel about a particular policy, or why they love or hate a celebrity, or whether or not they think a hotdog is a sandwich. They may also share information about behaviors they engage in, like eating lunch, meeting friends, or working with their child on homework. All of these attitudes and behaviors are potentially interesting to different researchers, and may therefore be collected in a useful way. However, there are obvious challenges regarding *who* uses social media and whether or not they post, let alone about a topic of interest. This creates questions about the representativeness of the information gleaned from social media in this way. Still, for some topics it may be an appropriate use of social media data collection and analysis.

Another way to deal with the issue of representation is to link social media users across multiple platforms, along with survey data, to gain a better picture of their use of social media and their background characteristics and attitudes. This would overcome the challenge of representativeness, because we would know the likelihood of inclusion for any given individual. It would also offer insight into how users make use of different platforms simultaneously. However, for this to be feasible at all, we would need permission from both the social media users and the social media companies to get their information and link it together. This would provide tremendous opportunities for the societal goods, academic research, business opportunities, and others. There are great challenges and risks to this, however, and expense, legal implications, privacy issues (Singh et al., 2015), cooperation from companies, and concerns about hacking of the information are all significant challenges that would need to be overcome. Despite these challenges, there are models we can look to for creating such a project, including the Federal Statistical Research Data Center, and some states that have centers to link individual records from their state agencies.

5.2. Partnering with Social Media Companies

Ideally, researchers and corporations that capture social media data would work as partners to use the data to support advancements in research. Together, they would develop a mechanism for obtaining consent for different research studies and empower users to make decisions about which data they feel comfortable sharing. Unfortunately, companies view these data as proprietary and use them for a competitive advantage or as an additional revenue stream. Until companies begin viewing these data as owned by users and as a way to support research, progress on this front will be limited. However, without this collaboration, many of the methodological questions that we need to answer about the samples generated will remain unknown.

5.3. Cataloging Platform Content and Algorithms

In the absence of these types of partnerships, we need to develop up-to-date catalogs of user information collected by each platform, marking which pieces of information are public and which are private. The data are continually evolving. Different fields are available through APIs and these continually change. This means that study designs must be flexible - perhaps they now need to be more iterative. It is also worthwhile to note recent efforts that approach data collection as an optimization problem, applying reinforcement learning techniques to bring much needed rigor to this important problem (Li et al., 2016).

We also need to create social media specific repositories of algorithms that are used for text mining and machine learning, and exact information on what datasets they are applied on, what the parameters are, the accuracy of the algorithms, and measures for fairness, reliability, and robustness. Computer scientists do not effectively communicate the assumptions and/or limitations of algorithms well. Social scientists sometimes use algorithms without understanding how they work or when they will not work well. When designing a study, we need to make sure that the algorithms that will be used throughout the study are specified so that study designs can be replicated even if data are not shared.

5.4. Rethinking Research Questions and Designs

In general, we encourage researchers to think about what questions are best suited to being answered with social media data. In particular, what are NEW question types that can be asked with social media data that have not been asked before or cannot easily be asked using traditional observational methods? One promising area might be questions about human behavior or beliefs for which spontaneous, unsystematically prompted answers are better - more accurate, more externally valid - than any we could ascertain from a more structured researcher intervention. For example, psychologists may want to examine how parents define “good parenting” - what actions or ideas are associated with that concept? Arguably, those definitions are more accurate and less biased if they are spontaneously generated rather than prompted by researchers. We cannot get this information for a representative sample of parents, but we can know what definitions *exist* in a more valid way, for those who do choose to engage on the question (that is, who produce relevant data) on social media.

Social media data are also potentially excellent for understanding human communication on a large scale. Linguists analyze conversations and conversational turns, but using social media data we can understand how various aspects of written language are interconnected. Does a post about a public policy that has a positive sentiment produce more similar posts in surrounding networks, a contagion effect, than posts about a public policy with a negative sentiment? With a combination of social media data and real-world observation, researchers might even uncover when and how social media content and use impacts human behavior outside of social media. So, the broader question of the effect of social media use on human interaction outside of social media is a valid one for which social media data are essential.

6. Conclusions

This group of social scientists, data scientists, and computer scientists, came together to try and gage how existing design and methods could be used to understand the large quantities of complex data on human behavior that are being generated in less than a second every day across multiple social media platforms and search engines. We started by considering how questions on beliefs and behaviors have commonly been examined in the social sciences and computer sciences and what we can draw on to create a “convergence” on methods that will allow for valid and low biased estimates of human behavior. What is clear is the value of having these conversations and working through the differences in the fields to try and understand the next steps that need to be taken. There is great value in combining the rigor of methods of the social sciences and the ability to manipulate and manage large, complex data of the computer scientists. There are many challenges to overcome, but we are beginning to move forward on thinking about how to harmonize the approaches to research and the methods used by potentially combining the knowledge of social scientist researchers about who has been sampled from the population and ethical considerations when using social media data with the knowledge of computer science researchers about algorithmic design, large-scale processing, and analysis of large amounts of data from social media. It is likely that new hybrid study designs and methods will result from these types of collaborative efforts. These new innovations will ultimately expand the methodological toolkits of both research communities.

Acknowledgements

We would like to thank the National Science Foundation and the Massive Data Institute (MDI) at Georgetown University for supporting this collaborative meeting. This white paper is an output of that meeting and is co-authored by those who participated in the meeting. A special thanks to Rebecca Vanarsdall from MDI for her support in planning the meeting and helping prepare this white paper. We also want to thank all the students and staff who helped facilitate and take notes during the meeting. Those notes were very valuable when putting this document together.

References

Agrawal, D., Budak, C. & El Abbadi, A. (2011). Information diffusion in social networks: Observing and affecting what society cares about. In *Proceedings of the ACM International Conference on Information and Knowledge Management* (pp. 2609-2610). Glasgow, Scotland, UK.

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

Anderson, M., Toor, S., Rainie, L., & Smith, A. (2018). Public attitudes toward political engagement on social media. *Pew Research Center.*

Anderson, M. & Toor, S. (2018). How social media users have discussed sexual harassment since #MeToo went viral. *Pew Research Center.*

Bode, L., Budak, C., Ladd, J., Newport, F., Pasek, J., Singh, L., Soroka, S., & Traugott, M. (2020). *Words that matter.* Brookings Institution Press.

Bowman Williams, J., Singh, L., & Mezey, N. (2020). #MeToo as catalyst: A glimpse into 21st century activism. *University of Chicago Legal Forum.*

Budak, C. (2019). What happened? The spread of fake news publisher content during the 2016 US presidential election. In *Proceedings of the World Wide Web Conference* (pp. 139-150). San Francisco, CA, USA.

Budak, C., Garrett, R. K., Resnick, P., & Kamin, J. (2017). Threading is sticky: How threaded conversations promote comment system user retention. In *Proceedings of the ACM on Human-Computer Interaction* (pp. 1-20). New York, NY, USA**.**

De Choudhury, M., Gamon, M., Counts, S., Horvitz, E. (2013). Predicting depression via social media. In *Proceedings of the International Conference on Weblogs and Social Media.* (pp. 128-137). Boston, MA, USA.

Dredze, M. (2012). How social media will change public health. *IEEE Intelligent Systems, 27*(4), 81-84.

Earle, P., Bowden, D., & Guy, M. (2012). Twitter earthquake detection: Earthquake monitoring in a social world. *Annals of Geophysics, 54*(6).

Ernala, S. K., Burke, M., Leavitt, A., & Ellison, N. B. (2020). How well do people report time spent on Facebook? An evaluation of established survey questions with recommendations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp 1-14). Honolulu, HI, USA.

Henderson, M., Jiang, K., Johnson, M., & Porter, L. (2019). Measuring Twitter use: Validating survey-based measures. *Social Science Computer Review*, 0894439319896244.

Hughes, A., Jones, B., Tyson, A., Remy, R. & Smith, A. W. (2019). National politics on Twitter: Small share of U.S. adults produce majority of tweets. *Pew Research Center.*

Gallego, J., Martínez, J. D., Munger, K., & Vásquez-Cortés, M. (2019). Tweeting for peace: Experimental evidence from the 2016 Colombian Plebiscite. *Electoral Studies*, *62*, 102072.

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. Boczkowski, & K. Foot (Eds.), *Media technologies, essays on communication, materiality and society* (pp. 167-194). The MIT Press.

Guess, A., Munger, K., Nagler, J., & Tucker, J. (2019). How accurate are survey responses on social media and politics?. *Political Communication*, *36*(2), 241-258.

Jensen, D. D., Fast, A. S., Taylor, B. J., & Maier, M. E. (2008, August). Automatic identification of quasi-experimental designs for discovering causal knowledge. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining* (pp. 372-380). New York, NY, USA.

Kanich, C., Kreibich, C., Levchenko, K., Enright, B., Voelker, G. M., Paxson, V., & Savage, S. (2008). Spamalytics: An empirical analysis of spam marketing conversion. In *Proceedings of the ACM Conference on Computer and Communications Security* (pp. 3-14). New York, NY, USA.

King, G., & Persily, N. (2019). A new model for industry–academic partnerships. *PS: Political Science & Politics*, 1-7.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. In *Proceedings of the National Academy of Sciences* (pp. 8788-8790). Princeton, NJ, USA.

Ledford, H. (2020). How Facebook, Twitter and other data troves are revolutionizing social science. *Nature*, *582*(7812), 328-330.

Lessig, L. (2000). Code is law: On liberty in cyberspace. *Harvard Magazine*.

Li, C., Resnick, P., & Mei, Q. (2016). Multiple queries as bandit arms. In *Proceedings of the ACM International on Conference on Information and Knowledge Management* (pp. 1089-1098). New York, NY, USA.

Menon, S., Berger-Wolf, T., Kiciman, E., Joppa, L., Stewart, C. V., Parham, J., J. Crall, J. Holmberg, & Van Oast, J. (2017). Animal Population Estimation Using Flickr Images. In *Proceedings of 2nd International Workshop on the Social Web for Environmental and Ecological Monitoring* (pp. 1-4). Troy, NY, USA.

Mneimneh, Z.N., McClain, C., Bruffarets, R., Altwaijri, A. Y. (2020). *Evaluating survey consent to social media linkage in three international health surveys. Invited special issue in research in social & administrative pharmacy* [Under Review].

Munger, K. (2019). The limited value of non-replicable field experiments in contexts with low temporal validity. *Social Media+ Society*, *5*(3), 2056305119859294.

Oktay, H., Taylor, B. J., & Jensen, D. D. (2010). Causal discovery in social media using quasi-experimental designs. In *Proceedings of the Workshop on Social Media Analytics* (pp. 1-9). Washington, DC, USA.

Ortega, F., Gonzalez-Barahona, J. M., & Robles, G. (2008). On the inequality of contributions to Wikipedia. In *Proceedings of the IEEE Hawaii International Conference on System Sciences* (pp. 304-304). Waikoloa, HI, USA.

Perrin, A., & Anderson, M. (2019). Share of US adults using social media, including Facebook, is mostly unchanged since 2018. *Pew Research Center.*

Ramakrishnan, N., Butler, P., Muthiah, S., Self, N., Khandpur, R., Saraf, P., Wang, W., Cadena, J., Kumar, A., Korkmaz, G., Kuhlman, C. J., Marathe, A., Zhao, L., Hua, T., Chen, F., Lu, C. T., Huang, B., Srinivasan, A., Trinh, K., . . . Mares, D. (2014). 'Beating the news' with EMBERS: Forecasting civil unrest using open source indicators. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining* (pp. 1799-1808). New York, NY, USA.

Schneider, D., & Harknett, K. (2019). Consequences of routine work-schedule instability for worker health and well-being. *American Sociological Review*, 84(1), 82-114.

Singh, L. (2016). Data ethics - Attaining personal privacy on the web. In J. Collmann & S. A. Matei (Eds.), *Ethical reasoning in big data: An exploratory analysis*. Springer.

Singh, L., Traugott, M., Bode, L., Budak, C., Davis-Kean, P. E., Guha, R., Ladd, J., Mneimneh, Z., Nguyen, Q., Pasek, J., Raghunathan, T., Ryan, R., Soroka, S., Wahedi, L. (2020). *Data blending: Haven't we been doing this for years?* [White paper]. Georgetown Massive Data Institute Report. https://live-guwordpress-mccourt.pantheonsite.io/wp-content/uploads/2020/05/MDI-Data-Blending-White-Paper-April2020.pdf

Singh, L., Wahedi, L., Wang, Y., Wei, Y., Kirov, C., Martin, S., Donato, K., Liu, Y., & Kawintiranon, K. (2019). Blending noisy social media signals with traditional movement variables to predict forced migration. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining* (pp. 1975-1983). Anchorage, AK, USA.

Singh, L., Yang, H., Sherr, M., Hian-Cheong, A., Tian, K., Zhu, J., and Zhang, S. (2015). Public information exposure detection: Helping users understand their web footprints. In the *Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 153–161). Paris, France.

Sloan, L. (2017). Who tweets in the United Kingdom? Profiling the Twitter population using the British social attitudes survey 2015. *Social Media+ Society*, *3*(1), 2056305117698981.

Wojcik, S. & Hughes, A. (2019). Sizing up Twitter users. *Pew Research Center.*

Yeung, C. (2019, August 19). *Social media usage statistics around the world*. Synthesio. https://www.synthesio.com/blog/social-media-usage-statistics/

Zhang, A. F., Livneh, D., Budak, C., Robert, L., & Romero, D. (2017a). Shocking the crowd: The effect of censorship shocks on Chinese Wikipedia. In *Proceedings of the International AAAI Conference on Web and Social Media* (pp. 367-376). Montreal, QC, Canada.

Zhang, A. F., Livneh, D., Budak, C., Robert Jr, L. P., & Romero, D. M. (2017b). Crowd development: The interplay between crowd evaluation and collaborative dynamics in Wikipedia. In *Proceedings of the ACM on Human-Computer Interaction (*pp. 1-21). New York, NY, USA.