Measurement Considerations for Quantitative Social Science Research Using Social Media Data

Jonathan Ladd[1], Rebecca Ryan[1], Lisa Singh[1], Leticia Bode[1], Ceren Budak[2], Frederick Conrad[2], Elizabeth Cooksey[3], Pam Davis-Kean[2], Keenan Dworak-Fisher[4], Deen Freelon[5], Dan Hopkins[6], Brad Jensen[1], Ken Kelley[7], Renee Miller[8], Zeina Mneimneh[2], Josh Pasek[2], Trivellore Raghunathan[2], Carole Roan Gresenz[1], Sudeepa Roy[9], Stuart Soroka[2], and Michael Traugott[2]

December 23, 2020

[1] Georgetown University
[2] University of Michigan
[3] Ohio State University
[4] Bureau of Labor Statistics
[5] University of North Carolina, Chapel Hill
[6] University of Pennsylvania
[7] University of Notre Dame
[8] Northeastern University
[9] Duke University

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[10]

In 2019, a group of computer and social scientists began a project to converge these disciplines, with the aim of harnessing data from social media to improve our understanding of human behavior. People all over the world have started using social media, search engines, smart devices, and other technologies that record their moment-to-moment behaviors (often called, "digital traces"). Social media, in particular, provides a truly massive amount of information on the everyday activities, opinions, thoughts, emotions, and behaviors of individuals, groups, and organizations in near real-time among those who use these services. Today, most adults in the US use some form of social media (Perrin & Anderson, 2019) either to consume information or to share and discuss topics as wide-ranging as politics, employment, parenthood, leisure activities, travel, sports, and health, to name only a few, making these platforms potentially excellent sources of information on constructs plausibly relevant to most if not all social science fields.

Expanding the availability and utility of this extremely rich, but underutilized data source in the social sciences, however, requires addressing fundamental differences between more traditional social science datasets and social media data. Unlike social science data from surveys and experiments, social media datasets are not usually designed by the researcher to answer specific hypotheses or questions. And unlike the vast majority of social science datasets, they are often not in a rectangular format, and have no set structure. Instead, the data are available "as is," which often means they are both complex and highly dense in nature. Moreover, these data involve unique bias concerns not typically encountered in traditional social science methods, including often not knowing who generated the data – or even if it is a person – or what population the "sample" of participants represent. Such design and sample differences, along with the magnitude and complexity of these data, require carefully considered methods to manage, structure, and generally make sense of the data to create useful measures for social scientific inquiry. Appropriate methods for doing so are most commonly found within the toolbox of computer scientists, making a convergence of computer science and social science methods potentially very fruitful.

While employing computer science methods to wrangle social media and digital trace data in order to answer social science questions has enormous potential, it also presents a number of challenges. First, neither field is well versed in the others' methods: before social scientists can begin using ideas and algorithms from computer science, they need to learn the approaches computer scientists use, including how to work with large-scale (often unstructured) data and assess the quality of products created by computer scientists. Likewise, computer scientists need to understand the customs and methods of social science research that may not traditionally apply to their discipline. Second, it is often unclear who or what are behind the accounts that produce the data that appear in social media datasets, and how the entities creating trace data 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 provide information about 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 may require collaboration across these and other disciplines.

[10] We note that the overview section of each of the white papers in this series is fairly similar.

Understanding what we need to know about the data that are gathered, what other supplementary data are needed to answer central research questions, and how to do so responsibly is critical for trace data to live up to their full potential.

To address these issues, bringing together social scientists who try to understand human behavior and computer scientists who design and deploy algorithms to solve computational problems is a first step towards addressing these challenges. In fact, we believe that the convergence of social science and computer science is the only way to unlock the potential of social media data such that the research community can benefit and understand this increasingly important window into human behavior and the nature of societies. The advantages to social science in effectively harnessing these data are clear. But for computer scientists too, this union holds great opportunity. Designing algorithms with a new set of structures and optimizing existing algorithms for this large-scale, real-time domain, all while addressing privacy, bias, and algorithmic fairness concerns, will not only advance computer science research, but will also make their data and models more usable by social science researchers, and more applicable to solving societal challenges.

To initiate this convergence, our group planned a set of topical meetings to bring together social scientists from multiple disciplines, data scientists, information scientists, computer scientists, and ethicists/philosophers with the goal of creating a common set of understandings for how to study complex human behaviors using social media data. The topics of these meetings address each stage of the research process as we have defined it: study design; data acquisition and sampling; measurement and feature engineering; model construction; analyses and storytelling. At each meeting, we also discuss criteria for the responsible conduct of research with social media data. This is the second in a series of white papers designed to provide a summary of our discussions and suggested future directions. The first white paper focused on study designs for social media (Bode et al., 2020). This paper focuses on issues of measurement in studying social media -- both methods for creating measures from social media data and methods for assessing their quality for use in social science research. To the extent that issues of study design, sampling strategy, and modeling relate to the processes of measurement and feature engineering, they will be discussed, but not in depth, for these topics are addressed at length in other papers in this series (see our website: http://smrconverge.org/).

2. Why Use Social Media Data in Social Science Research?

Social scientists use multiple methods to collect data on humans but rely most often on either self-report measures (i.e., survey questions) or researcher interventions designed to elicit specific types of responses. In both cases, the measures employed correspond to specific research questions and involve researcher intervention as critical parts of the research design. In other words, the data would not exist without the researcher, who plans their collection and (to varying degrees) their structure. Survey research relies heavily on sampling theory to reach conclusions about broad populations, and practitioners of survey research are carefully attuned to the representativeness of the samples they draw and thereby the extent to which their conclusions are externally valid. In contrast, observational data collection is typically designed to understand relations between variables of interest and is largely derived from small samples that frequently forgo generalizability in the service of reaching internally valid conclusions (Brito et al., 2015). Hence, social scientists largely adopt methods based on whether they are generating estimates of population parameters or are estimating causal effects.

Both survey and observational methods depend on researchers selecting in advance the constructs -- whether a behavior, opinion, emotion, state or trait -- they wish to study and thus the measures they will use to capture them. This means that neither approach is well positioned to capture emergent events, for instance, which researchers could not have forecast in advance. Similarly, concepts that may be salient to ordinary individuals, but that were not initially considered when designing the study are often inaccessible to researchers. At the same time, a series of administrative challenges raise questions about the future of traditional methods. In the survey context, added costs associated with the emergence of mobile phones and declining survey response rates threaten to undermine this workhorse of the social sciences' ability to generalize what is measured in the sample to the larger population (Dutwin et al., 2014; Dutwin & Lavrakas, 2016; Meyer et al., 2015). Traditional survey or observational data can also be potentially subject to social desirability bias, as people sometimes aim to present ideal versions of themselves to researchers (Tourangeau & Yan, 2007). And a growing body of work questions the extent to which causal claims from observational data are both replicable across samples and generalizable to the population (e.g., Stanley et al., 2018).

By contrast, measures drawn from social media can capture people's attitudes, emotions, and interests on topics typically studied using large surveys, as well as behaviors and reactions typically captured using observations, without many of the above listed drawbacks. For example, Lee and colleagues (2020) used Twitter data to identify positive beliefs about, and reports of using, corporal punishment among parents at a time when the vast majority of parents report neither engaging in nor supporting spanking when asked on surveys (Ryan et al., 2016). Social media can also capture constructs at a scale that standard social science data sources cannot: behavior, opinions, emotions, and attitudes of a vast number of people on a wide range of topics. As an example, Bode et al. (2020) used automated text analysis of Twitter posts and newspaper coverage about presidential candidates in the 2016 election to compare those texts with what thousands of people recalled hearing about candidates when asked in open-ended survey responses. Data from social media, like digital trace data broadly, can also offer insights into human behavior by providing a window into behavioral phenomena that may simply not be accessible with other methods. For example, Singh and colleagues (2019) combined Twitter posts, newspaper articles, and traditional movement variables to predict the forced migration of residents in Iraq by creating indirect measures of current movement to an extent not possible using surveys given the violence and political instability in the region. Most importantly, perhaps, because social media posts represent naturally-occurring conversations about, and reflections on, people's everyday lives without reference to any predetermined study topic, they can be used to answer questions researchers would have liked to ask in surveys had they known about events in advance, something standard survey design cannot accomplish. Common to all these applications is the use of social media data to create measures of important phenomena within a time frame and at a scale simply not possible with traditional survey data.

3. The Variety of Shared Data Type

Social media platforms have a number of differences that need to be factored in when thinking about the measurement properties of social media data. Today, 72% of adults in the US use some form of social media, including Facebook (69%), Twitter (22%), Instagram (37%), and Reddit (11%) (Perrin & Anderson, 2019). While researchers tend to focus on the largest social media platforms, there are also hundreds of thousands of smaller, custom social media sites. One

of the distinguishing features of different social media platforms is the dominant type of data shared. Social media platforms initially focused on text posts. Over time, we have seen a pronounced shift toward images, audio, and video. If we consider the most popular platforms, the dominant data type varies from text (Twitter and Reddit) to video (YouTube and TikTok) to images (Instagram, Pinterest, and Snapchat) to audio (HearMeOut) to a combination of data type (Facebook). Other more traditional forms of online content, e.g. blogs and newspapers, tend to be text-based as well.

Each data type has different challenges and idiosyncrasies. Text can be grammatically incorrect, misspelled, or abbreviated; images may be blurry or have different effects added; audio and text can be in different dialects or include slang that varies over time; video can vary in quality, even within the same video, and video can even be fake. The size of these data sources also varies considerably, with much greater processing power required to process and clean audio and video than, for example, text. Finally, data storage differs by format: images are pixel-based and typically stored as bitmaps whereas text is character-based and typically stored in that form. This necessitates different computational tools to process, clean, and analyze each data source.

Because the predominant type of data on the web remains text, our focus in this white paper is on measurement using the text modality. However, much of our discussion is applicable to other forms of data, and when it makes sense, we will highlight some of the concerns and possible solutions for these other types of data.

4. Measurement Challenges

The terms and approaches that computer scientists and social scientists use to evaluate the quality of their measures differ. Becoming familiar with these differences, and bridging gaps in terminology and approach, are key objectives of this stage of the convergence project. While we are developing a full glossary of misunderstood terminology that arise across these meetings, there are a handful of terms that we want to explicitly define here: “validity,” “reliability,” “bias,” “precision” and “construct.” In the social sciences, these terms are especially prominent in survey research and the social science disciplines that rely on survey data. We had conversations at the Measurement meeting about whether these concepts, as defined in the social sciences, especially validity and reliability, apply to social media data in the same way as they do in the social sciences, or whether they need to be at least partially reimagined for this context.

Measurement terms that are more prominent in computer science are “correctness,” “accuracy” (which can include concepts of precision and recall), “efficiency,” and “reliability” (all as defined in computer science). Because of the levels of noise and scale of social media data, these concepts as defined by computer science are most relevant to algorithms, models, and systems designed for social media. Reliability has a similar definition for both computer scientists and social scientists. What differs is the application - social scientists focus on the reliability of a measure for a construct. Computer scientists focus on the reliability of an algorithm or a system design. Other concepts such as correctness, efficiency, and accuracy are important, but likely more relevant to future white papers.

One area within computer science where there is more obvious overlap with social science measurement concepts is what computer scientists call *feature engineering*. Feature engineering is the process of using domain knowledge (subject matter context) to identify and

extract features from data. These features can be viewed as different representations of the data that are determined to be useful inputs into an algorithm or model. Sometimes features are generated based on domain knowledge and sometimes they are generated based on the properties of the data themselves. While features can be viewed as a similar concept to social science variables, they are rarely the measure of interest. They should represent some aspect of the underlying data well, but each feature does not need to be completely relevant or interpretable by itself. As an example, when building a model to predict opinion on an issue like gun ownership from text, the sentiment of posts related to guns may be an important subject or domain specific feature, while linguistic characteristics of the text may be useful as opinion features irrespective of the subject, e.g. sentence structure, punctuation usage, etc., but they would not be equivalent to measures generated for different constructs. While similar issues around reliability, validity, precision, and bias arise when computer scientists are engineering features, these issues are looked at with a different lens because the goals of computer scientists and social scientists differ. Because computation power is relatively cheap, computer scientists will tend to err on the side of constructing any feature that might be relevant to ensure that important ones are not missed. While we will discuss ideas related to feature engineering throughout, particularly with regards to feature construction, more discussion about the role of features within models will be presented in future papers.

The central challenge in converging social science and computer science for the purpose of constructing measures from social media data is to blend the criteria used to evaluate traditional social science measures with those used to assess sound feature engineering. One way to initiate that blending is to retain a distinction between a measure and the construct it captures. In the vast majority of social science projects, there is a theoretical construct -- be it a behavior, opinion, emotion, state or trait -- that is captured by, but is conceptually different from, the measure itself. The reason that "validity," "reliability," and "precision" are such dominant criteria for social scientists to use when evaluating their measures is that they are trained to determine what their theoretical concepts are and define those concepts as separate from their measures. Defining construct and measure separately allows them to think about "validity," "reliability," and "precision" as well as "bias," because each criterion involves evaluating the relationship between concepts and their measures in different ways.

Below we describe the challenges social media data pose with regard to each of these criteria when examining the relationship between constructs and measures and suggest best practices for improving on these criteria if possible. In doing so, we define these criteria and assess measures used in social media research that use them, and also note how evaluating the construction of features is similar or different. We also introduce core concepts that computer scientists use, with the caveat that they will be discussed more extensively in future meetings when modeling and scale become more central issues.

4.1. Reliability

One of the main methodological issues social scientists face when creating measures from social media data is assessing the measure's reliability. In the dominant social science definition of reliability, it captures the *consistency* with which a measure taps the construct in question. Stated more precisely, if a measure is unreliable, its variance will be far larger than the true variance of the construct. Thus, another way to conceptualize reliability is in terms of measurement error: reliability equals the true variance of the construct divided by the total

variance of the measure. A reliable measure is one that yields the same score, classification, or other metric each time it assesses the same phenomena.

Because the true variance of a construct is unknown, reliability must be estimated using different methods for different measurement types. Examples include interrater reliability for human coding or labeling, internal consistency for multi-item scales, or test-retest reliability for assessments. Some of these strategies, including internal consistency reliability, may not apply to unstructured social media data for which the idea of multiple interrelated "items" "questions" does not exist. Given that many measures drawn from social media involve labeling or classifying posts, among the most frequent strategies for establishing reliability with social media data is interrater reliability. This typically involves training human coders to label social media data points -- for instance, posts or users -- along a dimension until labeling of the same data is sufficiently consistent across coders (i.e. interrater reliability is reached). Conventional content analysis typically focuses on the analysis of dozens or hundreds of units. While this method has been scaled up to handle thousands of posts (e.g., Sadeque et al., 2019), crowdcoding is now a popular alternative. Distributing a coding task across a large number of coders by using platforms such as Amazon's Mechanical Turk allows the researcher to scale up their content analysis task. However, scaling up in this way creates problems from a social science perspective. Coders on platforms such as Amazon Mechanical Turk are rarely trained which is a standard practice in social science, nor are they rigorously selected, also a standard practice in social science. Instead, their codes are compared to expert-provided ground truth labels to detect more versus less accurate coders. This mismatch reflects in part the trade-off researchers must consider when using large-scale social media data. Either, they need to invest more time to hand-code data themselves or allow for more variability in those doing the labeling and rely on ground truth labels to detect unreliable coders. This entire process - including which of these approaches is used to establish reliability, and best practices for implementing them - is not yet standardized in social media research. However, given the need to label large quantities of data, and the limits of human coding capability, some middle ground standards need to be established.

Another way to approach measurement reliability is to acknowledge that when using social media data, traditional concepts of reliability should be balanced by evaluation of a measure's relative precision -- a common criterion in computer science. While the term precision has different meanings across disciplines, in this paper, it refers to the specificity of the measurement, i.e. how precisely an object being studied is measured (Salkind, 2010). For example, when we measure age, do we measure it at the month level, perhaps for babies, or at the year level? Depending upon the analysis, a "blunter" measure may suffice, e.g. five-year windows/bins, or bins based on the stage of life (adolescent, young adult, etc.). In the context of social media, the level of precision that is obtainable varies considerably depending upon the variable or feature being measured. For example, if we are interested in understanding the timeline of posts related to a specific topic, most social media platforms give researchers access to timestamps associated with each post. Those timestamps can be aggregated to any level needed for the analysis (minute, hour, day, month, year). On the other hand, location has much more variability. The most precise way to determine the location of a post is with precise latitude and longitude information. If this is not available, however, then the best level of precision available might be at the city, state, or country level. Such variability in precision for most measures may require researchers to adjust their research questions to coincide with the level of precision available to them.

Precision relates to a measure's reliability because one way to reduce the level of measurement uncertainty, or unreliability, is to reduce its precision. Perhaps ironically, measuring constructs with less precision often leads to greater reliability. For example, when trying to classify users by age, accurately classifying or labeling age in specific years, or even decades can yield a highly unreliable measure. Human coders -- or an automated classifier -- may have more reliability when classifying users into larger age bins. Thus, the measurement question pivots from how reliable a measure is to how precise it can reliably be. This can be a hard tradeoff for researchers to make, because decreasing precision, even if it means increased reliability, can reduce the explanatory power of the variable in the analysis in which it is included. In the context of feature engineering, it is very common to have features with a large number of missing values to maintain precision, and see if the more precise values that are present help with the machine learning task. It is also common to aggregate to the lowest level in which there is enough representation in each category or group. For example, we may group age data to make sure each age bin (group) has a minimum number of users in each group, so that our models are not biased against one subgroup. While the precision of the feature is reduced, similar to measuring a construct, it may be a more reliable feature for the machine learning task than the original more precise feature.

4.2. Validity

In the social sciences, the term validity applied to measurement is short for "construct validity." It refers, in general, to whether our measures correspond to the things we intend to measure, our constructs. This is generally assessed qualitatively by subject matter experts. It is not something for which there is a single mathematical test. A measure may be very reliable and score highly on quantitative reliability scores, but simply be measuring a different construct than the one the researcher intends to measure. For example, an instrument that measures weight instead of density may still always read 10 pounds when measuring the same object, making it reliable -- but it would still not provide a valid measure of density. Thus, a measure must be reliable to be valid, but a reliable measure may not be valid with regard to the construct in question.

There are several potential validity issues that arise when using social media data for social science research. For instance, several of our groups talked about validity issues in measuring the sentiment of social media posts. Consider the implications of measuring the sentiment of posts in response to Ruth Bader Ginsburg's death. A computer scientist might view such a measure as simply what it is -- an estimate of the sentiment, positive, negative, or neutral perhaps, of a post, but a social scientist may view the measure as a proxy for how people actually feel about Ginsburg's death or their psychological state about her death. The key question is whether our only goal is to measure the sentiment of the post or document, in which there is very little or no difference between the concept and measure, or whether our goal in doing this is to measure the author's underlying sentiment about the topic. If the goal is the latter, then concepts such as construct validity, reliability, and bias necessarily come into play.

In short, if we use social media to determine how people that post in a particular environment really feel, think or behave, versus simply asking about people's social media behavior, then there are likely threats to measurement validity that we must address. For example, there is some evidence that people's expressed sentiment is more extreme online than in person (Keeter et al., 2015). This could increase the variance in population samples and

increase bias in some circumstances. There is also the danger that, with regard to certain topics or within certain online communities, there might be social encouragement to express a mostly positive or mostly negative sentiment -- a phenomenon that could introduce serious social desirability bias in social media measures of sentiment. In these examples, people's social media personas may not always match their true psychological states, and this mismatch threatens measurement validity.

The standard approach among computer scientists and social scientists to establishing the validity of a social media measure is to compare it to a "ground truth" source -- a data source outside of social media that is considered a true or valid assessment of the construct. For example, a measure of political affiliation created from social media posts could be compared to survey responses from the same users, automated classification of sentiment could be compared to expert human coding of a single set of posts, or automated age classification could be compared to self-reported age on a platform among the subsample who report age. The ground truth is not necessarily free from error itself, but provides a benchmark against which to gauge the validity of a social media measure. This process also applies to feature engineering as computer scientists define it, although it is not typically used to assess feature quality. As computer scientists move more toward building models that map to different social science constructs, validity of features needs to become a more central consideration. Currently, when computer scientists build features, they focus less on validity because they assume the model will ignore the feature if it is irrelevant. However, spurious relationships can exist and more care may be needed during feature construction to ensure certain minimal levels of validity.

4.3. Bias

Bias occurs when a measure misrepresents, in its central tendency the true nature of the construct in a systematic way. Bias could be present in social media measures in several ways. One way is through social desirability. We know that, in survey research, people are more likely to admit to socially sensitive behaviors when they answer anonymously. For example, Tourangeau & Smith (1996) found people were more likely to admit to various types of sex acts when survey questions were asked in ways that were more anonymous and reduced personal interaction with the interviewer. By this rationale, it is possible that platforms with anonymity could allow more honest expressions of people's true opinions and behaviors.

However, even when people can be anonymous, other things may swamp the effect of anonymity. For example, because people are not being asked questions about their behavior, as in a survey, they may only express opinions or mention behaviors that the forum they are in or their group of friends/followers lead them to think is appropriate. Physiologically, many social media platforms are also designed so that getting interactions on one's posts produces a small dopamine release which leads people to come back to the site repeatedly (Turel et al., 2014; Elhai et al., 2017; He et al., 2017; Doucleff & Aubrey, 2018; Parkin, 2018). For all these reasons, the number and content of posts may be biased by people's desire to elicit engagement with their posts, especially but not exclusively positive engagement.

In some other circumstances, it is possible that an unrealistic lack of social constraints may cause people to express themselves on social media in more extreme and uncivil ways than elsewhere, producing a different form of bias. Pew Research Center (Keeter et al., 2015) ran a large experiment in which they randomly assigned half of the sample to be asked political

questions on the phone, and the other half to answer the same questions via a self-administered web survey. The study found that questions answered on the web produced more extreme and more negative political opinions across a range of questions, and more denial of the existence of racism, sexism, and homophobia. This is consistent with a long literature in survey research that finds that people give less socially desirable answers when they are expressing themselves in more anonymous platforms and in modes where they are not anonymous but still more distant from the audience (Traugott & Katosh, 1979; Tourangeau & Smith, 1996; Groves et al., 2009). It is not clear, in these circumstances, whether we should view how people express themselves more or less online as the truer measures of their views and behaviors. Possibly, whether the more anonymous or more personal version of opinion or sentiment is the truer expression varies across topics. With some concepts, there may not be one truth: although people may express themselves differently on social media than in other forms of communication, both versions contain a "truth". Even still, the differences between social media expression and other types of expression must be considered when constructing measures using social media data -- and should continue to be studied.

These same types of biases may exist when computer scientists construct features for machine learning algorithms. The difference is the goal. Because computer scientists are interested in building accurate models, they may use features that contain bias. In some cases, depending on the objective function, that may be reasonable. However, in other cases, these biases may lead to discriminatory outcomes. For example, Amazon built a machine learning model to identify applicants they should interview using existing employees as their training data. Because their hires were mostly male, however, the features learned discriminated against women (Dastin, 2018). A new subfield of fairness in machine learning and algorithmic bias has emerged in the last few years in response to these biases. It is a large concern and will be an important discussion during the Modeling meeting.

4.4. Coverage

Measurement problems that stem from issues of coverage represent a special case of measurement bias. Coverage is a concept traditionally used in survey research that indicates how well the sampling frame (i.e. the group of people from which a sample was drawn) maps to a target population, with a dearth of certain types of individuals in a sample producing undercoverage and an excess of other types producing overcoverage (Salkind, 2010). Coverage is also a prevalent concept in computer science within the information retrieval, machine learning, and software engineering communities. In the context of information retrieval and machine learning, coverage has to do with how well an algorithm identifies the relevant examples or instances. For example, suppose we use a search engine to identify all the webpages related to coronavirus. If one algorithm can identify a larger number of relevant webpages than another, it has better coverage. This type of coverage is measured using the concept of recall, where recall is defined as the number of relevant examples found divided by the number of relevant examples that exist (Manning et al., 2008).

When using social media data, problems with sample coverage (which itself is the focus of our next meeting and white paper) can produce a particular type of measurement bias: measurement coverage. That is, poor sample coverage can undermine how well a particular measure from social media, such as sentiment on a topic, topic variation, or stance on an issue, describes or covers the range of values of that construct that exist on the platform (or, in the

world). If a social media sample reflects only those who score, say, very high or very low on a measure and not those who would score at the other end of the distribution, because its users or those who post on a topic are systematically different from the general population, the measurement scale itself will be biased. This problem has important implications for social media data collection methods and will be part of the discussion in the forthcoming Data Acquisition and Sampling white paper.

In addition to issues that arise from sampling, measurement coverage issues can arise from insufficiently identifying relevant social media content. If we think about the task of identifying all tweets related to a social issue like gun violence, how should the researcher proceed to identify all relevant social media posts? Social media platforms have restrictive limits for free data collection through their Application Programming Interface (API) (see Data Acquisition and Sampling white paper, forthcoming). APIs yield an unranked mix of relevant and irrelevant results (e.g. tweets about someone's bicep muscles (guns) and gun shooting will be returned for an API call for the keywords "gun" or "guns"). Furthermore, the language on these platforms changes over time, making it difficult to pinpoint the set of queries to use to get the best coverage. This general approach typically employs a dictionary of relevant terms which can be generated by creating a predefined set of keywords manually (Barberá et al., 2015; Bozarth et al., 2020), or creating a more dynamically adjusting dictionary using machine learning (Linder, 2017; Magdy & Elsayed, 2014).

4.5. Measuring the Quality of Algorithms in General

The primary mechanism for creating, or engineering, measures from unstructured digital traces in computer science is the construction of algorithms for classification. At a basic level, computer science has standard measures for evaluating algorithms that assess both their correctness and efficiency. Correctness is typically proven mathematically through an assessment of the extent to which for all possible inputs, the output is correct. For example, if we have an algorithm that sorts a set of numbers, we want to show that for any set of numbers, the output will always be an accurate set of sorted numbers. There are many techniques for proving correctness, including proof by induction, proof by contradiction, and proof by example. Efficiency is measured by determining the execution time and memory usage of an algorithm. An algorithm that is both correct and efficient is considered a good algorithm.

Along with these fundamental measures of algorithms, computer scientists construct measures to understand the nature of systems. These include the behavior of systems (e.g. how do we measure the impact on the network traffic of machines in the network being attacked or randomly shutting down?), of users interacting with a system (e.g. how do we measure the reliability of our system given the user usage patterns?), and of users (e.g. how do we measure information spread when we consider different broadcast methods?). In a sense, correctness of algorithms is a parallel concept to reliability of measurements in social science, i.e. always guaranteeing the same output. However, these are distinct concepts. Algorithms may be designed for different purposes, and depending on the use of the algorithm, computer scientists would not necessarily further evaluate the algorithm in terms of reliability, validity, or bias in the way social scientists define those terms. For example, if an algorithm is designed to measure a construct, then these concepts are very important. However, if an algorithm is designed to execute a process that is not connected to social constructs, like generating random numbers or identifying vulnerabilities in a system, these concepts may not apply.

4.6. Bots, Fake Information, and Posers

A final issue to address in creating measures from social media data is that not all users are actual people. Thus, not all posts or views reflect real human behavior. Organizations create accounts that are automated -- often called "bots" -- to mimic real people and in doing so achieve other ends, such as driving traffic to a website, influencing social media conversation, or even changing attitudes through the spread of misinformation (Bessi & Ferrara, 2016). There are also "posers" who are not bots but rather people using real posts to manipulate others' attitudes or behavior and/or spread misinformation. To address these authenticity problems, researchers need to use validated methods to distinguish bots and posers (Davis et al., 2016) from authentic posts and views. A number of methods have been developed to identify bots. The identification and spread of misinformation have also received a growing amount of research attention (Briones et al., 2012; Broniatowski et al., 2018; Dredze et al., 2016; Guidry et al., 2015; Oyeyemi et al., 2014; Sharma et al., 2017; Singh et al., 2020), and this misinformation can be spread by bots and humans. Misinformation is an important concern across different online domains, such as politics (e.g., Budak, 2019) and health (e.g., Gunaratne et al., 2019). The presence of inauthentic data undermines the reliability, validity, and unbiasedness of any measure researchers might create.

It is important to note that the challenges associated with measurement described earlier are just as relevant when identifying bots, misinformation, and posers. There is significant disagreement in what makes a piece of content false information and what makes a content provider a producer of false information. Take, for instance, the numerous lists of web domains classified as fake news producers (e.g. Poynter Institute, 2019; Zimdars, 2016). These lists are produced by reputable news organizations and/or scholars. Yet, there is significant disagreement among them (Bozarth et al., 2020). Our scholarship depends on these lists to carry out important tasks such as building automated fake news trackers (e.g. Shao et al., 2016), assessing agenda-setting powers of fake and traditional news sites (e.g. Vargo et al., 2018), assessing the impact of disinformation on election outcomes (e.g. Allcott & Gentzkow, 2017) and examining fake news trends (Allcott et al., 2019). Therefore, it is important to acknowledge this potential shortcoming of social media and work towards solutions to address measurement errors resulting from bots, posers, and fake information spreaders.

5. Levels of Measurement in Social Media Data

An important and innovative feature of social media data is their ability to support the measurement of constructs at multiple levels of analysis. Every post shared provides information about the post itself, as well as characteristics of the post's author. When aggregated, posts can provide information about the group within which the posts and their authors are situated and the population or subpopulation of the platform as a reflection of the population at large. Thus, one dataset drawn from social media can provide measures of constructs at four levels: post, user, group, and population. And they can do so in real-time, over time. A social science project using traditional surveys or observations would typically need to conduct multiple studies with fundamentally different designs to yield longitudinal measures at multiple levels. We now briefly define each of these levels of measurements and provide some examples of different types of measures that can be constructed at each level.

5.1. Post-level Measurement

First, consider post-level measurement. At the level of a post (or tweet or share), measures can provide a close examination of the content of the text or image. Measures could include the *topic* of the post, the *emotion* or *sentiment* conveyed in the post, the *stance* or opinion expressed, and metrics about the post such as the time of day it was issued and the frequency with which it was liked (upvoted, etc.) and reshared (retweeted, etc.). A lot of research from both computer and social science uses post-level measures to understand social phenomena. For example, the topic of posts has been used to study the relationship between online conversation and protests related to the #BlackLivesMatter movement (Williams et al., under review), the emotion of posts about vaccination has been used to assess how anti-vaccination groups influence online conversations (Massey et al., 2016; Salathé & Khandelwal, 2011), while the stance or opinion in a post has been used by our own political communication group to identify and understand pro-Trump and pro-Biden tweets during the 2020 election cycle (Singh et al. 2020). Recently, researchers in political science and communications have paid increased attention to subtler dimensions of deliberation relevant to the nature of online discourse, creating measures of toxicity (Wulczyn et al., 2017), civility (Borah, 2014), reciprocity, trolling (Cheng et al., 2015) harassment (Blackwell et al., 2017), and hate speech (Mondal et al., 2017; Mossie & Wang, 2020). These initial post-level measures can then be used to create supplementary measures about the posts that can answer novel questions about the flow or influence of different post types -- amplifying the utility of the topic, sentiment, stance, or deliberation measure to answer novel social science questions.

5.2. User-level Measurement

User-level measures are also essential to create when social media data are used in social science research in part because many relevant questions require knowing the demographic characteristics of those generating the data. The way most current research does this is by constructing variables/features using some combination of user profile information, post content, and images to estimate users' demographics. Given their importance to social science research, inferring demographic characteristics such as age (Schler et al., 2006; Rosenthal & McKeown, 2011; Al Zamal et al., 2012; Chen et al., 2015), gender (Chen et al., 2015; Al Zamal et al., 2012; Sakaki et al., 2014; Taniguchi et al., 2015), race (Rao et al., 2011; Culotta et al., 2016) and education level (Culotta et al., 2015; Culotta et al., 2016) has received a lot of attention. For example, social scientists have used various crowdsourcing platforms, such MTurk or Appen (Appen, 2020; Amazon Mechanical Turk, 2020), as well as image recognition algorithms, such as Face++ (Face^{++}, 2020) to classify users by age and sex (Vikatos et al., 2017; Zagheni & Weber, 2015). Researchers have also classified users' occupation and age on Twitter using pattern matching algorithms that tracked salient words related to pre-specified occupations and phrases related to birth dates (e.g., Sloan et al., 2015; Sloan et al., 2013; Schler et al., 2006; Rosenthal & McKeown, 2011; Nguyen et al., 2011; Al Zamal et al., 2012; Zagheni & Weber, 2015). Although much of the research on user-level measures has addressed demographic characteristics, social media data have also been used to generate reliable measures of subtler individual characteristics such as political affiliation (e.g., Barberá, 2015; Conover et al., 2018), mental health (e.g., De Choudhury et al., 2013), and opinion (e.g., Lee et al., 2020), by combining information from the text of posts, the users' network, and the users' usage/posting behavior on a specific platform.

5.3. Group-level Measurement

The third level of measurement is "the group." A novel and potentially useful feature of social media is the ease and frequency with which people sort into identifiable groups. These groups can be constructed in different ways. For example, Facebook and Twitter allow users to create public groups and anyone on these platforms can join and post on group channels. These are explicit groups, where a group page exists, and membership can be determined based on a list or posting on the group page. Example groups include sports team fan groups (e.g. Baltimore Ravens fans), health-related groups (e.g. cancer survivors), and social issue groups (e.g. BlackLivesMatter). Other ways to construct groups are based on shared demographic (e.g. age), location, real-world groups (e.g. students from the same high school, or people who are members of a specific church), or shared behavioral characteristics (e.g. length of participation in an online group). While explicit groups are clearly identifiable on social media, inferred groups must be constructed by researchers.[11] Inferred groups that are not manually constructed can be constructed mathematically based on the connectivity structure of users on a social media platform. For example, we could represent social media connectivity as a network, where each user is a node and each connection to a friend or follower is an edge. Clustering or graph partitioning algorithms (Clauset et al., 2004; Girvan & Newman, 2001; Tang & Liu, 2010; Adamic & Glance, 2005) could be used to find dense regions in the network or similar regions in the network, and groups could be defined based on these mathematical clusters. Some examples of inferred groups created using network structures include basic friendship networks, shared content networks (reposts/retweets), opinion of shared content networks (post like networks), and shared purpose networks (adversarial, bots).

There are validity, reliability, and coverage issues specific to group construction. When the group is inferred, there is no ground truth to determine if the inferred members of the group are in fact members of the group. In explicit groups, membership is continually changing and structural changes may not be part of data collection. Finally, as researchers, we are only getting information from those who explicitly join the group and participate in the conversation, thereby potentially missing others who are reading the posts and engaging in discussion about the posts outside of the group structure.

Using the profiles, posts, shares, and activity of the members of a group, researchers can generate measures relevant to the groups' behavior, attitudes or internal dynamics. Examples of these group level measures include: (a) static and dynamic group characteristics, e.g. the number of group members (total or per day), the demographics of group members, and the average participation for members, (b) static and dynamic group content characteristics, e.g. the overall number of posts (total or per day), the most frequent words used, and the topics of posts, (c) group network characteristics, e.g. the type of network (small world/random), network clusters, and network centrality measures (degree, betweenness, eigenvector), and (d) information network/flow characteristics, e.g. the method of information spread (broadcast or peer to peer), the speed of information spread, and the key information spreaders. These measures serve as the basis for answering broader questions about group opinion, the changing or constant properties

[11] It may be the case that the inferred group is determined manually or computationally. But the key difference between the explicit and the inferred groups is that the researcher must determine how to construct the inferred one.

of groups, interaction among group members, influencers within a group, and the types of information shared and how they change through time, to name a few.

There is a substantial body of work about group level constructs. For example, Woolley and colleagues analyzed the discussion of political groups that were created during the 2008 Obama versus McCain election, creating measures for group membership, activity level, and the sentiment and use of profanity towards the candidates (2010). Bender and colleagues characterized the types of conversation taking place in 620 breast cancer groups on Facebook to understand the purpose and use of these groups (2011). Mishori and colleagues characterize the community structure and information flow of four medical associations, e.g. American Medical Association, and their followers using medical association accounts, followers, and who the accounts follow (Mishori et al, 2014). Finally, Rajadesingan and colleagues investigate how levels of toxicity are maintained in political subreddits (Rajadesingan et al., 2020).

Studying groups with social media data allows researchers to examine substantially larger groups in far more detail than traditional social science methods that would typically rely on qualitative methods like ethnographies or small-scale surveys to capture group dynamics. As previously mentioned, group dynamics are hard to capture using traditional survey methodologies. Structured observation may not capture a group's dynamic because the identity of the group itself is not initially known. Also, while survey participants may know explicit groups they participate in regularly, they are not likely to know the extent of their own inferred group memberships as reflected by clustering among the people they follow or the extent to which their content includes a given hashtag. Social media data is well suited to understand group level dynamics.

5.4. Population-level Measurement

Social media data are particularly well-positioned to identify and track population-level phenomena. Because people post in real time and at very high frequency, the data can be used to identify trends, turning points, and new events with regard to any topic discussed online, including economic activity, political movements and opinions, cultural attitudes, and even well-being at the population level. Population measures include both the *prevalence* of an existing or known phenomenon and the *emergence* of new or unknown events. Prevalence measures would indicate the salience of an attitude, topic or behavior throughout a platform and be used to gauge population level metrics. For example, Twitter data has been used to create labor market indexes, including job loss, job search, and job posting, by deriving signals from job-related phrases in tweets such as "lost my job" (Antenucci et al., 2014). These real time indicators have then been used to study the immediate economic impact of events such as Hurricane Sandy and the 2013 government shutdown.

Social media data can also reveal shifting topic salience in a population. For example, using a large sample of tweets containing words related to homeschooling and distance learning, our team was able to detect and track spikes in discussion about homeschooling and distance learning throughout the novel coronavirus. Much work has been done to understand the impact of online social movements. For example, Williams and colleagues studied the role of the #MeToo as a catalyst to broader social change by measuring the volume of specific keywords and hashtags and correlating them to events like strikes, legislation, and EEOC claims (Williams et al., 2019). While, in theory, it might be possible to track these phenomena with survey data, it

would require large high-response-rate surveys constantly in the field, asking a large number and changing mix of questions about recent events. Measuring these societal-level variables with social media data is far more practical, if its challenges can be overcome.

Social media data may also be powerful as a data source for identifying leading indicators or the emergence of phenomena that would be difficult to capture using traditional measurement methods. When possible, data gathering to capture emerging phenomena can be conducted using qualitative or open-ended methods, but this work requires a large sample and thus poses problems for identifying phenomena. Moreover, traditional methods, especially closed-ended survey questions but also open-ended questions and other qualitative research techniques, all to varying degrees require the researcher to know in advance what they are looking for and design measures to capture it. This is less of a problem when using social media data. Emerging concepts in social media data can be identified inductively, by detecting novel patterns. Once identified, emergence can be tracked using additional social media data, more traditional social science measures, or even other metrics. Ramakrishnan and colleagues use social media, blogs, and other data sources to forecast civil unrest across ten countries in Latin America. Their work is an example of the power of social media data to capture an emerging trend, particularly in places where it can be dangerous to collect data (Ramakrishnan et al., 2014).

As with measures at other levels, population level measures drawn from social media can be biased by the select population of users on any particular platform. When constructing population measures, however, the high frequency and volume of the data yield unique benefits for tracking trends over time, if not precise population estimates of phenomena at one time. For example, Antenucci and colleagues (2014) found that although the signals for job loss identified in Twitter posts were rare, they were able to gather so many tweets that the signals still provided a rich dataset that could be used as an indicator of weekly job losses when linked with official, weekly Unemployment Insurance claims. In each of these population level examples, the power of social media data lies in blending them either directly or indirectly with information from outside social media as a benchmark, and then using them to glean unique insight into important events and trends that affect entire populations.

5.5. Measurement Level in Computer and Social Science

One distinction that emerged during our meeting was the different emphases computer and social scientists place on measures at each level of analysis. Although social scientists ask questions about measures at all four levels described above, much social science research employs social media data measures at the post- or user-level. By contrast, most computer scientists use social media data with measurements typically at the group or population level. Computer scientists tend to think about how the measures they construct can be applied at scale, whereas social scientists often ask questions about individual behavior that would ideally, although not necessarily, be applied to a large sample. Of course, there are times when computer scientists think about individual level measurements. For example, anomaly detection and security breach identification are application areas where individual level measures may be designed to understand more unique characteristics of systems and users. Likewise, there are times when social scientists want to measure group or population level phenomena, such as labor market flows or civil unrest. The different emphases of computer and social sciences, however, offer in themselves the opportunity to expand the measurement methods, and research possibilities, of both disciplines.

6. Methodology

A central challenge in using social media data for social science research is converting unstructured data that were not generated for that purpose into structured measures (or features) that are more useful for analysis. There are multiple ways to approach that task. At our meeting, we discussed them in terms of four overarching methods: dictionary-based methods, machine learning methods, measurements describing social media system dynamics, and aggregate measures with external benchmarks. This section presents an overview of each method, provides examples of measures using these methods, and discusses issues and recommendations regarding reliability, validity, bias, and measurement coverage that arise when using each.

6.1. Dictionary Based Methods

Overview.

Many constructs social scientists want to capture using social media data rely on lexical analysis. Among the most frequently used methods for this are dictionary-based methods, in which a list of words and terms indicative of a construct (e.g., positive and negative word lists for emotion or sentiment measures) is compiled and used to label or classify a set of posts, users, etc. As this suggests, dictionary-based methods can be used at several different levels of measurement, including the post-level, user-level, group-level, and population-level. Dictionary-based methods are already fairly widely used to measure a variety of different constructs in social media posts. It is reasonable to consider their measurement quality, in order to compare more computation-heavy methods against them. Given their current prominence, they serve as a baseline for evaluating whether or not more automated methods can succeed in providing superior measures of key concepts.

Examples.

Two related constructs that researchers often want to measure in social media posts at different levels of aggregation are emotion and sentiment, which typically employ dictionary methods. We gave substantial attention to discussions of measuring both at the meeting. When measuring emotion, researchers usually want to detect a set of different emotions. At other times, researchers need to measure sentiment, which involves classifying text on a one-dimensional sentiment scale running from positive-to-neutral-to-negative.

From popular text-based social media platforms (such as Facebook and Twitter) it is common to use a “prevalidated” dictionary such as LIWC (Linguistic Inquiry and Word Count) to classify posts according to the “emotional” category of the words in them (Tausczik & Pennebaker, 2010). In video-based social media platforms such as YouTube and TikTok, an existing strategy for emotion measurement of posts is using facial recognition software to evaluate the “emotions" conveyed in the videos. Another approach, and one that is used in both text-based and video-based social media platforms, is to measure reactions that the post provokes from users on the platforms: i.e, the types of reactions generated by impressions (that is, viewing the post). In doing so, a researcher can look at both the nature of the reactions (the text of comments), or the volume of reactions, or even the ratio of negative comments to likes.

Both LIWC and facial recognition software face validity and reliability challenges that need to be addressed and overcome, however. First, both dictionary and facial recognition algorithms are developed by validating them with pre-existing datasets. Sometimes when we

apply them to new circumstances, they no longer measure emotions and sentiment accurately. For instance, this issue came up in our research group when measuring sentiment of media content and survey responses during the fall 2020 election campaign. The word "positive" in posts was leading them to be coded with a positive sentiment score. However, many of these mentions were people discussing President Trump or other prominent people testing positive for COVID-19, which we think we can reasonably infer was not intended to convey positive sentiment. So, in our case, the dictionary used needed to be modified for the specific circumstances of the COVID-19 pandemic and the new, more frequent, connotations of the words "positive" and "negative".

To some degree, this problem may not be a limitation of certain dictionaries or even of certain automated text analysis methods, but an inherent limitation of language, whose usage and meaning changes over time and across different contexts. Put more simply, words may not convey consistent emotions, presenting a potential problem for all dictionary methods. Our own example of adjusting our sentiment dictionary during the COVID-19 pandemic is an illustration of how sometimes a potential fix for validity in new circumstances requires manual or semi-automated monitoring and adjustment. For example, our team is building tools to identify the contexts of different phrases, thereby allowing scientists to see when contexts are changing. We can imagine using a tool containing these methods as a way to highlight possible context shifts, enabling social scientists to adjust dictionaries without having to manually code large volumes of posts continually. Considering interactive, semi-automated methods that use automation when it is useful is an important interdisciplinary direction. Of course, social scientists have a long history of using human coders to classify media content. In these methods, validity and reliability were tested by checking the inter-coder reliability scores (i.e., the correlations of codes across human coders).

Recommendations and Considerations.

While current methods for coding emotion from text are fairly widely used and usually capture emotion as intended by the researcher, substantial challenges remain for these methods for coding emotion in social media posts. In our meeting, we developed the following suggestions for improvement. First, when humans get involved in coding decisions, it is useful to have multiple people review coding decisions. When all coding of media content was done by humans in decades past in the social sciences, it was, as mentioned, common to calculate interrater reliability. This principle should extend to decisions by the researchers themselves when they guide or adjust dictionaries and algorithms throughout the data analysis process. It is useful for additional researchers to offer their opinions on adjustments "blind" to the adjustment decision preferences of the researchers on the project, to get independent verification of decisions made. Second, it may be necessary in some circumstances to reduce a high dimensional emotional coding system to fewer emotions---i.e., reduce the coding to coarser groups. For instance, you may decide, in some circumstances where you originally intended to code 5 or more emotions, to instead drop down to a one-dimensional sentiment analysis if the higher dimensional analysis is producing low validity and reliability. And finally, when a dictionary is transferred to a very different time or circumstance, it is always important for a subsample of the coded data to be checked by a human for validity.

6.2. Machine Learning Methods Involving Human Coding

Overview.

Many measures that social scientists want to extract from social media data do not map clearly onto a dictionary of words or terms for classification. These constructs may tap the intent, meaning or nature of a post or user that require complex judgments. The first step to creating these types of measures is to hand code a data set with the target values. If an analytic sample is small enough, hand coding may be the only step required. However, if researchers aim to code a large data set, the hand or human coded dataset can be used to train a machine learning (ML) algorithm. ML is a branch of artificial intelligence that uses previous observations in the form of labeled data (training data) to build a model that can classify or predict an outcome or one of a set of values (Kelleher & Tierney, 2018). The training dataset contains both the features of the data that are inputs (features of a post or user), and the output of those features (the label or score applied by coders). The algorithm can then use these examples to build a model that can be used to classify unlabeled data. As a reminder, the raw data is used to construct the set of features that are used to train the model. The features describe the structure of the data set. Good features represent the data well. Sometimes, even with good features, there is not enough information to produce a strong model. In these cases, it is necessary to find other features. Computer scientists tend to approach this problem by considering external information (e.g. contextual dictionaries or ontologies), considering different dimensionality reduction techniques to build features with different mathematical properties (principal component analysis or wavelets), or labeling more data. All of these are ways to improve feature coverage. The relationship between features and models will be discussed more in the modeling meeting; however, if the features created are not leading to accurate classification results, then the researcher needs to consider adjusting the set of features, the model, or both.

Examples.

As an example of using machine learning, consider how we might create a measure of stance or opinion on states' stay-at-home orders (SAHOs) in response to the COVID-19 pandemic. We could start by obtaining a sufficiently large sample of posts on a platform, say Twitter, about SAHOs using a list of keywords that would denote a post about SAHOs. This initial step requires ensuring the keyword list is reasonable enough to get enough examples of each position (for, against, neutral). These examples are hand-coded in terms of their stance toward SAHOs -- for example being for, against, or neutral to SAHOs. Standard practice in social science would be to first create a codebook defining the labels (what constitutes being for, against, or neutral about SAHOs) to ensure that the human coders are labeling both reliably and validly. Second, multiple coders must label a common set of posts until they reach an acceptable interrater reliability, because establishing that the construct *can be measured reliably* is essential to establishing the validity of the resulting measure. Once there are a reasonable number of hand-coded examples for each stance category, these examples can be used as training data to build an ML stance classifier. Once built, the classifier is then tested for validity -- or accuracy -- on an additional set of hand coded cases (a "test" data set) before using it to classify a large sample.

Computer scientists and social scientists often approach this general process quite differently given their orientations toward establishing measurement reliability and validity. For example, consider how each has approached measuring the quality of deliberation on social

media. They are looking for posts that contain elements like toxicity (Wulczyn et al., 2017), incivility (Borah, 2014), trolling (Cheng et al., 2015), harassment (Blackwell et al., 2017), or even hate speech. Computer scientists have been developing machine learning models to determine toxicity or incivility at scale (e.g. Chandrasekharan et al., 2017; Wulczyn et al., 2017). One of the most commonly used automated approaches to identifying toxicity at the post level is Perspective API by Google (Wulczyn et al., 2017). This classifier is trained on data that was gathered from human judges who were asked one simple question without a detailed codebook or any training. Although computer scientists often use multiple coders, and in some sub-disciplines compute interrater reliability, their focus is less on the reliability of the coders and more on the correctness and accuracy of the classifier.

Contrast this with the more typical social science approach. In the social sciences, researchers studying toxicity or incivility in deliberation have introduced various codebooks that share some similarities but also have important distinctions (Coe et al., 2014; Gervais, 2015; Stryker et al., 2016), highlighting the importance of defining elements of deliberation before assigning codes regarding the constructs. The typical process is then to train coders on a particular coding scheme until they reach a predetermined level of reliability, generally measured through interrater reliability measures. There are also approaches that establish validity by calibrating human codes to expert judgments, often called the "gold standard" coders, and evaluating performance with respect to that gold standard or "ground truth" dataset. Even when relying on workers from crowdsourcing platforms such as Amazon Mechanical Turk to code posts, social scientists typically include a sufficient number of cases for which there are ground truth labels to detect coders who are making mistakes to filter such coders out.

Recommendations and Considerations.

There are a number of important measurement issues and considers that arise when using machine learning methods to determine measurement values. We begin with level of precision human versus automated coding can reliably achieve. For instance, researchers studying SAHO stance might hypothesize that attitudes toward SAHOs follow a more varied distribution than for, against, or neutral, such as a 5-point scale with "strongly" for and against as anchors and "somewhat" for and against as less extreme codes. A survey could easily capture these distinctions. A machine learning classifier could as well with a large sample of training data, particularly if the data were less noisy than posts. However, given the noisiness and incompleteness of social media posts, both humans and algorithms have more difficulty labeling users. This means that sometimes more "blunt" measures (and noisy features) need to be used as opposed to more detailed ones. For example, considering only three categories for stance instead of five.

Another measurement issue is the performance of a classifier. Any one classifier can vary substantially across contexts, samples, and time periods—so ongoing validation of the classifier's accuracy is essential to establish both the reliability and validity of these supervised learning methods (Grimmer & Stewart, 2013). For example, would the algorithm trained on a sample drawn from the first few months of the pandemic perform consistently over time in labeling the stance of SAHO tweets given that words and lexical features associated with SAHOs shifted with events? When using a trained algorithm to classify attributes, it is often important to re-establish the model's accuracy and reliability regularly to account for shifts in language over time. In some domains these shifts do not occur often, but in others, like an election or a

pandemic, they may be frequent. Likewise, it is important to recalculate reliability regularly throughout the measurement and feature engineering process so that any changes in the phenomenon under study can be reflected in the measure.

Bias is also a concern with this methodology. Any bias in the population of people who talk about this topic on social media may also bias the ML algorithm and resulting measure of SAHO stance. For instance, those with very extreme opinions, either for or against SAHOs, may be far more likely to tweet about SAHOs than the typical Twitter user. This response bias would impact the generalizability of the estimated stances on SAHOs generated from the Twitter data -- within the population of Twitter users and outside of it -- and it could also bias the nature of the classification process itself if more extreme words, phrases or lexical features are used by those who post often than those who post seldom or not at all on the topic. Strategies to address this bias include comparing the post content of users who post often about SAHOs to that of those who post infrequently in terms of language and other lexical features or limiting the number of posts by a single user in the training data set of the classifier. Another alternative, but more costly approach is to create a ground truth data source that could be drawn from a survey of a subsample of users who are asked directly about their stance toward SAHOs and linked to their social media account. Then the post content of those with different stances could be compared to identify sources of bias in the ML classifier.

It is also important to consider that the quantity of data analyzed can make it challenging to provide human guidance to automated methods. With the types of massive datasets produced by social media, human supervision usually involves human audits of small samples of the automated coding. But it is always possible that the conclusions drawn by human researchers from this small sample might not be indicative of how the coding works in the entire dataset, leading the researcher to change the coding method in ways that do not improve the measurement. Finally, it is always possible that, when humans adapt measurement algorithms to new circumstances, they will end up overfitting the data---that is, capitalizing on chance and not producing an algorithm that would produce valid coding (of sentiment or stance, for instance) on new text, even if it was not a new situation where the meanings of words had changed. Computer scientists do work on algorithms and machine learning models that are more adaptive, adjusting for changing concepts or changing behaviors. There are well established techniques for identifying and adjusting machine learning models for concept drift in the underlying data (Wang et al., 2018). However, the changing dynamics of language can be difficult to identify and adjust for, making this a challenge for short, social media posts.

Measures that use hand coded datasets to train automated classifiers provide ripe opportunities for a convergence of computer science and social science methods. Particularly when the aim is to code posts or users at a large scale, social scientists need to develop sound practices for coding far more data than humans reasonably can manually. When using machine learning classifiers, it is essential that they are developed using reliable and valid training datasets. Finally, both computer and social scientists need to include processes to ensure ongoing reliability and validity, as social media content shifts over time (if they are going to continue to use the same classifiers), and to identify bias in the initial corpus of posts or users that might, when used as training data, invalidate the model built by the classifer.

6.3. Social Media System Dynamics

Overview.

A third way to generate measures is to look at the dynamics of larger groups and subpopulations and how information flows through these populations. We could view this as a system dynamics type of analysis. Data mining is a subdiscipline in computer science that is focused on developing methods and algorithms that identify patterns in large-scale data. Classic examples of data mining algorithms that have been developed for large-scale data include clustering, anomaly/burst detection, frequent pattern detection, and event detection to name a few (Tan et al., 2016). These ideas may be applied to classic well-structured data, text data, spatial data, temporal data, sequence data, and graph/network data. While it is not possible to go through all these techniques for computing different system level dynamics, all of these types of algorithms are useful for social media data. Some are more readily transferable than others.

Examples.

Instead of going through one or two specific algorithms, we focus on a specific example, and show how topic modeling and network clustering can be useful for understanding the sharing/reposting behavior of a group. What are the most popular types of information shared by group members through time? Because we are interested in the flow of information, i.e. information sharing, we need to construct dynamic content measures. We can view this system dynamics analysis as a multi-stage process that involves constructing post level and possibly user level measures and aggregating them to understand the system. Therefore, we begin by thinking about posts.

In order to understand what information is shared, we need to know the topic of each post. This may be done using a dictionary method, where domain experts identify the salient topics using frequently occurring words, and then manually identify words and phrases that are associated with each topic. Topics can also be constructed using machine learning methods if a set of labeled data exists. Finally, they can be constructed using topic modeling algorithms. Topic modeling algorithms generally assume that documents and words in documents conform to an underlying distribution and use the joint co-occurrence of words to generate sets of words and phrases that group together because they appear frequently together in documents within a document collection. The most commonly used topic models in the social sciences are Latent Dirichlet Allocation (LDA) (Blei et al., 2001) when no covariates are being used to generate topics and Structural Topic Models (Roberts et al., 2013) when there are covariates to help with the modeling. These algorithms do not require any labeled data, but are severely impacted by the presence of noise and the size of the vocabulary.

Once the topics are created, each post can be labeled with a topic or a distribution of topics based on the fraction of words that match different topics. A daily time series can be created by combining the individual topic distributions of each post. While that alone can give insight into the changing content, it is also interesting to see which posts are shared more than others, and what the topics of the posts are. We can use a normalized repost count or a normalized like count as a proxy for sharing. If the platform being used provides information about who shared the post with whom, we can create a network of the members in the system. If the system is a group, each node can be a member and each edge a link between members who repost or share the post. The edge can be weighted by the number of posts shared between pairs

of members, or the number of posts of a specific topic shared between members. Having this network structure allows researchers to see the clusters of nodes that share content with each other (friendship network) or the type of content that are being reposted or liked most often by subgroups in the group (reposting/like network). For example, are there subgroups that are more focused on specific conversation topics.

Once a network has been created, clusters can be identified using network clustering algorithms (Adamic & Glance, 2005; Clauset et al., 2004; Girvan & Newman, 2001; Tang & Liu, 2010). These algorithms partition a graph based on different optimization strategies. For example, the Girvin Newman edge betweenness algorithm removes edges with the highest betweenness values until the number of clusters desired is obtained (Girvan & Newman, 2001). In contrast, the Newman modularity algorithm looks for groups with a high modularity score, where the score is determined by measuring the difference between the actual number of edges in the cluster and the expected number of edges (Newman, 2006). We highlight these two algorithms because they are widely used. However, many algorithms have been developed by the computer science, physics, and sociology communities for clustering that are optimized using different network properties or designed for networks with specific connectivity structures (see Chakraborty et al., 2017 for a survey).

Recommendations and Considerations.

There are a number of validity challenges that are unique to computing system dynamic measures, many of which are not unique to social media. For example, when clusters are constructed, there is no ground truth. What may be more unique for social scientists is the scale of these clusters, making them harder to validate. For example, millions of users have engaged in social media conversations related to BlackLivesMatter and MeToo. Validation of those clusters may be confined to ensuring certain structural properties of the clusters, e.g. having clusters with high levels of cohesion, or by connecting the clusters identified to known external information about the groups or theoretical frames about online movements. Another example is with topic models or any large-scale system-level grouping of words. While the words associated with each topic can be manually validated by researchers, it is highly likely that there are words or subgroups of words that have been missed. There are millions of words used on social media and manually going through all of them is not possible for researchers. Also, words that are part of topics may have multiple meanings. All the meanings of the word may not fit with the topic. One way to alleviate that issue is to use phrases or synonyms or word vectors/embeddings to help validate the words associated with a topic. Once the topics are constructed, posts can be labeled with topics using the topic model and the accuracy of the topic model can be measured.

Because data are always changing on social media, the members of the system are also changing. This instability can lead to coverage issues. Therefore, to ensure that researchers know the membership properties, researchers may need to download the members daily. If that is not practical, it makes sense to measure them at the beginning and end of the time period being studied. This is important for understanding the fraction of the subpopulation that is participating in the conversation. Finally, if we are not careful about our measurement constructs, we can have issues with bias. For example, suppose we are measuring subpopulation properties, but only 5% of the subpopulation is participating in the conversation. We must be sure that we understand how broadly applicable the findings are to the entire system vs. a small fraction of the system.

6.4. Combining Social Media Measures with External Data Sources

Overview and Examples.

A fourth way to generate measures of interest to social scientists with social media data is to combine multiple social media features to create an aggregate measure or to combine social media traces with other data sources. In both cases, the goal is to predict complex social phenomena. Measures can be constructed in this way at the post, user, group or population levels provided that the data sources external to social media provide benchmarks at those levels. For example, at the user level, De Choudhury and colleagues (2013) were interested in assessing whether a person's behavior on social media could be used to detect and diagnose their mental health, specifically their experience of major depression. To do so, they quantified multiple features of users' social media postings over a year that were theoretically related to depression, including post frequency, emotion, language and linguistic styles, ego network, and mentions of antidepressant medications. They then used these and other social media features to test multiple ML classifiers to predict depression, benchmarked with survey data gathered via crowdsourcing that asked users whether and when they had ever been diagnosed with clinical depression, based on a standard psychometric instrument. The best classifier was built using a Support Vector Machine. Note, the modeling strategy for building an accurate classifier in this scenario resembles those described in Section 6.2 for use with hand coded training datasets, the key difference being the use of non-social media data to benchmark the outputs.

A particularly exciting application of this methodology is creating measures of population-level phenomena that would be logistically or financially difficult to gather using traditional social science methods. For example, Singh and colleagues (2019) used words and phrases on Twitter and in newspapers in both English and Arabic as indirect indicators of forced migration in Iraq. They focused on identifying online conversation topics as signals for specific displacement factors. For example, death count is traditionally a very reliable lagging indicator of forced displacement. By linking posts on violence to administrative data on deaths, the authors identify a leading indirect indicator of displacement. Specifically, they computed the frequency and emotion of posts about topics that are traditionally factors for predicting migration -- like deaths or conflict -- by determining the number of posts and sentiment associated with that topic in a particular location of interest, and the event volume by counting the number of events identified in newspapers, social media, and Wikipedia that also map to these factors. They compute features for both source and destination locations using geotagged tweets and mentions of locations in tweets. To validate their measures of migration, they compared these social media features to a common traditional variable: monthly conflict-related deaths, curated by Iraq-BodyCount.com. Once they identified meaningful features, they used them within a Hierarchical Bayesian model to predict migration to different locations in Iraq. They found that these predictions of large displacement were far more accurate -- that is, valid -- when social media data were used in combination with data from the World Bank, IOM, and UNHCR, then when they were not.

An example of combining social media data with an external benchmark comes from Antenucci and colleagues' (2014) use of Twitter data to create a measure of labor market flow, including job loss, job search and job posting. To do so, they gathered many repeated cross sections of tweets, or *k-grams* of a certain size, aggregated first to days and then weeks. They then identified *k*-grams about labor market flow by selecting signals that they believed were

indicative of each phenomena based on expert knowledge of the relevant terms and phrases. These signal lists then needed to be expanded to include variations on the phrasing or spelling of initial terms and then refined to exclude those terms that appeared so frequently (e.g., "let go" as a signal for job loss) that they were unlikely to identify the topic specifically. This process of expansion and then refinement demonstrates how some amount of coverage may need to be sacrificed to ensure a valid -- that is, specifically about the job market -- set of tweets about the phenomena. They then assessed how their measure of job market flow related to a standard measure of economic activity -- initial weekly claims for unemployment insurance (UI). They found that their measure of job loss and weekly UI claims moved strongly together both in the general trend and in some notable spikes, such as right after Hurricane Sandy in October 2012 and the October 2013 government shutdown. With this validation in hand, the authors were able to use the association between their social media job loss indicator and the official UI claims to build a prediction model of each week's new claims using social media posts with virtually no lag -- a powerful population level measure.

Recommendations and Considerations.

One advantage of measures created using an external benchmark is that the benchmarks themselves provide an assessment of any bias inherent in using a social media measure. As noted repeatedly, Twitter users are not representative of the US population -- or any other countries' population. Even within Twitter, many users will not post about topics of interest, such as job market flows or migration and conflict. As a result, any population level measure generated using social media data must be benchmarked against an accurate population metric. Thus using benchmarks in combination with social media data to generate more reliable measures or as ways to validate social media measures is a potentially powerful direction for future research. Further, with declining survey response rates, combining these data with smaller, cheaper surveys may lead to a more accurate understanding of a behavior or attitude. Finally, for these complex assessments, it is likely that a large number of features are being used within any machine learning algorithms. In some of the described examples, hundreds of variables are used within the learning process. We reiterate the need to view measures of constructs differently from features used within models. While features do need to have certain properties, we expect them to be noisier and less precise than measures.

6.5. Final Methodology Considerations

A final point to consider about all of these methods is that the main goal of this project is to improve social science research by blending computer science and social science methods in a purposeful way. Because the volume of data generated from social media is so large, we would like to make use of intelligent systems, while of course still iteratively improving models as analyses are being conducted. Within computer science, many systems have been designed using the principle of "humans in the loop". The premise of these systems is that humans must validate different decisions an intelligent system is making, giving feedback about what decisions are good and which ones are bad. This principle applies to all of the approaches above, as researchers review and update the dictionaries, coding, networks and benchmarks and use that information to refine future decisions. We think that as a complement to the idea of "humans in the loop" common in computer science, perhaps social scientists can consider the opposite - "computers in the loop" - identifying where computation would be most beneficial throughout the research lifecycle. In the context of measurement, this means identifying tasks that make it

easier to tackle large texts, and make the process less labor intensive. Encouraging social scientists to view computers in this way may also help social scientists use a broader range of methodologies in creating measures, outside commonly used manually constructed measures like dictionaries.

7. Ethical & Privacy Challenges

While there are a number of general ethical and privacy concerns in using social media data for research (Özkula, 2020 or Singh et al., in press), we focus here on those specific to measurement. The first ethical concern arises when selecting measures to construct. Should researchers construct measures that could inadvertently cause harm? For example, computer scientists have developed algorithms for identifying individuals exhibiting extremist behavior (Wei & Singh, 2017) or suicidal tendencies (De Choudhury et al., 2016). Should IRBs be involved in the decision to develop certain types of algorithms, even if the computer scientist does not intend to use it for traditional social science research? In other words, how should computer scientists grapple with understanding the potential for harm?

Just as unbiased, accurate measures of sensitive information could cause harm to users on social media in certain cases, errors in modeling and machine learning algorithms may ultimately cause harm because they promote unfairness or injustice For example, an algorithm that predicts race may be very accurate if the race category is *White*, but not accurate if it is *Black* or *Asian* or *mixed race*. This is an algorithmic bias issue that social scientists need to be aware of when integrating algorithms into measurement construction. A new subdiscipline has emerged in computer science around algorithmic fairness (Lepri et al., 2018). Most work in this space has focused on well-structured data sets and more traditional data applications. However, when using social media data, all the measurement issues we have identified throughout this paper, including missing data, measurement coverage, data quality, and data types, represent possible issues that can cause algorithmic bias. Suresh and Guttag have developed a framework for describing the biases related to the creation and use of machine learning models (Suresh & Guttag, 2019). This work and others like it can be used as a foundation for a broader look at ethical considerations for research involving social media data.

Another concern is the potential for privacy violations. While various measures may lead to privacy concerns, one that stands out is reidentification of individuals in a network. When constructing a network, the network structure may be unique and lead to identification of individuals. If the individuals did not consent to be in the network analysis, this is an example of a privacy breach. Even if the original network does not contain unique components, networks can be mapped to other data sets, like voter information, to reidentify individuals. In other words, if a measure is released and anonymized, it is possible to combine the anonymized data with other external data to re-identify individuals (Lepri et al., 2018; Hay et al., 2008). Network structures are not the only concern when considering privacy. Any measure that has unique properties for some individuals may lead to reidentification risks.

Additional privacy issues arise when considering how measures and features constructed from social media data are stored. As described above, many sensitive characteristics can be derived from someone's social media activity, such as their political affiliation, sexual orientation, health status, or support of or membership in extremist groups. If these characteristics, constructed from social media data, are stored alongside users' raw posts, or

stored in any way that could allow raw posts and user information to be linked, the researcher risks exposing users to deductive disclosure of sensitive information. This is particularly relevant for largely public platforms like Twitter with which it is quite easy to link a particular post, if entered verbatim, with its author. Researchers and IRBs should consider best practices for storing and securing raw and constructed social media data with an eye toward limiting to the extent possible the risk of deductive disclosure.

8. Strategies for Accelerating Convergent Research

In order to accelerate research involving social media data, it is important to identify outputs or artifacts, including lists of best practices, data sets, and pilot studies that are beneficial to the research community. It can be difficult and time-consuming to build everything from scratch. With respect to measurement, there are a number of artifacts we have identified as useful for those new to using social media data within their research.

First, having more shared dictionaries and code available for different tasks is important. We could imagine shared dictionaries for both sentiment and salient topics. While these may need to be customized for different projects, they can serve as a starting point for many researchers. If they were all stored in common shared data portals, researchers could also analyze the similarities and differences associated with dictionaries generated for salient topics. Next, social scientists prefer to program in R instead of python and other programming languages designed to handle large-scale data. Because of this difference, it is important to share example code in languages traditionally used in both social science and computer science to cross-fertilize programming techniques. Just as important is to help social scientists better understand the programming languages that computer scientists use for large scale data. While we are not advocating for every social scientist to learn programming in a particular language, we do advocate having a level of comfort programming in one language. That foundation will help researchers more readily translate between programming languages like R and python. We also note that while computer scientists have developed packages testing machine learning algorithms broadly, more work is needed to develop packages focused on algorithms for social media data or specific measures of interest to social scientists, including topic modeling, and stance detection, event detection.

As a general concern, many social scientists may not know where to go to find data sets, code, and example applications for different types of tasks. We hope this project will eventually be the place to start, but currently, there are many methodological resources posted in disparate locations across a variety of fields. The computational linguistics community has done a better job of aggregating artifacts and sharing them more broadly than the data mining and machine learning communities. For example, they have released labeled data for a large number of natural language processing tasks, and have released dictionaries and word embeddings to improve relevant machine learning models and promote replicability. They also share pre-prints through arXiv to get new methods out to researchers as soon as possible. Finding ways to expand these types of artifacts by creating enclaves of non-public data available to researchers and subject specific portals with simple ways to construct measures is vital to engaging more researchers in this space.

Similarly, the network science community has done a great deal to generate code bases and example data sets. However, there are very few examples and datasets available that are

related to social media. Expanding what they have already done to show how to construct networks from social media data, and how to measure different types of information in these networks, would enable researchers to innovate new ways to measure validity and reliability in the context of social media data.

Finally, it would be useful to create a working document of different ways researchers have measured different constructs. This document could also link to code and papers. Most importantly, however, it should highlight strengths and weaknesses associated with construct measurement that have been identified. Social media is not going to be a reasonable place for constructing all measures of interest. Seeing both the successes and failures are important. Our group will be hosting panels each semester about different methodological challenges and different applications. These types of meetings can also be a way to populate blogs and other informal documents about how different constructs have been created.

9. Conclusions

Harnessing social media data for social science research entails creating measures out of the largely unstructured, noisy data that users generate on different platforms. This harnessing, particularly of data at scale, requires using methods developed in computer science. But it also typically requires integrating these methods with assessments of measurement quality along social science criteria -- reliability, validity and unbiasedness. When measures capture a construct of interest that is defined outside social media, or broadly test a hypothesis about human behavior, opinion, etc. using social media data, researchers must evaluate how reliable, valid and unbiased the measures are with respect to the theoretical construct.

In this paper, we outlined certain best practices for assessing these criteria, drawn from standard social science practices but applied to social media data. Strategies for ensuring reliability include formalizing procedures and metrics for calculating interrater reliability on codes or labels, selecting a level of measure precision that maximizes power but also reliability, and revising measurement procedures repeatedly throughout a project to account for changes in users and platforms over time. Best practices for establishing validity center on the use and selection of appropriate ground truth data sources. Many sources of ground truth involve human judgments, such as human coded labels or codes for constructs like sentiment or stance, but not all. The latter include self-reports of characteristics like age or gender on platforms, geotagged locations, or administrative data. Ground truth data sources are also used to identify bias in measurement, including bias resulting from under or over coverage. In short, for many social science projects using social media, data sources from outside social media even in small amounts can be important to assess quality of social media measures.

Another consideration that emerged as important to measurement is identifying and addressing the existence of bots or posers on social media. Social scientists might not always want to exclude data generated by bots from analyses or measurement. For example, if researchers are measuring the experience of social media users, the role of bots in content, discussion and behavior might be important to include. Likewise, some bots have large followings (e.g., Lil Miquela on Instagram) and are thus worthy of study to gauge dynamics of interest such as how followers react to content. But if we want to understand how humans behave, think or feel, as we often do in social science, then distinguishing human and non-human

sources of data is crucial. We also stress that current methods for detecting bots on social media are imperfect; it is an active research area within computer science.

Finally, our discussions helped us identify the difference in goals between creating measures for constructs and creating features for algorithms. While some features may in fact be measures, many are not. Instead they are meant to be a representation of the data that may help a data mining algorithm highlight an important pattern in the data or may help a ML classifier improve their model. This distinction is important to keep in mind when considering measurement. While all features should have a clear purpose with clear measurement properties, it will not always be the case that the bar for reliability and validity needs to be as high.

Our group found it most useful to think about these practices in terms of both the level of analysis a measure captures -- post, user, group, or population -- and the methodology employed to create the measure -- be it dictionary-based, supervision machine learning, descriptive system or network analyses, or the linking of social measures and features to other data including external benchmarks. Much of the existing work that has measured constructs in social media data has imported automated methods that are used to evaluate other types of systems and other types of texts. For instance, automated topic models and dictionary methods for sentiment and emotion measurement were previously mainly used for coding article-length texts. Repurposing them for analyzing social media posts, which tend to be much shorter, structured less formally, and use vocabulary in different ways, has required researchers employing these methods to make adaptations.

The need for regular adaptations is one obvious reason that all measurement methods for social media data have required some level of human involvement to work well. At this point, we would rarely recommend “fully automated” methods for analysis of social media data. Volume analyzes are the obvious exception. Typically, computer methods should be semi-automated, in which the researcher still plays an active role in some way (depending on the method) reviewing the results and modifying the method when necessary. All these methods have the potential to produce valid, reliable, precise and relatively unbiased measures if used in the appropriate circumstances and when guided carefully by conscientious researchers who are familiar with the social media contexts that they are studying.

With these methods, practices, and guidelines in hand, we believe that social media holds great promise as a source of novel, insightful data for social science researchers. Social media data can capture attitudes, emotions, and interests on topics typically studied using large surveys, as well as behaviors and reactions typically captured using observations, spontaneously and in real time. Social media can capture constructs at a scale that standard social science data sources cannot: behavior, opinions, emotions, and attitudes of a vast number of people on a wide range of topics. Data from social media can also offer value for questions about human behavior because they provide a window into behavioral phenomena that may simply not be accessible with other methods.

Most importantly, perhaps, because social media posts represent naturally-occurring conversations about and reflections on people’s everyday lives without reference to any predetermined study topic, they can be used to answer questions researchers would have liked to ask in surveys but did not know about in advance, something standard social science designs cannot accomplish. There are real challenges with the potential unrepresentativeness of users of social media, which will be the core focus of our next meeting on “Data Acquisition and

Sampling." Still, in the ability to measure things that happened before the researcher knew to begin studying the topic, social media has a significant advantage over survey or laboratory research. In this way, social media data offer an unparalleled insight into new and emerging phenomena. The challenge of computer science and social science convergence is to capitalize on this promise in ways that build on and blend the best practices of both disciplines.

Acknowledgements

We would like to thank the National Science Foundation and the McCourt School's 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. 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. We would like to thank Professor Kobbi Nissim for his insightful comments during the measurement 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 invaluable when putting this document together.

References

Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the International Workshop on Link Discovery* (pp. 36-43). Chicago, IL, USA.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211-36.

Allcott, H., Gentzkow, M., & Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research & Politics*, *6*(2).

Al Zamal, F., Liu, W., & Ruths, D. (2012). Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Amazon Mechanical Turk. (2020). Retrieved December 17, 2020, from https://www.mturk.com/

Antenucci, D., Cafarella, M., Levenstein, M., Ré, C., & Shapiro, M. D. (2014). *Using social media to measure labor market flows* (No. w20010). National Bureau of Economic Research.

Appen. (2020). Retrieved December 17, 2020, from https://appen.com/figure-eight-is-now-appen/

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis, 23(1), 76-91.*

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531-1542.

Bender, J. L., Jimenez-Marroquin, M. C., & Jadad, A. R. (2011). Seeking support on Facebook: A content analysis of breast cancer groups. *Journal of Medical Internet research, 13*(1), e16.

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 US Presidential election online discussion. *First Monday*, *21*(11-7).

Blackwell, L., Dimond, J., Schoenebeck, S., & Lampe, C. (2017). Classification and its consequences for online harassment: Design insights from heartmob. In *Proceedings of the ACM on Human-Computer Interaction,* (pp. 1-19).

Blei, D., Ng, A., & Jordan, M. (2001). Latent Dirichlet allocation. *Advances in Neural Information Processing Systems*, *14*, 601-608.

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

Bode, L., Davis-Kean, P., Singh, L., Berger-Wolf, T., Budak, C., Chi, G., … Traugott, M. (2020, July 8). Study Designs for Quantitative Social Science Research Using Social Media. *PsyArXiv*: https://psyarxiv.com/zp8q2/

Borah, P. (2014). Does it matter where you read the news story? Interaction of incivility and news frames in the political blogosphere. *Communication Research, 41*(6), 809-827.

Bozarth, L., Saraf, A., & Budak, C. (2020). Higher ground? How groundtruth labeling impacts our understanding of fake news about the 2016 US presidential nominees. In *Proceedings*

of the International AAAI Conference on Web and Social Media (Vol. 14, pp. 48-59). Virtual.

Briones, R., Nan, X., Madden, K., & Waks, L. (2012). When vaccines go viral: an analysis of HPV vaccine coverage on YouTube. *Health Communication*, *27*(5), 478-485.

Brito, N., Ryan, R., & Barr, R. (2015). Methods for assessing parent-child interactions in large-scale studies. In O. N. Saracho (Ed.), *Contemporary perspectives in early childhood education. Handbook of research methods in early childhood education: Review of research methodologies, Vol. 2* (p. 147–189). IAP Information Age Publishing.

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.

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., ... & Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health*, *108*(10), 1378-1384.

Chakraborty, T., Dalmia, A., Mukherjee, A., & Ganguly, N. (2017). Metrics for community analysis: A survey. *ACM Computing Surveys (CSUR)*, *50*(4), 1-37.

Chandrasekharan, E., Samory, M., Srinivasan, A., & Gilbert, E. (2017). The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 3175-3187). Denver, CO, USA.

Chen, X., Wang, Y., Agichtein, E., & Wang, F. (2015). A comparative study of demographic attribute inference in Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media, 15*, (pp. 590-593).

Cheng, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2015). Antisocial behavior in online discussion communities. *arXiv preprint* arXiv:1504.00680.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E, 70*(6), 066111.

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication, 64*(4), 658-679.

Conover, M. D., Goncalves, B, Ratkiewicz, J., Flammini, A., & Menczer, F. (2018). Predicting the political alignment of Twitter users. In *Proceedings of the International Conference on Computer and Technology Applications* (pp. 1-8). Aizu-Wakamatsu, Japan.

Culotta, A., Kumar, N. R., & Cutler, J. (2015). Predicting the demographics of Twitter users from website traffic data. In *the Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 15, pp. 72-8).

Culotta, A., Ravi, N. K., & Cutler, J. (2016). Predicting Twitter user demographics using distant supervision from website traffic data. *Journal of Artificial Intelligence Research, 55,* 389-408.

Dastin, J. (2018, October 9). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). Botornot: A system to evaluate social bots. In *Proceedings of the International Conference Companion on World Wide Web* (pp. 273-274). Québec, QC, Canada.

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.

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 098-2110). San Jose, CA, USA.

Doucleff, M., & Aubrey, A. (2018, February 2018). Smartphone detox: How to power down in a wired world. *National Public Radio*.

Dredze, M., Broniatowski, D. A., & Hilyard, K. M. (2016). Zika vaccine misconceptions: A social media analysis. *Vaccine*, *34*(30), 3441.

Dutwin, D., Loft, J. D., Darling, J., Holbrook, A., Johnson, T., Langley, R. E., ... & Triplett, T. (2014). Current knowledge and considerations regarding survey refusals. *AAPOR Task Force on Survey Refusals.*

Dutwin, D., & Lavrakas, P. (2016). Trends in telephone outcomes, 2008-2015. *Survey Practice, 9*(3), 1.

Elhai, J. D., Dvorak, R. D., Levine, J. C., & Hall, B. J. (2017). Problematic smartphone use: A conceptual overview and systematic review of relations with anxiety and depression psychopathology. *Journal of Affective Disorders*, *207*, 251-259.

Face^{++}. (2020). Retrieved November 18, 2020, from https://www.faceplusplus.com/

Gervais, B. T. (2015). Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology & Politics*, *12*(2), 167-185.

Girvan, M., & Newman, M. E. (2001). Community structure in social and biological networks. In *Proceedings of the National Academy of Science* USA, 99(cond-mat/0112110), 8271-8276.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267-297.

Groves, R. M., Fowler Jr., F. J., Couper, M. C., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology, Second Edition*. Hoboken, NJ: Wiley-Interscience.

Guidry, J. P., Carlyle, K., Messner, M., & Jin, Y. (2015). On pins and needles: how vaccines are portrayed on Pinterest. *Vaccine*, *33*(39), 5051-5056.

Gunaratne, K., Coomes, E. A., & Haghbayan, H. (2019). Temporal trends in anti-vaccine discourse on twitter. *Vaccine*, *37*(35), 4867-4871.

Hay, M., Miklau, G., Jensen, D., Towsley, D., & Weis, P. (2008). Resisting structural re-identification in anonymized social networks. In *Proceedings of the VLDB Endowment* (pp. 102-114). Auckland, New Zealand.

He, Q., Turel, O., & Bechara, A. (2017). Brain anatomy alterations associated with Social Networking Site (SNS) addiction. *Scientific Reports*, *7*(1), 1-8.

Keeter, S., McGeeney, K., Igielnik, R., Mercer, A., & Mathiowetz, N. (2015). From telephone to the web: The challenge of mode of interview effects in public opinion polls. *Pew Research Center.*

Kelleher, J. D., & Tierney, B. (2018). *Data science.* MIT Press.

Lee, J. Y., Grogan-Kaylor, A. C., Lee, S. J., Ammari, T., Lu, A., & Davis-Kean, P. (2020). A qualitative analysis of stay-at-home parents' spanking tweets. *Journal of Child and Family Studies*, *29*(3), 817-830.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, *31*(4), 611-627.

Linder, F. (2017). Improved data collection from online sources using query expansion and active learning. *SSRN. 3026393*.

Magdy, W., and Elsayed, T. (2014). Adaptive method for following dynamic topics on Twitter. In *Proceedings of the ACM Conference on Hypertext and Social Media*, (pp. 335-345). Ann Arbor, MI, USA.

Manning, C. D., Schütze, H., & Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge University Press.

Massey, P. M., Leader, A., Yom-Tov, E., Budenz, A., Fisher, K., & Klassen, A. C. (2016). Applying multiple data collection tools to quantify human papillomavirus vaccine communication on Twitter. *Journal of Medical Internet Research*, *18*(12), e318.

Meyer, B. D., Mok, W. K., & Sullivan, J. X. (2015). Household surveys in crisis. *Journal of Economic Perspectives, 29*(4), 199-226.

Mishori, R., Singh, L. O., Levy, B., & Newport, C. (2014). Mapping physician Twitter networks: Describing how they work as a first step in understanding connectivity, information flow, and message diffusion. *Journal of Medical Internet Research*, *16*(4), e107.

Mondal, M., Silva, L. A., & Benevenuto, F. (2017). A measurement study of hate speech in social media. In *Proceedings of the ACM Conference on Hypertext and Social Media* (pp. 85-94). Prague, Czech Republic.

Mossie, Z. & Wang, J. (2020). Vulnerable community identification using hate speech detection on social media. *Information Processing & Management,* 57(3).

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of The National Academy of Sciences*, *103*(23), 8577-8582.

Nguyen, D., Smith, N. A., & Rose, C. (2011). Author age prediction from text using linear regression. In *Proceedings of the ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 115-123). Portland, OR, USA.

Özkula, S. M. (2020). The issue of "context": Data, culture, and commercial context in social media ethics. *Journal of Empirical Research on Human Research Ethics*, *15*(1-2), 77-86.

Oyeyemi, S. O., Gabarron, E., & Wynn, R. (2014). Ebola, Twitter, and misinformation: a dangerous combination? *British Medical Journal*, *349*, g6178.

Parkin, S. (2018, March 4). Has dopamine got us hooked on tech? *The Guardian*, *4*.

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

Poynter Institute. (2019). *International Fact Checking Network.* Retrieved November 18, 2020, from https://www.poynter.org/

Rajadesingan, A., Resnick, P., & Budak, C. (2020). Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 14, pp. 557-568). Virtual.

N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, and et al. (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 (KDD).*

Rao, D., Paul, M. J., Fink, C., Yarowsky, D., Oates, T., & Coppersmith, G. (2011). Hierarchical Bayesian models for latent attribute detection in social media. *Proceedings of the International Conference on Weblogs and Social Media, 11*, 598-601.

Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013). The structural topic model and applied social science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation* (Vol. 4).

Rosenthal, S., & McKeown, K. (2011). Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 763-772). Portland, OR, USA.

Ryan, R. M., Kalil, A., Ziol-Guest, K. M., & Padilla, C. (2016). Socioeconomic gaps in parents' discipline strategies from 1988 to 2011. *Pediatrics*, *138*(6).

Sadeque, F., Rains, S., Shmargad, Y., Kenski, K., Coe, K., & Bethard, S. (2019). Incivility Detection in Online Comments. In *Proceedings of the Joint Conference on Lexical and Computational Semantics* (pp. 283-291). Dublin, Ireland.

Sakaki, S., Miura, Y., Ma, X., Hattori, K., & Ohkuma, T. (2014). Twitter user gender inference using combined analysis of text and image processing. In *Proceedings of the Workshop on Vision and Language* (pp. 54-61). Dublin, Ireland.

Salathé, M., & Khandelwal, S. (2011). Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLoS Comput Biol*, *7*(10), e1002199.

Salkind, N. J. (Ed.). (2010). *Encyclopedia of research design (Vol. 1).* Sage.

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006). Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (Vol. 6, pp. 199-205).

Shao, C., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2016). Hoaxy: A platform for tracking online misinformation. In *Proceedings of the International Conference Companion on World Wide Web* (pp. 745-750). Geneva, Switzerland.

Sharma, M., Yadav, K., Yadav, N., & Ferdinand, K. C. (2017). Zika virus pandemic—analysis of Facebook as a social media health information platform. *American Journal of Infection Control*, *45*(3), 301-302.

Singh, L., Budak, C., Kawintiranon, K. & Soroka, S. (2020, November 4). Real- time analysis shows that the first debate shifted attitudes among Twitter users towards Biden, and the second solidified them. *The London School of Economics US Centre's daily blog on American Politics and Policy*. https://blogs.lse.ac.uk/medialse/2020/11/04/real-time-analysis-shows-that-the-first-debate-shifted-attitudes-among-twitter-users-towards-biden-and-the-second-solidified-them/

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., Bode, L., Budak, C. *et al.* (2020) Understanding high- and low-quality URL Sharing on COVID-19 Twitter streams. *Journal of Computational Social Science* 3, (pp. 343–366).

Singh, L., Polyzou, A., Wang, Y., Farr, J., & Gresenz, C. R. (2020). *Social Media Data - Our Ethical Conundrum* [In Press].

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., and Rana, O. (2013). Knowing the tweeters: Deriving sociologically relevant demographics from Twitter. *Sociological Research Online* 18(3): 7.

Sloan, L., Morgan, J., Burnap, P., and Williams, M. (2015). Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PloS ONE* 10(3): e0115545.

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin,* 144(12), 1325.

Stryker, R., Conway, B. A., & Danielson, J. T. (2016). What is political incivility? *Communication Monographs*, *83*(4), 535-556.

Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*.

Tan, P. N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education India.

Tang, L., & Liu, H. (2010). Community detection and mining in social media. *Synthesis Lectures on Data Mining and Knowledge Discovery, 2*(1), 1-137.

Taniguchi, T., Sakaki, S., Shigenaka, R., Tsuboshita, Y., & Ohkuma, T. (2015). A weighted combination of text and image classifiers for user gender inference. In *Proceedings of the Workshop on Vision and Language* (pp. 87-93). Lisbon, Portugal.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24-54.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly, 60*(2), 275-304.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859.

Traugott, M. W., & Katosh, J. P. (1979). Response validity in surveys of voting behavior. *Public Opinion Quarterly, 43*(3), 359-377.

Turel, O., He, Q., Xue, G., Xiao, L., & Bechara, A. (2014). Examination of neural systems sub-serving Facebook "addiction". *Psychological Reports*, *115*(3), 675-695.

Vargo, C. J., Guo, L., & Amazeen, M. A. (2018). The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society*, *20*(5), 2028-2049.

Vikatos, P., Messias, J., Miranda, M., and Benevenuto, F. (2017). Linguistic diversities of demographic groups in Twitter. *Proceedings of the ACM Conference on Hypertext and Social Media* (pp. 275–284). Prague, Czech Republic.

Wang, S., Minku, L. L., & Yao, X. (2018). A systematic study of online class imbalance learning with concept drift. *IEEE transactions on neural networks and learning systems, 29*(10), 4802-4821.

Wei, Y., & Singh, L. (2017). Using network flows to identify users sharing extremist content on social media. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 330-342). Jeju, South Korea.

Williams, J. B., Mezey, N., & Singh, L. (2020). *#BlackLivesMatter—Getting from Contemporary Social Movements to Legal Change* [Under review].

Woolley, J. K., Limperos, A. M., & Oliver, M. B. (2010). The 2008 presidential election, 2.0: A content analysis of user-generated political Facebook groups. *Mass Communication and Society*, *13*(5), 631-652.

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the International Conference on World Wide Web* (pp. 1391-1399). Perth, Austrailia.

Zagheni, E. and Weber, I. (2015). Demographic research with nonrepresentative internet data. *International Journal of Manpower* 36(1): 13–25.

Zimdars, M. (2016, November 18). My "fake news list" went viral. But made-up stories are only part of the problem. *Washington Post*.