16. Big data analytical methods

Hai Liang

INTRODUCTION

Social media are among the most popular sources of big data in contemporary computational social science studies. Social media platforms do not just provide rich data for social sciences; they also raise new research questions regarding the relationships between social media and society. Although different approaches have been developed to answer these questions, big data analytics (and computational methods in general) is the genetic method for social media data.

Big data as a concept has been defined in several ways. Originally, it was a technical term referring to data that could not be processed with traditional relational databases (Manovich, 2011). In this sense, big data has been defined according to its volume, variety, and velocity (Laney, 2001; Monroe, 2013). This concrete meaning evolved when the term was adopted by computational social science. Lazer and Radford (2017) have conceptualized big data as the data generated by the digitization of social life. By definition, big data includes all types of digital archives about social life. Social media platforms, such as Twitter and Facebook, are the major sources of such archives and include intrinsically digital human behaviors.

Given the various definitions of big data, there are also different meanings of big data analytics that depend on context. When 'big data' is used as a technical term, 'big data analytics' refers to the tools and methods used to cope with the practical problems related to big data collection, storage, and analysis. In social science, big data analytics is often equivalent to employed computational methods such as text mining, social network analysis, and machine learning. Generally speaking, big data analytics can refer to any computational or statistical model used to solve social science problems with big data.

This chapter outlines big data analytical methods that are widely used for research into social media and society. Specifically, the chapter reviews the different types of big data on social media platforms and summarizes four strategies of big data analysis in social science (i.e., measurement, description, causal inference, and intervention). Finally, the chapter also highlights the challenges and opportunities for future studies.

SOCIAL MEDIA AS BIG DATA

Various types of data on social media are valuable for social science research. First, the most common form is text data, which include any textual content posted on

social media platforms. By leveraging text mining techniques, researchers can extract useful latent variables from texts to answer important questions. For example, King et al. (2013) monitored and analyzed the content of millions of social media posts originating from thousands of platforms in China. Based on this continuous tracking dataset, the authors were able to estimate the volume of censorship; they then classified the content into different topics and automatically detected the sentiment of each statement (criticize or support the state, leaders, or policies) by using text mining. Finally, they compared the censored and uncensored content and found that posts related to collective action were more likely to be censored than other topics, even criticism of the state, policies, or leaders.

Network data are another type of big data commonly used in social science. The 'follow' function on most social media platforms provides unique data for social scientists to study the formation and influence of social networks, which are difficult to collect at a large scale using traditional methods. One of the earliest studies on this type of big data was conducted by Lewis et al. (2012). Based on panel data of Facebook friendship networks and the user profiles of a cohort of college students over four years, the authors used a formal network coevolution model developed by Snijders et al. (2010) to distinguish the effects of social selection (homophily: people are more likely to be friends if they are similar) and social influence (people are more likely to be similar because they are friends) on social networks. Lewis et al. (2012) found that students who share certain tastes in music and movies are likely to befriend one another. However, there is no evidence to support the social influence hypothesis. In addition to the friendship network, other interaction networks, such as the retweeting (reposting) network on Twitter, have been used to measure online fragmentation and echo chambers. Barberá (2015) found that Twitter users primarily retweeted messages from those with similar ideological preferences in the case of political issues but not in many other topic areas.

Behavioral data are usually collected by monitoring user behaviors on social media. The behaviors can be publicly available online or stored privately on web servers. The most accessible behavior data on social media are posting activities, which can be followed over time. Previous studies have found that posting behavior remains constrained by day–night and weekday–weekend schedules (Dodds et al., 2011; Golder & Macy, 2011); there are more posting activities during users' daytimes and on weekdays. Abnormal activities have been used to identify political astroturfing and other computational propaganda strategies (e.g., Keller et al., 2020). Social media activities have also been widely used to track and analyze the patterns of movements and collective action on social media (e.g., González-Bailón et al., 2011). For example, studies have found that peripheral users in online networks are critical for mobilizing participants in large-scale social movements (Jost et al., 2018; Steinert-Threlkeld, 2017). Private behavioral data, such as login and browsing activities, are not directly visible to general users or researchers. Nevertheless, researchers can collaborate with data owners (e.g., Facebook, Twitter) or purchase data from third-party vendors like Alexa, ComScore, and Nelson. For example, Taneja and Wu (2014) used a web browsing dataset from an online ComScore panel to study

how cultural proximity, rather than access blockage, shaped online user behavior in China.

Geospatial data include self-reported locations and geo-coordinates embedded in posts such as check-ins on Facebook. Location data are not only useful for marketing purposes but also facilitate large-scale, comparative studies across regions. For example, Liang et al. (2016) randomly sampled 3.3 million Twitter users and identified their country and territory information from the 'location' field. The country information was translated to cultural dimensions (e.g., individualism vs. collectivism), internet penetration, and other aggregate indicators at the country level. Finally, by comparing user behaviors from more than 100 countries, the study found that privacy settings in collectivist societies were more effective at encouraging self-disclosure.

Finally, social media are multimedia platforms including diverse multimedia data, chiefly audio, videos, and still images. Although these data are less popular in social science studies, their potential value has been noticed. For example, Zhang and Pan (2019) used a deep-learning approach to identify collective action events with images from Weibo. This method provides unique benefits in authoritarian regimes because information on protest is usually lacking in other ways.

BIG DATA ANALYTICS IN CONTEXT

Social media are not merely sources of big data but also the objects of study. Their data can be analyzed to answer a wide range of social questions. Big data analytics can be classified and summarized in different ways, depending on the data type to be analyzed. Methods that process text data are called text mining or computer-assisted content analysis; while social network analysis has been developed to deal with social network data, user behaviors are examined with user analytics, location data are usually processed with spatial models, and image processing is employed to analyze image data. Although this classification is useful for summarizing techniques, it provides little insight into how big data analytics can be incorporated into and facilitate social science research.

Another disadvantage of this typology is that researchers generally employ multiple types of data and analytics in a single study, and it is difficult to separate them in real-world studies. In practice, many studies have incorporated different types of data into a single regression or machine learning framework. For example, machine learning has been adopted to extract variables from and to model social media data, including texts, images, behaviors, locations, and even networks. Like regression models in traditional statistics, machine learning is not one model but a collection of models under the same framework. 'Machine learning' refers to techniques that use an automatic system (i.e., computer algorithms) to learn from past observations to classify new observations or make predictions about future occurrences. In order to teach computers to correctly classify and make predictions, sufficient (usually manually) labeled examples that are presumably correct must be supplied as training

data. Computer algorithms are developed and employed to learn the latent patterns in the training data; these learned patterns are then used to make classifications and predictions in a separate test dataset. If those classification or prediction results in the test data are insufficiently accurate, researchers adjust or change the computer algorithms to improve the final model. Training a good model is not the ultimate goal of social science research, which is focused on applying well-trained models to new datasets. Many studies have used pre-trained models (e.g., the SentiStrength model developed by Thelwall et al., 2010) to classify large-scale tweets into positive or negative tones and then use the polarity score as an independent variable to explain other phenomena.

Given the complexity of and disadvantages in the current classification of big data analytics, the present chapter classifies those analytics according to the purposes of their application in empirical social science research. As explained below, empirical studies have used computational methods and social media data to detect and measure key variables (e.g., incivility and political ideology), to describe the prevalence of various social media phenomena (e.g., fake news), to make causal inferences using observational data (e.g., social influence and social selection in network dynamics), and to enable interventions to improve society (e.g., voter mobilization and reducing incivility on social media).

Measurement

Measurement is the fundamental element in social science research. Detecting and measuring social constructs using big data analytics is one of the most common applications in computational social science. For this purpose, big data analytics are used to detect and measure socially related constructs from social media data so researchers can describe their prevalence at a large scale and relate them to other concepts to investigate relationships. Machine learning suits this purpose very well. For example, Theocharis et al. (2016) detected uncivil comments to politicians on Twitter using machine learning methods and relying only on textual data. They manually coded a random sample of 7,000 tweets according to two dimensions: politeness vs. impoliteness and whether a tweet contained a reference to moral or democratic issues. They selected regularized logistic regression to train the models for the two dimensions separately, eventually achieving an accuracy above 0.80. They then applied the models to classify the rest of the tweets – nearly 800,000 comments to politicians – according to their textual features into different categories. If a tweet was classified as impolite and related to morality or democracy, the tweet was classified as uncivil. Finally, they found that more engaging tweets sent by the politicians were associated with more uncivil tweets received.

It is not only textual data that are used to construct social measures but also other types of social media data. One interesting application is to use network data (e.g., following relationships) to infer the political leanings of online users. The basic assumption of these models is that users are more inclined to interact with other users with similar political leanings on social media platforms by following and liking.

Given the observed interaction networks, the political leanings of the users are considered latent variables and can be estimated statistically. These models have been successfully applied to estimate users' ideologies on Twitter (Barberá, 2015) and Facebook (Bond & Messing, 2015). Other researchers have attempted to combine different types of social media data to measure social concepts using machine learning. Many studies have demonstrated that personality traits can be accurately predicted by combining textual, behavioral, and image data from social media (see Azucar et al., 2018). However, measuring social constructs does not always require advanced methods. For example, without formal models, King et al. (2013) successfully detected online censorship on Chinese websites by continuously monitoring web content and comparing versions over time.

Description

Big data analytics can be used in social science to estimate and scale up the detailed descriptions of social processes and relationships. Instead of describing trivia or anecdotal phenomena on social media platforms, well-designed descriptive studies can offer deep insights into society. For example, social media have unique advantages to track the information diffusion process in social networks. Anecdotal accounts in viral marketing claim that social media can facilitate information to spread in a viral (person-to-person) manner, in contrast to the broadcast approach that dominated the mass media age. Goel et al. (2016) formally quantified the degree of viral diffusion (i.e., structural virality) by using Twitter data; they found that the broadcast model remains dominant. Furthermore, cascade size (the number of total retweets) has no clear positive relationship with the degree of viral diffusion. On the other hand, by using the measure of structural virality, Vosoughi et al. (2018) found that false news was more likely to spread virally than true news.

In addition to depicting social processes, descriptive statistics related to the prevalence of certain critical phenomena on social media platforms are especially important. In big data studies, researchers have investigated these phenomena, which include ideological segregation and misinformation. In response to the criticism of personalization algorithms and the creation of filter bubbles on social media platforms, Bakshy et al. (2015) examined how 10.1 million US Facebook users interact with socially shared news. In particular, they examined the extent to which heterogenous friends could expose users to cross-cutting content (echo chamber) and the extent to which users encounter diverse content while interacting via Facebook's algorithmically ranked news feed (filter bubble). The results showed that users' choices played a stronger role in limiting exposure to cross-cutting content than algorithmic ranking. The prevalence of 'fake news' is another phenomenon that has been overestimated by the public. By focusing on a sample of registered US voters on Twitter, Grinberg et al. (2019) estimated the composition of each panelist's news exposure from a random sample of tweets posted by their followees. They found that engagement with fake news sources was highly concentrated: only 1 percent of users accounted for 80 percent of fake news exposures, and 0.1 percent accounted

for 80 percent of fake news sources shared during the 2016 US presidential election. Similarly, Guess et al. (2019), by linking a survey with respondents' sharing activity recorded in Facebook profile data, found that sharing fake news content was relatively rare on Facebook.

Causal Inference

Big data analysis was initially focused on correlation (Mayer-Schönberger & Cukier, 2013), as causality and theories would not even matter in industrial and engineering applications. But simple correlation works in certain applications might cause serious problems in both social science and science (Pearl & Mackenzie, 2018), which primarily emphasize causal explanations. Causal inferences are the statistical strategies used to identify whether a variable *X* causes another variable *Y* and, if so, by how much. Researchers have long conducted randomized control experiments to make causal inferences. However, this is less feasible when researching big data on social media. Even before the age of big data and computational social science, several strategies were developed to make causal inferences with observational data in social science. Social media data are usually collected unobtrusively and are thus observational. Researchers can treat social media data with conventional causal inference strategies.

Social media data are usually continuously collected; given this always-on status, interrupted time series (ITS) design is one of the most popular strategies that use social media to make causal inferences. ITS involves tracking a long-term period before and after an intervention to assess the effect of the intervention. If the interruption is generated randomly, the ITS design is a natural experiment design. Salganik (2019) has summarized the strategy as follows: random variation + always-on data = natural experiment. Finally, the treatment effects are estimated by changes in the level and slope trends of the time series. Under the ITS framework, Penney (2016) found that traffic to privacy-sensitive Wikipedia articles, such as those related to terrorism, declined in both level and slope after the US mass surveillance program was leaked by Edward Snowden. This finding supports the findings regarding the chilling effects of online surveillance on internet use. Using a similar design, Hobbs and Roberts (2018) found a gateway effect that the sudden blocking of Instagram in China increased the use of Facebook and Twitter (both of which had previously been blocked) by Chinese netizens.

Causal treatment effects can also be estimated by matching, which controls for confounding factors that are observable and measured. The goal of matching is to find one or more non-treated units for every treated unit with similar observable characteristics. In this way, the treated and non-treated groups are fairly comparable, and any difference is thus an estimation of the treatment effect. If the observed characteristics for matching are sufficient to account for all confounding factors, the treatment effect will be estimated unbiasedly. Social media data, as big data, provide massive cases for researchers to find matched samples, which is difficult using traditional methods. For example, in order to study the effects of auction starting price on the

probability of a sale and sale price, Einav et al. (2015) found hundreds of thousands of matched cases (the same products with different starting prices but similar other characteristics) from the massive logs of eBay.

Although social media provide many cases for matching, the data usually have limited variables. Nevertheless, as mentioned above, text is the most common data format on social media platforms and could yield many variables using text mining techniques. Roberts et al. (2020) proposed a method to adjust for confounding with text matching. The assumption is that the latent variables extracted from texts are sufficient to block confounding. For example, to examine how the experience of being censored affects their future online experience (e.g., being censored again and posting less), researchers need to block the confounding effects of the content posted by the users. Certain types of content may be associated with both users' likelihood of being censored and future behaviors. By matching the content features between censored and uncensored posts, the authors found that being censored increased the likelihood of being censored in the future but did not decrease posting rates.

Instrumental variable (IV) estimation has long been employed in econometrics to estimate causal relationships when random experiments are not feasible. To estimate the causal effect of X on Y, an ideal IV is supposed to be highly correlated with X but have no direct effect on Y. Theoretically, random assignment in controlled experiments determines X – which condition to be assigned – in its entirety but has no direct relationship with Y and thus is a perfect IV. As long as IVs are found, estimators such as the two-stage least squares model can be used to estimate causal effects. The IV approach not only solves the problems caused by omitted confounding variables in many observational studies but also situations involving reverse causation, where Y causes X.

In social network analysis, the reciprocal relationship between homophily and social influence is a well-known problem in explaining many social behaviors in social networks. For example, happiness and other emotions are correlated between socially connected individuals both online and offline, such that the friends of happy individuals are also happy. However, it is difficult to ascertain whether this clustering phenomenon is caused by social influence (emotions spread to individuals from their friends) or by choosing social contacts with similar emotions. Coviello et al. (2014) employed the IV approach to solving the problem by collecting network and text data from millions of Facebook users. To estimate the causal effects of social connections (Facebook friends) on emotional similarity (expressing similar emotions in Facebook statuses), the authors proposed rainfall as an IV; the dependent variable was an individual's emotion, and the independent variable was that individual's friends' emotions. Rainfall outside the individual's city may negatively influence the emotional expressions of the individual's friends where it is raining but would not influence that individual's emotion directly. The authors found that rainfall outside an individual's city is correlated with the individual's emotional expression, indicating the existence of an emotional contagion effect.

While panel data are difficult to collect using traditional methods, they are relatively easy to collect from social media platforms. Panel data help social scientists

observe changes and make causal influences by tracking individual users, known as panelists, over the long term. Different models are available to work with panel data. The fixed-effects model controls for all variables, whether or not they are observed, as long as they remain constant within individual users over time. The underlying logic is that causal effects could be evaluated by within-individual variation since it would be possible to control for all confounding variables at the individual level. A special type of fixed-effects model is the difference-in-differences (DID) model, which is conditioned on a group-level rather than an individual-level effect. DID assumes that the outcomes of the treated and control group units would have evolved in a parallel way in the absence of the treatment. Therefore, the difference between the observed and assumed values is the treatment effect. For example, Zhang (2016) used geo-located posts from Weibo to construct treated groups who were physically close to one of the protests in Hong Kong when they occurred, whereas the control groups were those who were also close to those locations but had left Hong Kong before the protests occurred. The authors then used DID estimators and found that witnessing protests had a significant causal impact on civic engagement.

In addition to these classic causal inference strategies, special methods have been developed to make causal inferences with network data. Currently, the most sophisticated model is the stochastic actor-based model for network dynamics (see Snijders et al., 2010). The model adopts a causal mechanical explanation approach and assumes that the network evolves as a stochastic process driven by the actors. Various influences (e.g., the tendencies of popularity, reciprocity, and transitivity) on network changes could be modeled and estimated within this model. More recently, machine learning has also been incorporated into the estimation of causal effects, especially to estimate heterogeneous treatment effects and to deal with high-dimensional variables (see Grimmer et al., 2021).

Intervention

Although it is not always feasible for social scientists, field experiments are increasingly conducted on popular social media platforms. Field experiments are randomized control experiments conducted in natural settings and can involve more representative participants than laboratory experiments. Field experiments are generally thought to have better external validity and a stronger basis for causal inferences than other methods. However, conducting field experiments in offline conditions is technically difficult and expensive. By comparison, implementing randomization on social media platforms is more realistic and less obtrusive because the interventions might not be noticed by the participants. Field experiments on social media are usually referred to as digital field experiments. They have been employed to examine the causal relationships between social influence and the unpredictability of the success of cultural products (Salganik et al., 2006) and between social networks and political mobilization (Bond et al., 2012) and emotional contagions (Kramer et al., 2014); see Salganik (2019) for more examples.

In addition to establishing causal relationships by conducting digital field experiments, social scientists can leverage interventions to 'solve' notable social problems on social media, such as political incivility and polarization. Munger (2017) conducted a digital field experiment on Twitter to reduce the use of anti-black racist slurs by white men. The author collected a sample of Twitter users who had harassed other users by automatic keyword searching (using the word 'n****r'). The sampled users were randomly assigned to the treatment conditions, where the harassers received sanction messages from researcher-controlled accounts known as Twitter bots. The study shows that even simple interventions can be effective at reducing uncivil behaviors. Bail et al. (2018) designed a large-scale experiment on Twitter to examine whether exposure to opposing views could reduce polarization. They surveyed a large sample of Democratic and Republican Twitter users; then, some of the Republican (Democratic) users were randomly assigned to the treatment conditions, where they were encouraged to follow a liberal (conservative) Twitter bot for a month. The results were not encouraging: Republicans (not liberals) who followed the liberal bot became significantly more conservative, indicating that exposure to cross-cutting views might actually increase political polarization.

CHALLENGES AND OPPORTUNITIES

Big data analytics studies have demonstrated that social media data can facilitate and advance research into social media and society. Despite numerous studies having been published, many problems are unsolved, and new challenges are emerging. The first major challenge involves data quality, which is also related to data access, ethics, and privacy. Although social media data are pervasive in computational social science studies, by considering social media as big data sources, a few theoretical and methodological assumptions are usually implicit in empirical studies. Computational social scientists, unlike computer scientists, usually have to assume a parallel universe on social media to generalize their findings (Lazer & Radford, 2017). In most cases, this is not a major problem for engineering purposes. A good recommender system on Amazon works well even if it does not work at all on other platforms. On the contrary, generalizability is a core requirement in social science research. However, substantial differences between online and offline behaviors and across online platforms have been reported in previous studies (see Golder & Macy, 2014).

Theoretically speaking, bridging datasets online and offline (e.g., combining Twitter data with surveys) and from different platforms (e.g., combining Facebook and Twitter data) will improve the generalizability of big data studies. Nevertheless, the reality is that much of the most valuable social media data are controlled by a few social media companies, and those data are either inaccessible or only accessible by researchers with high status (Edelmann et al., 2020). Even when ideal datasets are accessible, linking datasets across sources can cause additional ethical problems. In addition, despite the richness of social media data, important data related to psychological states are lacking. Even if certain variables like personality and emotion can

be predicted using machine learning methods with social media data, many others, such as motivation (especially for inactive social media users), are difficult to predict. Furthermore, many social media studies are not conducted based on representative samples, even though some random sampling methods have been proposed for social media data collection (Liang & Zhu, 2017).

The second challenge is that existing big data analytics are mostly descriptive. It is relatively straightforward to use big data analytics for measuring social concepts and incorporating them into conventional models. Although descriptive studies are essential for social sciences, as mentioned above, many other strategies are available and should be strengthened in future studies. In particular, social media data provide unique benefits for causal influences, given their key features: big, unobtrusive, and always on (Salganik, 2019). In order to make causal inferences using social media data, researchers need to be active in research design instead of relying on purely data-driven approaches. Even though the data are already there, in a carefully designed study, it remains necessary to think about what data should be collected and how to analyze the data in order to make valid inferences. For example, in a panel design, sampling users is far more appropriate than sampling that uses keyword searching. If researchers need to estimate influences in a network using the stochastic actor-based model, a whole network (instead of ego networks) should be collected.

Intervention in big data studies, though not as common as other methods, is increasingly important to promote a more 'solution-oriented approach' in social science (Watts, 2017). Digital field experiments have enormous potential to develop solution-oriented social science. In this kind of research design, concrete interventions are usually required to solve certain practical, important, and pervasive social problems on social media platforms. If an intervention is not concrete and practical, it is unlikely that researchers could implement the experiment on an appropriately large scale or observe meaningful consequences. However, as long as an intervention is demonstrated to be effective, it could be applied directly to the real world. The outcomes should also be practical and pervasive social problems, such as political motivation, online polarization, incivility, and the diffusion of misinformation. Otherwise, it might be difficult to recruit enough qualified participants. The solution-oriented approach should be different from traditional applied research or solving engineering problems for specific social media platforms. For example, thousands of A/B tests that provide different interfaces are conducted every day on the leading social media platforms to optimize usability and revenues. Solution-oriented social science would be equally concerned with the advancement of social theories. As in the examples mentioned above, by reducing uncivil comments using Twitter bots, researchers also know more about the role of norms in political deliberation (Munger, 2017). By increasing the likelihood of voting using Facebook networks, researchers have demonstrated how social networks can influence human behaviors (Bond et al., 2012).

Finally, since many social scientists emphasize the advancement of theories over data and methods, researchers using big data analytics are inevitably required to answer the question of how big data analytical methods could contribute to the devel-

opment of (novel) social theories. First, big data has been used to revisit long-standing social theories or hypotheses that were once thought impossible to test. For example, the availability of social network data and the development of large-scale computational network analytics have enabled the formal examination of network formation dynamics and the impacts of social networks on human behaviors (e.g., Lewis et al., 2012). Leveraging large-scale text mining, Liang and Fu (2019) were able to test the enduring assumption in social network theories that structural redundancy is positively correlated with information redundancy. In addition to testing pre-existing theories, big data analytics provide new opportunities for exploratory studies and, therefore, the development of new theories. Nelson (2020) proposed a methodological framework for theory development using computational methods, which she calls a computational grounded theory. More specifically, computational methods (especially unsupervised methods) could help researchers detect novel patterns in big data; those patterns can be refined through an interpretive engagement with data and finally confirmed using further computational models.

REFERENCES

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, *124*, 150–159. https://doi.org/10.1016/j.paid.2017.12.018.

Bail, C.A., Argyle, L.P., Brown, T.W., Bumpus, J.P., Chen, H.H., Hunzaker, M.B.F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(37), 9216–21. https://doi.org/10.1073/pnas.1804840115.

Bakshy, E., Messing, S., & Adamic, L.A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160.

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, *23*(1), 76–91. https://doi.org/10.1093/pan/mpu011.

Bond, R., & Messing, S. (2015). Quantifying social media's political space: Estimating ideology from publicly revealed preferences on Facebook. *American Political Science Review*, *109*(1), 62–78. https://doi.org/10.1017/S0003055414000525.

Bond, R.M., Fariss, C.J., Jones, J.J., Kramer, A.I., Marlow, C., Settle, J.E., & Fowler, J.H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295–98. https://doi.org/10.1038/nature11421.

Coviello, L., Sohn, Y., Kramer, A.D.I., Marlow, C., Franceschetti, M., Christakis, N.A., & Fowler, J.H. (2014). Detecting emotional contagion in massive social networks. *Plos One*, *9*(3), e90315. https://doi.org/10.1371/journal.pone.0090315.

Dodds, P.S., Harris, K.D., Kloumann, I.M., Bliss, C.A., & Danforth, C.M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *Plos One*, *6*(12), e26752. https://doi.org/10.1371/journal.pone.0026752.

Edelmann, A., Wolff, T., Montagne, D., & Bail, C.A. (2020). Computational social science and sociology. *Annual Review of Sociology*, *46*, 61–81. https://doi.org/10.1146/annurev-soc-121919-054621.

Einav, L., Kuchler, T., Levin, J., & Sundaresan, N. (2015). Assessing sale strategies in online markets using matched listings. *American Economic Journal: Microeconomics*, *7*(2), 215–47. https://doi.org/10.1257/mic.20130046.

Goel, S., Anderson, A., Hofman, J., & Watts, D.J. (2016). The structural virality of online diffusion. *Management Science*, *62*(1), 180–196. https://doi.org/10.1287/mnsc.2015.2158.

Golder, S.A., & Macy, M.W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, *333*(6051), 1878–81. https://doi.org/10.1126/science.1202775.

Golder, S.A., & Macy, M.W. (2014). Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology*, *40*, 129–52. https://doi.org/10.1146/annurev-soc-071913-043145.

González-Bailón, S., Borge-Holthoefer, J., Rivero, A., & Moreno, Y. (2011). The dynamics of protest recruitment through an online network. *Scientific Reports*, *1*, Article 197. https://doi.org/10.1038/srep00197.

Grimmer, J., Roberts, M.E., & Stewart, B.M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, *24*(1), 395–419. https://doi.org/10.1146/annurev-polisci-053119-015921.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 US presidential election. *Science*, *363*(6425), 374–78. https://doi.org/10.1126/science.aau2706.

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1). https://doi.org/10.1126/sciadv.aau4586.

Hobbs, W.R., & Roberts, M.E. (2018). How sudden censorship can increase access to information. *American Political Science Review*, *112*(3), 621–36. https://doi.org/10.1017/S0003055418000084.

Jost, J.T., Barberá, P., Bonneau, R., Langer, M., Metzger, M., Nagler, J., Sterling, J., & Tucker, J.A. (2018). How social media facilitates political protest: Information, motivation, and social networks. *Political Psychology*, *39*, 85–118. https://doi.org/10.1111/pops.12478.

Keller, F.B., Schoch, D., Stier, S., & Yang, J. (2020). Political astroturfing on Twitter: How to coordinate a disinformation campaign. *Political Communication*, *37*(2), 256–80. https://doi.org/10.1080/10584609.2019.1661888.

King, G., Pan, J., & Roberts, M.E. (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review*, *107*(2), 326–43. https://doi.org/10.1017/S0003055413000014.

Kramer, A.D.I., Guillory, J.E., & Hancock, J.T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(24), 8788–90. https://doi.org/10.1073/pnas.1320040111.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, *6*(70), 1.

Lazer, D., & Radford, J. (2017). Data ex machina: Introduction to big data. *Annual Review of Sociology*, *43*, 19–39. https://doi.org/10.1146/annurev-soc-060116-053457.

Lewis, K., Gonzalez, M., & Kaufman, J. (2012). Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(1), 68–72. https://doi.org/10.1073/pnas.1109739109.

Liang, H., & Fu, K.W. (2019). Network redundancy and information diffusion: The impacts of information redundancy, similarity, and tie strength. *Communication Research*, *46*(2), 250–272. https://doi.org/10.1177/0093650216682900.

Liang, H., Shen, F., & Fu, K.W. (2016). Privacy protection and self-disclosure across societies: A study of global Twitter users. *New Media & Society*, *19*(9), 1476–97. https://doi.org/10.1177/1461444816642210.

Liang, H., & Zhu, J.J.H. (2017). Big data, collection of (social media, harvesting). In J. Matthes, C.S. Davis, & R.F. Potter (Eds.), *The International Encyclopedia of Communication Research Methods*. Wiley Press.

Manovich, L. (2011). Trending: The promises and the challenges of big social data. In M.K. Gold (Ed.), *Debates in the Digital Humanities* (Vol. 2, pp. 460–475). University of Minnesota Press.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt.

Monroe, B.L. (2013). The five Vs of big data political science: Introduction to the virtual issue on big data in political science political analysis. *Political Analysis*, *21*(V5), 1–9. https://doi.org/10.1017/S1047198700014315.

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 629–49. https://doi.org/10.1007/s11109-016-9373-5.

Nelson, L.K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, *49*(1), 3–42. https://doi.org/10.1177/0049124117729703.

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Penney, J.W. (2016). Chilling effects: Online surveillance and Wikipedia use. *Berkeley Technology Law Journal*, *31*(1), 117–82. http://www.jstor.org/stable/43917620.

Roberts, M.E., Stewart, B.M., & Nielsen, R.A. (2020). Adjusting for confounding with text matching. *American Journal of Political Science*, *64*(4), 887–903. https://doi.org/10.1111/ajps.12526.

Salganik, M.J. (2019). *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.

Salganik, M.J., Dodds, P.S., & Watts, D.J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, *311*(5762), 854–6. https://doi.org/10.1126/science.1121066.

Snijders, T.A.B., Van de Bunt, G.G., & Steglich, C.E.G. (2010). Introduction to stochastic actor-based models for network dynamics. *Social Networks*, *32*(1), 44–60. https://doi.org/10.1016/j.socnet.2009.02.004.

Steinert-Threlkeld, Z.C. (2017). Spontaneous collective action: Peripheral mobilization during the Arab Spring. *American Political Science Review*, *111*(2), 379–403. https://doi.org/10.1017/S0003055416000769.

Taneja, H., & Wu, A.X. (2014). Does the Great Firewall really isolate the Chinese? Integrating access blockage with cultural factors to explain web user behavior. *Information Society*, *30*(5), 297–309. https://doi.org/10.1080/01972243.2014.944728.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment in short strength detection informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544–58. https://doi.org/10.1002/asi.21416.

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S.A., & Parnet, O. (2016). A bad workman blames his tweets: The consequences of citizens' uncivil Twitter use when interacting with party candidates. *Journal of Communication*, *66*(6), 1007–31. https://doi.org/10.1111/jcom.12259.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–51. https://doi.org/10.1126/science.aap9559.

Watts, D.J. (2017). Should social science be more solution-oriented? *Nature Human Behaviour*, *1*(1), Article 15. https://doi.org/10.1038/s41562-016-0015.

Zhang, H. (2016). *Causal Effect of Witnessing Political Protest on Civic Engagement*. SSRN. https://papers.ssrn.com/abstract=2647222.

Zhang, H., & Pan, J. (2019). CASM: A deep-learning approach for identifying collective action events with text and image data from social media. *Sociological Methodology*, *49*, 1–57. https://doi.org/10.1177/0081175019874760.