Digital Footprints: Opportunities and Challenges for Online Social Research

Scott A. Golder and Michael W. Macy

Department of Sociology, Cornell University, Ithaca, New York 14853;
email: mwm14@cornell.edu, sag262@cornell.edu

Annu. Rev. Sociol. 2014. 40:129–52

First published online as a Review in Advance on June 16, 2014

The *Annual Review of Sociology* is online at soc.annualreviews.org

This article's doi:
10.1146/annurev-soc-071913-043145

Keywords

Internet, Web, social media, online networks, social networks

Abstract

Online interaction is now a regular part of daily life for a demographically diverse population of hundreds of millions of people worldwide. These interactions generate fine-grained time-stamped records of human behavior and social interaction at the level of individual events, yet are global in scale, allowing researchers to address fundamental questions about social identity, status, conflict, cooperation, collective action, and diffusion, both by using observational data and by conducting in vivo field experiments. This unprecedented opportunity comes with a number of methodological challenges, including generalizing observations to the offline world, protecting individual privacy, and solving the logistical challenges posed by "big data" and web-based experiments. We review current advances in online social research and critically assess the theoretical and methodological opportunities and limitations.

[T]hust as the invention of the telescope revolutionized the study of the heavens, so too by rendering the unmeasurable measurable, the technological revolution in mobile, Web, and Internet communications has the potential to revolutionize our understanding of ourselves and how we interact . . . . [T]hree hundred years after Alexander Pope argued that the proper study of mankind should lie not in the heavens but in ourselves, we have finally found our telescope. Let the revolution begin.

—Duncan Watts (2011, p. 266)

INTRODUCTION

Scientific disciplines make revolutionary advances not only through new discoveries, theories, and paradigms but also because of the invention of new tools and methodologies (Kuhn 1962). The electron microscope, space telescope, particle accelerator, and magnetic resonance imaging have allowed scientists to observe the world at greater scale or finer resolution, revealing previously obscured details and unexpected patterns and experiencing the eureka moments of scientific breakthroughs. In this review, we argue that newly developed tools for observing online activity are having a similar transformative effect on the social and behavioral sciences. These studies show how digital footprints collected from online communities and networks enable us to understand human behavior and social interaction in ways we could not do before.

Although the societal impact of electronic communication is widely recognized, its impact on social and behavioral science is also profound, providing global yet fine-grained observational data and a locus for population-scale experimentation. A 2001 *Annual Review of Sociology* article on the "Social Implications of the Internet" (DiMaggio et al. 2001) assessed the Internet as a transformational phenomenon in the reproduction of social inequality (Hargittai 2010, L. Robinson 2011), community mobilization (Hampton & Wellman 2003, Rainie & Wellman 2012), and the use of leisure time (J.P. Robinson 2011). This review turns the tables: Rather than address the societal implications of the Internet, we survey studies that use online data to advance knowledge in the social sciences. A 2004 *Annual Review of Sociology* article entitled "The 'New' Science of Networks" surveyed recent advances in the mathematics of networks, including biological and mechanical as well as social systems (Watts 2004). Although network analysis is clearly an important application of online data, the transformative research opportunities opened up by new sources of empirical data include but also extend beyond network analysis. The rapidly growing field has become too large for a comprehensive review, and we therefore reference only a limited number of studies that illustrate the theoretical and methodological opportunities and challenges, with a slight bias toward papers authored by sociologists or published in sociological journals. Although we include studies that examine online purchases to study social behavior, we primarily focus on studies in which people interact directly with one another, such as social networking sites.

HARD SCIENCE

Over the past century, there has been no shortage of social theory, but there have been severe constraints on access to data. The reason is simple: Social life is very hard to observe. For example, it is much easier to ask an isolated individual about their friends than to observe the ongoing interactions and exchanges that are the stuff of friendship. Ethnographic participant-observation studies and surveys of complete networks make it possible to fully document social interactions, but at costs that can be prohibitively expensive to implement except in very small groups. The need to collect relational data through direct contact has therefore generally limited studies of social interactions to small bounded groups such as clubs (Zachary 1977) and villages (Entwisle et al. 2007). Lengthy time-series data on nation-level populations, such as the Framingham Heart Study (**http://www.framinghamheartstudy.org**) or

the National Longitudinal Study of Adolescent Health (**http://www.cpc.unc.edu/projects/addhealth**), are enormously expensive logistical challenges and are usually undertaken by multiple cooperating institutions in government and academia. Attempts to measure network structure at the population level by surveying egocentric networks (a randomly chosen person and their network neighbors) can be useful for studying the attributes of network nodes (such as degree), and edges (such as tie strength), but this methodology has serious limitations (Flynn et al. 2010, Marsden 1990), including the inability to measure essential network attributes (e.g., distances, clustering, connectivity, and centrality) or social interactions (e.g., diffusion and polarization).

Because of the difficulty of observing social interactions at population scale, most surveys rely on random samples composed of observations that are selected to be independent and to provide an unbiased representation of the underlying population distribution. However, independent observations preclude the ability to directly measure influence from a respondent's friends. We know that people do not entirely "think for themselves," but when we study opinion formation using random samples, we are left with little choice but to assume that a respondent's opinions are shaped entirely by his or her other traits, such as demographic background, material self-interest, or personal experience. As a result, we cannot rule out the possibility that demographic differences in opinions (e.g., the social liberalism of college graduates) are spuriously generated or exaggerated by the unmeasured effects of peer influence (Della Posta et al. 2013, McPherson 2004, Salganik et al. 2006). Conversely, snowball sampling makes it possible to obtain relational data among network neighbors with which to measure demographic differences in beliefs and behavior net of the similarity between network neighbors, but the path dependence in selecting respondents makes it more difficult to obtain an unbiased representation of the population distribution.

Long-standing limitations on the ability to observe social interaction are rapidly disappearing as people all over the globe increasingly choose to interact using devices that provide detailed relational records. Data from online social networks—email archives, phone logs, text messages, and social media postings—allow researchers to relax the atomistic assumptions that are imposed by reliance on random samples. In place of path analytic models of social life as relationships among variables that measure individual traits (Duncan 1966, Wright 1934), data from online social networks allow researchers to model social life as relationships among actors (Macy & Willer 2002).

The rapid increase in the use of digital technologies that generate time-stamped digital footprints of social interactions, from email (Kossinets & Watts 2006), to mobile telephones (Eagle et al. 2010), to social media (Lewis et al. 2008), affords unprecedented opportunities for the collection of both experimental and observational data on a scale that is at once massive and microscopic—massive in the sense that the people under study can number into the millions and the data grow into the terabytes, and microscopic in the sense that individual microinteractions are recorded. In place of retrospective reports about respondents' behavior and interactions, online data can provide a detailed record of daily activities and the frequency and intensity of social relationships. These methods greatly expand our ability to measure changes in behavior, not just opinion; to measure these changes at the individual level yet on a global scale; to observe the structure of the underlying social network in which these individuals are embedded; to travel back in time to track the lead-up to what later becomes an event of interest; and to find the non-events and failed outcomes that escape the attention of publishers, editors, and authors.

This research strategy is not new. For many decades, social and behavioral scientists have acquired data collected as a by-product of the administrative or record-keeping processes of governments and organizations. Organizations

track their membership lists, firms track the purchases of customers and the performance of employees, and banks collect massive data from credit card transactions.

What is new is the macroscopic global scale and microscopic behavioral extensiveness of the data that are becoming available for social and behavioral science. The web sees everything and forgets nothing. Each click and key press resides in a data warehouse waiting to be mined for insights into behavior, to enable useful functions from spam detection to product recommendations to targeted advertising. Our mobile phones, tablets, and laptops report every web page we visit and every link we click and can even report our precise location and movements. Our social interactions are mediated through email, Skype, instant messaging, Facebook, and Twitter. Our photographs are identity-tagged, geo-tagged, and time-stamped, creating a who-when-and-where recording of everything we upload. Social media platforms like Facebook and online labor markets like Amazon Mechanical Turk enable controlled experiments using thousands of participants from all over the world.

The emerging field of computational social science (Lazer et al. 2009) is concerned with computational methods to collect, manipulate, and manage massive amounts of data, as well as with employing the appropriate techniques to derive inferences, such as automated content classification and topic modeling, natural language processing, simulation, and statistical models for analyzing nonindependent observations (Anderson et al. 1999). The rapid growth of computational social science reflects the growing recognition that these new tools can be used to address fundamental puzzles of social science, including the effects of status competition, trust, social influence, and network topology on the diffusion of information, the dynamics of public opinion, the mobilization of social movements, and the emergence of cooperation, coordination, and collaboration.

Computational social science has also reinvigorated social network analysis, one of the historical specialties in sociology that has long been concerned with understanding the processes behind the formation of social ties and their consequences for and constraints on the actions and efforts of individuals and entire communities. Until recently, social network analysis has been limited to very small groups by the requirements of direct observation of interpersonal interactions. We can now obtain detailed measures of network structure and network processes at the population level. The challenge of analyzing massive amounts of online data has pushed social network analysis into the forefront of computational social science, as these techniques have been applied to the inherently relational data created from online interaction.

WHEN IT RAINS IT POURS

The Social Telescope

The ability to observe hundreds of millions of people means we can measure differences with small effect sizes that might otherwise be swamped by random variability. Just as an enormous antenna like the Arecibo Observatory is required to detect the low-frequency radiation emitted from neutron stars (Lovelace & Tyler 2012), online networks comprise a massive antenna for social science that makes visible both the very large [e.g., global patterns of communication densities between countries (State et al. 2012)] and the very small [e.g., hourly changes in emotional affect and microbehaviors such as doing homework, getting drunk, or getting a headache (Golder & Macy 2011; see also **http://timeu.se**, which provides an interactive tool for plotting the prevalence of keywords over the course of the day and week)].

Online behavior is recorded in real time rather than retrospectively. In social network studies, when individuals are given "name generators" and surveyed about their communication patterns, they are subject to a variety of potential biases. Question wording and ordering can cause respondents to artificially limit or otherwise vary the individuals they report, leading to underestimates of network size (Fischer

2009, Pustejovsky & Spillane 2009) or even to measures of some other network (Burt 1997) when survey questions mistakenly elicit a report of a social tie outside the researcher's intended scope. Online behavior—time-stamped and passively recorded—provides an unambiguous recording of when, and with whom, each individual communicated.

When activities are recorded via mobile devices, real-time mosaic accounts of collective behavior become possible that otherwise could not be reconstructed. As smartphone use increases in prevalence, the offline context of online behavior becomes available, such as common participation in a public event. For example, sampling a corpus of tweets (brief messages posted on Twitter) that occurred during a certain time range and within a limited radius of a given event can reconstruct how online activity complemented a parade or demonstration or add a geographic variable back into an analysis that is otherwise blind to spatial location.

Relatedly, online behavior is observed unobtrusively, limiting the potential for Hawthorne-type effects in which researcher-induced desirability bias could inhibit normatively inappropriate behaviors (e.g., expressions of racial and ethnic prejudice) that participants might self-censor in surveys and laboratory studies (Zizzo 2010). Observing behavior unobtrusively ensures that the social pressures and normative constraints on individuals are exerted by their peers rather than by the researchers. For example, online dating sites provide an unprecedented opportunity to study the effects of racial and ethnic preferences on mate selection choices. Using a sample of 6,000 online profiles from Yahoo! personals, Robnett and Feliciano (Feliciano et al. 2009, Robnett & Feliciano 2011) found that, among those Whites who stated racial preferences in their online profiles, men were more likely to exclude Blacks than other racial categories, whereas women were more likely to exclude Asians. Similar results were reported by the online dating site OK Cupid (**http://blog.okcupid.com**), which showed that Black women received replies at lower rates, and women of several races preferred White men over men of other races. In another dating-related study, Taylor et al. (2011) found support for the "matching hypothesis" that people seek partners whose perceived social desirability matches their own self-assessment.

Moreover, online interactions have been characterized as "persistent conversations" (Erickson 1999) that can be observed in real time, even if the observation is taking place after the fact. Unlike ephemeral in-person conversations, online conversations are recorded with perfect fidelity and can persist forever. Although we must take care when analyzing documentary evidence out of its original historical context, long after perspectives and circumstances have changed, the conversations themselves can be largely reconstructed, allowing retrospective analyses to be far more complete and exact than in most archival research.

The task for the researcher is to see online behavior as social behavior, the kind that might occur in any field site, be it a remote village, a law office, or a high school cafeteria. Some researchers explicitly conceptualize online sites as field sites in the ethnographic sense (Lyman & Wakeford 1999). Relatedly, online behavior in social media represents social action in the Weberian sense—action that is oriented toward others (Weber 1922), involving what Weber called "*verstehen*"—the subjective meaning for the actors involved. Paccagnella (1997) noted the multiple ways one might interpret the purpose, use, and limitations of technology, and hence the need not to conflate the meaning to the researcher with the meaning for users (Pinch & Bijker 1984).

The Virtual Laboratory

Although most research using online data has been observational, a growing number of studies use the web as a virtual lab for controlled experiments. Experiments address a key limitation of all observational studies, online or off: the inability to measure a phenomenon free from potentially confounding unmeasured

factors. For example, it has been difficult to distinguish between contagion on a social network and common exposure of network neighbors to some unobserved source of similarity. An outbreak of sneezing, for example, could indicate a spreading virus or common exposure to seasonal allergens. Aral et al. (2009) reviewed numerous statistical techniques that have been proposed to tease these processes apart using observational data and concluded that none are sufficient, a conclusion also reached by Shalizi & Thomas (2011).

Controlled experiments with random assignment are one solution. In a path-breaking pair of studies, Centola (2010, 2011) created a web-based health information community in which the levels of clustering and homophily in users' social networks could be manipulated. By randomly assigning participants to conditions, Centola removed shared environment and homophily as sources of network autocorrelation, leaving only the possibility for contagion as an explanation. He found that the rate and extent of contagion were higher in the clustering condition than in the random condition (Centola 2010), consistent with the predictions of theoretical models (Centola & Macy 2007) of the spread of simple and complex contagions on small-world networks in which complex contagions benefit from the social reinforcement provided when multiple neighbors become infected. This social reinforcement is more likely when the network is highly clustered. Adoption was also greater in the homophilous condition (Centola 2011) than in the random condition, with no variation in network structure. Although these findings advance our knowledge of social influence and diffusion, there is an even larger methodological message in these studies about the possibilities for randomized trial experiments in virtual labs.

Experimentation online offers several advantages as well as challenges compared with traditional offline experiments in laboratory settings. An obvious advantage is the greater economy of scale. For example, Centola's (2010) online experiment with repeated involvement of up to 144 participants per iteration would be logistically prohibitive in the lab, but once an online system is built for a few users, the marginal cost of scaling it up to hundreds or even thousands of users is relatively minimal. Larger numbers of participants not only increase statistical power but also allow new research opportunities. For example, it becomes possible to test hypotheses about changes in collective behavior, in which groups rather than individuals are the units of analysis.

Less obvious but arguably more important theoretically, scalable experiments allow multiple simultaneous realizations of the same starting conditions. This enables researchers to test the possibility that highly nonrandom patterns may nevertheless have very limited predictive value if the patterns observed in one world vary widely (and perhaps entirely randomly) from another owing to processes that are path dependent or that confer cumulative advantage. This possibility was demonstrated for the first time by Salganik and colleagues (Salganik & Watts 2009, Salganik et al. 2006) in an experiment that has become an instant classic. They varied the level of social influence on a music download site that they created for the research. When participants were subject to influence, the researchers found that music preferences were highly nonrandom in each world, making it possible to predict what would be downloaded simply by knowing how many others in that world had downloaded the same song. The surprising result was that this information was not very helpful for predicting what songs would be downloaded in another world. Their findings are a telling reminder to academic researchers, marketing departments, campaign managers, and epidemiologists that statistically significant patterns can be randomly generated by highly path-dependent processes such as social diffusion (Watts 2011).

Other online experiments have used existing websites rather than creating their own. Bond et al. (2012) tested the effects of social influence on voter turnout by manipulating whether Facebook users were exposed to information about the number of their friends who had voted. Although this experiment required the

cooperation of Facebook, such cooperation is reinforced by the widespread use of online experiments by industry. Web practitioners are already familiar with "A/B testing," in which multiple versions ("A" and "B") of a website are created and visitors are randomly assigned to one version or another to test the effects of different layouts, colors, or content on user engagement, retention, click-throughs, and so on. In many cases, studies motivated by theoretical questions can piggyback on the practical needs of industry to better understand user behavior.

Researchers conducting otherwise traditional laboratory studies may now turn to an online labor pool. Amazon's Mechanical Turk is an online labor market with a vast global user base that is culturally, geographically, and demographically far more diverse than the undergraduate psychology majors that populate most offline participant pools. Touted by Amazon as "artificial artificial intelligence," Mechanical Turk is designed to be a programmatic means of having humans complete tasks for which artificial intelligence is inferior, such as summarizing a document or choosing the best of five photographs. Typical compensation per task ranges from a few cents to a dollar, depending on the time required. Mason & Suri (2011) provide a review of methodological issues arising in the use of Mechanical Turk for online experiments. Rand (2012) points to a number of inherent limitations in nearly all online experiments, such as the inability to maintain consistency in and control over participants' immediate physical surroundings, with the associated risk that results may be contaminated by distractions or by outside sources of information. Additionally, so-called Turkers sometimes click mindlessly simply to complete the task, which requires steps to detect random clicking and failure to follow instructions. Incentives may also not operate as intended because Turkers appear to anchor on payment levels so that paying more makes them believe they deserve more, producing a greater quantity of work but not at a greater level of quality (Mason & Watts 2009).

RESEARCH APPLICATIONS

Social Networks, Contagion, and Diffusion

Social network analyses have been among the earliest studies to use online data. Although numerous social networking sites exist, researchers have focused on two of the largest, Twitter and Facebook, with over 300 million and 1 billion worldwide users, respectively. Facebook profiles contain rich demographic data, including full names; dates of birth; geo-location; affiliations with friends, organizations, and political and social movements; and cultural tastes. Though less demographically rich, Twitter data are much easier to obtain via a more open API (application programming interface; see the section entitled "Methods, Skills, and Training" below). Private data from Facebook are not generally available for research purposes, but several strategies exist for researchers to use Facebook data. First, researchers may build apps, or add-on applications, that, when adopted by users, allow researchers access to users' demographic and behavioral data. These apps can be narrowly targeted to just those users with the desired demographic traits, network properties, or cultural or political preferences. However, researchers need to keep in mind that reliance on self-selection means that the result is a nonrandom convenience sample whose results cannot be generalized even to the targeted subpopulation. Second, researchers may invite participants into the lab the way they might for any other lab experiment, who then log into their Facebook account (Gilbert & Karahalios 2009). Several studies have leveraged a Facebook policy that allowed people affiliated with the same university to see a more detailed user profile than is otherwise generally available (Lewis et al. 2008, 2011; Wimmer & Lewis 2010). These studies examined a complete university cohort to study homophily patterns in race as well as cultural tastes.

Some researchers have arranged with Facebook staff to gain access to anonymized

private user data for research purposes. For example, Golder et al. (2007) showed that private messaging by nonfriends took place primarily at late-night hours, Traud et al. (2010) compared the network structures of multiple universities, and Mayer & Puller (2008) modeled tie formation within one university. Some researchers have collaborated with Facebook's own internal research team to analyze private data as well as conduct large-scale experiments. Das & Kramer (2013) examined inhibition in self-expression, but this was only possible because of the internal logging that takes place on messages that users write but ultimately choose not to post. Bond et al. (2012) isolated the effects of social influence from mass-media influence in increasing likelihood to vote by conducting a massive experiment on 61 million Facebook users.

boyd & Ellison (2007) identified distinctive structural aspects of social networking sites: a personal profile and a publicly visible list of network neighbors (who share a tie). They note that the visibility of others' egocentric networks varies by site and as the sites themselves change over time. For example, LinkedIn makes some profile aspects visible only to paid users (viewer and viewed). Twitter allows users to view indirectly the content received by those they follow only if the user also follows those same people.[1] Facebook requires symmetric social ties (two friends must each indicated friendship with the other), whereas Twitter and most blogging platforms allow asymmetric ties, leading to an extremely long-tailed degree distribution (e.g., celebrities often have many thousands of followers). Some demand a clear tie to one's offline identity (e.g., Google Plus and Facebook), while most do not, though even among the latter, participants often choose to establish a verified identity, especially on blogs and online dating sites where credibility is needed. These varying requirements impact users' behavior, helping some spaces to flourish and elicit trust and cooperation, while others exhibit distrust and hostility. These differences in turn open up important research opportunities for understanding how variations in structure, norms, cultural protocols, and incentives affect individual and collective outcomes.

Borgatti & Halgin (2011) distinguish two types of network ties based on their persistence over time—states (e.g., kinship and friendship) and events (e.g., exchanges and conversations). A further distinction can be made between ties of affiliation (e.g., participation in the same event) and interactions (e.g., discussing the event). Ties can also be positively signed (attraction, friendship, cooperation) or negatively signed (repulsion, antipathy, conflict), and they can be directed (listening, liking) or undirected (marriage, kinship, partnership). Online social networks share these properties. Leskovec et al. (2010) examined tie formation in online networks including Epinions, Slashdot, and Wikipedia and found that undirected ties are formed as predicted by structural balance theory (the product of signs in a balanced triad must be positive), but when ties are directed, status effects appear to play the larger role (e.g., if A defers to B and B defers to C, then C is unlikely to defer to A).

A number of studies have used online networks to confirm two classic findings on the importance of ties that span large network distances, Granovetter's (1973) "The Strength of Weak Ties" and Burt's (1992) *Structural Holes*. For example, Eagle et al. (2010) used national telephone logs among 65 million subscribers (about 90% of the population) to show that diversity in the networks of the members of a community was positively related to economic development, confirming the offline network results reported by Granovetter and by Burt. Gilbert & Karahalios (2009) studied the relationship between tie strength and connectivity using data from Facebook. They employed an innovative approach to developing a metric for online tie strength. Whereas some studies often

[1] That is, if A follows B, then A can see all of B's messages, but if B and C engage in a conversation, this is visible to A only if he follows both B and C. The purpose of this is ostensibly due more to preventing cluttering A's message stream with irrelevant conversations than to protecting the privacy of B's and C's conversation.

rely on the number of messages exchanged as a metric for tie strength, Gilbert & Karahalios used multiple indicators, including exchange of photos and public and private messages. In their lab study, they instrumented a web browser to collect participants' Facebook activity and compared this with participants' ratings of the strength of their ties to various friends. A similar study compared the volume and direction of messages, retweets, and @mentions among Twitter followers with the same users' offline friends and discovered a close correspondence (Xie et al. 2012).

Other research has replicated Milgram's classic investigation of the small-world phenomenon in which letters traveled through the mail through a chain of acquaintances until a target unknown to the originator was reached, which revealed the celebrated "six degrees of separation" (Milgram 1967, Travers & Milgram 1969). Dodds et al. (2003) found a similar average path length in an experimental study of search on global email networks, and Leskovec & Horvitz (2008), using a global instant messenger network of 240 million users, observed a mean path length of 6.6 steps, comparable to Milgram's 5.2. A similar analysis of the global Facebook network (Ugander et al. 2011) found that the number of steps separating users had declined from 5.3 in 2008 to 4.7 in 2011 as the network grew in size.

Massive network data have also enabled the study of how structural conditions affect the spread of social contagions, including the decision to join a group, adopt a convention, or spread information. Bakshy et al. (2012) used news feed posts for 250 million Facebook users to show that novel information spread primarily through weak ties. In contrast, using mobile phone call records for 4.6 million subscribers (about 20% of the national population), Onnela et al. (2007) found that although weak ties "held the network together" (in that disconnection of the network into isolated components was most vulnerable to deletion of these ties), most information traveled through ties of intermediate strength. The authors conclude that models of network structure typically rely on global characteristics such as betweenness, implicitly weighting all ties equivalently, but tie strength may play a larger role than the global characteristics.

Backstrom et al. (2006) investigated social influence in two social networks, LiveJournal (an online blogging community) and DBLP (a database of academic paper coauthors). The offline coauthorship network differs from the online blogging community in requiring far greater personal interaction and coordination. Nevertheless, the likelihood of joining a community (on LiveJournal) and attending a conference (evidenced by DBLP) both increased not only with the number of network neighbors who had joined, but more surprisingly, also with the number of closed triads among these neighbors. A possible explanation is that the ties between two neighbors are stronger when the triad is closed (the two neighbors of an actor are also neighbors of each other, as found by Van der Leij & Goyal 2011). In addition, closed triads may entail greater fear of exclusion and more closely synchronized communications, leading to stronger social influence than when the triad is open.

Ugander et al. (2012), using Facebook-internal data about users' and nonusers' email addresses, investigated how the probability that a user would accept an invitation changed with the number of Facebook neighbors and the number of "connected components" (connected only by links through ego). Contrary to the result reported by Backstrom et al. (2006), the authors found that the probability increased not with the number of Facebook neighbors but with the number of connected components, even after controlling for demographic diversity. A possible explanation is that invitees discount multiple invitations from friends who know one another, interpreting these invitations as conveying redundant information about the benefits of membership. Because the data are missing two potentially important social ties—Facebook friends who did not include ego in their contact lists and friends who are not on Facebook—there is no way to know if

distinct connected components in the observed ego network might actually be connected by these missing links.

Romero et al. (2011) found evidence to support the theory of complex contagions (Centola & Macy 2007) by examining the spread of the use of Twitter hashtags. Hashtags for controversial topics like politics were more likely to be adopted following exposure to multiple adopting neighbors, compared with topics like music or sports. More recently, Weng et al. (2013) used Twitter hashtags to confirm a key implication of the theory of complex contagions—that the spread of complex contagions depends on network structure, a result that is consistent with the experimental findings reported by Centola (noted above). Other studies have used online data to test long-standing theories about information diffusion, including the existence of well-connected influentials who initiate cascades. Popularized by Gladwell (2000) in *The Tipping Point*, the theory of these high-degree network nodes (or hubs) was earlier proposed by Katz & Lazarsfeld (1955), who referred to them as "opinion leaders" in a two-step model of the flow of influence. Billions of advertising dollars are targeted at so-called influentials based on this theory, but a growing number of studies cast serious doubts. Dodds et al. (2003) found that successfully completed chains in their replication of Milgram's "six degrees" study did not in fact leverage highly connected hubs. Cha et al.'s (2010, p. 10) study of 1.7 billion tweets found that hubs "are not necessarily influential in terms of spawning retweets or mentions," a result consistent with Kwak et al. (2010) that also casts doubt on the influence of widely followed users on Twitter. Similarly, Bakshy et al. (2011) identify cases in which actors with average degree are the source, and González-Bailón et al. (2011) point to the importance of random seeds as well as nodes with higher centrality.

Exchange, Cooperation, and Trust

A growing number of studies are using online data to address enduring problems of trust and cooperation in social exchange, in which the valued goods being exchanged are time, attention, information, and status. Research by State et al. (2012) is consistent with a basic principle of exchange theory (Homans 1958, 1961; Emerson 1962, 1972), that exchange relations tend to be reciprocally balanced. They found that "couchsurfers" (people who are part of an online community of budget travelers who stay in others' homes) compensate their hosts' hospitality by conferring status in the form of public comments.

Attention is also a valued resource in social exchange. Podolny (2001) suggests that attention is a prism or lens through which one is judged by others; having the attention of powerful others can, in turn, redound to one's financial benefit and is a signal to others about who is worth an investment of attention. Online experiments confirm that individuals are willing to exchange monetary compensation for praise and attention from peers, even when it is artificial (Huberman et al. 2004). More broadly, Huberman's research program centering on the "attention economy" created by online interactions addresses the puzzle created by the sheer volume of information available online, which makes attention scarce and valuable. Yet little is known about how attention is directed or attracted. Twitter users have been shown, for example, to rate others as more interesting to the extent that their own neighbors expressed interest in those others (Golder & Yardi 2010).

Like attention, trust is another resource that can be especially important in online interactions where identities can be ephemeral, limiting the reliability of reputational information and the ability to punish cheaters (Friedman & Resnick 2001). In response, users have evolved norms to regulate behavior, such as requiring newcomers to a community to be first to commit to the exchange. In a pioneering study of online interactions, Kollock (1999) observed this practice in a community of bootleg tape recording traders, who also collectively paid enforcement costs by maintaining a blacklist of people who should not be traded with due to perceived past transgressions. Other studies

have confirmed a principle originally proposed by Hechter (1988) that it is more effective to reward trustworthy behavior than to punish transgressions because the latter creates incentives to increase the costs of detection. Friedman & Resnick (2001) attribute the remarkable effectiveness of the eBay feedback system in part to the incentives the system creates to maintain one's identity (rather than changing names to hide negative feedback), an incentive that increases over time.

Exchange-theoretic analysis can also be applied to personal as well as business and organizational relationships. For example, Backstrom & Kleinberg (2014) randomly selected 1.3 million adult Facebook users to test the effect of network embeddedness (defined as the overlap in their friendship circles) on the formation and durability of romantic relationships. Surprisingly, they found that dispersion (or lack of overlap), not embeddedness, was conducive to successful relationships, a result that contradicts the theory of the strength of embedded ties but is consistent with a previously unexplored romantic implication of Burt's (1992) theory of structural holes—that people are attracted to partners who can fill in structural holes.

Online research on social exchange includes survey research as well as observational studies. Willer et al. (2012) administered surveys on Freecycle and Craigslist to compare the levels of solidarity reported by the sites' respective members. The results confirmed the exchange-theoretic hypothesis that the generalized exchange of Freecycle entails greater levels of solidarity than the negotiated exchange taking place on Craigslist.

Collective Action and Social Movements

Many online communities rely on voluntary contributions by large numbers of unrelated individuals, presenting researchers with a remarkable opportunity to address long-standing puzzles in the study of collective action: How do order and consensus emerge among loosely affiliated contributors, and what motivates them to contribute to this public good? Two prominent examples are open-source task groups like Wikipedia and Linux and massively multiplayer online games such as World of Warcraft and Everquest. An overview of these two areas is provided by Contractor (2013). Wikipedia is an openly editable collaborative encyclopedia written and edited by thousands of volunteers every day. Like many voluntary associations in the offline world, Wikipedia, Usenet, and many other online communities are self-governed almost entirely by the evolving normative obligations and limits collectively established and agreed to by their participants, but with the critical difference that the detailed evolutionary records are preserved for study by the scientific community.

As Wikipedia has grown, its community of editors has created a number of policies to guide contributors and to resolve disputes, such as policies requiring articles to be written from a neutral point of view and to include statements only if they can be supported by reference to a publicly available source (not firsthand research by the editor) (Kriplean et al. 2007, 2008). Although these institutional arrangements help to regulate and coordinate user behavior, they also make the motivation to contribute even more puzzling because there is less opportunity to exploit the community to promote a parochial point of view. Anthony et al. (2009) examined the quality of Wikipedians' contributions and pose the interesting puzzle that "anonymous Good Samaritans" contributed among the highest-quality content, whereas Welser et al. (2011) point out that Wikipedians self-organize into roles, focusing on "cleaning up vandalism," providing domain expertise, and so on.

Many of the challenges faced by formal organizations—recruiting a skilled labor force, defining roles and responsibilities, and monitoring and rewarding performance—also arise in massively multiplayer online roleplaying games, or MMORPGs. Players can take on a particular role (trolls, warriors, etc.), and they

can unite to form guilds (teams), work together to attack other guilds, and perform in-game tasks such as achieving quests. As with the communities in Wikipedia, guilds must overcome collective action and coordination problems in order to select, train, and retain members. Choi et al. (2008) found that a good fit between persons and tasks is associated with longer membership in a guild, whereas Wang et al. (2011) found that players' orientation toward performance and achievement displayed greater expertise, and those oriented toward having an immersive experience displayed less expertise.

To date, most of the research on these communities has been largely descriptive, and a vast opportunity remains for researchers to use data from user interactions to test hypotheses derived from the collective action, public goods, and game-theoretic literature. Data from Wikipedia are freely available for download,[2] and Sony has made Everquest data available for academic research.

Collective action and social movement mobilization have also been studied—by scholars as well as government agencies—using data from social media, particularly Twitter and Facebook. For example, data from Twitter have been used to provide digital traces of the spread of protest information and public sentiment in the Arab Spring (González-Bailón et al. 2011). Because information about protests reaches people through numerous channels besides social media, it is impossible to isolate the effects of social media net of other channels. However, users' messages can be used to measure the rate and extent of mobilization by tracking topic changes in user-generated content at a very fine-grained temporal level. Topic changes can be associated with changes in the users' social and spatial environment and considered in light of the locations represented in news accounts. For example, Weber et al. (2013) tracked secular versus Islamist postings by Egyptian Twitter users over the course of the Arab Spring.

Researchers have also used changes in the distribution of user-generated content not only to explain political outcomes but also to try to predict them. For example, Digrazia et al. (2013) showed that local election of Republicans was positively correlated with the number of times Republicans were mentioned in tweets. Nevertheless, a review of recent papers (Gayo-Avello 2012) concluded that predictive claims may be exaggerated. One important limitation on predictive power is that users of social media are not randomly selected in the way that is possible with survey research. Users preferentially choose to follow sources that conform to their existing worldviews (Sunstein 2001) and preferentially rebroadcast (retweet) conforming messages, as well (Conover et al. 2011). Boutyline & Willer (2011) showed that there is a valence effect to the formation of so-called echo chambers—those farther to the political right exhibited more ideological homophily in who they chose to follow on Twitter.

Krebs (2008) took a different approach to analyze the "red-blue" divide between Republicans and Democrats using online data. Krebs constructed a social network of the top 100 political books sold on Amazon in three time periods, 2003, 2004, and 2008. The network edges corresponded to copurchases ("customers who bought this book also bought ___"). In each year, almost all of the political books were tightly grouped into red and blue clusters, with only one or two books (e.g., *Ghost Wars* and *Rise of the Vulcans*) linking the two camps. This result is consistent with the red-blue ideological clustering reported by Adamic & Glance (2005) using data collected from blogs during the 2004 US electoral cycle.

These studies show that the use of social media to study opinion dynamics provides a potentially important complement to—not substitute for—traditional survey methods. Each can be used to obtain information that is missing in the other. Surveys provide more reliable estimates of the distribution of opinion

[2] **http://en.wikipedia.org/wiki/Wikipedia:Database_download**.

in the underlying population but typically provide only retrospective responses and lack network data with which to study the flow, diffusion, and clustering of opinion.

CHALLENGES

The Privacy Paradox

These new data confront researchers with imposing hurdles, ranging from validity of both the data and how it is sampled to the ethical issues regarding its use. Online data present a paradox in the protection of privacy: Data are at once too revealing in terms of privacy protection, yet also not revealing enough in terms of providing the demographic background information needed by social scientists. Online data often lack the detailed demographic profile information that is standard in survey research. For example, although Twitter data are public, many users provide sparse, invented, incomplete, or ambiguous profile information, making it difficult for researchers to associate the content of tweets or the attributes of network nodes with basic demographic measures such as age, gender, ethnicity, or location. Identity is slippery and poorly defined in some online communities where participants are known only by a self-chosen username that they may change at any time. In some cases, it is difficult to tell who is a human; the growing incidence of spam accounts (fake Twitter accounts created to send marketing or other unwelcome messages) is worrisome, and despite progress in spam-detection methods (Yardi et al. 2010), spammers manage to circumvent these methods and keep the "arms race" going. As spammers become more sophisticated, it becomes harder for social scientists to clean the data they collect without specialized technical training, a problem we explore in more detail below.

Nevertheless, rapid progress is being made to address these limitations. For example, Compton et al. (2014) showed how label-propagation algorithms can be adapted to potentially geo-locate most Twitter users to within a few kilometers, and Jernigan & Mistree (2009) showed how Facebook content can be used to infer a wide range of user attributes, including age, gender, sexual preference, and political party affiliation.

These advances illustrate the other side of the dilemma—that online data may not be private enough. These new sources of data raise challenging procedural, legal, and ethical questions about how to protect individual privacy that are beyond the scope of this review, but there is a growing body of research showing that anonymizing or encrypting data is not sufficient for protecting privacy, as this can sometimes be reverse engineered (Backstrom et al. 2007, de Montjoye et al. 2013) using the unique attributes of individuals' egocentric networks or physical mobility patterns.

Access to private data can be a significant challenge. Most online data are owned by private corporate entities who may restrict access in large part because of concern over protecting the privacy of their subscribers. These restrictions have raised concerns about reproducibility of results, corporate influence, and stratification in the research community between a small elite that is well connected to social media companies and everyone else (boyd & Crawford 2012, Huberman 2012). New protocols and institutional arrangements are needed to align the goals and needs of industry and the academic community. Online companies compete aggressively to attract academic talent (including social scientists), and several companies maintain university relations departments that can help to facilitate research collaboration. In addition, advanced programming and other technical skills are required to access and process large semistructured data sets, as we describe in more detail below.

Measurement Issues

Although advances in identifying sentiment and opinion from text are proceeding rapidly (Pang & Lee 2008), we can only measure inner states indirectly, through their behavioral expression. For example, psychological lexicons (Pennebaker et al. 2001) can be used to measure the expression of affective rhythms on

a global scale (Golder & Macy 2011), but these methods cannot account for temporal lags between expression and experience. Moreover, asynchronous communication allows users to introspect and revise what they write, such that what appear to be spontaneous expressions of an underlying mental state may instead by self-censored and deliberate (Das & Kramer 2013).

As noted above, an important limitation in all observational studies of network contagion, whether online or offline, is the difficulty distinguishing between homophily and contagion. Homophily refers to a variety of selection mechanisms by which a social tie is more likely between individuals with similar attributes and environmental exposures (McPherson et al. 2001). Contagion refers to influence mechanisms (e.g., imitation or peer pressure) by which traits diffuse along network edges. Homophily and contagion offer competing explanations for network autocorrelation, which refers to the greater similarity in the attributes of closely connected nodes. Based on simulated networks, Shalizi & Thomas (2011, p. 216) conclude that "there is just no way to separate selection from influence observationally" (see also Manski 1993). This does not mean that observational studies using online networks are useless, but researchers need to refrain from assuming that the observed network autocorrelation reflects contagion effects and to acknowledge that the similarity between adjacent nodes may reflect the mutually reinforcing effects of influence and selection whose separate contributions may be impossible to tease apart. For example, although Ugander et al. (2012) controlled for demographic similarity (sex, age, and nationality), there are countless other ways in which shared environments, affiliations, interests, and personality traits might cause two friends to join Facebook independently but not on the same day, making it look like the "early adopter" influenced the friend they invited who would have joined anyway.

One solution is to conduct controlled experiments that manipulate exposure to a possible contagion, as in the Facebook experiment by Bond et al. (2012) noted above. Where experimental methods are not feasible and the only data are observational, researchers can tease apart influence and selection by using an instrumental variable that is associated with alter's exposure to the contagion but not with ego's exposure to the contagion and then comparing the presence of the contagion among egos with and without exposed alters (Imbens & Angrist 1994).

Another fundamental problem in online as well as offline network studies is deciding what constitutes a social tie (Butts 2009). In survey-based research on ego networks, controversy has centered on how to ask respondents to nominate a friend. In studies of online communication networks based on telephone logs, email traffic, or Twitter messages, a key question is how to determine the type and number of exchanges (e.g., emails or wall postings) that are necessary to indicate the existence of an enduring social relationship (Borgatti & Halgin 2011). For example, in their network analysis of UK telephone logs, Eagle et al. (2010) required at least one call in each direction, which is the most widely used threshold in studies that use communication data to identify social networks. Other studies have examined robustness and changes of results across a range of thresholds (Romero et al. 2011, Adamic & Adar 2005, De Choudhury et al. 2010) both with and without a requirement of bidirectionality. Studies using Twitter data face the additional question of which kind of relation to include: follower relations, @mentions in which users refer/reply to others (Honeycutt & Herring 2009), or retweets in which users repost what others have written (boyd et al. 2010, Conover et al. 2011).

A related issue is whether the metric for establishing a link is consonant with the actors' conception of a social relationship. These questions arise as well in studies of offline networks, particularly affiliation networks, for which a number of heuristics have been proposed to determine whether the edge corresponds to an actual interaction, such as similarity (Flynn et al. 2010), regularity of structure and kinship terms (Brashears 2013), and indications of

instrumental versus sentimental ties (Freeman 1992). One study of Twitter networks confirmed that follower relations do not correspond to offline friendships. However, they developed new algorithms that detect offline friendships using a novel measure of user closeness (Xie et al. 2012).

A similar issue arises in deciding where to set the threshold for which users to include in the analysis, given that active participation in online environments is often highly skewed (Preece & Shneiderman 2009). Low-activity users may not represent committed members, but arbitrary thresholds may also have the effect of artificially excluding a large number, even a majority, of individuals from the analysis, with potentially misleading effects on network measures like density, degree distribution, and mean path length.

Is the Online World a Parallel Universe?

Researchers also face the challenge of generalizing from online to offline behavior. Interactions online differ in important and obvious ways from those offline, including the lifting of geographic and temporal constraints of face-to-face communication. For example, the ability to wait to answer an email or text message or respond to a status update affords the opportunity to introspect and be more deliberate and strategic about one's self-presentation (Goffman 1959). The anonymity permitted by some online platforms frees users to invent an entirely new persona, raising doubts about the credibility of demographic profile data. Anonymity can also permit or encourage the production of the vitriolic speech that pervades many online conversations but is generally unthinkable offline. Differences between online and offline modes of communication have been the subject of a number of studies focusing on their comparative richness, or the bandwidth available for the transmission of verbal and visual cues (Daft & Lengel 1986, Walther 2007). Although face-to-face interaction is richer visually, there are other aspects of online communication that can be much denser than their offline counterparts, such as the ready availability of persistent histories (Hollan & Stornetta 1992) and the opportunity to craft novel modes of expression such as the emoticon or Twitter @reply (Herring 1999, Honeycutt & Herring 2009, Menchik & Tian 2008).

Early studies also raised questions about possible distorting effects of online access. The displacement theory posited that Internet use was an asocial activity that took time away from family and friends (Nie & Hillygus 2002), and empirical research showed that these effects varied, depending on the type of online activity (Kraut & Kiesler 2003). However, this research predates the social media era, and more recent research (J.P. Robinson 2011) suggests that the displacement theory is less relevant today.

The digital divide raises additional concerns about generalizing from the online to offline populations. The online population tends to be younger, better educated, and more affluent than the general public, which also raises important questions about the potential for reproducing and even amplifying social stratification. Even where technological access is available, the skills to make use of that access remains unevenly distributed (DiMaggio et al. 2001, Hargittai 2010, L. Robinson 2011), likely leading to biased levels of participation across kinds of online spaces.

Nevertheless, these differences do not warrant the widely used distinction between the web and the real world, with the implication that users enter a metaphysical realm every time they open their browser. The online world is not identical to the offline, but it is entirely real. Users who desire status, admiration, social approval, and attention in their offline relationships will bring those desires with them to their online networks. Individuals must navigate many of the same social obstacles online that they do offline as they seek information, political support, friendship, romance, or consumer goods.

Although the activities and populations in the online world differ from their offline counterparts, the differences are rapidly declining as

Internet access and use of the web becomes increasingly universal and online interactions become more fully integrated with people's daily offline activities. Today, most US adults (64%) are Internet users, and home Internet access predominantly takes the form of high-speed, broadband connections (Horrigan 2009). Mobile phones are rapidly replacing desktop computers as the online portal, as 45% of US adults in 2013 reported having a smartphone. Paradoxically, cell phone use is increasing particularly fast in developing countries that lack the infrastructure for landline access.

Because of the network externality of communication media, as the numbers of users increases, online resources have become the primary mechanism by which people engage in many everyday activities, such as following the news; arguing about politics, sports, music, and movies; maintaining social ties with friends and family; shopping; dating; and even seeking employment. An early study (Wellman & Hampton 1999) found that online and offline networks were already merging as early as the 1990s, as neighborhoods and local communities began to use electronic communications tools to augment their existing modes of communication. Today mobile technologies and social media websites such as Facebook and Twitter provide a seamless transition between the offline and online worlds (Rainie & Wellman 2012, Xie et al. 2012). Although some online communities continue to permit pseudonyms, both Facebook and Google Plus require users to disclose (and verify) their offline identity, supplemented by a detailed profile that includes location, photograph, organizational affiliations, and interests and activities. Even in environments in which users remain pseudonymous, they often establish long-standing and cherished identities and reputations that they are reluctant to cast off.

The mobile web is particularly important in bridging the online and offline worlds. Survey respondents reported feeling an obligation to have their mobile phone on at all times so as to not miss out on a social interaction (Smith 2011, 2012). When Internet access took place primarily at a desk, the temporal patterns of electronic social interaction typically matched the temporal patterns of school and work (Golder et al. 2007, Grinter & Palen 2002). Those temporal and spatial constraints are loosening as Internet access becomes increasingly mobile, allowing users to interact online in the course of their offline activities in real time, independently of time and place.

Representativeness of research participants has long been a concern in lab experiments, especially when the subject pool has consisted largely of college sophomores (Sears 1986), recently dubbed "WEIRD," an acronym for Western, educated, industrialized, rich, and democratic (Henrich et al. 2010). Although some of these sample biases exist in online access as well, online communities in many cases span not only age and class ranges but also diverse global cultures.

Online communities may differ fundamentally in the demographic profile of their users, not only from the offline world but even from other online communities, but these differences can open up research possibilities. Just as offline social clubs, community groups, street gangs, firms, and specialized organizations can be opportunities for comparative case studies, so too can highly idiosyncratic online communities like couchsurfing or local message boards. In sum, online interaction is already deeply woven into the daily experience of millions of people worldwide, and the numbers are rapidly growing. Although differential levels of access, skills, and engagement persist, those differences are declining as usage becomes increasingly universal. For millions, socializing, dating, shopping, and learning take place in a digital environment that is second nature.

Methods, Skills, and Training

A primary obstacle to online research by social scientists is the need for advanced technical training to collect, store, manipulate, analyze, and validate massive quantities of semistructured data, such as text generated by hundreds of millions of social media users. In addition,

advanced programming skills are required to interact with specialized or custom hardware, to execute tasks in parallel on computing grids composed of hundreds of nodes that span the globe, and simply to ensure that very large amounts of data consume memory efficiently and are processed using algorithms that run in a reasonable amount of time. As a consequence, the first wave of studies of online behavior and interaction has been dominated by physical, computer, and information scientists who may lack the theoretical grounding necessary to know where to look, what questions to ask, or what the results may imply. In the short term, multidisciplinary collaborations can be highly fruitful, but the long-run solution is for graduate programs in the social sciences to adapt to the era of big data by providing training in skills that are needed for online research. The list includes

- Making use of programming interfaces. Many commercial services, in the interest of interoperating with other services as well as with third-party software developers, provide application programming interfaces (APIs) that allow data to be downloaded from the service in a structured and permissible way. To use an API, the researcher must typically first register for an API key, or unique access token, and then write a script to successively query the service and retrieve the desired information.
- Manipulating unstructured data and nested data structures. Data retrieved via APIs is often structured very differently from the flat files that social scientists are trained to work with. Online data are likely to have nested structures, as in XML or JSON documents, that cannot be directly imported into standard statistical packages. Learning to use regular expressions facilitates data transformation from human-readable to machine-readable format.
- Creating web pages and databases to collect and store surveys or online experiments. Online services like Survey Monkey and Amazon Mechanical Turk make it possible to conduct online surveys and experiments easily and inexpensively, but for studies that require specialized platforms, researchers may need to build a custom website.
- Manipulating and storing large data sets. Finding the degree distribution or average path length in a social network with hundreds of millions of individuals or the relative frequency of positive and negative emotion words in a large text corpus (Golder & Macy 2011) could be impractical or impossible on a single computer. However, the problem of computational load can be addressed by parallelizing the task on a computer cluster. Among the most important innovations in computing in the past decade has been the development of the MapReduce programming paradigm (Dean & Ghemawat 2004) and the availability of commodity cloud storage. Developed at Google to process the petabytes of web pages the search engine collects, MapReduce provides a convenient way to process data that are too large to process (or even fit) on a single computer. A series of transformations are performed in succession on subsets of a large data set, each of which resides on a different computer or processing node. Following these transformations, aggregations are performed so that summary statistics may be generated. Storing large data sets has similarly been made easier because of the availability of commodity cloud storage. For example, Amazon.com's Web Services (**http://aws.amazon.com**) and Microsoft's Azure (**http://windowsazure.com**) platform rent Internet-based computing resources, such as servers that can be used for pennies per hour or storage that costs pennies per gigabyte. Researchers faced with spending thousands of dollars of research funds on computer hardware may find this to be a cost-effective alternative because there

is no upfront cost, resources may be turned off when no longer needed, and information technology staff do not need to be hired to support or manage the equipment.

- Machine learning, sentiment analysis, and topic modeling. Machine learning refers to statistical techniques that use past observations to classify new observations or make predictions about the associated outcomes. These techniques may be useful when data have nonlinear relationships or a large number of variables that interact in a complex system in ways that cannot be modeled by traditional regression-based methods. Applications range from understanding natural human language to detecting which emails are spam. Researchers may code only a random sample from a massive set of observations and approximate the rest. Text analysis techniques such as Latent Dirichlet Allocation perform unsupervised topic modeling or automatic clustering of the words found in a body of texts into topical groups (Blei et al. 2003) by examining the co-occurrences of the words found within. Sentiment analysis uses a combination of statistical techniques and human-created lexicons to identify the valence and intensity of various emotional states expressed in a body of text. Libraries that perform some of these techniques are available in R or in standalone software packages such as University of Waikato's Weka (**http://www.cs.waikato.ac.nz/ml/weka**) or Stanford's Topic Modeling Toolbox (**http://nlp.stanford.edu/software/tmt/tmt-0.4**).

One reason these methods have not gained greater currency in the social sciences is that many current applications are deliberately atheoretical, placing higher value on the ability to predict future observations than on testing a theoretically motivated hypothesis. However, one should not throw out the methodological baby with the atheoretical bath water. After all, every research method, from linear regression to participant observation, can be applied descriptively, with little or no theoretical direction, or analytically, in a program of research that targets the underlying causal mechanisms. Online data open up transformative possibilities for both descriptive and analytical studies, but without the automated data management and coding tools developed by computer scientists, the analysis of massive unstructured data will remain beyond the reach of most social scientists, leaving the field to disciplines that are much better at building powerful telescopes than at knowing where to point them (Lazer et al. 2009, Watts 2011). Although few social science departments are currently able to incorporate these skills into graduate methods courses, interested students can be directed to computer and information science departments for specialized training.

CONCLUSION

In the earliest days of the field of information theory, Claude Shannon's (1956) "The Bandwagon" essay warned that the flurry of interest in the new field would generate a large amount of low-quality work but that this should not lead the research community to conclude that this was an inherent limitation. On the contrary, it should be taken as an exhortation to focus on producing more rigorous studies. Shannon's advice may apply as well to the coming era of online social science. The unprecedented opportunity to observe human behavior and social interaction in real time, at a microscopic level yet on a global scale, is attracting widespread interest among scientists with the requisite skills to mine these data but not always with the theoretical background needed to guide the inquiry. Studies that identify patterns of behavior or map social landscapes invite dismissal as atheoretic empiricism, but this may be shortsighted. These pioneering studies should instead be taken as evidence not of the most that can be learned from online research but of the vast opportunities that lie ahead for a new science of social life.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This research was supported by grants from the US National Science Foundation SES 1226483 and IIS 0910664, Google, and the National Research Foundation of Korea NRF-2013S1A3A2055285.

LITERATURE CITED

Adamic L, Adar E. 2005. How to search a social network. *Soc. Netw.* 27(3):187–203

Adamic LA, Glance N. 2005. The political blogosphere and the 2004 U.S. election: divided they blog. *Proc. 3rd Int. Workshop Link Discov.*, pp. 36–43. New York: ACM

Anderson CJ, Wasserman S, Crouch B. 1999. A p^* primer: logit models for social networks. *Soc. Netw.* 21(1):37–66

Anthony DL, Smith SW, Williamson T. 2009. Reputation and reliability in collective goods: the case of the online encyclopedia Wikipedia. *Ration. Soc.* 21(3):283–306

Aral S, Muchnik L, Sundararajan A. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl. Acad. Sci. USA* 106:21544–49

Backstrom L, Dwork C, Kleinberg J. 2007. Wherefore art thou R3579X? Anonymized social networks, hidden patterns, and structural steganography. *Proc. 16th Int. Conf. World Wide Web*, pp. 181–90. New York: ACM

Backstrom L, Huttenlocher D, Kleinberg J, Lan X. 2006. Group formation in large social networks: membership, growth and evolution. *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 44–54. New York: ACM

Backstrom L, Kleinberg J. 2014. Romantic partnerships and the dispersion of social ties: a network analysis of relationship status on Facebook. *Proc. 17th ACM Conf. Comput. Support. Coop. Work* (*CSCW*), *Baltimore, MD, Feb. 15–19*, pp. 831–41. New York: ACM

Bakshy E, Hofman JM, Mason WA, Watts DJ. 2011. Everyone's an influencer: quantifying influence on Twitter. *Proc. 4th ACM Int. Conf. Web Search Data Min.*, pp. 65–74. New York: ACM

Bakshy E, Rosenn I, Marlow C, Adamic L. 2012. The role of social networks in information diffusion. *Proc. 21st Int. Conf. World Wide Web*, pp. 519–28. New York: ACM

Blei DM, Ng AY, Jordan MI. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022

Bond RM, Fariss CJ, Jones JJ, Kramer ADI, Marlow C, et al. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489(7415):295–98

Borgatti SP, Halgin DS. 2011. On network theory. *Organ. Sci.* 22:1168–81

Boutyline A, Willer R. 2011. *The social structure of political echo chambers: Ideology leads to asymmetries in online political communication networks*. Work. Pap., Univ. Calif. Berkeley. **http://www.ocf.berkeley.edu/~andrei/downloads/echo.pdf**

boyd d, Crawford K. 2012. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Inf. Commun. Soc.* 15(5):662–79

boyd d, Ellison NB. 2007. Social network sites: definition, history, and scholarship. *J. Comput.-Mediat. Commun.* 13(1):210–30

boyd d, Golder SA, Lotan G. 2010. Tweet, tweet, retweet: conversational aspects of retweeting on Twitter. *Proc. 43rd Hawaii Int. Conf. Syst. Sci. (HICSS-43), Kauai, HI, Jan. 6*, pp. 1–10. Piscataway, NJ: IEEE

Brashears ME. 2013. Humans use compression heuristics to improve the recall of social networks. *Sci. Rep.* 3:1513

Burt RS. 1992. *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard Univ. Press

Burt RS. 1997. A note on social capital and network content. *Soc. Netw.* 19:355

Butts CT. 2009. Revisiting the foundations of network analysis. *Science* 325:414–16

Centola D. 2010. The spread of behavior in an online social network experiment. *Science* 329:1194–97

Centola D. 2011. An experimental study of homophily in the adoption of health behavior. *Science* 334(6060):1269–72
Centola D, Macy M. 2007. Complex contagions and the weakness of long ties. *Am. J. Sociol.* 113(3):702–34
Cha M, Haddadi H, Benevenuto F, Gummadi KP. 2010. Measuring user influence in Twitter: the million follower fallacy. *Proc. 4th Int. AAAI Conf. Weblogs Soc. Media*, pp. 10–17. Menlo Park, CA: AAAI
Choi B, Kraut R, Fichman M. 2008. *Matching people and groups: recruitment and selection in online games*. Work. Pap. No. 92, Human-Computer Interaction Inst., Carnegie Mellon Univ., Pittsburgh, PA. **http://repository.cmu.edu/hcii/92**
Compton R, Jurgens D, Allen D. 2014. *Geotagging one hundred million Twitter accounts with total variation minimization*. **http://arxiv.org/abs/1404.7152**
Conover MD, Ratkiewicz J, Francisco M, Gonçalves B, Flammini A, Menczer F. 2011. Political polarization on Twitter. *Proc. 5th Int. AAAI Conf. Weblogs Soc. Media*, pp. 89–96. Menlo Park, CA: AAAI
Contractor N. 2013. Some assembly required: leveraging Web science to understand and enable team assembly. *Philos. Trans. R. Soc. A* 371:1987
Daft RL, Lengel RH. 1986. Organizational information requirements, media richness and structural design. *Manag. Sci.* 32(5):554–71
Das S, Kramer A. 2013. Self-censorship on Facebook. *Proc. 7th Int. AAAI Conf. Weblogs Soc. Media*, pp. 120–27. Palo Alto, CA: AAAI
De Choudhury M, Mason WA, Hofman JM, Watts DJ. 2010. Inferring relevant social networks from interpersonal communication. *Proc. 19th Int. Conf. World Wide Web* (*WWW2010*), pp. 301–10. New York: ACM
de Montjoye Y-A, Hidalgo CA, Verleysen M, Blondel VD. 2013. Unique in the crowd: the privacy bounds of human mobility. *Sci. Rep.* 3:1376
Dean J, Ghemawat S. 2004. MapReduce: simplified data processing on large clusters. *Proc. OSDI'04: 6th Symp. Oper. Syst. Des. Implement.*, pp. 137–49. Berkeley, CA: USENIX Assoc.
Della Posta D, Shi Y, Macy MW. 2013. *Why do liberals drink lattes?* Work. Pap., Cornell Univ. Soc. Dyn. Lab., Ithaca, NY
Digrazia J, McKelvey K, Bollen J, Rojas F. 2013. More tweets, more votes: social media as a quantitative indicator of political behavior. *PLoS ONE* 8:e70449
DiMaggio P, Hargittai E, Neuman WR, Robinson JP. 2001. Social implications of the Internet. *Annu. Rev. Sociol.* 27:307–36
Dodds PS, Muhamad R, Watts DJ. 2003. An experimental study of search in global social networks. *Science* 301:827–29
Duncan OD. 1966. Path analysis: sociological examples. *Am. J. Sociol.* 72(1):1–16
Eagle N, Macy MW, Claxton R. 2010. Network diversity and economic development. *Science* 328:1029–31
Emerson RM. 1962. Power-dependence relations. *Am. Sociol. Rev.* 27(1):31–41
Emerson RM. 1972. Exchange theory, part I and II. In *Sociological Theories in Progress*, Vol. II, ed. J Berger, M Zelditch, B Anderson, pp. 58–87. Boston: Houghton Mifflin
Entwisle B, Faust K, Rindfuss RR, Kaneda T. 2007. Networks and contexts: variation in the structure of social ties. *Am. J. Sociol.* 112(5):1495–533
Erickson T. 1999. Persistent conversation: an introduction. *J. Comput.-Mediat. Commun.* 4(4). doi: 10.1111/j.1083-6101.1999.tb00105.x
Feliciano C, Robnett B, Komaie G. 2009. Gendered racial exclusion among white internet daters. *Soc. Sci. Res.* 38:39–54
Fischer CS. 2009. The 2004 GSS finding of shrunken social networks: an artifact? *Am. Sociol. Rev.* 74(4):657–69
Flynn FJ, Reagans RE, Guillory L. 2010. Do you two know each other? Transitivity, homophily and the need for (network) closure. *J. Personal. Soc. Psychol.* 99(5):855–69
Freeman LC. 1992. Filling in the blanks: a theory of cognitive categories and the structure of social affiliation. *Soc. Psychol. Q.* 55(2):118–27
Friedman E, Resnick P. 2001. The social cost of cheap pseudonyms. *J. Econ. Manag. Strategy* 10(2):173–99
Gayo-Avello D. 2012. No, you cannot predict elections with Twitter. *IEEE Internet Comput.* 16(6):91–94

Gilbert E, Karahalios K. 2009. Predicting tie strength with social media. *Proc. 27th Int. Conf. Hum. Factors Comput. Syst.*, pp. 211–20. New York: ACM
Gladwell M. 2000. *The Tipping Point: How Little Things Can Make a Big Difference*. New York: Little, Brown
Goffman E. 1959. *The Presentation of Self in Everyday Life*. New York: Anchor
Golder SA, Macy MW. 2011. Diurnal and seasonal mood vary with work, sleep and daylength across diverse cultures. *Science* 333:1878–81
Golder SA, Wilkinson D, Huberman BA. 2007. Rhythms of social interaction: messaging within a massive online network. In *Communities and Technologies 2007: Proceedings of the Third Communities and Technologies Conference*, ed. C Steinfield, BT Pentland, M Ackerman, N Contractor, pp. 41–66. London: Springer
Golder SA, Yardi S. 2010. Structural predictors of tie formation in Twitter: transitivity and mutuality. *Proc. 2nd IEEE Int. Conf. Soc. Comput., Aug. 20–22, Minneapolis, MN*, pp. 88–95. Piscataway, NJ: IEEE
González-Bailón S, Borge-Holthoefer J, Rivero A, Moreno Y. 2011. The dynamics of protest recruitment through an online network. *Sci. Rep.* 1:197
Granovetter M. 1973. The strength of weak ties. *Am. J. Sociol.* 78:1360–80
Grinter RE, Palen L. 2002. Instant messaging in teen life. *Proc. 2002 ACM Conf. Comput.-Support. Coop. Work*, pp. 21–30. New York: ACM
Hampton K, Wellman B. 2003. Neighboring in netville: how the Internet supports community and social capital in a wired suburb. *City Community* 2(4):277–311
Hargittai E. 2010. Digital na(t)ives? Variation in Internet skills and uses among members of the net generation. *Sociol. Inq.* 80(1):92–113
Hechter M. 1988. *Principles of Group Solidarity*. Berkeley: Univ. Calif. Press
Henrich J, Heine SJ, Norenzayan A. 2010. The weirdest people in the world? *Behav. Brain Sci.* 33:61–83
Herring SC. 1999. Interactional coherence in CMC. *J. Comput.-Mediat. Commun.* 4(4). doi: 10.1111/j.1083-6101.1999.tb00106.x
Hollan J, Stornetta S. 1992. Beyond being there. *Proc. SIGCHI Conf. Hum. Factors Comput. Syst. (CHI'92)*, pp. 119–25. New York: ACM
Homans GC. 1958. Social behavior as exchange. *Am. J. Sociol.* 63(6):597
Homans GC. 1961. *Social Behavior: Its Elementary Forms*. New York: Harcourt Brace Jovanovich
Honeycutt C, Herring SC. 2009. Beyond microblogging: conversation and collaboration via Twitter. *Proc. 42nd Hawaii Int. Conf. Syst. Sci.*, pp. 1–10. Washington, DC: IEEE
Horrigan J. 2009. *Home broadband adoption 2009*. Rep. Pew Res. Cent., June 17, Pew Internet and American Life Project, Washington, DC. **http://www.pewinternet.org/files/old-media/Files/Reports/2009/Home-Broadband-Adoption-2009.pdf**
Huberman BA. 2012. Sociology of science: Big data deserve a bigger audience. *Nature* 482:308
Huberman BA, Lock CH, Onculer A. 2004. Status as a valued resource. *Soc. Psychol. Q.* 67(1):103
Imbens GW, Angrist JD. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62(2):467
Jernigan C, Mistree BFT. 2009. Gaydar: Facebook friendships expose sexual orientation. *First Monday* 14(10)
Katz E, Lazarsfeld P. 1955. *Personal Influence*. New York: Free Press
Kollock P. 1999. The production of trust in online markets. In *Advances in Group Processes*, Vol. 16, ed. EJ Lawler, SR Thye, MW Macy, HA Walker, pp. 99–123. Greenwich, CT: JAI
Kossinets G, Watts DJ. 2006. Empirical analysis of an evolving social network. *Science* 311:88–90
Kraut R, Kiesler S. 2003. The social impact of Internet use. *Psychol. Sci. Agenda* Summer, pp. 8–10
Krebs V. 2008. *A network of books about recent US politics sold by the online bookseller Amazon.com*. Case Stud., Orgnet.com. **http://www.orgnet.com/divided.html**
Kriplean T, Beschastnikh I, McDonald DW. 2008. Articulations of WikiWork: uncovering valued work in Wikipedia through barnstars. In *CSCW 08: Proc. ACM 2008 Conf. Comput. Support. Coop. Work*, pp. 47–56. New York: ACM
Kriplean T, Beschastnikh I, McDonald D, Golder SA. 2007. *Community, consensus, coercion, control: CS*W or how policy mediates mass participation*. Presented at ACM Group 2007, Nov. 4–7, Sanibel Island, FL. **http://www.acm.org/conferences/group/conferences/group07**
Kuhn TS. 1962. *The Structure of Scientific Revolutions*. Chicago: Univ. Chicago Press. 3rd ed.
Kwak H, Lee C, Park H, Moon S. 2010. What is Twitter, a social network or a news media? *Proc. 19th Int. Conf. World Wide Web (WWW'2010)*, pp. 591–600. New York: ACM

Lazer D, Pentland A, Adamic L, Aral S, Barabási A-L, et al. 2009. Computational social science. *Science* 323:721–23

Leskovec J, Horvitz E. 2008. Planetary-scale views on a large instant-messaging network. *Proc. 17th Int. Conf. World Wide Web*, pp. 915–24. New York: ACM

Leskovec J, Huttenlocher D, Kleinberg J. 2010. Signed networks in social media. *Proc. ACM Conf. Human Factors Comput. Syst.* (*CHI 2010*), pp. 1361–70. New York: ACM

Lewis K, Gonzalez M, Kaufman J. 2011. Social selection and peer influence in an online social network. *Proc. Natl. Acad. Sci. USA* 109(1):68–72

Lewis K, Kaufman J, Gonzalez M, Wimmer A, Christakis N. 2008. Tastes, ties, and time: a new social network dataset using Facebook.com. *Soc. Netw.* 30:330–42

Lovelace RVE, Tyler GL. 2012. On the discovery of the period of the Crab nebula pulsar. *Observatory* 132(3):186–88

Lyman P, Wakeford N. 1999. Going into the (virtual) field. *Am. Behav. Sci.* 43(3):359–76

Macy MW, Willer R. 2002. From factors to actors: computational sociology and agent-based modeling. *Annu. Rev. Sociol.* 28(1):143–66

Manski CF. 1993. Identification of endogenous social effects: the reflection problem. *Rev. Econ. Stud.* 60(3):531–42

Marsden PV. 1990. Network data and measurement. *Annu. Rev. Sociol.* 16:435–63

Mason W, Suri S. 2011. Conducting behavioral research on Amazon's Mechanical Turk. *Behav. Res. Methods* 44:1–23

Mason W, Watts DJ. 2009. Financial incentives and the performance of crowds. *Proc. ACM SIGKDD Worksh. Human Comput.* (*HCOMP '09*), pp. 77–85. New York: ACM

Mayer A, Puller SL. 2008. The old boy (and girl) network: social network formation on university campuses. *J. Public Econ.* 92:329–47

McPherson M. 2004. A Blau space primer: prolegomenon to an ecology of affiliation. *Ind. Corp. Change* 13(1):263–80

McPherson M, Smith-Lovin L, Cook JM. 2001. Birds of a feather: homophily in social networks. *Annu. Rev. Sociol.* 27:415–44

Menchik DA, Tian X. 2008. Putting social context into text: the semiotics of e-mail interaction. *Am. J. Sociol.* 114(2):332–70

Milgram S. 1967. The small-world problem. *Psychol. Today* 1(1):61–67

Nie NH, Hillygus DS. 2002. The impact of internet use on sociability: time-diary findings. *IT Soc.* 1(1):1–20

Onnela J-P, Saramaki J, Hyvonen J, Szabó G, Lazer D, et al. 2007. Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. USA* 104(18):7332–36

Paccagnella L. 1997. Getting the seats of your pants dirty: strategies for ethnographic research on virtual communities. *J. Comput.-Mediat. Commun.* 3(1). doi: 10.1111/j.1083-6101.1997.tb00065.x

Pang B, Lee L. 2008. Opinion mining and sentiment analysis. *Found. Trends Inform. Retr.* 2(1–2):1–135

Pennebaker JW, Francis ME, Booth RJ. 2001. *Linguistic Inquiry and Word Count* (*LIWC*): *LIWC2001*. Mahwah, NJ: Erlbaum

Pinch TJ, Bijker WE. 1984. The social construction of facts and artefacts: or how the sociology of science and the sociology of technology might benefit each other. *Soc. Stud. Sci.* 14(3):399–441

Podolny JM. 2001. Networks as pipes and prisms of the market. *Am. J. Sociol.* 107(1):33–60

Preece J, Shneiderman B. 2009. The reader-to-leader framework: motivating technology-mediated social participation. *AIS Trans. Hum.-Comput. Interact.* 1(1):13–32

Pustejovsky JE, Spillane JP. 2009. Question-order effects in social network name generators. *Soc. Netw.* 31(4):221–29

Rainie L, Wellman B. 2012. *Networked: The New Social Operating System*. Cambridge, MA: MIT Press

Rand DG. 2012. The promise of Mechanical Turk: how online labor markets can help theorists run behavioral experiments. *J. Theor. Biol.* 299:172–79

Robinson JP. 2011. IT use and leisure time displacement. *Inf. Commun. Soc.* 14(4):495–509

Robinson L. 2011. Information channel preferences and information opportunity structures. *Inf. Commun. Soc.* 14(4):472–94

Robnett B, Feliciano C. 2011. Patterns of racial-ethnic exclusion by internet daters. *Soc. Forces* 89(3):807–28
Romero DM, Meeder B, Kleinberg J. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. *Proc. 20th Int. Conf. World Wide Web* (*WWW2011*), pp. 695–704. New York: ACM
Salganik MJ, Dodds PS, Watts DJ. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311:854–56
Salganik MJ, Watts DJ. 2009. Web-based experiments for the study of collective social dynamics in cultural markets. *Top. Cogn. Sci.* 1:439–68
Sears DO. 1986. College sophomores in the laboratory: influences of a narrow data base on social psychology's view of human nature. *J. Personal. Soc. Psychol.* 51(3):515–30
Shalizi CR, Thomas AC. 2011. Homophily and contagion are generically confounded in observational social network studies. *Sociol. Methods Res.* 40(2):211–39
Shannon C. 1956. The bandwagon. *IRE Trans. Inf. Theory* 2:3
Smith A. 2011. *13% of online adults use Twitter and half of Twitter users access the service on a cell phone*. Rep. Pew Res. Cent., June 1, Pew Internet and American Life Project, Washington, DC. **http://pewinternet.org/Reports/2011/Twitter-Update-2011.aspx**
Smith A. 2012. *The best* (*and worst*) *of mobile connectivity*. Rep. Pew Res. Cent., Nov. 30, Pew Internet and American Life Project, Washington, DC. **http://www.pewinternet.org/2012/11/30/the-best-and-worst-of-mobile-connectivity**
State B, Abrahao B, Cook K. 2012. From power to status in large scale online exchanges. *Proc. 4th ACM Conf. Web Sci.* (*WebSci*), *June 22–24.* New York: ACM
Sunstein CR. 2001. *Republic.com.* Princeton, NJ: Princeton Univ. Press
Taylor LS, Mendelsohn GA, Fiore AT, Cheshire C. 2011. "Out of my league": a real-world test of the matching hypothesis. *Personal. Soc. Psychol. Bull.* 37:942–54
Traud AL, Kelsic ED, Mucha PJ, Porter MA. 2010. Comparing community structure to characteristics in online collegiate social networks. *SIAM Rev.* 53:526–43
Travers J, Milgram S. 1969. An experimental study of the small world problem. *Sociometry* 32(4):425–43
Ugander J, Backstrom L, Marlow C, Kleinberg J. 2012. Structural diversity in social contagion. *Proc. Natl. Acad. Sci. USA* 109(16):5962–66
Ugander J, Karrer B, Backstrom L, Marlow C. 2011. *The anatomy of the Facebook social graph*. **http://arxiv.org/abs/1111.4503**
Van der Leij M, Goyal S. 2011. Strong ties in a small world. *Rev. Netw. Econ.* 10(2):1
Walther JB. 2007. Selective self-presentation in computer-mediated communication: hyperpersonal dimensions of technology, language, and cognition. *Comput. Hum. Behav.* 23:2538–57
Wang J, Huffaker DA, Treem JW, Fullerton L, Ahmad MA, et al. 2011. Focused on the prize: characteristics of experts in massive multiplayer online games. *First Monday* 16(8). **http://firstmonday.org/ojs/index.php/fm/article/viewArticle/3672/3028**
Watts DJ. 2004. The "new" science of networks. *Annu. Rev. Sociol.* 30:243–70
Watts DJ. 2011. *Everything Is Obvious: How Common Sense Fails Us.* New York: Crown Business
Weber I, Garimella VRK, Batayneh A. 2013. Secular versus Islamist polarization in Egypt on Twitter. *Proc. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Min.*, pp. 290–97. New York: ACM
Weber M. 1922. *Economy and Society*. Berkeley: Univ. Calif. Press
Wellman B, Hampton K. 1999. Living networked on and offline. *Contemp. Sociol.* 28(6):648–54
Welser HT, Cosley D, Kossinets G, Lin A, Dokshin F, et al. 2011. Finding social roles in Wikipedia. *Proc. 2011 iConf.*, pp. 122–29. New York: ACM
Weng L, Menczer F, Ahn Y-Y. 2013. Virality prediction and community structure in social networks. *Sci. Rep.* 3:2522
Willer R, Flynn FJ, Zak S. 2012. Structure, identity, and solidarity: a comparative field study of generalized and direct exchange. *Adm. Sci. Q.* 57(1):119–55
Wimmer A, Lewis K. 2010. Beyond and below racial homophily: ERG models of a friendship network documented on Facebook. *Am. J. Sociol.* 116(2):583–642
Wright S. 1934. The method of path coefficients. *Ann. Math. Stat.* 5(3):161–215

Xie W, Li C, Zhu F, Lim E-P, Gong X. 2012. When a friend in Twitter is a friend in life. *Proc. 3rd Annu. ACM Web Sci. Conf.—WebSci '12*, pp. 344–47. New York: ACM

Yardi S, Romero DM, Schoenebeck G, boyd d. 2010. Detecting spam in a Twitter network. *First Monday* 15(1). **http://firstmonday.org/ojs/index.php/fm/article/view/2793/2431**

Zachary WW. 1977. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* 33:452–73

Zizzo DJ. 2010. Experimenter demand effects in economic experiments. *Exp. Econ.* 13(1):75–98

Annual Review
of Sociology

Volume 40, 2014

Contents

Urban and Rural Community Sociology

Policy

Sociology and World Regions

Indexes

Errata

An online log of corrections to *Annual Review of Sociology* articles may be found at http://www.annualreviews.org/errata/soc