1-1-2013

Social Science in the Era of Big Data

Sandra González-Bailón
University of Pennsylvania, sgonzalezbailon@asc.upenn.edu

Social Science in the Era of Big Data

Abstract
Digital technologies keep track of everything we do and say while we are online, and we spend online an increasing portion of our time. Databases hidden behind web services and applications are constantly fed with information of our movements and communication patterns, and a significant dimension of our lives, quantified to unprecedented levels, gets stored in those vast online repositories. This article considers some of the implications of this torrent of data for social science research, and for the types of questions we can ask of the world we inhabit. The goal of the article is twofold: to explain why, in spite of all the data, theory still matters to build credible stories of what the data reveal; and to show how this allows social scientists to revisit old questions at the intersection of new technologies and disciplinary approaches. The article also considers how Big Data research can transform policymaking, with a focus on how it can help us improve communication and governance in policy-relevant domains.

Keywords
networks, complexity, interactions, social influence, public opinion, governance

Disciplines
Communication | Social and Behavioral Sciences

Social Science in the Era of Big Data

Sandra González-Bailón

Digital technologies keep track of everything we do and say while we are online, and we spend online an increasing portion of our time. Databases hidden behind web services and applications are constantly fed with information of our movements and communication patterns, and a significant dimension of our lives, quantified to unprecedented levels, gets stored in those vast online repositories. This article considers some of the implications of this torrent of data for social science research, and for the types of questions we can ask of the world we inhabit. The goal of the article is twofold: to explain why, in spite of all the data, theory still matters to build credible stories of what the data reveal; and to show how this allows social scientists to revisit old questions at the intersection of new technologies and disciplinary approaches. The article also considers how Big Data research can transform policymaking, with a focus on how it can help us improve communication and governance in policy-relevant domains.

KEY WORDS: networks, complexity, interactions, social influence, public opinion, governance

Introduction: The End of Theory?

The increasing availability of Big Data for research (in private or public institutions) and the design of interventions (from marketing to public administration) has divided public opinion into two camps: the skeptics, who question the legitimate use of that data on the basis of privacy and other ethical concerns (Morozov, 2013); and the enthusiasts, who focus on the transformational impact of having more information than ever before (Mayer-Schoenberger & Cukier, 2013). This latter camp is further divided into two groups. There are some who believe that Big Data will radically change the way in which we make sense of the world, mostly by making theory and interpretation less necessary: The data —so the argument goes—will speak for themselves and override the need for theoretical models. Alternatively, there are those who believe the exact opposite: That theory and interpretation are more necessary than ever before if we are to find the appropriate layer of information in what otherwise is an unnavigable sea. The main argument of those who proclaim the "end of theory" (Anderson, 2008) is that the most measured and recorded age in history demands a different approach to data: being able to track human behavior with unprecedented fidelity and precision is more powerful, they claim, than imperfect models of people behavior. The counterargument to this is that data-driven approaches underestimate the role played by researchers in the analytical process; disentangling signal from noise is still a subjective matter, as is providing the context that will help identify meaningful correlations and discard those that are unsubstantial (Silver, 2012). This article addresses the divide that separates these two views, arguing that those who adopt the latter view offer a more accurate picture of how social science can benefit from Big Data in order to advance its research agenda.

What follows is an overview of recent research that shows why theory still matters, and why the data, more often than not, do not "speak for themselves." It also discusses why old theoretical questions, formulated back in the day when data were small (if they existed at all), can help reset the sail necessary to catch the wind of Big Data. Prior to this, however, four caveats are

necessary. First, the adjective "big" refers here less to the size of the data sets and more to their spatial and temporal resolution, meaning that the data are richer than before and that they span several levels of analysis, from the individual to the collective. The "size" of digital data is significant because it requires the logistics and expertise to make their storage more efficient and manageable; but what makes Big Data unique is their higher level of detail and refinement in the quality of observations, not just the number of data points or the amount of memory that their storage takes.

Second, the argument that follows focuses mostly on data that keep track of communication dynamics and social interactions, and less on transactional records of atomized behavior, like the purchasing history of customers. If there is something that Internet technologies have highlighted like no other technology before is the importance of interdependence and the complexity that interactions add to social dynamics. Big Data can help illuminate that complexity—which lies at the core of social science research (Coleman, 1990; Schelling, 1978; Watts, 2011) —with an impressive level of detail, and promises to yield significant theoretical advances in the study of social change. The third caveat is that this is by no means a complete review of all the papers that deserve mention: those that are discussed should be seen as representative of what is a fast growing area of work, to which this article can barely make justice. Finally, although this paper is a review of how Big Data can change Social Science, implicit in the message is the assumption that the analysis of these large data sets is not the dominion of a single discipline or approach; it requires a joint effort where different research traditions and methods converge (Lazer et al., 2009; Watts, 2007). The research reviewed in this paper is a testament to how fruitful such multidisciplinary efforts can be.

The Communicative Web

In the introduction to a collection of his essays, Stanley Milgram, one of the most influential social psychologists to date, wrote:

> If the world were drained of every individual and we were left only with the messages that passed between them, we would still be in possession of the information needed to construct our discipline. For every truly socio-psychological phenomenon is rooted in communication. (Milgram, 1977, p. 317; original emphasis)

This collection of essays goes under the title "The Individual in a Communicative Web" where (this being the 1970s) "web" does not refer to the world wide web but rather to the structure of interactions that ties people together—a structure that, back when Milgram wrote his essays was abstract and intangible. We all now have clear pictures in our minds of what that structure looks like: we have all seen a version of our social graph in Facebook; LinkedIn tells us how many degrees away we are from a particular person; and Twitter allows us to quickly identify the most prominent users in terms of their broadcasting power by looking at their number of followers. But when Milgram was researching the channels of communication that tie us together, as he put it, these channels were invisible; there was not a clear idea of how people are connected to people, or of how the structure of those paths assembles at a societal level. There were some theoretical models that tried to assess this structure (de Sola Pool & Kochen, 1978) but not an empirical map charting the paths that allow us to navigate the connections of

friends and acquaintances in the "communicative web." This is why Milgram decided to design his famous small world experiment (Milgram, 1967; Travers & Milgram, 1969)—to measure the intangible, and test an old intuition that at the same time was very counterintuitive: With all the people that exist in the world, how is it possible that we are all just a few handshakes away from each other?

In his experiments he asked randomly selected people in Kansas and Nebraska to send a letter to a person living in Massachusetts; there was one constraint: The letter could only be sent to someone they knew on a first-name basis who they thought would be in a better position to take it closer to the final destination. Most initiated chains reached that destination in six steps, confirming the idea that we live in a small world, long held in popular wisdom and from which the concept of "six degrees of separation" arises (Watts, 2003). However, what was less perfect is that only about a quarter of all chains were completed (N= 44 in the first study, N = 64 in the second). The small N creates a problem because it undermines the generality of findings, and also because attrition is adding a bias to the data: it might be creating the illusion that the chains are shorter than they really are. In other words, the chains that were broken before they reached their destination might have changed the observed distribution had they been completed, and shifted the average to the right—hence making the world a bit larger than the experiment suggested.

But then the Internet came in, and the "communicative web" stopped being a metaphor. The Internet created a context where other researchers could revisit Milgram's experiments, capitalizing on the ability to recruit a large number of participants online (Dodds, Muhamad, & Watts, 2003). This time, the experiment was based on 18 target people in 13 different countries, and it involved more than 60,000 Internet users. Instead of a letter, the researchers asked participants to forward an email, coming up with about 24,000 distinct message chains. Again, they found that most chains were completed after a small number of steps (four, in this case, a bit less than what Milgram found in his experiments); they also found that attrition was once again a problem: only 384 chains out of the approximately 20,000 initiated reached the final destination. After estimating the length of the abandoned chains, however, the authors concluded that, had they continued to the destination, about half of them would be expected to reach their target in seven or fewer steps.

This study provided more robust evidence of the small world phenomenon, and qualified some of the original findings. It also suggested that social search is a nontrivial problem: it is one thing to claim that we all live in a small world; it is another to find the one shortcut that brings us closer to a random person (Kleinberg, 2000; Watts, Dodds, & Newman, 2002). The small world nature of social networks highlights a global feature of those structures: there might be nodes that, being better connected than average, create shortcuts in the network and make it shrink (much in the same way that some airports are more important routing hubs than others in the air transportation system); but when we navigate those networks, we usually just manage local information. We know who we know, and who is acquainted to those we know, but we cannot go much further than that in delineating the shortest path to a random unknown person. Finding that chain of connections requires global knowledge, but we often only have local information. Digital data are allowing us to understand better how we navigate networks. They might also be changing our mental maps of those networks, and our ability to navigate them.

The small world research just discussed offers the perfect example of how the web, and digital communication more generally, allows a scaling up of the data sets that we can use for research, which is itself a precondition to build better theories—in this case, of how we self-organize in social networks. But Milgram had other ideas that are also being revisited now with the help of digital technologies. For instance, he wrote several essays on the experience of living in cities, and the psychological maps that people build of the cities they live in. If we were asked to draw a map of our city, we would emphasize the areas that are most important to us, or those that are more familiar—not necessarily the best that the city can offer. Milgram's contention was that these personalized maps reveal a lot of information of a person's history, social class, and aspirations; so uncovering these maps, and identifying the patterns that are universal and those that are specific can help design better planning interventions (Milgram, 1977, chapters 7 and 8). A group of computer scientists have built an online game that borrows this idea and puts it to the test in the streets of London (Quercia, Pesce, Almeida, & Crowcroft, 2013). The game picks up random locations from Google Street View and presents them individually to users to see if they can identify the location by naming the borough, region, or closest Tube station. These researchers aim to test assumptions made in urban planning, and to see which areas of the city are more recognizable for different people—and thereby to contribute to the "smart city" agenda, which aims to use advances in computer infrastructure to manage the complexity of urban life. One of the advantages of this approach compared to Milgram's is that, again, the sample size could be increased, in this case, by an order of magnitude: from a couple of hundred subjects to a couple of thousand.

This experiment offers another example of how online technologies are being used to test and develop old social theories about how we interact with each other and with the spaces we inhabit. A classic urban theorist, Jane Jacobs, held in the 1960s that street life is what defines the character and wealth of cities. What she called "the many little public sidewalk contacts" fulfill a self-regulatory role that cannot be engineered from above. As she put it:

> Most of it is ostensibly utterly trivial but the sum is not trivial at all. The sum of such casual public contact at a local level—most of it fortuitous, most of it associated with errands, all of it metered by the person concerned and not thrust upon him by anyone—is a feeling for the public identity of people, a web of public respect and trust, and a resource in time. (Jacobs, 1961, 73)

These observations, insightful but anecdotal for the most part, can now be examined on a much larger scale with the help of location-based services and technologies. Only with large-scale data can we assess why the sum is less trivial than the parts, and how the "web of public respect and trust" is constructed. Researchers have analyzed large data sets drawn from mobile phone communication (González, Hidalgo, & Barabási, 2008) and location-based software (Noulas, Scellato, Lambiotte, Pontil, & Mascolo, 2012) to offer insights that were beyond reach at the time Jacobs wrote her observations. For instance, they can measure the predictability of individual mobility patterns, identify highly frequented locations, and analyze how social networks (the public web to which Jacobs referred) emerge out of those locations. Most importantly, the existence of these data is encouraging the development of tools and methods that allow researchers to tackle the complexity of urban life, which before could only be depicted

with impressionistic strokes. As important as these are in understanding the drivers and mechanisms of urban dynamics, general models can be more helpful in designing and planning interventions aimed at improving city life.

Communication, Effervescence, and Social Influence

Other areas where Big Data have allowed a revisiting of old theoretical questions include the study of communication and public opinion, collective effervescence, and social influence. As far back as the late nineteenth century, Gabriel Tarde wrote a series of essays on imitation, publics, and opinion formation (Clark, 1969). In them, he laid the foundations of what has become known as the two-step flow of information, which refers to the influence of opinion leaders as mediators between the media and the public (Katz, 1957). In an essay entitled "Opinion and Conversation" he wrote:

> If no one conversed, the newspapers would appear to no avail—in which case one cannot conceive of their publication because they would exercise no profound influence on any minds. They would be like a string vibrating without a sounding board. (Clark, 1969, chapter 17)

Tarde's idea is that communication is a strong agent of imitation, and the main channel through which sentiments, ideas, and actions spread. According to this view, newspapers—which back in his day were the main source of public information—are just part of the conversation. Unlike crowds, the public that arises out of communication (fed by the invention of printing) are not constrained by physical space; the implication is that the effects of imitation, and the ideas it helps spread, scale up with the growth of the public. As he put it:

> One should thus not be surprised to see our contemporaries so pliant before the wind of passing opinion, nor should one conclude from this that characters have necessarily weakened. When poplars and oaks are brought down by a storm, it is not because they grew weaker but because the wind grew stronger. (Clark, 1969, chapter 17)

Digital media, indeed, have made the wind of public opinion blow stronger. The question is: do they have any of the effects that Tarde assumed? The data resulting from online communication have been used to test his theoretical ideas with the benefit of measurements and metrics that were, again, out of reach in his time. The role of mainstream media in the spreading of news, and the mediating role played by opinion leaders, have been analyzed by tracking millions of Twitter accounts (Wu, Hofman, Mason, & Watts, 2011). The conclusions reveal that almost half the information that originates from the media is transmitted to the public indirectly, that is, through the mediation of users that fall in the category of opinion leaders. They also reveal that attention, measured as number of follower links and tweets received, is highly concentrated: about 0.05 percent of users account for almost half of public attention. Another article analyzed the flow of information between the media and the public by tracking millions of mainstream media sites and blogs (Leskovec, Adamic, & Huberman, 2007). Their findings suggest a typical lag of about 2.5 hours between the peak of attention in the news media around specific topics, and the corresponding peak in blogs; they also found that only about 3.5 percent of all topics analyzed tend to diffuse from blogs to news media; the vast majority of them travel

in the opposite direction. Put together, these and related studies reveal that public opinion is still monopolized by a minority of actors, and that these tend to be in the media or celebrity categories. So while digital media allow us to be part of the public conversation on a larger scale than ever before, some actors are still substantially more important than others in setting the agenda for that conversation.

Durkheim, a contemporary of Tarde, held a similar view on the effects of interdependence, despite their intellectual rivalry. He famously insisted that society is more than the sum of its parts, using the term "collective effervescence" to refer to the cumulative effects of individual interactions, often triggered by emotions. In his study of religious life he claimed:

> Within a crowd moved by a common passion, we become susceptible to feelings and actions of which we are incapable on our own. And when the crowd is dissolved, when we find ourselves alone again and fall back to our usual level, we can then measure how far we were raised above ourselves (…) Under the influence of some great collective upheaval, social interactions become more frequent and more active. Individuals seek each other out and assemble more often. The result is a general effervescence characteristic of revolutionary or creative epochs. (Durkheim [1912] 2008, pp. 157–58)

Of course, it is one thing to think that Durkheim is onto something, pointing at a significant dimension of social life (who has not been part of a crowd that made us feel greater than ourselves?); it is another to measure and model those dynamics. Again, this is something that Big Data enables researchers to do. The analysis of online interactions shows that human communication is characterized by long periods of tranquility and sudden spikes, or bursts, of activity (Barabási, 2005). Applied to social media, the analysis of temporal attention has allowed researchers to identify how popularity grows and fades overtime, which can help predict the overall dynamics of content popularity (Yang & Leskovec, 2011), and analyze how spikes of collective attention are related to information spreading (Lehmann, Goncalves, Ramasco, & Catuto, 2012). This might not capture the full extent of what Durkheim meant by "effervescence" but it is closer to an understanding of the empirics of collective upheavals that he had in mind. Research on political protests mediated by digital media (González-Bailón, Borge-Holthoefer, Rivero, & Moreno, 2011) shows that similar bursts or spikes of activity can be seen in protest-related communication, contributing to the global diffusion of information and the explosion of protest.

This research on spikes of attention and information diffusion is closely related to the study of social influence and to the reconstruction of chain reactions. The theory of interpersonal influence sketched out by Tarde was further developed in the 1950s in the context of electoral behavior and political communication (Katz & Lazarsfeld, 1955; Lazarsfeld, Berelson, & Gaudet, 1948) and recovered in more recent years as part of the research agenda linking personal networks with political behavior (Zuckerman, 2005). Online networks have allowed the testing of these ideas on a large scale, providing unparalleled evidence that social influence lies behind politically relevant behavior, like voting (Bond et al., 2012). Although field experiments had already suggested the effects of interpersonal influence on voting (Nickerson, 2008), what changes in the digital era is again the size of the samples: now it is possible to conduct experiments with millions of people and thus identify effects that would have been impossible to grasp with smaller data sets.

So the ideas may be relatively old, but digital data—and the methods and models such data allow us to develop—are helping to push those intuitions much further than would otherwise have been possible. For instance, the level of detail allowed by online data is helping us to shift the focus of attention from opinion leaders (or influential users) to the more unassuming category of "susceptible" people. Without them, no influence chain would reach far, and finding who they are and how they are connected to each other is as important as tracing the chains back to their origins (Aral & Walker, 2012). Digital data are also helping us to understand how the "communicative web" that Milgram aimed to map shapes the dynamics of influence and diffusion at the same time that it evolves itself; networks, after all, are nonstationary objects (Holme & Saramäki, 2012). In brief, social scientists have never been in a better position to think about the dynamics of interpersonal communication and the impact these networks have on society— not only because we have better data, but also because there are more people thinking about these issues, which have been at the core of social thinking for decades, if not centuries.

Interpretation and Context

This article has so far focused on the advantages of scaling up the size of data, and gaining resolution on the temporal and spatial dimensions, which helps to generalize findings and analyze the dynamics of social systems. This section will argue that what makes Big Data so interesting is often not their size but rather the way in which we can reduce them, either by applying filters that allow us to identify the relevant streams of information; or by aggregating them in a way that helps identify the right temporal scale or spatial resolution. Only when the data are assembled in the right way it is possible to build a story that makes substantive sense. Social theory can help discriminate noise from signal, and provide the right context for that interpretation. This is particularly important when so many of the models that are being applied to social systems were developed by mathematicians, physicists, or computer scientists to understand the behavior of other systems that have no agency—that is, no actors capable of processing their own information about the world they inhabit and that are free to react in accordance. True, some aspects might be generalizable, but these tend not to be the interesting ones, at least for a social science audience. The models we build about social systems with the help of Big Data should be consistent with what we know about human actors and their behavior. And filtering and aggregating the data in the right way is a necessary first step for the delivery of that goal.

The first immediate way in which online data can be filtered is sampling. When doing research with digital media, there are two main types: the first is the sampling the researcher chooses, for example, based on keywords or hashtags that identify the relevant streams of information, or a set of seed users from whom to snowball in reconstructing networks of communication. The second type is the sampling that the researcher does not choose, usually imposed by the application programming interfaces (APIs) that offer the main access channel to online databases. Usually, APIs do not give access to the full stream of information, and what users get is not necessarily a random sample of all activity. This bias has to be taken into account when results are interpreted—to the extent that the nature of the bias can be identified—because it could lead to incorrect conclusions (González-Bailón, Wang, Rivero, Borge-Holthoefer, & Moreno, 2012). For instance, the centrality of certain users in the flow of communication might

be overestimated simply because they are more active and have a greater chance of being picked up by the sample; but peripheral users (those who, in Tarde's terms, comprise the "sounding board"), and their role in the network of communication, might still be crucial to understanding the dynamics under consideration. Research has suggested that social media sampling strategy impacts the discovery of dynamic processes like diffusion (Choudhury et al., 2010). The computational challenges created by Big Data means that sampling is often a necessity, and systematic attention needs to be paid to its impact on analyses and findings.

The sampling frame chosen by researchers is also important because it is the main reason why the data can barely ever "speak by themselves." For instance, the answer as to who opinion leaders are—in line with the theories discussed in the previous section—very much depends on the information domain being analyzed (which is the researcher's choice), and on how "influence" is operationalized (again, the researcher's choice). If political protests are the focus of analysis, identifying the influential users requires analyzing the stream of information or communication that emerges around these protests. "Global" influential users, that is, those who are in a better position to spread their messages regardless of content, are not necessarily the most important users when it comes to spreading domain-specific information, as in a protest (Cha, Haddadi, Benevenuto, & Gummadi, 2010; González-Bailón, Borge-Holthoefer, & Moreno, forthcoming). Focusing on the relevant stream requires identification of the right set of keywords that can pull out the relevant messages from the full communication stream; and it requires putting those keywords in a context that gives them substantive sense.

Once the data are collected, the way in which they are assembled is also important. When the relevant messages have been sampled, for instance, the unique number identifying the authors can be used to reconstruct networks of communication, using the relationships that those messages create. In Twitter there are two conventions that allow reconstruction of these interactions: retweets (RTs), used to broadcast messages previously sent by other users; and @mentions, which are used to engage in direct communication with others. One article aiming to assess ideological polarization on the twitter sphere has reported opposite findings depending on whether the network was reconstructed using RTs or @mentions: the findings reveal clear polarization in the first case, but no evidence of it in the second (Conover et al., 2011). What this means is that users employ these two conventions to signal different motivations, although what these are or the reasons behind them cannot be clearly identified on the basis of large, aggregated data. It also means that we can reach very different conclusions depending on how the data are filtered, in this case to retain one channel of communication (RTs) or another (@mentions). Once again, the data cannot speak by themselves, because a lot of choices are made along the way to determine how best to analyze them—their interpretation very much depends on those choices, which are not data-driven but human.

The importance of taking into account the potential biases of the data, via sampling or filtering, is not specific to Big Data research; but the relevance of it is somehow overshadowed by the sheer volume of information that we now have at our disposal. And yet the fact is that the volume of information does not reduce the role of human interpretation, or the biases that we introduce in choosing the level, or resolution, for the analyses. One can conceive of scientific research as the institutionalization of the controls that put a bridle on those biases, or at least help us identify them. But for this, we still need theories and models, and cumulative research that works on the basis of improvements and readjustments. In other words, Big Data will not bring

about the end of theory; quite the contrary. And social science has a crucial role to play in the discovery of the biases that are intrinsic to digital data, as well as in the construction of convincing stories about what those data reveal.

Policy Implications

The already prominent, but continually increasing, research that is exploiting Big Data is the most powerful proof of their transformational impact. But evidence-based policymaking offers another area where the analysis of large data sets can have substantive effects. The way in which digital data can improve our understanding of urban dynamics, and help plan interventions to improve city life, has already been discussed. But there are many other avenues by which topdown approaches to city governance can be improved with the bottom-up information that Big Data often contain. Online platforms like FixMyStreet and SeeClickFix, for instance, aim to improve local governance by encouraging users to report street problems to the authorities. The data thus generated help raise attention when complaints cluster in a particular neighborhood or area. Residents offer more accurate and responsive information because they generate it locally,with the knowledge that only living in that area can provide.

Of course, digital data are also flawed by socioeconomic divides: not everybody is online and represented, and there are groups that are systematically excluded from these channels of communication (Graham, Hale, & Stephens, 2012). Their streets cannot be fixed if they do not have the means to communicate faults—even less so if they are not on the map. But these divides can also be better identified with Big Data research, which allows us to spot the holes in "Internet geographies" and thus offer evidence to help support and devise better inclusion plans. Privacy is another serious issue surrounding the use of Big Data; that many of the data generated online are publicly available does not mean that their analysis is legitimate (Boyd & Crawford, 2012). While these privacy concerns are important and should always be taken into account, the benefits of sharing data often offset the dangers. The tradeoffs that people are willing to make when disclosing personal information if in return they get something of value are not that well understood; And in any case, the benefits of analyzing Big Data depend, as with everything, on how they are used. Since large data sets that track our behavior are here to stay, it is probably best to start demanding responsible use of that information than to prevent its use.

The kind of self-government to which Jacobs (1961) referred in her study of urban life can be greatly expanded in the era of Big Data—if the data are used to illuminate how to make best use of decentralized networks for decision-making. Online initiatives like peer-to-patent, designed to crowd-source the patenting process in the United States (Noveck, 2009), already suggest that online technologies can help unblock the bottlenecks of public administration. But the design of those technologies would benefit greatly from a better understanding of how decentralized networks of information can feed into policymaking. For instance, research on the editing dynamics of collaborative platforms like Wikipedia help shed light on the underpinnings of conflict, and provide a rationale for policies that can prevent it (Török et al., 2013). Large volumes of search data have also been used to detect epidemics, which allows planning of more timely intervention policies (Ginsberg et al., 2009). We tend to think of public administration as orchestrated from the top-down by bureaucrats and councils; but most of the activity that is relevant for its functioning emerges from the daily interactions of residents and citizens, already

being transformed by online technologies (Margetts & Dunleavy, 2013). Analyzing the communication and mobility patterns of citizens as they make use of public spaces can make the regulation of these spaces more efficient and responsive, and policymakers more accountable for the interventions they design.

The use of online communication to extract public opinion indicators offers once such channel for increased accountability. The public conversation of which Tarde was already speaking at the end of the nineteenth century is now being recorded in the multiple online venues where it takes place today. This can be used to grant a louder voice to the people, and to amplify their reactions to political events and policy discussions. Researchers are working on methods to automate content analysis to best extract the meaning and affect of that communication (Liu, 2012). The same limitations of bias and representativeness discussed in the previous section affect the validity of these measurements; but methodological improvements and a more systematic assessment of the validity of the approach, compared to more standard ways of measuring public opinion (Converse, 1987), can increase the significance of online public discourse as a regulatory mechanism for policymaking—or at least as a barometer to measure the public mood, and the topics that are most salient in their minds (González-Bailón, Banchs, & Kaltenbrunner, 2012; Young & Soroka, 2012).

Big Data can never be a substitute for the main channels of political expression (such as voting, demonstrations, and strikes) but they offer an alternative source to assess what matters most to the public, and how they react to the decisions and actions of their representatives. Only by analyzing the opinions that people voluntarily express online can we assess the validity of this method to gauge what the public thinks; but this is an option that did not exist prior to the massive data sets that are constantly being fed with the contents of our online interactions. These data sets offer unprecedented richness, both in scale and breadth, to start examining policymaking through a different lens—that of the public—and hopefully to visualize patterns that were hitherto hidden.

Conclusions

Returning to Milgram's (1977) quote, it is fair to say that we are now, more than ever, "in possession of the information needed to construct our discipline" (p. 317). We have more information about communication patterns than ever before, and we have the tools and computational power to make sense of what those patterns reveal. But we need to confront challenges that Milgram probably never anticipated. The first one is that social scientists can no longer do research on their own: the scale of the data that we can now analyze, and the methods required to analyze them, can only be developed by pooling expertise with colleagues from other disciplines. Social science has the theoretical tradition to build a context for these data, point to the right mechanisms of the dynamics analyzed, and build credible interpretations—which is as important as having access to vast amounts of information and cutting-edge methods. This collaboration requires working on the grounds of a common language, which in turn demands making the analytical toolkit of social scientists compatible with that of other disciplines. The mathematical language of networks opens one such point of contact, as does coding and programming. Although all collaborations are based on a division of labor, a mutual understanding of the divided tasks is crucial for these synergies to work. Social scientists need to embrace Big Data and gain literacy in the research that it makes possible, even if it goes beyond traditional

disciplinary boundaries. The implications of this emerging research for how we understand the social world are huge—and it is part of our remit to help shape that understanding.

Sandra González-Bailón, Oxford Internet Institute, University of Oxford
[sandra.gonzalezbailon@oii.ox.ac.uk]

References

Anderson, C. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." Wired Magazine (June 23), http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.

Aral, S., and D. Walker. 2012. "Identifying Influential and Susceptible Members of Social Networks." Science 337: 337–41.

Barabási, A.L. 2005. "The Origin of Bursts and Heavy Tails in Human Dynamics." Nature 435: 207–11.

Bond, R.M., C.J. Fariss, J.J. Jones, A.D.I. Kramer, C.A. Marlow, J.E. Settle, and J.H. Fowler 2012. "A 61-Million-Person Experiment in Social Influence and Political Mobilization."Nature 489: 295–98.

Boyd, D., and K. Crawford. 2012. "Critical Questions for Big Data. Provocations for a Cultural,Technological, and Scholarly Phenomenon." Information, Communication & Society 15 (5): 662–79.

Cha, M., H. Haddadi, F. Benevenuto, and K.P. Gummadi. 2010. "Measuring User Influence in Twitter:The Million Follower Fallacy." Paper presented at the International AAAI Conference on Weblogs and Social Media (ICSWM), May 23_26, Washington, DC.

Choudhury, M.D., Y.-R. Lin, H. Sundaram, K.S. Candan, L. Xie, and A. Kelliher. 2010. "How Does the Data Sampling Strategy Impact the Discovery of Information Diffusion in Social Media?" Paper presented at the International AAAI Conference on Weblogs and Social Media (ICSWM), May 23_26, Washington, DC.

Clark, T.N. 1969. Gabriel Tarde. On Communication and Social Influence. Chicago, IL University of Chicago Press.

Coleman, J.S. 1990. Foundations of Social Theory. Cambridge, MA: Belknap Press of Harvard University Press.

Conover, M.D., J. Ratkiewicz, M. Francisco, B. Goncalves, A. Flammini, and F. Menczer. 2011. "Political Polarization on Twitter." Paper presented at the AAAI International Conference on Weblogs and Social Media (ICWSM), July 17_21, Barcelona, Spain.

Converse, P.E. 1987. "Changing Conceptions of Public Opinion in the Political Process." Public Opinion Quarterly 51: S12–24.

de Sola Pool, I., and M. Kochen. 1978. "Contacts and Influence." Social Networks 1 (1): 5–51.

Dodds, P.S., R. Muhamad, and D.J. Watts. 2003. "An Experimental Study of Search in Global Social Networks." Science 301 (5634): 827–29.

Durkheim, É. [1912] 2008. The Elementary Forms of Religious Life. Oxford: Oxford University Press.

Ginsberg, J., M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant. 2009. "Detecting Influenza Epidemics Using Search Engine Query Data." Nature 457: 1012–14.

González, M.C., C.A. Hidalgo, and A.L. Baraba´ si. 2008. "Understanding Individual Human Mobility Patterns." Nature 453: 779–82.

González-Bailón, S., E.R. Banchs, and A. Kaltenbrunner. 2012. “Emotions, Public Opinion and U.S. Presidential Approval Rates: A 5-Year Analysis of Online Political Discussions.” Human Communication Research 38: 121–43.
González-Bailón, S., J. Borge-Holthoefer, and Y. Moreno. Forthcoming. “Broadcasters and Hidden Influentials in Online Protest Diffusion.” American Behavioral Scientist DOI: 10.1177/0002764213479371
González-Bailón, S., J. Borge-Holthoefer, A. Rivero, and Y. Moreno. 2011. “The Dynamics of Protest Recruitment Through an Online Network.” Scientific Reports 1 (197). DOI: [10.1038/srep00197].http://www.nature.com/srep/2011/111215/srep00197/abs/srep00197.html#supplementaryinformation.
González-Bailón, S., N. Wang, A. Rivero, J. Borge-Holthoefer, and Y. Moreno. 2012. “Assessing the Bias in Communication Networks Sampled from Twitter.” Working Paper. http://papers.ssrn.com/sol3/papers.cfm?abstract_id¼2185134.
Graham, M., S.A. Hale, and M. Stephens. 2012. “Digital Divide: The Geography of Internet Access.” Environment and Planning A 44 (5): 1009–10.
Holme, P., and J. Sarama¨ki. 2012. “Temporal Networks.” Physics Reports 519 (3): 97–125.
Jacobs, J. 1961. The Death and Life of Great American Cities. London: Pimlico.
Katz, E. 1957. “The Two-Step Flow of Communication: An Up-to-Date Report on an Hypothesis.” Public Opinion Quarterly 21 (1): 61–78.
Katz, E., and P. Lazarsfeld. 1955. Personal Influence. The Part Played by People in the Flow of Mass Communications. New York, NY: Free Press.
Kleinberg, J.M. 2000. “Navigation in a Small World.” Nature 406: 845.
Lazarsfeld, P., B. Berelson, and H. Gaudet. 1948. The People’s Choice: How the Voter Makes Up His Mind in a Presidential Campaign. New York, NY: Columbia University Press.
Lazer, D., A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N.A. Christakis et al. 2009. “Computational Social Science.” Science 323: 721–23.
Lehmann, J., B. Goncalves, J.J. Ramasco, and C. Catuto. 2012. “Dynamical Classes of Collective Attention in Twitter.” Paper presented at the World Wide Web Conference, April 16_20, Lyon, France.
Leskovec, J., L. Adamic, and B.A. Huberman. 2007. “The Dynamics of Viral Marketing.” ACM Transactions on the Web 1 (5). DOI: 10.1145/1232722.1232727
Liu, B. 2012. Sentiment Analysis and Opinion Mining. Chicago, IL: Morgan & Claypool.
Margetts, H., and P. Dunleavy. 2013. “The Second Wave of Digital-Era Governance: A Quasi-Paradigm for Government on the Web.” Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371 (1987). DOI: 10.1098/rsta.2012.0382
Mayer-Schoenberger, V., and K. Cukier. 2013. Big Data. A Revolution That Will Transform How We Live,Work and Think. London: John Murray.
Milgram, S. 1967. “The Small World Problem.” Psychology Today 2: 60–7.
Milgram, S. 1977. The Individual in a Social World: Essays and Experiments. London: Pinter & Martin.
Morozov, E. 2013. To Save Everything, Click Here: The Folly of Technological Solutionism. New York, NY: Public Affairs.
Nickerson, D.W. 2008. “Is Voting Contagious? Evidence from Two Field Experiments.” American Political Science Review 102 (1): 49–57.
Noulas, A., S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. 2012. “A Tale of Many Cities: Universal Patterns in Human Urban Mobility.” PloS ONE 7 (5): e37027.

Noveck, B.S. 2009. Wiki-Government. How Technology Can Make Government Better, Democarcy Stronger, and Citizens More Powerful. Washington, DC: Brookings Institution.
Quercia, D., J.P. Pesce, V. Almeida, and J. Crowcroft. 2013. “Psychological Maps 2.0: A Web Engagement Enterprise Starting in London.” Paper presented at the World Wide Web Conference, May 13_17, Rio de Janeiro, Brazil.
Schelling, T.C. 1978. Micromotives and Macrobehavior. London: Norton.
Silver, N. 2012. The Signal and the Noise: Why So Many Predictions Fail—But Some Don’t. New York, NY: Penguin Press.
Török, J., G. Iñiguez, T. Yasseri, M.S. Miguel, K. Kaski, and J. Kertész. 2013. “Opinions, Conflicts, and Consensus: Modeling Social Dynamics in a Collaborative Environment.” Physical Review Letters 110 (8): 088701.
Travers, J., and S. Milgram. 1969. “An Experimental Study of the Small World Problem.” Sociometry 32 (4): 425–43.
Watts, D.J. 2003. Six Degrees. The Science of a Connected Age. London: William Heinemann.
Watts, D.J. 2007. “A Twenty-First Century Science.” Nature 445: 489.
Watts, D.J. 2011. Everything is Obvious.Once You Know the Answer. New York, NY: Crown Business.
Watts, D.J., P.S. Dodds, and M.E.J. Newman. 2002. “Identity and Search in Social Networks.” Science 296: 1302–5.
Wu, S., J.M. Hofman, W.A. Mason, and D.J. Watts. 2011. “Who Says What to Whom on Twitter.” Paper presented at the World Wide Web Conference, March 28_April 1, Hyderabad, India.
Yang, J., and J. Leskovec. 2011. “Patterns of Temporal Variation in Online Media.” In Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. Hong Kong, China: ACM, 177–86.
Young, L., and S. Soroka. 2012. Affective News: The Automated Coding of Sentiment in Political Texts. Political Communication 29: 205–31.
Zuckerman, A.S., ed. 2005. The Social Logic of Politics: Personal Networks as Contexts for Political Behavior. Philadelphia, PA: Temple University Press.