Computational Social Science: Exciting Progress and Grand Challenges

Duncan J. Watts
Microsoft Research

Introduction

The past 15 years have witnessed an incredible increase in both the scale and also scope of social and behavioral data available to researchers. Over the same period, and driven by the same explosion in data, the study of social phenomena has increasingly become the province of computer scientists, physicists and other "hard" scientists. Papers on social networks and related topics appear routinely in top science journals and computer science conferences; network science research centers and institutes are sprouting up in top universities; and funding agencies from DARPA to NSF have moved quickly to embrace what is being called "computational social science."

Against these exciting developments stands a stubborn fact—that in spite of many thousands of published papers, surprisingly little progress has been made on the "big" questions that motivated the field—from systemic risk in financial systems, to problem solving in complex organizations, to the dynamics of epidemics or social movements.

There are many reasons for this state of affairs, but in this talk I will concentrate on three. First, social scientific problems are almost always more difficult than they seem. Second, the data required to address many problems of interest to social scientists remain hard to assemble. And third, thorough exploration of complex social problems often requires the complementary application of multiple research traditions—including statistical modeling and simulation, but also social and economic theory, lab experiments, surveys, ethnographic fieldwork, historical or archival research, and practical experience—many of which will be unfamiliar to any one researcher. Meeting these challenges, I claim, will require both new platforms for collecting the appropriate data, and also new institutions for conducting social science research.

Why is social science hard?

Almost by definition, "social" phenomena are less about the behaviour of individuals, than of collections of individuals like groups, crowds, organizations, markets, classes, and even entire societies, all of which interact with each other via networks of information and influence, which in turn change over time. As a result, social systems—like other complex systems in physics and biology—exhibit "emergent" behaviour, meaning that the behaviour of entities at one "scale" of reality is not easily traced to the properties of the entities at the scale below (Anderson 1972).

Firms, for example, can exhibit highly stable identities and cultures even as the particular employees who work in them change completely over time, just as you remain "you" even as the cells in your body turn over during the course of your lifetime. Conversely, the stock market, the economy, or a political regime can collapse suddenly and unexpectedly even as the various players and background conditions remain the same.

Complicating matters further, emergent properties can be both the cause and the effects of social change. Sometimes, that is, the decisions of corporations or even governments may depend critically on the personal interests of a handful of executives, whereas at other times the behaviour of those same individuals may be powerfully constrained by the corporate or political culture to which they belong. Nor is emergence as simple as one "scale" of reality aggregating to another. Rather, in many problems of interest to social scientists, the actions of individuals, firms, regulatory and government agencies, markets, and political institutions may all play important roles. Moreover, because these different types of actors not only exist at different scales—firms comprise individuals, markets comprise firms and individuals, etc.—but may also interact with each other in important ways, problems of this type require one to consider events, actors, and forces across multiple scales simultaneously.

Given the unavoidably multi-scale, complex, and emergent nature of social phenomena, it is not surprising that theories of social behaviour and change have been difficult to work out in any realistic detail. Compounding this theoretical difficulty is an empirical one, or rather two separate but related empirical difficulties. First, observational data on the scale of hundreds of millions, or even tens of thousands, of individuals has been impossible collect historically. Second, because cause and effect can be difficult to infer from observational data alone, experimental studies are also necessary. Yet experiments involving, say the performance of an organization with a particular structure, or the popularity of songs in a single instance of a cultural market—represents the collective behavior of hundreds or even thousands of individuals, designs that are clearly impossible to implement in a physical lab (Zelditch 1969).

The emergence of computational social science

In light of these three interrelated difficulties, (a) the complexity of the theoretical issues confronting social science, (b) the difficulty of obtaining the relevant observational data, and (c) the difficulty of manipulating large-scale social organizations experimentally, it is hardly surprising that progress in social science has been slow relative to that in the physical, engineering, and biological sciences, in particular over the past century. Correspondingly, in lifting some of the constraints the computing revolution of the past two decades—a revolution that has not only dramatically increased the speed and memory of computers themselves, but also the scale and scope of social data that can now be analyzed—has the potential to

revolutionize traditional social science, leading arguably to a new paradigm of "computational social science"[1] (Lazer et al. 2009).

The most prominent strand of research in computational social science leverages the communication technologies—including email, social networking and microblogging services, cell-phones, as well as online games, e-commerce sites, and other internet enabled services—all generate signals, often referred to as "digital exhaust" or "digital breadcrumbs", from which inferences can be made about individual and/or collective behavior. In this way, it is increasingly possible to observe the actions and interactions of hundreds of millions of individuals in real time, and also over time.

Data derived from instant messaging services and social networking sites, for example, have been used to construct networks of hundreds of millions of nodes have been analyzed (Leskovec and Horvitz 2008; Ugander et al. 2011), confirming earlier conjectures about the topology of large social networks (Newman 2003; Watts and Strogatz 1998). Other studies have mined email data to estimate the micro-level rules describing new tie formation (Kossinets and Watts 2006), or used blog networks to measure the propensity to join new groups (Backstrom et al. 2006). Others still have mapped the diffusion of online content (Bakshy, Karrer and Adamic 2009; Dow, Adamic and Friggeri 2013; Goel, Watts and Goldstein 2012; Leskovec, Adamic and Huberman 2007; Sun et al. 2009), or conducted massive randomized field experiments to estimate the causal effects of social influence on adoption (Aral and Walker 2011), voting turnout (Bond et al. 2012), or likelihood to share content (Bakshy et al. 2012).

A less well explored but also important strand of work comprises research that uses the web to create "virtual labs": controlled environments within which lab-style macro-sociological experiments can be conducted (Hedstrom 2006). Although early efforts relied on volunteers (Dodds, Muhamad and Watts 2003; Salganik, Dodds and Watts 2006), an important recent development in this field has been the use of crowdsourcing sites such as Amazon's Mechanical Turk to recruit and pay subjects analogous to the longstanding tradition in behavioral science of recruiting from college student populations (Mason and Watts 2009). One important advance due to crowdsourced virtual labs has been to resolve the "synchronicity problem"—namely assuring that N subjects will arrive contemporaneously and remain engaged in the experiment for its duration (Abell 2001; Abraham 2008), thereby allowing for networked experimental designs. Another has been that experiments can be

[1] Computational social science is a contested label, referring in some quarters to simulation of agent-based models (see, for example, http://computationalsocialscience.org/), and in others strictly to the analysis of computationally challenging datasets (http://research.microsoft.com/en-us/groups/cssnyc/). Here I follow Lazer et al (2009), who invoke the term somewhat liberally to refer to the emerging intersection of the social and computational sciences an intersection that included analysis of web-scale observational data, virtual-lab style experiments, and computational modeling.

designed, launched and executed on a much shorter timescale than has been historically feasible, and on a lower cost basis (Wang, Suri and Watts 2012). Finally, by shrinking the "hypothesis-testing cycle"—the lag between analyzing one set of experimental results and running the next set of experiments—from on the order of months or years to days or even hours, crowdsourced virtual lab experiments can dramatically expand the range of conditions that can be studied experimentally.

Challenges for computational social science

As impressive as its recent accomplishments have been, computational social science faces a number of pressing challenges if it is to address the important questions of social science—organizational and inter-organizational problem solving; collective action problems; influence, adoption and information diffusion; collective decision making; deliberation, segregation, and polarization; technology, governance, and democracy; predicting sudden cultural change, outbreaks of political conflict, or the emergence of disruptive technologies—in a meaningful way.

A Social Supercollider. First among these challenges, the dominant "digital exhaust" model of data collection imposes important limitations on the type of research questions that can be answered. To illustrate, consider the problem of measuring how friends influence each other's purchase behavior, a question that is of great interest both to social scientists and also to marketers, policy makers and other change agents. Although simple to ask, answering such a question requires being able to observe both the complete friendship network — already a difficult task — and also everyone's shopping behavior. Using existing systems, one might obtain an approximation of the friendship network by using Facebook data, or mining email logs, while e-commerce sites or retailer databases may show how much individuals are spending on particular products. Currently, however, it is extremely difficult to combine even two such sources of data, and of course there are many different modes of communication, and many different places to make purchases.

Many questions of interest to social scientists encounter a similar problem, in that they require studying the relation between different modes of social action and interaction—for example, search data to infer intent, network data to infer relationships, e-commerce data to infer choices, and social media data to infer opinions—yet these "modes" are all recorded and stored separately, often by different companies. A major breakthrough for computational social science, therefore, would be a proverbial "social supercollider:" a facility to combine multiple streams of data, creating richer and more realistic portraits of individual behavior and identity, while retaining the benefits of massive scale.

Against this considerable promise stands the equally pressing concern of protecting individual privacy. Privacy is already an important issue for all industries that collect digital information about their consumers; however, for exactly the same reason that the social supercollider would be so powerful a scientific tool — namely

that it would put all the pieces together — it raises far more serious questions about individual privacy even than are posed by existing commercial platforms. Precisely these questions, in fact, have already been raised by recent revelations of the NSAs Prism project, which also appears to be an attempt to combine data from multiple sources. Construction and management of anything like a social supercollider would therefore have to proceed under the strictest scrutiny, both with respect to the governance of the data itself, and also its end uses.

Expanding Virtual Labs. A second challenge for computational social science concerns the continued development of experimental macrosociology. Perhaps surprisingly the major limitation to existing experimental designs has been the difficulty of recruiting large numbers of subjects in a reliable and cost-effective manner. For example, the largest synchronous virtual lab experiments to-date have not exceeded N = 36, largely because of the practical difficulty of recruiting more than that number at any single time.

One potential solution to this problem would be to construct a large, persistent, and well-documented panel of subjects—potentially hundreds of thousands of individuals, who may participate in many experiments over months or years—it would be possible to increase the scale of synchronous experiments to involve hundreds, or even thousands of contemporaneous subjects. By allowing researchers to specify their sampling frame in advance, moreover, a large persistent panel of this type would also facilitate investigations of how behavior varies by demographic, national, or racial group. Such a panel would also allow for entirely novel questions about the connections between individual attributes and behavior, as well as between different elements of behavior itself. Do people who contribute generously to public good games behave in any characteristic way when participating in a collaborative problem solving exercise, or in an exchange network? Finally, beyond virtual lab experiments, a panel of this scale and duration could also be of great value for survey research and also for randomized field experiments.

*Putting the "social" in computational social science***.** A final challenge for computational social science is highlighted by the observation that in spite of the many thousands of papers that have been published on topics related to social networks, financial crises, crowdsourcing, influence and adoption, group formation and so on, relatively few of these papers are published in traditional social science journals, or even attempt to engage seriously with the existing social scientific literature. The result is that much of computational social science to date has effectively evolved in isolation of the rest of social science, largely ignoring much of what social scientists have to say about the same topics, and largely being ignored by them in return.

One can argue about who to blame for this state of affairs—computer scientists for being presumptuous or social scientists for being defensive—and also whether or not it is even a bad thing. Perhaps all interdisciplinary fields start out as ugly

ducklings and have to become swans on their own, not by making friends with existing fields but by outcompeting them. My view, however, is that meaningful progress on important problems will require serious engagement between the communities, each of which has much to offer the other. On the one hand, that is, computer scientists and physicists have technical capabilities that are of great potential benefit to social scientists, while on the other hand deep subject matter knowledge is essential in order to ask the right questions, and to formulate even simple models in ways that address these questions.

Harnessing the complementary strengths of both communities, however, is easier said than done. Consider, for example, the problem of modeling systemic risk in financial systems. On the one hand, a recent networks literature on financial crises has sprung up around the analogy of crises as cascades of failures in interbank networks (Delli Gatti et al. 2012; Gai and Kapadia 2010; May and Arinaminpathy 2010; Nier et al. 2007). Unfortunately, while the analogy of contagion in networks is appealing, these models turn out to omit certain features of real banking systems—for example, that banks "create" money by expanding their balance sheets, that shocks can propagate non-locally via asset prices, or that prices must adjust in order that markets will clear—that are critical to understanding recent crises. On the other hand, however, descriptively accurate accounts of real financial crises also tend to be so complex and multifaceted (Brunnermeier 2009; Commission 2011; Gorton 2012; Hellwig 2009) that it is difficult even for experts to agree on which mechanisms are the most important, and therefore what features are critical to include in even a simple model.

Prima facie, it is not obvious how, or even if, these approaches can be reconciled, but any attempt to do so would likely take months, or even years—if only to become familiar with the multiple, often incommensurate literatures on the topic—and is likely beyond the ability of any one individual to undertake alone. Meanwhile, encouraging interdisciplinary teams of researchers to engage in long-term, intensive collaborations involves not only creating actual places where the right mix of people are physically proximate, but also ensuring that they have the time and the incentives to invest in high-risk projects. Deep and significant progress in computational social science, in other words, will require not only novel and creative approaches to data collection and creation—the overwhelming focus of enthusiasm so far—but also novel and creative approaches to designing research institutions.

Abell, Peter. 2001. "Causality and Low-Frequency Complex Events: The Role of Comparative Narratives." *Sociological Methods and Research* 30(1):57--80.

Abraham, Magid. 2008. "The Off-line Impact of Online Ads." *Harvard Business Review* April:28.

Anderson, P. W. 1972. "More is different." *Science* 177(4047):393-96.

Aral, Sinan, and Dylan Walker. 2011. "Creating Social Contagion Through Viral Product Design: A Randomized Trial of Peer Influence in Networks." *Management Science* Forthcoming.

Backstrom, L., D. Huttenlocher, J. Kleinberg, and X. Lan. 2006. "Group formation in large social networks: membership, growth, and evolution." Pp. 44-54: ACM.

Bakshy, Eytan, Brian Karrer, and Lada Adamic, A. 2009. "Social Influence and the Diffusion of User-Created Content." in *10th ACM Conference on Electronic Commerce*. Stanford, California: Association of Computing Machinery.

Bakshy, Eytan, Itamar Rosenn, Cameron Marlow, and Lada Adamic. 2012. "The role of social networks in information diffusion." Pp. 519-28 in *Proceedings of the 21st international conference on World Wide Web*: ACM.

Bond, Robert M, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. 2012. "A 61-million-person experiment in social influence and political mobilization." *Nature* 489(7415):295-98.

Brunnermeier, Markus K. 2009. "Deciphering the Liquidity and Credit Crunch 2007-2008." *The Journal of Economic Perspectives* 23(1):77-100.

Commission, United States. Financial Crisis Inquiry. 2011. *Financial crisis inquiry report: final report of the national commission on the causes of the financial and economic crisis in the United States*: Government Printing Office.

Delli Gatti, Domenico, Mauro Gallegati, Bruce Greenwald, Joseph Stiglitz, and Stefano Battiston. 2012. "Liaisons Dangereuses: Increasing Connectivity, Risk Sharing and Systemic Risk." *Journal of Economic Dynamics and Control* 36(8):1121-41.

Dodds, P. S., R. Muhamad, and D. J. Watts. 2003. "An experimental study of search in global social networks." *Science* 301(5634):827-29.

Dow, P Alex, Lada A Adamic, and Adrien Friggeri. 2013. "The Anatomy of Large Facebook Cascades."

Gai, Prasanna, and Sujit Kapadia. 2010. "Contagion in financial networks." *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science* 466(2120):2401-23.

Goel, Sharad, Duncan J Watts, and Daniel G Goldstein. 2012. "The structure of online diffusion networks." Pp. 623-38 in *Proceedings of the 13th ACM Conference on Electronic Commerce*: ACM.

Gorton, Gary B. 2012. *Misunderstanding Financial Crises: Why We Don't See Them Coming*: Oxford University Press.

Hedstrom, Peter. 2006. "SOCIOLOGY: Experimental Macro Sociology: Predicting the Next Best Seller." *Science %R 10.1126/science.1124707* 311(5762):786-87.

Hellwig, Martin F. 2009. "Systemic risk in the financial sector: an analysis of the subprime-mortgage financial crisis." *De Economist* 157(2):129-207.

Kossinets, Gueorgi, and Duncan J. Watts. 2006. "Empirical Analysis of an Evolving Social Network." *Science* 311(5757):88-90.

Lazer, D., A. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, and M. Gutmann. 2009. "Social Science: Computational Social Science." *Science* 323(5915):721.

Leskovec, Jure, Lada Adamic, A., and Bernardo Huberman, A. . 2007. "The dynamics of viral marketing." *ACM Trans. Web* 1(1):5.

Leskovec, Jure, and Eric Horvitz. 2008. "Planetary-Scale Views on a Large Instant-Messaging Network." in *17th International World Wide Web Conference*. Beijing, China.

Mason, W., and D. J. Watts. 2009. "Financial incentives and the performance of crowds." *Proceedings of the ACM SIGKDD Workshop on Human Computation*:77-85.
May, Robert M, and Nimalan Arinaminpathy. 2010. "Systemic risk: the dynamics of model banking systems." *Journal of the Royal Society Interface* 7(46):823-38.
Newman, M. E. J. 2003. "The structure and function of complex networks." *Siam Review* 45(2):167-256.
Nier, Erlend, Jing Yang, Tanju Yorulmazer, and Amadeo Alentorn. 2007. "Network models and financial stability." *Journal of Economic Dynamics and Control* 31(6):2033-60.
Salganik, Matthew J., Peter Sheridan Dodds, and Duncan J. Watts. 2006. "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market." *Science* 311(5762):854-56.
Sun, Eric Sun , Itamar Rosenn, Cameron A. Marlow, and Thomas M. Lento. 2009. "Gesundheit! Modeling Contagion through Facebook News Feed." in *International Conference on Weblogs and Social Media*. San Jose, CA: AAAI.
Ugander, J., B. Karrer, L. Backstrom, and C. Marlow. 2011. "The anatomy of the facebook social graph." *Arxiv preprint arXiv:1111.4503*.
Wang, J., S. Suri, and D.J. Watts. 2012. "Cooperation and assortativity with dynamic partner updating Supporting Information."
Watts, D. J., and S. H. Strogatz. 1998. "Collective dynamics of 'small-world' networks." *Nature* 393(6684):440-42.
Zelditch, Morris. 1969. "Can you really study an army in the laboratory." *A sociological reader on complex organizations*:528-39.