Published in final edited form as:
Science. 2009 February 6; 323(5915): 721–723. doi:10.1126/science.1167742.

Life in the network: the coming age of computational social science

David Lazer,
Harvard University

Alex (Sandy) Pentland,
MIT

Lada Adamic,
University of Michigan

Sinan Aral,
NYU

Albert Laszlo Barabasi,
Northeastern University

Devon Brewer,
Interdisciplinary Scientific Research

Nicholas Christakis,
Harvard University

Noshir Contractor,
Northwestern University

James Fowler,
UCSD

Myron Gutmann,
University of Michigan

Tony Jebara,
Columbia University

Gary King,
Harvard University

Michael Macy,
Cornell University

Deb Roy, and
MIT

Marshall Van Alstyne
Boston University

We live life in the network. When we wake up in the morning, we check our e-mail, make a quick phone call, walk outside (our movements captured by a high definition video camera), get on the bus (swiping our RFID mass transit cards) or drive (using a transponder to zip through the tolls). We arrive at the airport, making sure to purchase a sandwich with a credit card before boarding the plane, and check our BlackBerries shortly before takeoff. Or we visit the doctor or the car mechanic, generating digital records of what our medical or automotive problems are. We post blog entries confiding to the world our thoughts and feelings, or maintain personal

social network profiles revealing our friendships and our tastes. Each of these transactions leaves digital breadcrumbs which, when pulled together, offer increasingly comprehensive pictures of both individuals and groups, with the potential of transforming our understanding of our lives, organizations, and societies in a fashion that was barely conceivable just a few years ago.

The capacity to collect and analyze massive amounts of data has unambiguously transformed such fields as biology and physics. The emergence of such a data-driven "computational social science" has been much slower, largely spearheaded by a few intrepid computer scientists, physicists, and social scientists. If one were to look at the leading disciplinary journals in economics, sociology, and political science, there would be minimal evidence of an emerging computational social science engaged in quantitative modeling of these new kinds of digital traces. However, computational social science is occurring, and on a large scale, in places like Google, Yahoo, and the National Security Agency. Computational social science could easily become the almost exclusive domain of private companies and government agencies. Alternatively, there might emerge a "Dead Sea Scrolls" model, with a privileged set of academic researchers sitting on private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest in the accumulation, verification, and dissemination of knowledge.

What potential value might a computational social science, based in an open academic environment, offer society, through an enhanced understanding of individuals and collectives? What are the obstacles that stand in the way of a computational social science?

From individuals to societies

To date the vast majority of existing research on human interactions has relied on one-shot self-reported data on relationships. New technologies, such as video surveillance, e-mail, and 'smart' name badges offer a remarkable, second-by-second picture of interactions over extended periods of time, providing information about both the structure and content of relationships. Consider examples of data collection in this area and of the questions they might address:

Video recording and analysis of the first two years of a child's life (1)

Precisely what kind of interactions with others underlies the development of language? What might be early indicators of autism?

Examination of group interactions through e-mail data

What are the temporal dynamics of human communications—that is, do work groups reach a stasis with little change, or do they dramatically change over time (2,3)? What interaction patterns predict highly productive groups and individuals? Can the diversity of news and content we receive predict our power or performance (4)?

Examination of face-to-face group interactions over time using sociometers

Small electronics packages ('sociometers') worn like a standard ID badge can capture physical proximity, location, movement, and other facets of individual behavior and collective interactions. What are patterns of proximity and communication within an organization, and what flow patterns are associated with high performance at the individual and group levels (5)?

Macro communication patterns

Phone companies have records of call patterns among their customers extending over multiple years, and e-Commerce portals such as Google and Yahoo collect instant messaging data on global communication. Do these data paint a comprehensive picture of societal-level communication patterns? What does the "macro" social network of society look like (6), and how does it evolve over time? In what ways do these interactions affect economic productivity or public health?

Tracking movement

With GPS and related technologies, it is increasingly easy to track the movements of people (7,8). Mobile phones, in particular, allow the large scale tracing of people's movements and physical proximities over time (9), where it may be possible to infer even cognitive relationships, such as friendship, from observed behavior (10). How might a pathogen, such as influenza, driven by physical proximity, spread through a population (11)?

Internet

The Internet offers an entirely different channel for understanding what people are saying, and how they are connecting (12). Consider, for example, in this political season, tracing the spread of arguments/rumors/positions in the blogosphere (13), as well as the behavior of individuals surfing the Internet (14), where the concerns of an electorate become visible in the searches they conduct. Virtual worlds, by their nature capturing a complete record of individual behavior, offer ample opportunities for research, for example, experimentation that would be impossible or unacceptable (15). Similarly, social network websites offer an unprecedented opportunity to understand the impact of a person's structural position on everything from their tastes to their moods to their health (16), while Natural Language Processing offers increased capacity to organize and analyze the vast amounts of text from the Internet and other sources (17).

In short, a computational social science is emerging that leverages the capacity to collect and analyze data with an unprecedented breadth and depth and scale. Substantial barriers, however, might limit progress. Existing ways of conceiving human behavior were developed without access to terabytes of data describing their minute-by-minute interactions and locations of entire populations of individuals. For example, what does existing sociological network theory, built mostly on a foundation of one-time 'snapshot' data, typically with only dozens of people, tell us about massively longitudinal datasets of millions of people, including location, financial transactions, and communications? The answer is clearly "something," but, as with the blind men feeling parts of the elephant, limited perspectives provide only limited insights. These emerging data sets surely must offer some qualitatively new perspectives on collective human behavior.

There are significant barriers to the advancement of a computational social science both in approach and in infrastructure. In terms of approach, the subjects of inquiry in physics and biology present different challenges to observation and intervention. Quarks and cells neither mind when we discover their secrets nor protest if we alter their environments during the discovery process (although, as discussed below, biological research involving humans offers some similar concerns regarding privacy). In terms of infrastructure, the leap from social science to a computational social science is larger than from, say, biology to a computational biology, in large part due to the requirements of distributed monitoring, permission seeking, and encryption. The resources available in the social sciences are significantly smaller, and even the physical (and administrative) distance between social science departments and engineering or computer science departments tends to be greater than for the other sciences. The availability of easy-to-use programs and techniques would greatly magnify the presence

of a computational social science. Just as mass-market CAD software revolutionized the engineering world decades ago, common computational social science analysis tools and the sharing of data will lead to significant advances. The development of these tools can, in part, piggyback on those developed in biology, physics and other fields, but also requires substantial investments in applications customized to social science needs.

Perhaps the thorniest challenges exist on the data side, with respect to access and privacy. Many, though not all, of these data are proprietary (e.g., mobile phone and financial transactional data). The debacle following AOL's public release of "anonymized" search records of many of its customers highlights the potential risk to individuals and corporations in the sharing of personal data by private companies (18). Robust models of collaboration and data sharing between industry and the academy need to be developed that safeguard the privacy of consumers and provide liability protection for corporations.

More generally, properly managing privacy issues is essential. As the recent NRC report on GIS data highlights, it is often possible to pull individual profiles out of even carefully anonymized data (19). To take a non-social science example: this past Summer NIH and the Wellcome Trust abruptly removed a number of genetic databases from online access (20). These databases were seemingly anonymized, simply reporting the aggregate frequency of particular genetic markers. However, research revealed the potential for de-anonymization, based on the statistical power of the sheer quantity of data collected from each individual in the database (21).

A single dramatic incident involving a breach of privacy could produce a set of statutes, rules, and prohibitions that could strangle the nascent field of computational social science in its crib. What is necessary, now, is to produce a self-regulatory regime of procedures, technologies, and rules that reduce this risk but preserve most of the research potential. As a cornerstone of such a self-regulatory regime, Institutional Review Boards (IRBs) must increase their technical knowledge enormously to understand the potential for intrusion and individual harm because new possibilities do not fit their current paradigms for harm. For example, many IRBs today would be poorly equipped to evaluate the possibility that complex data could be de-anonymized. Further, it may be necessary for IRBs to oversee the creation of a secure, centralized data infrastructure. Certainly, the status quo is a recipe for disaster, where existing data sets are scattered among many different groups, with uneven skills and understanding of data security, with widely varying protocols.

Researchers themselves must tackle the privacy issue head on by developing technologies that protect privacy while preserving data essential for research (22). These systems, in turn, may prove useful for industry in managing privacy of customers and security of their proprietary data.

Finally, the emergence of a computational social science shares with other nascent interdisciplinary fields (e.g., sustainability science) the need to develop a paradigm for training new scholars. A key requirement for the emergence of an interdisciplinary area of study is the development of complementary and synergistic explanations spanning different fields and scales. Tenure committees and editorial boards need to understand and reward the effort to publish across disciplines (23). Certainly, in the short run, computational social science needs to be the work of teams of social and computer scientists. In the longer run, the question will be: should academia be building computational social scientists, or teams of computationally literate social scientists and socially literate computer scientists?

The emergence of cognitive science in the 1960s and 1970s offers a powerful model for the development of a computational social science. Cognitive science emerged out of the power of the computational metaphor of the human mind. It has involved fields ranging from

neurobiology to philosophy to computer science. It attracted the investment of substantial resources to establish a common field, and it has created enormous progress for public good in the last generation. We would argue that a computational social science has a similar potential, and is worthy of similar investments.

References

1. Roy, D.; Patel, R.; DeCamp, P.; Kubat, R.; Fleischman, M.; Roy, B.; Mavridis, N.; Tellex, S.; Salata, A.; Guiness, J.; Levit, M.; Gorniak, P. The Human Speechome Project. Twenty-eighth Annual Meeting of the Cognitive Science Society; 2006.
2. Eckmann JP, Moses E, SergI D. Entropy of dialogues creates coherent structures in e-mail traffic. Proceedings of the National Academy of Sciences of the United States of America 2004;101:14333–14337. [PubMed: 15448210]
3. Kossinets G, Watts D. Empirical Analysis of an Evolving Social Network. Science 2006;311(5757): 88–90. [PubMed: 16400149]
4. Aral, S.; Van Alstyne, M. Network Structure & Information Advantage. Proceedings of the Academy of Management Conference; Philadelphia, PA. 2007.
5. Pentland, A. Honest Signals: how they shape our world. MIT Press; Cambridge, MA: 2008.
6. Onnela, J-P.; Saramäki, J.; Hyvönen, J.; Szabó, G.; Lazer, D.; Kaskil, K.; Kertész, J.; Barabási, A-L. Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences of the United States of America; 2007.
7. Shaw, B.; Jebara, T. Minimum Volume Embedding. Proceedings of the Conference on Artificial Intelligence and Statistics; 2007.
8. Jebara, T.; Song, Y.; Thadani, K. Spectral Clustering and Embedding with Hidden Markov Models. Proceedings of the European Conference on Machine Learning; 2007.
9. González MC, Hidalgo CA, Barabási AL. Understanding individual human mobility patterns. Nature 2008;453:779–782. [PubMed: 18528393]
10. Eagle, N.; Pentland, A.; Lazer, D. Inferring friendships from behavioral data. HKS working paper; 2008.
11. Colizza V, Barrat A, Barthelemy M, Vespignani A. Prediction and predictability of global epidemics: the role of the airline transportation network. Proceedings of the National Academy of Sciences of the United States of America 2006;103:2015–2020. [PubMed: 16461461]
12. Watts D. Connections A twenty-first century science. Nature 445:489. [PubMed: 17268455]
13. Adamic, L.; Glance, N. The Political Blogosphere and the 2004 U.S. Election Divided They Blog. LinkKDD-2005; Chicago, IL: 2005.
14. J. Teevan. 2008. “How People Recall, Recognize and Re-Use Search Results,” To appear in ACM Transactions on Information Systems (TOIS) special issue on Keeping, Re-finding, and Sharing Personal Information.
15. Bainbridge W. The scientific research potential of virtual worlds. Science 2007;317(5837):472–476. [PubMed: 17656715]
16. Lewis K, Kaufman J, Gonzalez M, Wimmer A, Christakis N. Tastes, Ties, and Time: A New (Cultural, Multiplex, and Longitudinal) Social Network Dataset Using Facebook.com. Social Networks. 2009in press
17. Gardie C, Wilkerson J. Text annotation for political science research. Journal of Information Technology and Politics 2008;5:1–6.
18. Barbarao M, Zeller T Jr. A Face Is Exposed for AOL Searcher No. 4417749. New York Times. 2006 August 9;
19. National Research Council. Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data. In: Myron, P., editor. Gutmann and Paul Stern. Washington: National Academy Press; 2007.
20. Felch, J. DNA databases blocked from the public. LA Times; August 29. 2008

21. Homer N, Szelinger S, Redman M, Duggan D, Tembe W. Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. PLoS Genetics 2008;4(8):e1000167.10.1371/journal.pgen.1000167 [PubMed: 18769715]

22. Backstrom, L.; Dwork, C.; Kleinberg, J. Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography. Proc. 16th Intl. World Wide Web Conference; 2007.

23. Van Alstyne M, Brynjolfsson E. Could the Internet Balkanize Science? Science 1996;274:1479–1480.

24. Image courtesy of Sense Networks.

25. We will supply animation in supporting online materials.

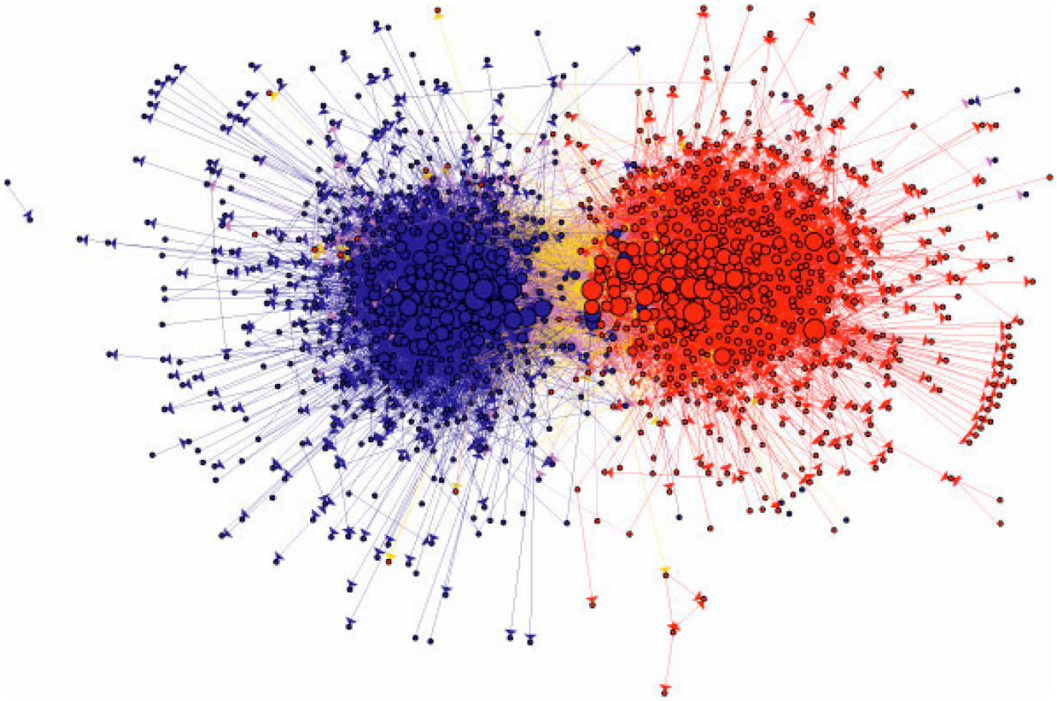

Figure 1.
This figure summarizes the link structure within a community of political blogs (from 2004), where red nodes indicate conservative blogs, and blue liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it (10).

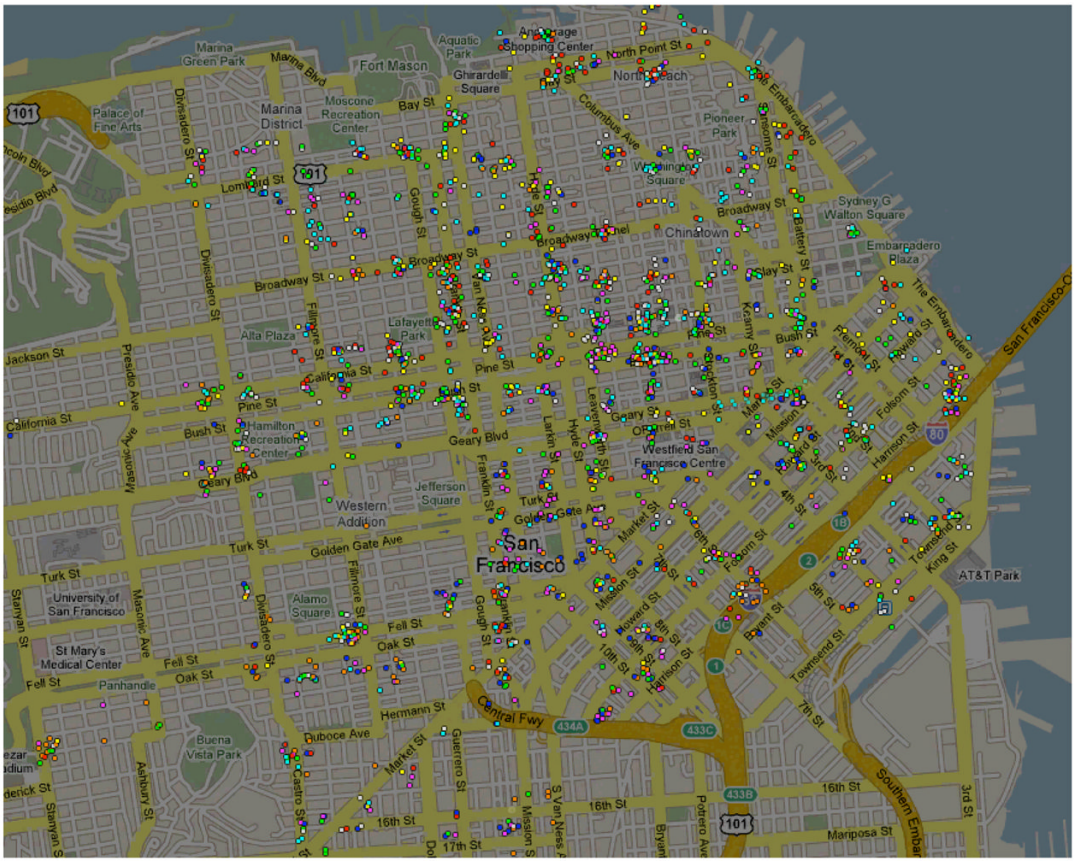

Figure 2.

The location (after adding randomized synthetic noise) of several hundred mobile devices in the city of San Francisco. Each location is color coded to indicate which of 8 “tribes” (or social clusters) each user belongs to. Tribes are computed by clustering (otherwise anonymized) users according to how similar their movement patterns are over a few weeks. The movement analysis is performed using the Minimum Volume Embedding algorithm (7,8,24)

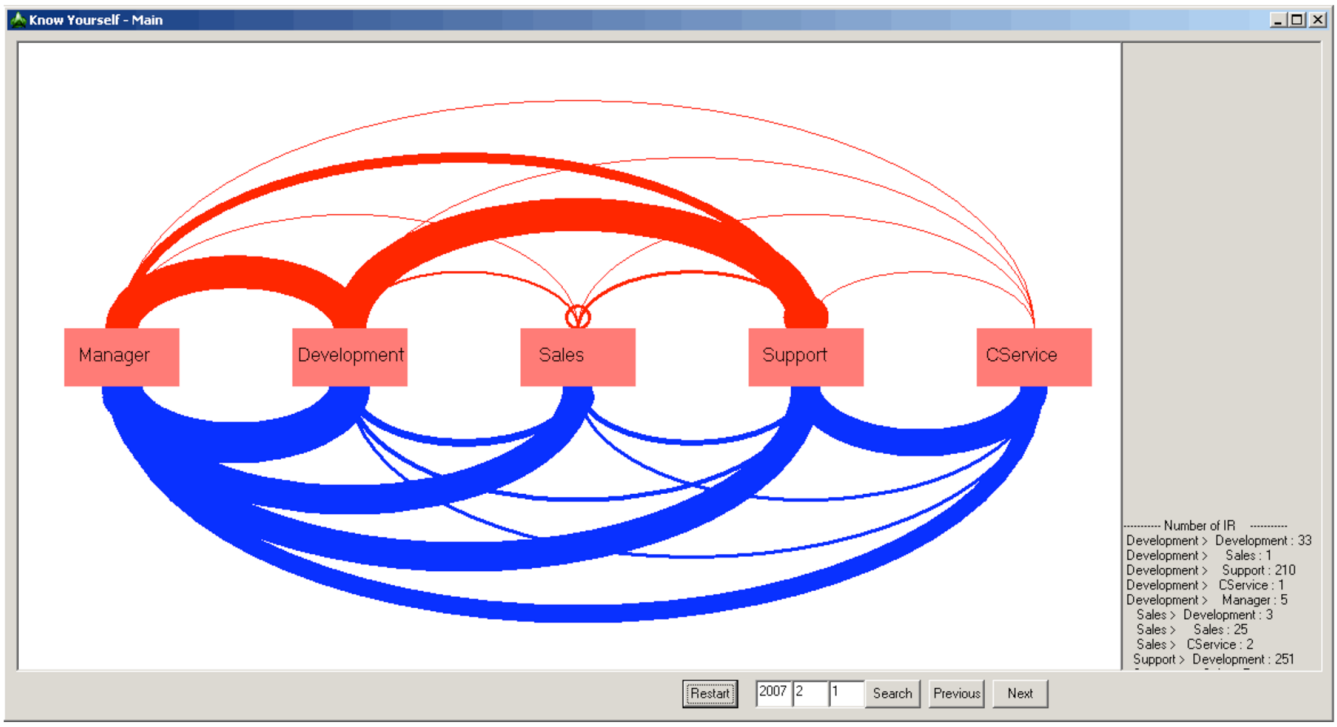

Figure 3.
Patterns of email (blue) and face-to-face communication (read) within a German bank over a period of one month. Productivity and information overload is correlated with the sum of both types of communication, but not with either alone (25)