

Using New Data Analysis Techniques to Understand Information Flows
White Paper for “SBE 2020: Future Research in the Social, Behavioral & Economic Sciences”

Sean Aday
School of Media & Public Affairs
George Washington University

Henry Farrell
Marc Lynch
John Sides
Department of Political Science
George Washington University

Cosma Shalizi
Department of Statistics
Carnegie Mellon University
External Faculty
Santa Fe Institute

Abstract: Information flows motivate key questions in the major social sciences. Yet scholars have had great difficulties in studying them directly. The movement of social activity to the Internet means that it is now possible to study information flows directly in a much more systematic fashion than before – data on many forms of social interaction is readily available in machine-readable format. Yet properly studying this new data will require new tools and new techniques. This White Paper proposes a two-stage program to develop new tools in conjunction with pilot initiatives studying information flows, and then apply them more broadly. It then outlines how these methods and data might be applied to three major problems spanning different social sciences – collective cognition, frames and mobilization, and political polarization – and concludes by discussing the policy benefits of better analysis.

Abstract Word Count: 135 words.

White Paper Word Count: 1,940 words.

This White Paper is made available under a Creative Commons Version 3.0 Attribution Non-Commercial Share Alike License. See <http://creativecommons.org/licenses/by-nc-sa/3.0/legalcode> for details and conditions.

How does information move within and between societies, and what are the consequences of different patterns of flow? These are important questions for political science, sociology, economics, psychology, and the cognitive sciences. We have good reason to believe that information flows help shape important explaining political, economic, and social outcomes. Authoritarian regimes invariably seek to secure their power by controlling the spread of information among their subjects. Advanced industrialized democracies impose strict rules on information disclosure by businesses listed on stock markets to prevent market manipulation. Networked information flows shape social phenomena as different as the creativity of Broadway musical writers and the likelihood of liberal arts colleges to adopt professional degree programs.

Yet social scientists have not been able to study information flows systematically and on any large scale. They have had to use various inadequate proxies, leading to many valuable findings but also many frustrations. There is now a major new opportunity to change this. The movement of social activity to the Internet creates huge pools of potential data on how individuals communicate with each other, all of it in machine-readable formats and much (though not all) of it publicly available.

These new data can illuminate how information flows work in society, and facilitate hitherto impracticable forms of analysis. For example, large-scale, continuous, persistent and automatic data collection from many sources allows researchers to “roll back the tape” when an important and unexpected event occurs and see how information flow might have helped make this event happen. This kind of data-gathering also allows detailed analysis of who received specific information, from whom, and when.

While computer scientists and others are developing new techniques to analyze these data, social science has lagged behind in applying them for several reasons. First, and most obviously, social scientists are unfamiliar with the relevant techniques and require further training to use available tools. Second, social scientists require a better understanding not only of information flows themselves, but also of their role in political, economic and social phenomena. Estimating the consequences of information flows will need new kinds and sources of data so as to capture those effects.

What is needed is a cross-disciplinary initiative, funding the development of basic research tools alongside multiple projects by a variety of scholars using these tools. The next section lays out the parts of such an initiative. The next section details three major research questions that this initiative could help address. The final section addresses the practical benefits of a better understanding of information flows.

An Initiative to Study Information Flows

We propose a two-stage initiative.

The first stage would focus on *developing tools* for social scientists, and *providing training* in their use. Rather than creating entirely new tools, the emphasis would be on adapting existing tools so as to make them better fitted to the needs of social scientists. Two existing initiatives illustrate the opportunities and challenges involved. MemeTracker (<http://www.memetracker.org>), developed by computer scientists at Cornell University and Stanford University, tracks how specific bits of

information (such as short phrases) travel across different media (newspapers with online RSS feeds, blogs, Twitter) and mutate in their travels. MediaCloud (<http://www.mediacloud.org>), developed by researchers at Harvard University's Berkman Center, can compare patterns of appearance or conjunction across a wide array of media sources.

There are (at least) two modifications which would make such tools far more useful to social scientists. First, these tools need application programming interfaces and user-friendly front ends that would let social scientists easily customize them, formulate queries, and extract and analyze data. Second, for cross-national research, the tools must work across multiple languages, not just (as at present) English.

An initial request for proposals would therefore encourage cross-disciplinary proposals from computer and social scientists to develop these tools in ways that would benefit the social sciences. Any funded software should be released under a relevant open-source or Creative Commons license. Proposals should also include pilot projects that would apply the tools to social scientific puzzles, obliging applicants to think carefully about the scientific utility of the tools, rather than just technical or data-driven issues. Such pilot projects could help guide the second stage of the program.

The second stage would build on the first stage in two ways. First, it would train social scientists in these new techniques and tools, with an emphasis on training graduate students through workshops under the auspices of disciplinary associations and institutions, such as the Santa Fe Institute. These training sessions would also address ethical issues, as some data on information flows (e.g., across Facebook or Twitter) can expose information that ought to stay private.

Second, this stage would include a broader request for proposals, encouraging the use of these techniques and tools to address questions of real social-scientific interest. Along the lines of other recent NSF initiatives, this request would be explicitly cross-disciplinary, encouraging scholars to collaborate with social scientists from other disciplines, with statisticians, and with computer scientists. The result would be distinct yet complementary investigations of how information flows shape politics, the economy, and society.

Outstanding Research Questions

We suggest three major research questions that cut across the social sciences, that resist traditional forms of analysis, and that can be addressed using the new data and analytic techniques. Our list is far from exhaustive but merely illustrate the potential.

Processes of Collective Cognition

Fundamental debates in political science (over democratic theory and the political consequences of markets), economics (over social choice and public choice theory), and sociology (disputes in organizational and economic sociology) concern the relative merits of different forms of social organization, such as bureaucratic hierarchy, decentralized decision-making, and democratic deliberation. Recently, these debates focused on the cognitive merits of these different forms in organizing knowledge, making it usable, and structuring decisions. Economists posit that the process of exchange itself adaptively finds and implements optimal allocation decisions. Political scientists suggest that democratic deliberation can effectively improve policies piecemeal, in light of the information and ideas of many participants. Similar remarks apply to bureaucracies, such as corporations, and to scientific disciplines.

We can consider such modes of social organization as forms of *collective cognition*, processing knowledge according to different logics and institutional schemes. Data on online behavior in collaborative filtering services (Reddit, Digg), knowledge production sites (Wikipedia), and political organizations (community blogs such as the Daily Kos) will then show how people engage in collective cognition, how they draw on different mixtures and forms of hierarchical and non-hierarchical decision-making as well as quasi-deliberation, and how collective cognition depends on the external circumstances and goals of the group.

We can thus begin to build and test explanations of the relative success of different forms of collective cognition. Current work in machine learning (“ensemble” learning systems) and the cognitive science of cooperative problem-solving provides initial hypotheses regarding, for example, the role of diversity in collective cognition. By combining comprehensive datasets with new means of measuring information flows, scholars could put bounds on the performance of ensemble-learners and group problem-solvers, see how close actual social information processing systems come to those bounds, and how these systems could be improved.

The Effects of Framing on Social Movements

Social movement theory has identified mechanisms that facilitate large-scale contentious political action. Many of those mechanisms can be affected by the broader information environment. For example, Internet-based media may make it easier to organize protest activities and communicate with like-minded citizens, in turn making contentious politics more prevalent.

Framing is another commonly cited mechanism, and one particularly amenable to these new tools. Framing involves competitive attempts to impose meaning upon a set of actors, issues, or events and thereby empower a particular mode of political action. Understanding why some efforts at framing succeed more than others requires tracing the emergence and diffusion of different frames: who initiates them, how and where; how they spread through networks; and why some die out while others become widely shared collective understandings.

In addition, these new data analysis tools allow scholars to test the relative influence, diffusion, and dynamics of frames emanating from various sources, thus providing far more complex and rich empirical and theoretical models across a globalized media environment than previously possible. For instance, a tool such as MemeTracker could in theory allow us to better understand the nature and effects of varying media frames (e.g., Western vs. Arab media coverage; blog vs. mainstream news media coverage) of a given protest movement.

The Underlying Mechanisms of Group Polarization

Sociologists and political scientists have long wanted to know how social contact affects identities and beliefs. Political scientists are especially interested in polarization: the tendency of individuals or groups to assume diverging political positions. Scholars argue that contact with like-minded individuals as well as selective exposure to and acceptance of information drive polarization.

Data on information flows will illustrate how, and how much, selective exposure occurs. For example, there is a vigorous academic debate over whether polarized patterns of link exchange between political websites mean that readers of left-leaning and right-leaning websites inhabit radically inconsistent universes of political information. We need, however, to measure the content of these universes, to determine whether and how particular information actually crosses over from left to right, to characterize the temporal dimension of information flows. With these data, we will know exactly how much selectivity is possible.

These data can then be used to answer causal questions, in tandem with other forms of data (e.g., on network structures or individual attitudes). For example, a burgeoning literature in sociology, building both on sociological network theory and on physicists' studies of the structure of large networks, argues that different network topologies produce different patterns of information diffusion. Data on information flows, plus tools to extract the patterns from these data, would let us test these hypotheses. Measuring individual attitudes, together with data on information flows—leveraged through observational and experimental research—would allow researchers to grasp the consequences of information flows for attitudes and polarization.

Conclusions – The Practical Benefits to Understanding Information Flows

The intellectual case for better research on information flows is clear. Yet if we had better answers to the questions discussed above (let alone the many other questions that we could address with better tools), there would be clear pay-offs for policy makers as well as academics.

First, political institutions, businesses and non-profit organizations all experiment continuously with different organizational forms aimed at improving the quality of information processing and decision making. However, it is extremely hard to determine the respective merits of such organizational form given the lack of useful data. Better tests of how different schemes of collective cognition work 'in the wild' would have immediate practical benefits for organizational designers.

Second – a better understanding of the causes and consequences of group polarization would have practical significance both for US politics and for international relations. The US political system appears to be moving towards greater polarization. By understanding the role of information flows in polarization, we can better gauge the likely impact of these changes, and of institutional reforms designed to accommodate them or mitigate their negative impact as appropriate. We can also understand better the mechanisms of group polarization in unstable societies, and how best to mitigate them before conflict erupts.

Third, by understanding better the dynamics of framing, we could contribute both to public diplomacy, and the better understanding of global information dynamics. Political debate both within the US and outside it is frequently structured by pernicious myths (e.g. the rumors surrounding September 11 2001, and the recent mosque controversy). Combating these myths requires a detailed understanding of the structures and dynamics of how information flows affect

efforts to frame and reframe specific political debates.