David Karpf

SOCIAL SCIENCE RESEARCH METHODS IN INTERNET TIME

This article discusses three interrelated challenges related to conducting social science research in 'Internet Time'. (1) The rate at which the Internet is both diffusing through society and developing new capacities is unprecedented. It creates some novel challenges for scholarly research. (2) Many of our most robust research methods are based upon ceteris paribus *assumptions that do not hold in the online environment. The rate of change online narrows the range of questions that can be answered using traditional tools. Meanwhile, (3) new research methods are untested and often rely upon data sources that are incomplete and systematically flawed. The paper details these challenges, then proposes that scholars embrace the values of* transparency *and* kludginess *in order to answer important research questions in a rapidly-changing communications environment.*

Introduction

The Internet has dramatically changed through the past decade. Consider: in 2002, streaming video was rare, short and choppy. Wireless 'hotspots' were a novelty. Mobile phones were primarily used for (gasp!) phone calls. Commercial GPS applications were still in the early stages of development. Bloggers could be counted by the handful. Social networking sites like Friendster, Myspace and Facebook were still confined to Bay Area networks and technologists' imaginations (boyd & Ellison 2007).[1] A 'tumbler' was a type of drinking glass, a 'tweet' was a type of bird call. Simply put, the Internet of 2012 is *different* from the Internet of 2002. What is more, there is little reason to suppose this rapid evolution is finished: the Internet of 2022 will likely be different from the Internet of 2012.

   In a 2001 article, Barry Wellman playfully suggested that 'an Internet year is like a dog year, changing approximately seven times faster than normal human

time' (Wellman 2001). Alongside Wellman's claim, technologists routinely make reference to 'Moore's Law', the 1965 prediction by Intel founder Gordon Moore that transistor capacity would grow exponentially, doubling once every 18−24 months (Moore 1965). Moore's Law has proven surprisingly (albeit approximately) resilient over the past 45 years, and has become synonymous in popular technology literature with the rise of abundant and cheap computational capacity. It is an oversimplified concept, usually associated with naïve strains of techno-optimism, which tends to overlook the multi-layered 'hourglass architecture' associated with the Internet (Zittrain 2008, pp. 67−71).

But if technology writers have relied too much on Moore's Law as a concept, social scientists have all but ignored it. Tour through the indexes of the leading political science and sociology texts and you will find nary a mention of Moore's Law, or of the continuing evolution of the medium (examples from my bookshelves and syllabi include Sunstein 2001; Bimber 2003; Bimber & Davis 2003; Howard 2006; Sunstein 2006, 2007; Perlmutter 2008; Hindman 2008; Davis 2009; Kerbel 2009). 'Internet Time' is a subject grudgingly acknowledged in our research designs, rather than incorporated within them. Members of the interdisciplinary research community are aware that the suite of technologies collectively referred to as 'the Internet' continues to develop. But that awareness rarely becomes a feature of our research methods.

This feature of 'Internet time' creates a substantial hurdle. The Internet is unique among Information and Communications Technologies (ICTs) *specifically because* the Internet of 2002 has important differences from the Internet of 2005, or 2009, or 2012. It is a suite of overlapping, interrelated technologies. The medium is simultaneously undergoing a social diffusion process and an ongoing series of code-based modifications. Social diffusion brings in new actors with diverse interests. Code-based modifications alter the technological affordances of the media environment itself (Lessig 1999; Goldsmith & Wu 2006). What was costly and difficult in 2004 is cheap and ubiquitous in 2008. That leads, in turn, to different practices.[2] The Internet's effect on media, social, and political institutions will be different at time X from that at time $X + 1$, because the suite of technologies we think of as the Internet will itself change within that interval.

Standard practices within the social sciences are not well suited to such a rapidly changing medium. Traditionally, major research endeavors move at a glacial pace. Between grant applications, data collection, writing, peer review, and publication, it is not uncommon for a research project to consume five or more years between conception and publication. Several books and articles have been published in 2011 on the basis of 2006 data. Scholars openly acknowledge the limitations imposed by the system. It is a common topic for commiseration at conference hotel bars. In the time it takes to formulate, fund, conduct, revise, and publish a significant research question, we all are left to worry that changes in the media environment will render our work obsolete.

The last two stages of the process – peer-review and publication – have become topics of robust conversation. There has been ample innovation already, with online journals, automated components of peer-review, and academic blogging being three obvious examples. The conversation surely must continue, but I for one am satisfied with how it has begun. When it comes to the 'tools of the trade', collective introspection among social scientists has been far less forthcoming. Everyone can agree that it would be nice to see publication speeds improve. Exploring the limitations of panel and survey data, when applied during the early lifecycle of a changing media environment, is a steeper climb.

This article is an effort to broach that methodological conversation. I offer a methodological prescription for conducting social science research that takes 'Internet Time' seriously. It is rooted in methodological pluralism – the belief that social science researchers ought to be question-driven, letting the nature of their inquiry determine their methods, rather than the other way around (Smith 1997). To be clear, I do not argue that traditional methods have been rendered obsolete by the Internet. That is too strong a claim, and there are several counter-examples readily at hand. Surveys of mass behavior have led to key findings regarding political participation (Schlozman *et al*. 2010) and the digital divide, to name only two (Hargittai 2010). Rather, I argue that an expanding range of interesting and important questions cannot be answered using the most refined tools in our toolbox. The new media environment demands new techniques. Those techniques carry risks – they have not undergone the years of seasoning and sophistication that dominant methods have. But they also carry the promise of expanding the scope of our inquiry and applying intellectual rigor to topics of broad social significance.

This methodological prescription is based heavily upon my own experiences dealing with novel topic areas, facing the hurdles imposed by messy data, and creating novel datasets for assessing the questions that interest me most. Different branches of the Internet research community define their research questions and appropriate designs differently. I offer this article as a political scientist who has migrated into the interdisciplinary field of Internet research. Researchers with different academic lineages will surely define a few key concepts differently (what constitutes a dominant method, or a relevant theory, or an acceptable form of data?). Yet I suspect the challenges and solutions I present will be broadly familiar to all readers.

The argument proceeds in the form of four brief essays. I try my very best to be provocative throughout. The first essay highlights changes in the online environment itself. Key areas of inquiry – from political blogs to websites and mobile applications – are in a state of maturation. In the course of a few years, Twitter has moved from the lead-adopter stage to the late-majority stage of diffusion. The social practices and technical architecture of social networking sites like Facebook have substantively changed as they have moved through the diffusion process. The boundaries that once clearly defined the

blogosphere have become irredeemably fuzzy – today, I will argue, there is no such thing as 'the blogosphere'. The rate at which these sociotechnical systems evolve is something new for social scientists to deal with.

The second essay turns to Internet Time and social science research. Our most robust research techniques are based upon *ceteris paribus* assumptions that are routinely violated in this fast-developing media environment. Online behavior at Time X only predicts online behavior at Time $X + 1$ if (1) the underlying population from which we are sampling remains the same and (2) the medium itself remains the same. A changing media environment, under early adopter conditions, violates both (1) and (2). *Ceteris* is not *paribus*. All else cannot be assumed equal. One obvious consequence is that research findings are rendered obsolete by the time they have been published. The terrain we can explore with traditional social science techniques has narrowed in scope.

The third essay focuses upon endemic problems associated with online data quality. The online environment is lauded for its abundant data. But anyone who has worked closely with this data is well aware of its limitations. Spambots, commercial incentives, proprietary firewalls, and noisy indicators all create serious challenges for the researcher. We can rarely be sure whether our findings are an artifact of a flawed dataset, particularly when working with public data. Online data abundance offers the promise of computational social science, but only to the extent that we can bridge the data divide.

The fourth and final essay offers some hopeful examples of how social science research can be augmented by the dynamic, changing nature of the online environment. By treating Moore's Law and Internet Time as a feature of our research designs, rather than a footnote to be acknowledged and dismissed, well-theorized, rigorously-designed studies can gain a foothold on an expanded array of research questions. Doing so requires some new habits among the research community, however – or at least a renewed commitment to some old habits that have been lost. I argue for embracing the messiness, promoting design transparency, and supporting 'kludginess', or the use of hacks and workarounds that allow for messy-but-productive solutions where more elegant choices would prove unworkable.

Internet time and sociotechnical systems, or, there is no such thing as the blogosphere

My own perspective on 'Internet Time' has been driven by longtime observation of the political blogosphere. Writing in 2003, Clay Shirky offered the following observation: 'At some point (probably one we've already passed), weblog technology will be seen as a platform for so many forms of publishing, filtering, aggregation, and syndication that blogging will stop referring to any particularly coherent activity' (Shirky 2003). Notice that Shirky is still referring to blogs as

'weblog technology'. In 2003, blogging was still new enough that writers had to explain that *blog* is short for *weblog*.

Today, Shirky's statement appears prescient.

There used to be a single 'blogosphere'. A few lead adopters had taken up the blogger.com software produced by Pyra Labs, or had latched on to similar platforms. They shared a common style. They relied on common software architectures, and those architectures offered the same technological affordances. Bloggers in 2003 were mostly pseudonym-laden. They engaged in an opinionated 'citizen-journalism'. They made use of in-text hyperlinks and passive (blogroll) hyperlinks, networking together with peer blogs. They offered comment threads, allowing for greater interactivity than traditional online journalism. They were, by and large, critics of existing institutions of power – be they journalistic, political, or commercial. A 'blogger' was, simply, 'one who blogged'. And that was a small-but-growing segment of the population.[3]

The blogosphere grew. In so doing, it underwent a common pattern of simultaneous *adoption and adaptation*. We see this pattern repeated in other online channels as well. It is predictable as the tides. First, consider adoption. As the 'blogosphere' grew from under a thousand blogs to over a million blogs, the new adopters used it towards different ends. The uses and gratifications driving a lead adopter are not the same as those driving a late adopter. The one is experimenting, often because they *enjoy* experimenting with new media. The other is applying a well-known technology toward their existing goals. Different users, with different motivations and interests, applied the technology of online self-publication to different ends.

This adoption pattern holds true for all ICTs. It is not unique to the suite of technologies that make up the Internet. Early adopters of the radio, the telegraph, and the automobile all made use of those technologies in different ways from the early- and late-majority adopters. Robust literatures in science and technology studies, history of technology, and diffusion-of-innovation studies all provide valuable insights into the development of communications media, as well as the formative role that government policy can play in their development (Mueller 2004, 2010; Gillespie 2009; John 2010; Johns 2010; Wu 2010). The sheer speed of adoption, and the numerous technologies that are all considered to be 'the Internet' or 'the Web' has hastened confusion on this matter, however. We still tend too easily to assume that the interests and motivations of early 'bloggers' will be replicated among later 'bloggers', even though we know that this population is far from static.

The adaptation pattern is even more pernicious. As the blogosphere grew, the underlying software platform underwent key modifications. The community blogging platform offered by Scoop, for instance, added user diaries to certain blogs, allowing them to operate as gathering spaces for online communities-of-interest. Sites like DailyKos developed interest group-like qualities, using the blog to engage not in modified citizen journalism, but in modified citizen

activism (Karpf 2008b; Benkler & Shaw 2010). DailyKos endorses candidates, engages in issue mobilization, and even holds an annual in-person convention. It has more in common with traditional advocacy groups than it does with the solo-author political blogs of 2003. Likewise, existing institutions of authority added blogging features to their web offerings. The *New York Times* now hosts blogs. So does the Sierra Club. Unsurprisingly, the blogging activity on these sites advances their existing goals. A blog post at *NYTimes.com* has more in common with an article on *NYTimes.com* than it does with a blog post at mywackythoughts.blogspot.com. Counting Krugman.blogs.nytimes.com, dailykos.com/blog/sierradave, and (more problematically) Huffingtonpost.com as part of a single, overarching 'blogosphere' is clearly problematic. They are not members of a singular underlying population any longer. The inclusion of new software code means that there is no longer a categorical distinction to be drawn between 'bloggers' and other web readers. To rephrase Shirky, at *this* point, blogging *has stopped* referring to any particularly coherent activity. Blogging is simply writing content online.

   Today, it is no longer analytically useful to conduct research on '*The* Blogosphere'. I say this as the proprietor of an open data resource called the Blogosphere Authority Index (BAI). Maintained since 2008, the index offers a ranked tracking system for researchers interested in elite political blogs in America. The BAI tracks two blog *clusters*, and it is a methodology that can be used for ranking other blog clusters (Karpf 2008a). But it does not actually track or measure the entire blogosphere, because there is no such thing anymore. Speaking of the blogosphere as though it is a single, overarching component of the world wide web only encourages faulty generalizations – leading researchers and public observers alike toward inaccurate claims about the quality of 'citizen journalism', the trustworthiness of 'bloggers', and the goals and effectiveness of 'bloggers'. All of these generalizations misinform rather than enlighten. They create fractured dialogues and regrettable cul-de-sacs in the research literature. 'Blogging', today, is a boundary object, with different meanings and implications among different research communities.

   From a methodological perspective, it is of little specific importance that there is no longer any such thing as the blogosphere. Internet-related social science research covers a much larger terrain than the blogosphere. What *is* important is that *this was a predictable sequence of events*. As the technology diffuses, it attracts different adopters with different interests. Many of these adopters will come from existing institutions of authority, creating a process of 'political normalization' (Margolis & Resnick 2000). Also as the technology diffuses, it changes. Blog software today includes key modifications and variations on the basic 1999 Pyra Labs platform. The software modifications provide variant technological affordances (Kelty 2008; Reagle 2010). The same is true for social networking, peer-to-peer file sharing, micro-blogging, and online video creation. It is a feature that makes Internet innovations a bit trickier than previous periods of

ICT innovation. A television set in 1940 was largely the same as a television set of 1950. The intersection of new adopters, government regulators, and commercial interests drove the development of television, but the ICT itself remained basically the same thing until the development of cable in the 1970s and 1980s (as such, cable television is often treated as a separate innovation). Blogging and other developments at the social layer of the Internet diffuse very fast, and also acquire key code-based adaptations along the way.

This makes for a messy research landscape. The most robust traditional techniques available to social scientists – sophisticated quantitative methods that isolate and explore causal claims – were not designed with such a landscape in mind. As such, a few common problems routinely crop up.

Internet time and social science research

Nearly every US election since 1996 has been labeled 'the Year of the Internet'. Important milestones have indeed been reached in each of these elections, with 1996 marking the first campaign website, Jesse Ventura's Internet-supported 1998 victory in the Minnesota Governor's race, John McCain's online fundraising in the 2000 Presidential Primary, Howard Dean's landmark 2004 primary campaign, the netroots fundraising and Senator George Allen's YouTube 'Macaca Moment' in 2006, and Barack Obama's historic 2008 campaign mobilization (Lentz 2002, Bimber & Davis 2003, Cornfield 2004, Trippi 2005, Foot & Schneider 2006; Karpf 2010a, 2010b; Kreiss Forthcoming). Were claims that 2000 was the 'Year of the Internet' premature? Were claims that 2008 was the 'Year of the Internet' lacking in historical nuance? I would suggest an alternate path: that both were accurate, but the Internet itself changed in the interim. The Internet of 2008 is *different* from the Internet of 1996, 2000, or 2004, and this is a recurrent, ongoing pattern.

Consider the following trivia question: 'What was John Kerry's YouTube strategy in the 2004 election?'

YouTube is a major component of the Internet today. The video-sharing site is the third most popular destination on the Internet, as recorded by Alexa.com. Political campaigns now develop special 'web advertisements' with no intention of buying airtime on television, simply placing the ads on YouTube in the hopes of attracting commentary from the blogosphere and resultant media coverage (Wallsten 2010; Barzilai-Nahon *et al*. 2011). The medium is viewed as being influential that an entire political science conference and a special issue of the *Journal of Information Technology and Politics* were devoted to 'YouTube and the 2008 election'. Yet no social scientist has ever looked at John Kerry's use of the site in the prior election cycle. The absence should be utterly baffling (if it were not so easy to explain). How could we focus so much attention on YouTube in 2006 and 2008 while ignoring it completely in earlier cycles?

The answer, of course, is that John Kerry *had* no YouTube strategy. YouTube, founded in 2005, did not exist yet. The Internet of the 1990s and early 2000s featured smaller bandwidth, slower upload times, and less-abundant storage. The technical conditions necessary for YouTube to exist were not present until approximately 2005. To the extent that the sociotechnical practice of video-sharing, and the capacity of individuals to record, remix, and react to video content without relying on traditional broadcast organizations, impacts American politics, it is an impact that makes the Internet of 2004 *different* from the Internet of 2008.

Social science observational techniques were not developed with such a rapidly-changing information environment in mind. Bruce Bimber and Richard Davis, earned the McGannon Center Communication Policy Research Award for *Campaign Online*, a rigorously detailed study of the Internet in the 2000 election. After clearly demonstrating that political websites were predominantly visited by existing supporters, they concluded that the new medium would prove relatively ineffective for persuading undecided voters. As such, they came to the conclusion that the Internet would have a relatively minimal impact on American politics, offering only 'reinforcement', of existing beliefs, rather than persuasion. Their book is a standard feature of the Internet politics canon today.

Bimber and Davis's finding about candidate websites remains nominally accurate today. In 2012, we can safely believe that most of the visitors to the Republican candidate's website will be existing supporters. Low-information, undecided voters (by definition) are not seeking out such political information. But the implications of their findings are overshadowed by further development in the online landscape. Candidate website engagement is no longer synonymous with online political activity.

As it happens, Bimber and Davis's book was published in November 2003, just as the Internet-infused Howard Dean primary campaign had reached its zenith. The Dean campaign was using the Internet to mobilize supporters with overwhelming effectiveness, drawing large crowds and setting online donation records. To this day, the Dean campaign is synonymous with the birth of Internet campaigning; the contrast between scholarly wisdom and current events could not have been much more stark. The 'reinforcement' of 2000 had morphed into resource mobilization. Yet it would be patently absurd to criticize Bimber and Davis for not foreseeing the Dean phenomenon. The Internet of 2004 had features not present in the Internet of 2000 (such as large-scale adoption of online credit card payments[4]). Those features leveraged different social practices. A network of political actors was simultaneously learning to use the new media environment and also helping to change the boundaries of the environment itself. New software code and participatory sites, supported by increasingly cheap bits and bytes, afforded new practices with alternate results.

Zysman and Newman (2007) have helpfully described the Internet as a 'sequence of revolutions'. While I myself try to avoid the loaded term 'revolution' in my work, it has a certain utility in this setting. The medium keeps changing, apace Internet Time. The Internet is in a state of *ongoing* transformation. As a result, academic pronouncements about the Internet's impact on politics in any given year face a limited 'shelf life'. As the medium changes, the uses it supports change as well. *Ceteris paribus* is violated at the outset – we can assume to begin with that all other things will not be equal.

This feature of Internet Time is particularly problematic for some of our most robust, traditional research methods that seek to impute the political behavior of an entire population based on a sample. Randomly sampling the US population in 2000 is of limited use in determining online political behavior in 2004 or 2012. Changes in the media environment produce new technological affordances. Social actors then leverage these affordances to alternate ends (Earl & Kimport 2011). Meanwhile the 'gold standard' in behavioral research is the biennial American National Election Study, with questions that can be traced back 60 years. The power of such time series studies is diluted when sociotechnical systems are undergoing rapid, ongoing change. We can look for statistically significant findings in the large datasets, but only while embracing *ceteris paribus* assumptions that are clearly unsupportable.

This is not a call for abandoning survey instruments and behavioral studies. I am suggesting that we broaden our methodological toolboxes, not discard tools. And indeed, behavioral studies from sources like the Pew Internet and American Life project have been a tremendous resource for the Internet research community. But Pew is an exceptional example – a large, well-funded research shop that continually modifies its poll questions in light of changes in the new media environment. For university-based researchers attempting to construct multi-year panels or comparable survey questions while navigating Institutional Review Boards, grant timelines, and peer-reviewed publication cycles, it becomes imminently likely that our best methods will yield research that is systematically behind-the-times. Large-N quantitative studies are, as a rule, slow-moving affairs. If we limit ourselves to the set of research questions that can be profitably explored through our most robust techniques, we substantially narrow the research landscape. We make it more likely that studies of the blogosphere or online campaigning will accurately explain a phenomenon only after it has substantially changed in an unexpected manner.

The Internet has also provided a wealth of new data, however. The digital traces of political behavior are now evident wherever we look, and these have spurred new efforts to collect and analyze novel datasets. Some of these new methods – in particular the trend toward 'computational social science' that combines social science and computer science to leverage large datasets nearly in real-time – hold great promise (Lazer *et al*. 2009). Massive datasets, rapidly compiled, appear at first blush to offer a solution to the problems described above. But they

also face the lurking threat of GIGO (Garbage In, Garbage Out). Publicly available online data often has deep, systematic flaws which too often goes ignored.

Data everywhere, and not a drop to drink

Internet research offers the siren-song promise of abundant data. Indeed, many research techniques have been augmented by online data. Web-based survey instruments are cheaper, faster and easier than postal service-based surveys. Blog and webpage hyperlinks provide ample traces of network ties, enhancing social network analysis. Several scholars have begun experimenting with Amazon's Mechanical Turk as a low-cost, rapid service for content analysis and other tasks that are simple-in-theory but difficult-at-scale (Zittrain 2010; Shaw 2011). Qualitative case analysis and process-tracing research also find fertile ground in the online environment, as time-stamped archives lower the barriers to constructing timelines of events and identifying how participants discussed a case as it unfolded. There is plenty about the online environment for the intrepid researcher to appreciate.

Yet online data abundance often proves to be a mirage. Three problems commonly confront academic researchers: (1) the ephemeral nature of open data, (2) the noisy influence of spammers and (3) the limitations of public data. These three problems do not affect all research methods equally. But, to date, they have gone completely unacknowledged in the research community. As a result, we rarely see any frank discussion of what can and cannot be effectively measured online. 'Big data' faces its own Internet Time-related problems.

The ephemeral World Wide Web: a call for more 'lobster traps'

The 'Wayback Machine', hosted by the Internet Archive (web.archive.org), is an *exceptional* research resource. The site takes frequent snapshots of the changing web, and makes them available for free. Want to conduct a study of how Slate.com, Facebook.com, or Democrats.org have evolved over time? The Wayback Machine is the only publicly-available source. It is exceptional in the narrow sense, though: as an exception to a general rule. The rule is that many publicly available types of content go unsaved and disappear forever. Researchers are limited to whatever the Wayback Machine happens to capture. Some popular sites will be thoroughly canvassed. Others are rarely-if-ever catalogued. If one is interested in the development of a website that did not happen to receive frequent Internet Archive attention, or received it during the wrong intervals of time, the research endeavor will meet an abrupt and unfortunate end. The Wayback Machine is the 'only game in town', so to speak.[5]

We can conceive of the Wayback Machine as a 'lobster trap', of sorts. Lobster traps sit passively in the ocean, placed in areas of strategic interest.

From time to time, one can check the traps and see if anything interesting has come up. The Internet is similarly awash in data that may be of interest to researchers. We often want to make across-time comparisons. But without lobster traps, we are bound to go hungry, so to speak.

In several other areas of interest, no such Wayback Machine-analog exists. In particular, web traffic is publicly traced through a handful of services (Alexa, Sitemeter, Comscore, Quantcast, and Hitwise are all frequently used by scholars in my discipline). This data is publicly available,[6] but it has a shelf life. If one is interested in (for instance) how a political blog's traffic has changed over the past 5 years, none of these services can tell you.[7] Online fundraising, e-mail, web traffic, online video, and twitter traffic are all areas of abundant data that goes generally un-archived. In the sea of 'big data', we find surprisingly often there is not a drop to drink.

This particular challenge has a simple solution – simple, at least, when compared to the other two data challenges discussed below. Like the Wayback Machine, it is relatively simple to develop lobster traps for specific data sources of interest. Both of the open dataset I operate – the BAI and the Membership Communications Project (MCP) – follow this principle. The BAI collects ephemeral hyperlink and traffic numbers for two clusters of elite political blogs. Saving that data provides an invaluable resource for political blog researchers – in 2009, when a research team at the University of Washington wanted to study the role of elite blogs in disseminating viral videos, they found that the BAI was the *only* resource that could provide information from the heavily documented 2008 election cycle. Likewise, the MCP gathers mass membership e-mails from a cluster of 70 large-scale progressive advocacy groups. For scholars who want to know how advocacy groups mobilized around Obama's Health Care Reform bill or reacted to the Gulf Oil Spill, the MCP provides valuable data that was public at the time, but otherwise would have disappeared from the web.[8]

Privacy advocates are fond of claiming that, once something is published online, it is available forever (Solove 2008; Mayer-Schonberger 2011). This is only approximately true, though. Particularly for researchers seeking contextual information, including researchers interested in virality, blog/webpage influence, and new forms of contention, much of the data that we fail to consciously save is lost forever.

GIGO: Spambots and data noise

It is time to acknowledge a hard truth that will always confront the research community online. Our data is likely never going to be all that good. Measuring online clicks and hyperlinks is technically hard enough. Distinguishing between real traffic and fake traffic is, with a few rare exceptions, impossible.

Again, let us use the blogosphere as an example. Hyperlinks are frequently used to identify top blogs. As a result, there is a financial incentive for talented

code-writers to game the system. One computer science study suggested that up to 10−20 percent of all blogs are 'splogs', or spam-blogs, set up as phantom sites to artificially boost hyperlink levels (Kolari *et al*. 2006). Similarly, Web traffic measures can be gamed by computer programs. On Twitter, former House Speaker Newt Gingrich was discovered to have *purchased* 80 per cent of his 1.3 million Twitter followers (Cook 2011). These were fake accounts, but they served the very real purpose (for a time) of propping up claims that his campaign had mass support. The media attention his large follower list attracted was well worth the price (and likely cheaper than running a traditional media campaign).

These fake links, clicks, and followers are ghosts. They do not read, click-through, take action, forward, or donate. They serve to distort our metrics, and there is good money in doing so. Let me brazenly label this 'Karpf's Rule of Online Data:' There is an inverse relationship between the reliability of an online metric of influence and its financial or political value. *Any metric of digital influence that becomes financially valuable, or is used to determine newsworthiness, will become increasingly unreliable over time*. When financial value or public attention is determined by an online metric, an incentive is created for two industries of code-writers: spammers/distorters, who falsely inflate the measure, and analytics professionals, who algorithmically separate out the spam/noise to provide a proprietary value-added. In the endless battle between these two industries, the public data that most social science researchers must rely upon will always suffer.

Incidentally, it is for this reason that I make use of blog comments as a measure of influence in the BAI. Bloggers have an interest in pruning spam comments from their sites, and there has never been a financial incentive attached with having high comment levels. Regardless of web 2.0 discussions of the value of 'user-generated content', there has never been much financial gain in artificially inflating comment levels. As such, specifically *because* it is an obscure measure of influence, comment levels provide a much clearer signal of blog community activity than the more popular, but more-noisy hyperlink and site traffic metrics. Blog comments are not worth much, so they have not been subjected to Karpf's Rule.

Note that this Rule of Online Data operates as a function of Internet Time. As a new online platform moves through the diffusion process and attracts wider adoption levels, it also becomes 'valuable online real estate' (Karpf 2011). At the very same time that scholars are likely deciding that a new element of the online environment is worthy of study, the core metrics we would make use of are attracting spammers and becoming less trustworthy.

The opposite of public is not private, its proprietary

In September 2010, I found myself in an online conversation with a digital campaign professional. She was curious about the state of academic research on digital

political advocacy – wondering, in effect, whether all their hard work was making any difference. As we discussed the state of the field, I speculated on some of the fascinating questions that were yet to be answered. One question was 'what are the differential effects of Facebook and Google ads?' Campaigns use both, and we know that Facebook and Google serve different functions for the end-user. In all the discussion of Facebook as quasi-online-public-sphere, we did not have an answer to such basic questions as 'which one receives more clicks?'

The campaign professional responded in surprise, 'The analytics are super good, actually. Google adwords make it very easy to track . . . People click on Facebook ads far more than on Google ads. At least for us'. The proprietary data provided to her electoral campaign made this common knowledge for insiders. The public data available to academics like myself rendered the question intractable.[9]

Likewise, practitioners will occasionally reveal major insights into the political blogosphere in a matter-of-fact manner. In a 2005 blog post titled 'I'm Not Going to Blogroll You', Chris Bowers of MyDD.com explained that 'as a serious student of blog traffic, I'm here to tell bloggers of all shapes and sizes that linking, especially blogrolling, is neither the main engine for building website traffic nor is it the main way for validating that what a website is doing is productive' (Bowers 2005). Drawing upon the analytics provided to MyDD, Bowers notes that 'blogrolls account for less than 2% of all blog traffic [to MyDD, at least]'. He goes on to offer detailed insights into where (MyDD's) web traffic comes from, and to offer actionable advice on how an upstart blogger can build a following. His data, of course, is limited to his own site. It might not be generalizable. But it is *good* data – much better than academic researchers generally have access to. The public data available through sitemeter, alexa, quantcast, comscore and hitwise provides no such detail. Bowers's eyeball-level analysis offers stronger findings than anything we could hope to produce through regression analysis and modeling based on public data, GIGO.

In the MCP, the same trend is present. Organizations routinely run 'A/B tests' to see which subject lines, issue topics, and action requests are most popular with their membership. MoveOn.org has records of every click that every member has ever taken. They are not particularly interested in sharing this data with academics, and for good reason. The marginal added value that sophisticated modeling techniques could offer would be offset by publicizing their best practices to opposing organizations[10] Likewise, some of the essential forms of network-based communication occur on 'network backchannel' Google-Groups (Karpf 2012). These network backchannels are off-the-record lists. Their communication is not public; their very existence is rarely acknowledged in the research literature.

Academics can occasionally partner with organizations, developing complicated legal agreements for the use and publication of proprietary data. These are

fruitful partnerships, and worthy of pursuit. But they are also rare, and slow. By the time our research is published, the underlying phenomenon may have changed in an important way.

My point is not that the online data environment is hopeless. It is simply to note that Internet Time poses just as many challenges for new research methods as it does for older ones. We must begin from a perspective of methodological skepticism. The glimmering promise of online data abundance too often proves to be fool's gold.

A partial solution: embracing transparency and kludginess

So there we have it. The online communications environment changes apace with Internet Time. New elements of the landscape are modified even as they diffuse through the population. The Internet of 2006 is different from the Internet of 2011 in important and occasionally unexpected ways. Internet Time unsettles some of the *ceteris paribus* assumptions built into our most robust research methods. In light of traditional publishing timelines, our classical methods of analysis can be relied upon to answer a narrowing set of research questions. Meanwhile, the abundant, public data that social scientists can get our collective hands on turns out not to be very good. There is no 'magic bullet' solution to conducting quality research today. And yet, to borrow another phrase from Clay Shirky, we have entered a time period where 'Nothing will work, but everything might' (Shirky 2009).

Shirky was referring to the news business when he made that comment. The old model for producing, funding, and distributing journalism is broken. No single model clearly replaces it. So, he argues, we have entered into a decade of rampant experimentation. The tried-and-true is suspect. The suspect-and-new is, well, suspect as well. But it also has potential. I would argue that we ought to approach digital research methods from a similar perspective. In particular, there are two characteristics that we should attempt to foster in the face of Internet Time: *transparency* and *kludginess*.

By transparency, I specifically mean that researchers should be up-front about the limitations of our data sets and research designs. This has always been a good habit, but it takes on additional importance in the context of Internet Time. Studies of campaign websites should make clear what role these sites play in the broader campaign context, so that future readers can evaluate (1) whether the role has changed and (2) if so, whether that change affects the implications of the research findings.[11] Studies of Twitter usage should highlight that the medium is still in its early adoption stage, and discuss what this implies about demographic and user experience changes as it gains broader adoption. In the places where we are making the best of a bad data situation, we should be

forthcoming about this. Doing so will help later scholars effectively periodize the findings, helping us avoid a useless back-and-forth in which new studies attempt to 'disprove' later studies, when in fact they are describing new changes to the underlying phenomena. Likewise, frank discussions of the limits of public data can help spur partnerships with proprietary data houses. We will only improve the awful data situation if we are honest with ourselves about it.

Again, I will offer the BAI as an example. The BAI is a *tremendously* limited tracking system. Its initial design was (literally) formulated on the back of a napkin. I convert four types of data into ordinal rankings, then merge those rankings while dropping the lowest rank. It artificially smoothes out some forms of system instability. It treats hyperlinks, comments, and site traffic as equal, without even attempting to explore how the three measures influence one another. To develop ordinal rankings for site traffic, I engage in a messy triangulation process, starting with sitemeter data where available,[12] then estimating other site ranks by eyeballing comparative quantcast or alexa scores.[13] I have never met a proper methodologist who did not cringe at the decision to drop the lowest rank.

Yet, despite its many flaws, the BAI is also used by several research institutions. The reason, I would like to think, is that these are *transparent* limitations. I drop the lowest rank because many political blogs do not provide one form of data or another. Instapundit – once the largest conservative blog – has no comment function. Glenn Greenwald's blog at Salon is very influential within the progressive community. But it is impossible to estimate his traffic based on public data, and Salon has no interest in sharing their internal traffic numbers. Bridge blogs like the *Huffington Post*, which are essentially news organizations at this point, receive systematically low blogroll (Network Centrality) scores, because some bloggers view them as non-blogs. The BAI is a system of reasonable compromises, based on the limitations of a complicated and changing, messy data environment. And all of these limitations are made public, so that the research community knows exactly what the BAI can and cannot do. If it eventually encourages another researcher to build something better, that would be a good thing. In the meantime, my little 'lobster trap' provides a useful resource where none would otherwise exist. The other reason why many researchers make use of the BAI is because, like the Wayback Machine, it turns out to be the only existing data source for many research efforts. In the face of a rapidly changing media environment, it is better to have transparently flawed data than none at all. Without these lobster traps, it becomes nearly impossible to demonstrate that underlying phenomena (like 'the blogosphere') have changed.

I would also suggest that transparency is an area where peer-reviewers should play an active role. Authors will always feel pressure to talk up the elegance of their methodology, burying limitations in footnotes or brief comments. If we want our interdisciplinary fields to take transparency seriously, we will need peer reviewers to call for it, and reward it, in their recommendations.

'*Kludginess*' is a term borrowed from hacker culture. Wikipedia (appropriately) provides a definition: 'A kludge (or kluge) is a workaround, a quick-and-dirty solution, a clumsy or inelegant, yet effective, solution to a problem, typically using parts that are cobbled together'.[14] The essence of a kludge is that it is inelegant, but usefully solves a problem. In the face of Internet Time, kludgy design choices become particularly attractive.

I will offer the MCP as an example of kludgy design. One of the defining challenges within interest group scholarship lies in population estimation. Identifying the total population of organizations seeking to influence politicians, even in the pre-Internet era, proves to be almost impossible. And without a clear population, of course, we cannot be sure that our findings about partisanship, activity, bias, etc within the interest group system are not exhibiting selection bias. As a result, tremendous attention has been paid to crafting elegant solutions to the population definition problem (Walker 1991; Baumgertner *et al.* 2009) Those solutions do not account for the upwelling of Internet-mediated (quasi-) organizations, however. Nothing, at present, does. And while population definition has always been problematic, interest group formation in Internet Time renders sophisticated counting strategies outdated. If it takes 2 years to estimate the online advocacy population, it is entirely possible that new technologies and organizational forms will have rendered that estimation obsolete in the meantime.

The MCP conveniently sidesteps this problem. Rather than attempting to develop an elegant solution to population definition in Internet Time, it instead focuses on a smaller, intrinsically interesting cluster. I used the organizations listed in *The Practical Progressive*, a book by Democracy Alliance co-founder Erica Payne (2008). The Democracy Alliance is an influential community of wealthy progressive donors. The organizations they fund form a network, and it is a network well worthy of study. So I used that practitioner list to set my convenience sample, transparently explained the limitations, and set up a 'lobster trap' for studying advocacy group e-mail usage. The results clearly demonstrated a major empirical claim: that newer progressive advocacy groups use e-mail in very different ways than their legacy counterparts (Karpf 2010b, 2012). They also have proved to be the first academic study of advocacy group e-mails – and I have had several researchers approach me and note that they started work on the problem, but gave up because of the population definition hurdle.

Kludgy solutions tend to be simple, reasonable workarounds. Elegant solutions tend to have longer timelines and grander aspirations. In the long term, the social sciences should rightly be driven towards elegant solutions. But in the short term, facing the ongoing disruption of Internet Time, kludginess holds a lot of promise. It should be embraced. Many of the most interesting phenomena are in flux. We can study them, but only if we begin to embrace 'hacks and workarounds'. This goes beyond the textbook advice to remain open to diverse methods and techniques. In the face of Internet Time, untried, messy solutions are often *preferable* to robust, elegant ones.

In particular, it appears to me that there is outstanding potential for welding together computational social science with qualitative case analysis. The problem for computational social science is that the data is messy, the relevant metadata often changes as sociotechnical systems evolve, and we thus have trouble setting categories and definitions. We can track massive data flows, but that only becomes meaningful if we know what the underlying phenomena signify. Meanwhile, qualitative case analysis offers the benefits of analytic depth. A good critical case analysis (see Chadwick 2011 for one nice example, dealing with Twitter) can outline relevant categories (metadata) and testable hypotheses for massive computational analysis.

Case studies and computational research are odd, kludgy neighbors. They also hold great potential for addressing one another's shortcomings. This is not the sole research pairing that ought to be attempted ('nothing will work, but everything might'), it is just the one that strikes me as particularly promising for those questions that I choose to pursue. Combined with ample 'lobster traps', the vagaries of Internet Time will at least become easier to identify through scholarly research endeavors.

Transparency and kludginess must be paired together. Kludges fail in computer science when their weaknesses are not transparent enough. They then can be applied in the wrong setting, and they simply break as a result. Likewise, elegant methodological solutions are usually complex, requiring years of specialized training. Under those circumstances, transparency is necessarily obscured. These two values have long histories in the social sciences – they are not novel within themselves. Heightened attention to these values becomes more important in the face of Internet Time. If we are to understand the changing digital landscape, we will need to get our hands dirty and then take note of the dirt under our fingernails, so to speak.

Conclusion

The field of Internet research is cross-disciplinary and interdisciplinary. Researchers come from different academic lineages, and approach puzzles with a multitude of assumptions. The goal of the article has been not to provide a definitive solution to a single research problem, but rather to serve as a conversation-starter regarding a shared challenge: the problems introduced into our research by Internet Time. The first problem is that underlying concepts, such as 'The Blogosphere' are not fixed in place. They change even while we study them. The second problem is the impact that these changes can have on the *ceteris paribus* assumptions of our most robust research methods. The third problem is the set of limitations posed by publicly-available online data. Novel research techniques that make use of 'Big Data' are often hobbled by the flaws embedded in that data. The essay has been reflective in nature, expounding upon a few research instincts that I have

developed in the course of my work. Those instincts are now embodied in a pair of research artifacts (the BAI and the MCP). My goal here has been to offer up these instincts for (transparent) review, to provoke a conversation within the Internet research community.

Internet research is messy-but-promising. That is a claim that I believe all participants in the Oxford Internet Institute Symposium and readers of this *ICS* special issue can rally behind. The question that faces us as an interdisciplinary research community is 'what should we do with this promising mess?' My preferred solution is focused on selecting appealing questions, hacking together designs for systematic evaluation, and being transparent about our assumptions and limitations so we can observe as the medium continues to change and surprise. 'Lobster trap'-style data collection tools are inelegant, but they can also prove tremendously useful. As Internet Time marches on, public data quality degrades further, and the next wave of Internet innovations takes us by surprise, hacks and workarounds can help us gain intellectual purchase on an expansive range of socially important and intellectually engaging research questions. No single, overarching solution will answer the challenges posed by Internet Time. But promoting the values of transparency and kludginess – embracing the messiness of the data, rather than leaning on the reliability of older techniques – can help guide the way.

Notes

1   Particularly with the 2001 closing of SixDegrees.org – the initial social networking site – demonstrating that being a first-mover is not *always* such an advantage.

2   My favorite example of this phenomenon is technology writer Gleick's (1995) *New York Times Magazine* essay, 'this is sex?' Therein, he argues that the Internet will never become a popular medium for pornography, because the content is too grainy and slow. . . . Suffice it to say, as bandwidth, storage capacity, and processor speed all grew, the online market for prurient information developed some new dynamics.

3   I am borrowing terminology and concepts from the diffusion-of-innovation literature here. See Rogers (2003) and Von Hippel (2005) for further discussion.

4   Trippi (2005) refers to this development as 'snow plowing'. He argues that before the Dean campaign could set online fundraising records, they needed commercial giants like Amazon.com and Ebay to effectively acclimate citizens-as-consumers to online purchasing habits.

5   As anecdotal evidence, in the course of preparing this article for submission, I have had conversations with two doctoral students who began with promising research questions, but had to abandon or

deeply modify their projects when the Wayback Machine turned out to lack the relevant archived pages.

6   Except Hitwise. And Sitemeter is an opt-in decision for individual blogs. Alexa, Comscore, and Quantcast only provide traffic measures for larger sites, and each has systematic flaws. The data landscape is really quite a headache.

7   The sole option is to contact their webmaster, hope they chose to save this data themselves, and then ask very, very nicely.

8   A parallel challenge faces large-scale text classification. Most content-scrapers are built around RSS feeds. Provided with realtime data, they can offer sophisticated analysis. Provided with archived data, they frequently get bogged down or require time-consuming cleaning. Enterprising computational social scientists would be well-advised to set up several 'lobster traps'.

9   What is more, it would take months to gather this data, and likely multiple years to publish. At which point the finding would likely be dated because of some new development in the two platforms.

10   To be clear, I do not mean 'opposing' in the Exchange Theory sense (Salisbury 1969), where peer groups are viewed as competitors. MoveOn regularly shares tips and best practices with ideological peers, and even occasionally raises money for them. They do not, however, share with conservatives.

11   To their credit, Bimber and Davis (2003) do provide such a transparent discussion. I am calling for an increased emphasis on this practice, rather than suggesting previous scholars have failed in it entirely.

12   Sitemeter is an opt-in system that measures unique visitors per day.

13   Quantcast measures unique visitors per month. It operates at the domain name level, meaning political blogs hosted by large online news sites (Slate, NYTimes, Salon) simply cannot be estimated. Alexa and quantcast work well for large websites, but poorly for mid-tier and smaller blogs.

14   http://en.wikipedia.org/wiki/Kludgy

References

Barzilai-Nahon, K., Helmsley, J., Walker, S. & Hussain, M. (2011) 'Fifteen minutes of fame: the place of blogs in the life cycle of viral political information', *Policy & Internet*, vol. 3, no. 1, [Online] Available at: http://www.psocommons. org/policyandinternet/vol3/iss1/art2/ (29 August 2011).

Baumgertner, F., Berry, J., Hojnacki, M., Kimball, D. & Leech, B. (2009) *Lobbying and Policy Change*, University of Chicago Press, Chicago, IL.

Benkler, Y. & Shaw, A. (2010) *A Tale of Two Blogopheres: Discursive Practices on the Left and Right*, Berkman Center for Internet & Society, [Online] Available at:

http://cyber.law.harvard.edu/publications/2010/Tale_Two_Blogospheres_ Discursive_Practices_Left_Right (29 August 2011).

Bimber, B. (2003) *Information and American Democracy*, Cambridge University Press, Cambridge.

Bimber, B. & Davis, R. (2003) *Campaigning Online*, Oxford University Press, Oxford.

Bowers, C. (2005) 'I'm not going to blogroll you', Blog Post, MyDD.com, [Online] Available at: http://mydd.com/2005/9/22/im-not-going-to-blogroll-you (29 August 2011).

boyd, d. & Ellison, N. (2007) 'Social network sites: definition, history, and scholar-ship', *Journal of Computer-Mediated Communication*, vol. 13, no. 1, article 11, [Online] Available at: http://jcmc.indiana.edu/vol13/issue1/boyd.ellison. html (29 August 2011).

Chadwick, A. (2011) 'The political information cycle in a hybrid news system: British Prime Minister and the "Bullygate" affair', *International Journal of Press/Politics*, vol. 16, no. 1, pp. 3–29.

Cook, J. (2011) 'Most of Newt Gingrich's twitter followers are fake', *Gawker.com*, 1 August, [Online] Available at: http://gawker.com/5826645/most-of-newt-gingrichs-twitter-followers-are-fake (29 August 2011).

Cornfield, M. (2004) *Politics Moves Online: Campaigning and the Internet*, Century Foundation Press, Washington, DC.

Davis, R. (2009) *Typing Politics*, Oxford University Press, New York.

Earl, J. & Kimport, K. (2011) *Digitally Enabled Social Change*, MIT Press, Cambridge, MA.

Foot, K. & Schneider, S. (2006) *Web Campaigning*, MIT Press, Cambridge, MA.

Gillespie, T. (2009) *Wired Shut: Copyright and the Shape of Digital Culture*, MIT University Press, Cambridge, MA.

Gleick, J. (1995) '*This is Sex*?' Republished in Gleick, J. (2002) *What Just Happened?*, Pantheon, New York.

Goldsmith, J. & Wu, T. (2006) *Who Controls the Internet?: Illusions of a Borderless World*, Oxford University Press, New York.

Hargittai, E. (2010) 'Digital na(t)ives? Variation in internet skills and uses among members of the net generation', *Sociological Inquiry*, vol. 80, no. 1, pp. 92–113.

Hindman, M. (2008) *The Myth of Digital Democracy*, Princeton University Press, Princeton, NJ.

Howard, P. N. (2006) *New Media and the Managed Citizen*, Cambridge University Press, New York.

John, R. (2010) *Network Nation: Inventing American Telecommunications*, Belknap Press, Cambridge, MA.

Johns, A. (2010) *Piracy: The Intellectual Property Wars from Gutenberg to Gates*, University of Chicago Press, Chicago, IL.

Karpf, D. (2008a) 'Measuring influence in the political blogosphere', *Politics and Technology Review*, George Washington University's Institute for Politics, Democracy, & the Internet, March, pp. 33–41, [Online] Available at:

http://www.asiaing.com/politics-and-technology-review-march-2008.html (9 December 2008).

Karpf, D. (2008b) 'Understanding blogspace', *Journal of Information Technology and Politics*, vol. 5, no. 4, pp. 369–385.

Karpf, D. (2010a) 'Macaca moments reconsidered: electoral panopticon or netroots mobilization?', *Journal of Information Technology and Politics*, vol. 7, no. 2, pp. 143–162.

Karpf, D. (2010b) 'Online political mobilization from the advocacy group's perspective: looking beyond clicktivism', *Policy & Internet*, vol. 2, no. 4, [Online] Available at: http://www.psocommons.org/policyandinternet/vol2/iss4/art2/ (29 August 2011).

Karpf, D. (2011) 'Open source political community development: a five-stage adoption process', *Journal of Information Technology and Politics*, vol. 8, no. 2/3, pp. 323–345.

Karpf, D. (2012) *The MoveOn Effect: The Unexpected Transformation of American Political Advocacy*, Oxford University Press, New York.

Kelty, C. (2008) *Two Bits: The Cultural Significance of Free Software*, Duke University Press, Durham, NC.

Kerbel, M. (2009) *Netroots: Online Progressives and the Transformation of American Politics*, Paradigm Publishers, Boulder, CO.

Kolari, P., Java, A. & Finin, T. (2006) 'Characterizing the Splogosphere', WWW 2006, 22–26 May, Edinburgh.

Kreiss, D. (Forthcoming) *Taking Our Country Back: The Crafting of Networked Politics from Howard Dean to Barack Obama*, Oxford University Press, New York.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. & Van Alstyne, M. (2009) 'Computational social science', *Science*, vol. 323, no. 5915, pp. 721–723.

Lentz, J. (2002) *Electing Jesse Ventura: A Third-Party Success Story*, Lynne Rienner, Boulder, CO.

Lessig, L. (1999) *Code and Other Laws of Cyberspace*, Basic Books, New York.

Margolis, M. & Resnick, D. (2000) *Politics as Usual: The Cyberspace 'Revolution'*, Sage Press, New York.

Mayer-Schonberger, V. (2011) *Delete: The Virtue of Forgetting in the Digital Age*, Princeton University Press, Princeton, NJ.

Moore, G. (1965) 'Cramming more components onto integrated circuits', *Electronics*, vol. 38, no. 8, pp. 114–117.

Mueller, M. (2004) *Ruling the Root: Internet Governance and the Taming of Cyberspace*, MIT University Press, Cambridge, MA.

Mueller, M. (2010) *Networks and States: The Global Politics of Internet Governance*, MIT University Press, Cambridge, MA.

Payne, E. (2008) *The Practical Progressive: How to Build a Twenty-first Century Political Movement*, Public Affairs, New York.

Perlmutter, D. (2008) *Blog Wars*, Oxford University Press, New York.

Reagle, J. (2010) *Good Faith Collaboration: The Culture of Wikipedia*, MIT Press, Cambridge, MA.

Rogers, E. (2003) *Diffusion of Innovations*, 5th edn, Free Press, New York.

Salisbury, R. (1969) 'An exchange theory of interest groups', *Midwest Journal of Political Science*, vol. 13, no. 1, pp. 1–32.

Schlozman, K. L., Verba, S. & Brady, H. (2010) 'Weapon of the strong? Participatory inequality and the internet', *Perspectives on Politics*, vol. 8, no. 2, pp. 487–509.

Shaw, A. (2011) 'Crowdsourced content analysis', paper presentation, *International Communication Association Annual Meeting*, Boston, MA, May.

Shirky, C. (2003) 'POWER laws, weblogs, and inequality' [Online] Available at: http://www.shirky.com/writings/powerlaw_weblog.html (29 August 2011).

Shirky, C. (2009) 'Newspapers and Thinking the Unthinkable', [Online] Available at: http://www.shirky.com/weblog/2009/03/newspapers-and-thinking-the-unthinkable/ (29 August 2011).

Smith, R. (1997) 'Still blowing in the wind: the American quest for a democratic, scientific political science', *Daedalus*, vol. 126, no. 1, (Winter, 1997), pp. 253–287.

Solove, D. (2008) *The Future of Reputation: Gossip, Rumor, and Privacy on the Internet*, Yale Univesity Press, New Haven, CT.

Sunstein, C. (2001) *Republic.com*, Princeton University Press, Princeton, NJ.

Sunstein, C. (2006) *Infotopia: How Many Minds Produce Knowledge*, Oxford University Press, New York.

Sunstein, C. (2007) *Republic.com 2.0*, Princeton University Press, Princeton, NJ.

Trippi, J. (2005) *The Revolution Will Not Be Televised: Democracy, the Internet, and the Overthrow of Everything*, Harper Collins, New York, NY.

Von Hippel, E. (2005) *Democratizing Innovation*, MIT Press, Cambridge, MA.

Walker, J. L. (1991) *Mobilizing Interest Groups in America: Patrons, Professions, and Social Movements*, University of Michigan Press, Ann Arbor, MI.

Wallsten, K. (2010) 'Yes we can: How online viewership, blog discussion, campaign statements, and mainstream media coverage produced a viral video phenomenon', *Journal of Information Technology and Politics*, vol. 7, nos 2–3, pp. 163–181.

Wellman, B. (2001) 'Computer networks as social networks', *Science*, vol. 293, no. 5537, pp. 2031–2034.

Wu, T. (2010) *The Master Switch: The Rise and Fall of Information Empires*, Knopf, New York.

Zittrain, J. (2008) *The Future of the Internet (And How To Stop It)*, Yale University Press, New Haven, CT.

Zittrain, J. (2010) 'Minds for Sale', Public Presentation, Berkman Center for Internet and Society, 22 February, 2010 [Online] Available at: http://www.youtube.com/watch?v=cnEL4aAAjgo (29 August 2011).

Zysman, J. & Newman, A. (2007) *How Revolutionary Was the Digital Revolution?* Stanford University Press, Stanford, CA.

David Karpf is an Assistant Professor in the School of Communication and Information at Rutgers University, as well as a Faculty Associate at the Eagleton Institute of Politics. His research on the Internet and American political associations has appeared in several journals, and can be found at www.davidkarpf.com. His first book, *The MoveOn Effect: The Unexpected Transformation of American Political Advocacy*, was published by Oxford University Press in May 2012. *Address*: Journalism and Media Studies, Rutgers University, 4 Huntington St, New Brunswick, NJ 08901, USA. [email: davekarpf@gmail.com]