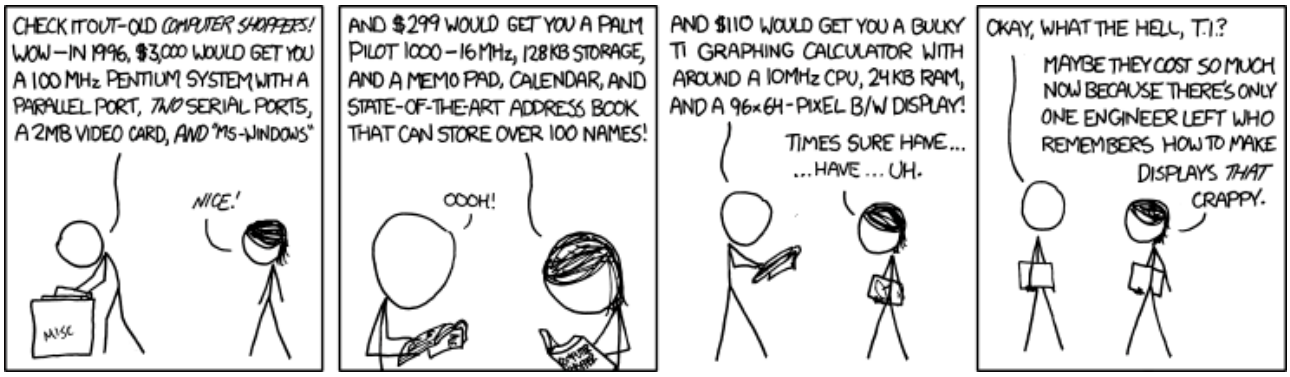
Social Science Research Methods In Internet Time

Dave Karpf
Assistant Professor
Rutgers University School of Communication and Information
davekarpf@gmail.com

[Working Paper, an obvious work-in-progress]



Abstract

This paper discusses three interrelated challenges related to conducting social science research in "Internet Time." The rate at which the Internet is both diffusing through society and also developing new capacities is unprecedented. It creates some novel challenges for scholarly research. 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, 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. In 2001, streaming video was rare, short and choppy. Wireless "hotspots" were a novelty. Mobile phones were used for phone calls and text messages. Commercial GPS applications were in the early stages of development. Bloggers could be counted by the handful, and social networking sites like Friendster, Myspace and Facebook were still confined to Bay Area networks and technologists' imaginations.[1] Simply put, the Internet of 2011 is *different* than the Internet of 2001. What's more, there is no reason to suspect this rapid evolution is finished. The Internet of 2021 will likely be different than the Internet of 2011.

[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.

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."[2] 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-to-24 months. 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, with limited application to most layers of the Internet's "hourglass architecture."[3]

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 the Moore's Law or the continuing evolution of the medium.[4] "Internet time" is a subject to be acknowledged, rather than incorporated into our research designs. Yet this creates a substantial hurdle. The Internet is unique among Information and Communications Technologies (ICTs) *specifically because* the Internet of 2001 has important differences from the Internet of 2005, or 2009, or 2011. It is a suite of overlapping, interrelated technologies. The medium is simultaneously undergoing a diffusion process and a series of transformations. Diffusion brings in new actors with diverse interests. Transformation alters the technological affordances of the media environment itself. What was costly and difficult in 2004 is cheap and ubiquitous in 2008. That leads, in turn, to different practices.[5] The Internet's effect on media and political institutions will be different at time X than at time X+1, because the suite of technologies we think of as the Internet will itself change within that interval.

The dominant methodologies 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, peer review, and publication, it is not uncommon for a research project to consume five or more years between conception and publication. Several books in articles have been published in 2011 on the basis of 2006 data. Scholars openly acknowledge the limitations imposed by the system. They commiserate over it at the conference hotel bar. In the last two stages of the process – peer-review and publication – there has been ample innovation, with online journals, automated components of peer-review, and academic blogging being three obvious examples. But when it comes to the "tools of the trade," collective introspection among

---

[2] Wellman 2001

[3] Zittrain 2008, chapter 4

[4] See Hindman 2008, Bimber 2003, Bimber and Davis 2003, Davis 2009, Howard 2006, Sunstein 2001, 2006, 2007, Kerbel 2009.

[5] My favorite example of this phenomenon is technology writer James 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.

social scientists has been far less forthcoming. Everyone can agree that it would be nice to see peer-review and publication turnaround improved. Exploring the limitations of panel and survey data conducted in the early lifecycle of a changing media environment is a steeper climb.

This paper is an effort to 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. To be clear, I do not argue that traditional methods have been rendered useless by the Internet. Rather, I argue that an expanding range of interesting and important questions cannot be answered using the best tools in our toolbox. The new media environment demands new tools and techniques. Those techniques carry risks – they haven't 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.

The argument proceeds in the form of four brief essays. I will 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 amorphous – today there is no such thing as "the blogosphere." The rate at which these sociotechnical systems evolve, I will argue, 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 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.

Internet Time and Sociotechnical Systems, or, There Is No Such Thing as the Blogosphere

My own perspective on "Internet Time" has been driven by early interest and ongoing observation of the political blogosphere. Writing in 2003, Clay Shirky offered the following prescient 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.[6]" 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*.

As is so often the case, Shirky called it.

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.

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 than the early- and late-

---

[6] Clay Shirky. 2003. "Power Laws, Weblogs, and Inequality." http://www.shirky.com/writings/powerlaw_weblog.html.

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. 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 perncious. 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. 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.[7] 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, or (more problematically) Huffingtonpost.com as part of a single, overarching "blogosphere" is clearly problematic. 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. 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.

---

[7] Karpf 2008, 2010a, forthcoming

From a methodological perspective, it is of little importance that there is no longer any such thing as 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."[8] Also as the technology diffuses, it changes. Blog software today includes key modifications and variations on the basic 1999 Pyra Labs platform. That provides variant technological affordances. 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 were not designed with such a landscape in mind. As such, a few common problems routinely crop up.

Internet Time and Social Science Research

Consider the following: 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.[9] 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* than the Internet of 1996, 2000, or 2004, and this is a recurrent, ongoing pattern.

Consider the following puzzle: "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 3rd most popular destination on the internet, as recorded by Alexa.com. Political campaigns now develop special "web advertisements" with no intention of buying

[8] Margolis and Resnick 2000

[9] See Foot and Schneider 2006, Lentz 2002, Trippi 2005, Kreiss Forthcoming, Bimber and Davis 2003, Cornfield 2004.

airtime on television, simply placing the ads on YouTube in the hopes of attracting commentary from the blogosphere and resultant media coverage.[10] The medium is viewed as so influential that an entire political science conference in 2009 was 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 (if it weren't so easy to explain) should be utterly baffling. 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 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, for instance, received multiple awards 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.

Bimber and Davis's finding about candidate websites remains 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) aren't seeking out such political information. But the implications of their findings are curtailed by further development in the online landscape.

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 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 forseeing 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[11]). Those features leveraged different social practices. A network

---

[10] See Wallsten 2010 and Barzilai-Nahon 2011 for empirical studies of viral video diffusion patterns.

[11] 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

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 with 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 may have 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. As the media environment itself changes, behavior is likely to change in unexpected ways as well. These samples are best conducted as a time series, such as those provided by Pew Internet, because they can then provide a snapshot of changing engagement levels. But Pew is an exceptional example – a large, well-funded research shop that continually modifies its poll questions. 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.

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 realtime – hold great promise. But they also face the lurking threat of GIGO (Garbage In, Garbage Out). Publicly available online data often has deep, systematic flaws which we must not ignore.

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

---

Amazon.com and Ebay to effectively acclimate citizens-as-consumers to online purchasing habits.

difficult-at-scale.[12] 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 spambots and (3) the limitations of public data. These three problems do not affect all research methods equally. But they too often go unacknowledged in the research community. As a result, we rarely see any frank discussion of what can and cannot be effectively measured online.

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. Want to conduct a study of how Slate.com or Democrats.org have evolved over time? The Wayback Machine is the only available source. It is exceptional in the narrow sense, though: as an exception to a general rule. And of course, researchers are limited to whatever the Wayback Machine chooses to capture. Some popular sites will be thoroughly canvassed. Others will be rarely-if-ever catalogued. If one is interested in the development of a website that didn't happen to receive frequent Internet Archive attention, the research will encounter an abrupt end.

We can think 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 US scholars). This data is publicly available[13], 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.[14] Online fundraising, e-mail, web traffic, online video, and twitter traffic are all

---

[12] Aaron Shaw presented about this at ICA 2011. It also appears in Jonathan Zittrain's public talks about his forthcoming book.

[13] 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.

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

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 Blogosphere Authority Index (BAI) and the Membership Communications Project (MCP) – follow this principle. The BAI collects ephemral 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 well-studied 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.[15]

Privacy advocates are fond of claiming that, once something is published online, it is available forever. That 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.

Garbage In, Garbage Out: Spambots and Data Noise

It's 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's 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% of all blogs are "splogs," or spam-blogs, set up as phantom sites to artificially boost hyperlink levels.[16] Similarly, Web traffic measures can be gamed by computer programs. On Twitter, former House Speaker Newt Gingrich was recently discovered to have

---

[15] 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."

[16] See http://uploadi.www.ris.org/editor/1135776405umbria_splog.pdf, or Kolari, Java, and Finin 2006.

purchased 80% of his 1.3 million Twitter followers.[17] These were fake accounts, but they served the 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, clickthrough, take action, forward, or donate. They serve to distort our metrics, and there is good money in doing so. Let me brazenly call this "Karpf's Law of Online Data:"[18] 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 are 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. Despite 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.

Note that Karpf's Law 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."[19] 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 untrustworthy.

The Opposite of Public Is Not Private, It's 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

---

[17] http://gawker.com/5826645/most-of-newt-gingrichs-twitter-followers-are-fake

[18] As far as I can tell, no one else has called dibs on this one yet. So… dibs!

[19] See Karpf 2011 for a discussion of the institutional development challenge posed by this valuable online real estate problem.

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 didn't 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.[20]

Likewise, practioners 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."[21] 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. 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. Garbage In, Garbage Out.

In the Membership Communications Project, 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[22] Likewise, some of the essential forms of network-based communication occur on backchannel GoogleGroups (what I call "Network Backchannels" in my book). These GoogleGroups are off-the-record lists. Their communication is not public, and their very existence is rarely acknowledged in the research literature!

---

[20] What's 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.

[21] Bowers 2005 http://mydd.com/2005/9/22/im-not-going-to-blogroll-you

[22] To be clear, I do not mean "opposing" in the Exchange Theory sense, 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.

Academics can occasionally partner with organizations, developing complicated legal agreements for the use and publication of proprietary data. These are fruitful partnerships, and should be pursued. But they are also rare. And they happen slowly. By the time our research is published, the state of the quick-moving field may have changed in an important way.

My point is not that the online data environment is hopeless. There is some tremendous, cutting edge research being done, and exciting opportunities for tracking and visualizing new forms of *activated public opinion*.[23] But we should always begin from a perspective of methodological skepticism. The most robust, traditional research methods can be poorly matched to the rapid modifications of Internet Time. Meanwhile, the glimmering promise of online data abundance too often proves to be fool's gold.

A Modest Suggestion: 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. Internet time unsettles some of the ceteris paribus assumptions built into our most robust research methods. Particularly when compared to the long, multistage process associated with academic publishing, our classical methods of analysis can be used 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 a phrase from Clay Shirky, we have entered a time period where "Nothing will work, but everything might."[24]

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 should 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

---

[23] Again, this is a term from my forthcoming book (chapter 7). Apologies to readers, this is what happens when you jump right into a new research endeavor immediately after you completed the last one.

[24] Shirky 2009: http://www.shirky.com/weblog/2009/03/newspapers-and-thinking-the-unthinkable/

that later readers can evaluate (1) whether the role has changed and (2) if so, whether that change affects the implications of the research findings.[25] 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 up front about it.

Again, I will offer the BAI as a model example. The BAI is a *tremendously* limited tracking system. It was (literally) initially created 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 makes for an artificially stable system. 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,[26] then estimating other site ranks by eyeballing comparative quantcast or alexa scores.[27] 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 low blogroll (Network Centrality) scores, because some bloggers view them as non-blogs. In other words, 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

---

[25] 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.

[26] Sitemeter is an opt-in system that measures unique visitors per day.

[27] 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.

the meantime, my little "lobster trap" provides a useful resource where none would otherwise exist.

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."[28] 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'll 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.[29] Those solutions don't account for the upwelling of internet-mediated (quasi-) organizations, however. Nothing, at present, does.

The MCP sidesteps this problem entirely. Rather than attempting to develop an elegant solution to population definition in Internet Time (which would likely be outdated by the time it was printed), it instead focuses on a smaller, interesting cluster. I used the organizations listed in *The Practical Progressive*, a book by Democracy Alliance co-founder Erica Payne. The Democracy Alliance is an influential community of wealthy progressive donors. The organizations they fund form a network, and it is a network well worth studying. 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.[30] 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.

---

[28] http://en.wikipedia.org/wiki/Kludgy

[29] See Baumgartner et al 2009 for a recent, elegant example. See Walker 1991 for the classic explanation of the underlying problem.

[30] Karpf 2010b, Karpf forthcoming.

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 righly 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.

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 will 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.

Transparency and kludginess must be paired together. Kludges fail in computer science when their weakness 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.

Conclusion

This paper has been a winding methodological essay of sorts. I have developed a few instincts about conducting research in the face of Internet Time, and those instincts have been embodied in particular technological 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 research community.

Internet research is messy-but-promising. That is a claim that I believe all participants in this Oxford Internet Institute Symposium 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. I look forward to discussing and debating alternate formulations to the problem.

Works Cited

Barzilai-Nahon, Karine, Jeff Helmsley, Shawn Walker, and Muzammil Hussain. 2011. "Fifteen Minutes of Fame: The Place of Blogs in the Life Cycle of Viral Political Information," *Policy & Internet*. Volume 3, No. 1.

Baumgertner, Frank, Jeffrey Berry, Marie Hojnacki, David Kimball, and Beth Leech. 2009. *Lobbying and Policy Change*. Chicago, IL: University of Chicago Press.

Bimber, Bruce. 2003. *Information and American Democracy*. Cambridge, UK: Cambridge University Press.

Bimber, Bruce and Richard Davis. 2003. *Campaigning Online*. Oxford: Oxford University Press.

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

Davis, Richard. 2009. *Typing Politics*. New York: Oxford University Press.

Foot, Kirsten and Steven Schneider. 2006. *Web Campaigning*. Cambridge, MA:MIT Press.

Gleick, James. 1995. "This is Sex?" Republished in Gleick, James 2002. *What Just Happened?* New York: Pantheon.

Hindman, Matthew. 2008. *The Myth of Digital Democracy*. Princeton, NJ: Princeton University Press.

Howard, Philip N. 2006. *New Media and the Managed Citizen*. New York: Cambridge University Press.

Karpf, David. 2008. "Understanding Blogspace." *Journal of Information Technology and Politics*. 5(4). Pp 369-385

Karpf, David. 2010a. "Macaca Moments Reconsidered: Electoral Panopticon or Netroots Mobilization?" *Journal of Information Technology and Politics*. 7(2), 143-162.

Karpf, David. 2010b. "Online Political Mobilization from the Advocacy Group's Perspective: Looking Beyond Clicktivism." *Policy & Internet*. Volume 2, Issue 4. Accessed online: http://www.psocommons.org/policyandinternet/vol2/iss4/art2/

Karpf, David. 2011. "Open Source Political Community Development: A Five-Stage Adoption Process." *Journal of Information Technology and Politics*. Volume 8, Issue 2/3

Karpf, David. Forthcoming. *The MoveOn Effect: The Unexpected Transformation of American Politcan Advocacy*. New York: Oxford University Press

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

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

Lentz, Jacob. *Electing Jesse Ventura: A Third-Party Success Story.* Boulder, CO: Lynne Rienner.

Margolis, Michael and David Resnick. 2000. *Politics as Usual: The Cyberspace 'Revolution*.' New York: Sage Press.

Shirky, Clay. 2003. "Power Laws, Weblogs, and Inequality." http://www.shirky.com/writings/powerlaw_weblog.html.

Shirky, Clay. 2009:"Newspapers and Thinking the Unthinkable." http://www.shirky.com/weblog/2009/03/newspapers-and-thinking-the-unthinkable/

Sunstein, Cass. 2001. *Republic.com*. Princeton, NJ: Princeton University Press

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

Sunstein, Cass. 2007. *Republic.com 2.0*. Princeton, NJ: Princeton University Press.

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

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

Wallsten, Kevin. 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*. 7(2-3), 163-181.

Wellman, Barry. 2001. "Computer Networks as Social Networks." *Science*. Volume 293, No 5537. 2031-2034.

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

Zysman, John and Abraham Newman. 2007. *How Revolutionary Was the Digital Revolution?* Stanford, CA:Stanford University Press.