On the cutting edge of Big Data: Digital politics research in the social computing literature

Deen Freelon

Forthcoming in the *Handbook of Digital Politics,* eds. Stephen Coleman and Deen Freelon

03/17/14

Most of this volume's chapters review studies rooted in political science, communication, and closely related disciplines. Indeed, many reference a small clique of foundational authors in agreement and/or disagreement, including Castells, Benkler, Hindman, Jenkins, Morozov, and Shirky. In the current chapter I diverge from this norm to examine a body of literature only rarely acknowledged by mainstream digital politics scholarship. This literature contains politically-relevant research by computer scientists and information scientists and is published under a variety of disciplinary labels, but will be referred to here as *social computing research.* As its name implies, social computing research's purview includes any aspect of human behavior involving both digital technology and more than one person (Parameswaran & Whinston, 2007; Wang, Carley, Zeng, & Mao, 2007). Politics accounts for a small but thriving subset of this literature, which also encompasses health, business, economics, entertainment, artificial intelligence, and disaster response, among other topics.

Social computing research on politics holds relevance for scholars of digital politics and political communication for two related reasons, the first methodological and the second theoretical. Social computing researchers have for many years led the vanguard in computational and "Big Data" methods (sometimes in combination with other methods), in which the disciplines of political science and communication have both expressed great interest of late.[1]

Reviewing how social computing researchers have applied such methods to politically-relevant datasets will help readers unfamiliar with their work to consider how the methods could be applied to their own research. The field’s methods and findings also hold a number of theoretical implications, but its researchers devote only sporadic attention to such concerns. For the benefit of those with a more theoretical scholarly orientation, and perhaps also for some social computing researchers with social science leanings, I explore major theoretical trends in the literature. I conclude with suggestions for future research, focusing on how digital politics researchers can best adapt the insights of social computing research to their own ends.

Before proceeding to these sections, it is necessary to more thoroughly describe social computing and its goals, which differ in key ways from those of the social science mainstream. The following section is devoted to this task.

Social computing: a brief introduction

A caveat before I begin: this section is written from the perspective of one who was trained in and still operates within a social-science-based research orientation that emphasizes abstract theory as a guide and justification for empirical work (Fink & Gantz, 1996). My participation in social computing research up to this point in my career has been minimal. Accordingly, the description of social computing I offer here is intended as an introduction for those of a similar scholarly orientation to me, which I imagine will include many if not most of this volume’s audience.

In a widely-cited overview article, Wang et al. (2007) define social computing as “computational facilitation of social studies and human social dynamics as well as the design and use of ICT technologies that consider social context” (p. 79). A similar characterization by

Parameswaran and Whinston (2007) establishes social computing as a highly ubiquitous activity of study:

> Social computing shifts computing to the edges of the network, and empower (sic) individual users with relatively low technological sophistication in using the Web to manifest their creativity, engage in social interaction, contribute their expertise, share content, collectively build new tools, disseminate information and propaganda, and assimilate collective bargaining power. (p. 763)

Both of the above quotes emphasize the two essential elements of social computing: digital tools ("computing" broadly construed) and social interaction. Of course, researchers in communication, sociology, anthropology, and other social-scientific disciplines have explored topics such as "computer-mediated communication" and "cyberculture" for decades. This similarity in subject matter invites the question of how social computing research differs from approaches with which we are more familiar.

The main difference between social computing research and research traditions grounded in social science is as paradigmatic as that between social science and critical theory (Fink & Gantz, 1996; Potter, Cooper, & Dupagne, 1993). Whereas social science's goals are to explain empirical outcomes while promulgating theory, and critical theory's is to foment social change, social computing research is devoted to the development of new techniques for organizing, analyzing, and improving the user experience of social computing software and its output. As such, social computing studies are usually published in highly technical articles that focus on methods, analysis, and evaluation at the expense of what we would consider "theory" (Freelon, 2014). The call for papers for the 2014 conference on Computer-Supported Collaborative Work

(CSCW), a prominent social computing publication venue, expresses this notion in its introduction: “We invite submissions that detail existing practices, inform the design or deployment of systems, or introduce novel systems, interaction techniques, or algorithms” (CSCW, n.d.).[2] Further evidence for this claim can be seen in the strong presence of employees of well-known tech companies such as Google, Microsoft, and Yahoo on major social computing conferences’ program committees. Of course, theory is not always entirely absent: some articles include a few theoretical references of relevance to the project at hand, but the discussions tend to be much shorter than in most social science fields. And articles are often accepted without referencing any social science theories at all.

In addition to downplaying theory, social computing research relies heavily on computational methods such as social network analysis, machine learning, computational linguistics, and algorithmic preprocessing of raw web data. These methods are common in computer science and information science, disciplines which many (though by no means all) social computing researchers call home. Programming serves at least two major purposes in social computing: 1) to develop and improve digital platforms for social interaction, and 2) to evaluate their performance efficiently and at scale. Qualitative methods such as ethnography and depth interviews are occasionally seen, often as part of a multi-method approach with one or more computational methods. However, such studies are fairly rare, as the next section will demonstrate. The field places the highest value on research techniques and metrics that can be implemented algorithmically. The ability to visualize quantitative results in intuitive and innovative ways is also highly prized.

The final characteristic of social computing research of relevance to the digital politics researcher may seem rather obvious: the field is not principally concerned with politics per se, but rather with social computer use. In other words, social computing researchers typically analyze political cases to make broader points about social computing systems and affordances rather than about politics. Matters of system development and algorithm optimization almost always come first, and broader implications for politics are discussed secondarily if at all. As a result, the results sections of social computing research papers often leave many implications of theoretical interest unexplored. Later in this chapter I will attempt to reclaim some of these implications in order to clarify their value for students of political science and communication. But first, I will examine in detail the most common methods social computing researchers employ.

Methods in social computing research on politics

Social computing research is sometimes published in journals, but many of the most relevant studies for our purposes are published in the proceedings of prominent conferences in computer science, information science, and human-computer interaction. Haphazardly selecting papers from these conferences would bias my discussion, so instead I chose them using a systematic and replicable method. First, I focused on conferences and publications sponsored by the ACM (Association of Computing Machinery) and the IEEE (Institute of Electrical and Electronic Engineers), the premier professional organizations in computer science and computer engineering. Using Google Scholar, I searched for the term “political” exclusively within such outlets. I then ranked the results in descending order by number of citations in order to capture the most widely-referenced articles. Being interested only in articles that address politics as a central concern, I qualitatively assessed the most-cited items in each list, flagging articles that empirically analyze political messages, opinions, attitudes, and/or other content *as their main focus*. (Thus, for example, I excluded articles that analyzed political content as only one of three or more other content categories.) I continued this process until I had flagged 20 articles within each group, for a total of 40 articles (see Table 1). These form the basis of the discussions in this section and the next.

[Table 1 here]

After finalizing the sample, I informally classified the articles based on the methods they employed.[3] Three methodological categories were used: traditional quantitative, qualitative, and computational. The traditional quantitative category included long-established quantitative methods in social science such as surveys, experiments, content analysis, and statistical analysis of secondary data. The qualitative category included depth interviews, field observations, and close readings of texts, among others. The computational category included any method that entailed the creation of original source code whose purpose was to collect, preprocess, or analyze data. The rest of the chapter will focus mainly on this last category, as the others are much more familiar to scholars of digital politics.

Unsurprisingly, the most prevalent methodological category throughout the sample was by far computational (29/40, 72.5%), followed by traditional quantitative (19/40, 47.5%) and then qualitative (5/40, 12.5%). All but one of the qualitative studies used a mixed methodology which combined either multiple qualitative methods or one qualitative method with one of the other types. The authors used a variety of traditional quantitative methods, with surveys and content analyses being the two most popular. A large minority of studies employing traditional quantitative methods complemented them with computational methods (8/19, 42.1%). Based on this highly-cited sample, it would seem that social computing publication venues welcome political research that is methodologically traditional, although computational methods are more common.

I classified an extremely heterogeneous collection of methods as "computational" in accordance with the operational definition given above. These fall into three general subcategories: data collection, preprocessing, and analysis.

Data collection

All major social media services, including Twitter, Facebook, and Youtube, offer application programming interfaces (APIs) through which large amounts of data can be harvested computationally. By far the easiest way to collect these data is by writing a script in the programming language of one's choice. Some articles that analyzed social media content briefly described their data collection process, including such details as the language and specific API used (Mascaro, Black, & Goggins, 2012; Ratkiewicz et al., 2011; Skoric, Poor, Achananuparp, Lim, & Jiang, 2012), while others did not (Diaz-Aviles, Orellana-Rodriguez, & Nejdl, 2012; A. Garcia, Standlee, Beckhoff, & Cui, 2009; Golbeck & Hansen, 2011; Jürgens, Jungherr, & Schoen, 2011; Vallina-Rodriguez et al., 2012). Interfacing with APIs to extract data is evidently such a routine activity in social computing research that documenting its details is optional. Studies that examined content from sources without APIs—blogs for example—usually used their own custom web-scraping scripts (Adamic & Glance, 2005; Nahon & Hemsley, 2011; Ulicny, Kokar, & Matheus, 2010).

For researchers in disciplines like political science and communication that are relatively new to computational methods, this lack of detail on data collection methods is unfortunate. I do not intend to imply that it is the responsibility of social computing researchers to educate outsiders on the elementary aspects of social media data collection, but only to observe that those interested in getting started researching social media content will not learn much about how to collect it from articles in the field. Textbooks on social media analysis (Leetaru, 2012; Russell, 2013) are more helpful in this regard, but their utility will inevitably decrease with time due to the rapid developmental pace of social media platforms. Some enterprising political science and communication researchers will be able to teach themselves effectively using such resources, but until computational methods become a disciplinary priority, social scientists' ability to collect and analyze social media data will remain marginal.

Preprocessing

Preprocessing encompasses a miscellany of techniques to convert raw text and other content collected from the web into research-grade data suitable for quantitative and qualitative analysis. Examples include manipulating social media posts into formats suitable for calculating descriptive statistics (Mascaro et al., 2012), social network analysis (Adamic & Glance, 2005; Conover, Goncalves, Ratkiewicz, Flammini, & Menczer, 2011; Jürgens et al., 2011; Ratkiewicz et al., 2011), simple time-series plots (Vallina-Rodriguez et al., 2012), statistical associations with non-social media data (Golbeck & Hansen, 2011; Skoric et al., 2012), automated content analysis (Diaz-Aviles et al., 2012; Stieglitz & Dang-Xuan, 2012), and analysis of metadata such as "likes" or star ratings (D. Garcia, Mendez, Serdült, & Schweitzer, 2012). Like data collection, computational preprocessing requires programming skills by definition, but while the former is a rote task that rarely changes substantially between projects, the latter is completely open-ended. Indeed, creativity in preprocessing determines the kinds of analyses that can be applied to one's data; as such it is more akin to an art than a science.

The articles in the sample furnish a number of examples of the dizzying range of choices researchers face when preprocessing their data. In using social network methods to analyze relationships between social media users, a preprocessing script may count @-mentions, retweets relationships, and/or follow relationships as tie indicators, among other features (Conover et al., 2011; Golbeck & Hansen, 2011; Jürgens et al., 2011; Ratkiewicz et al., 2011). The findings of the ensuing social network analysis will obviously differ based on which tie indicators were used. Similarly, most types of automated text analysis require some preprocessing to allow the algorithms to output intelligible results. In a sentiment analysis of political tweets, Stieglitz and Dang-Xuan (2012) imported Twitter-specific jargon and emoticons from their dataset into a dictionary of positively- and negatively-valenced terms which they used to classify tweets as positive or negative in tone. Using a similar dictionary-based technique, Diaz-Aviles, Orellana-Rodriguez, and Nejdl (2012) assembled "profiles" of tweets and blog posts mentioning 18 Latin-American presidents to analyze the online sentiments associated with each. More sophisticated automated techniques such as supervised and unsupervised learning require even more extensive preprocessing. After removing very common words that contain little informational value (called stopwords), raw documents are often disaggregated into clusters of one-, two-, or three-word phrases called *n-grams* which learning algorithms analyze directly. The choice of which stopwords, types of n-grams, and algorithms to use all influence the end results. For example, Fang et al. (2012) attempted to quantify the ideological distance between differing political opinions in newspapers and in statements by US senators. To prepare their data for analysis, they used verbs, adjectives, and adverbs as opinion descriptors and retained certain opinion-relevant terms such as "should" and "must" that would otherwise be considered stopwords. In a very different research context, Zhang, Dang, and Chen (2009) extracted unigrams and bigrams from an Islamic women's web forum to examine gender differences in content and writing style using supervised learning.

Analysis

Programming usually plays some part in the analysis phase of studies that use computational methods. Complex and creative visualizations produced using specialized code libraries often appear in the results. Most of these tools are applied to communication content—tweets, blog posts, video transcripts, news articles—that do not require direct interaction with participants. The most common computational analytical methods for texts among the sample are dictionary- (or corpus-) based approaches, unsupervised learning, supervised learning, and network analysis.[4]

Dictionary-based approaches use either predefined or custom word collections representing different concepts to classify texts. For example, a dictionary of positive emotions might include terms such as "love," "awesome," "happy," and "best," and the software might measure positivity as the number of such terms within each text. This technique was used in several articles to analyze social media users' positive and negative feelings toward political issues and politicians (Diaz-Aviles et al., 2012; D. Garcia et al., 2012; Sarmento, Carvalho, Silva, & de Oliveira, 2009; Stieglitz & Dang-Xuan, 2012). Unsupervised learning approaches attempt to detect latent structure in texts inductively and automatically; one of its applications to politics research is the identification of topics mentioned in political texts (Fang et al., 2012). Supervised learning, in contrast, is a deductive method whose goal is to identify pre-established content categories automatically. It begins with a traditional content analysis, the results of which the algorithm uses as exemplars to classify previously unexamined texts. Several social computing research teams have used supervised learning to predict the political leanings of social media users (Conover et al., 2011; Jiang & Argamon, 2008; Park, Ko, Kim, Liu, & Song, 2011). Finally, network methods have proven themselves quite versatile, with applications in the study of political spam (Ratkiewicz et al., 2011), communication patterns among political bloggers (Adamic & Glance, 2005; Nahon & Hemsley, 2011; Ulicny et al., 2010), and political gatekeeping in social media (Jürgens et al., 2011).

This very brief survey was intended to highlight some of the ways computational methods have been used to study political topics. The kinds of research questions social computing scholars pursue using these methods are limited by their field-specific concerns; thus, there are many opportunities for innovative work by enterprising scholars in other fields with different concerns. The following section substantiates this point more fully.

Theory in social computing research on politics

There is a great deal of variation in how social computing research addresses theoretical concerns. Two broad approaches to theory are apparent in the current sample. The first is an *explicit* approach that closely resembles the norm in social science: relevant theoretical contributions from prior research are explored in an in-depth literature review, and then empirical research questions and/or hypotheses are derived from them. The depth of these literature reviews varies widely, as we shall see. The second approach is *implicit* in that theoretical concerns about politics are not discussed at all, but the methods or findings could be integrated into theory-based research by innovative authors. This section will first discuss the theoretical implications of explicitly theoretical papers, and then offer suggestions as to how implicitly theoretical work can inform existing theoretical traditions.

Explicitly theoretical work

Social computing research that explicitly incorporates theory does so in a similar fashion to social science. In fact, some such papers are theoretically rigorous enough to be published in a traditional political science or communication journal (Munson & Resnick, 2010; Nahon & Hemsley, 2011; Wei & Yan, 2010). However, others mention theoretical concerns only in passing: these will typically cite a small number of classic theoretical pieces without exploring much or any of the recent empirical work they have inspired (e.g. Adamic and Glance 2005; Baumer et al. 2009; Kaschesky and Riedl 2011; Weber, Garimella, and Borra 2012). I do not intend to fault the less theoretical pieces here—as explained earlier, social computing and social science have different goals. But observing trends in how the former field uses prior research is important for social scientists who may be interested in building on its studies or in submitting papers to social computing publication venues.

Only one cluster of theories attracted attention from more than one or two papers: online political polarization, homophily, and selective exposure. The research on this topic fell into two categories: studies of online content and evaluations of design interventions. The content-based research analyzed text and metadata from YouTube, the American political blogosphere, Twitter, online newspaper comments, and Yahoo!'s search query logs. Most of these studies found clear evidence of online homophily, e.g. that the blogosphere is divided in terms of hyperlinking patterns (Adamic & Glance, 2005), liberal blogs tend to link primarily to liberal election videos and *mutatis mutandis* for conservatives (Nahon & Hemsley, 2011), the Twitter followers of media outlets tend to skew liberal or conservative (Golbeck & Hansen, 2011), and liberals and conservatives tend to use ideologically distinctive queries in search engines (Weber et al., 2012). The design intervention studies evaluated the effects of human interaction with systems designed to promote exposure to opinion-challenging content (Munson & Resnick, 2010) and critical thinking about politics (Baumer et al., 2009; Baumer, Sinclair, & Tomlinson, 2010). Unsurprisingly, all three of these studies reported some degree of success in their stated goals.

The remaining explicitly theoretical pieces covered a hodgepodge of theoretical concerns. Kaschesky and Riedl (2011) justified their research examining how opinions form and diffuse online partly by reference to the public sphere and deliberation. Along somewhat similar lines, Wei and Yan (2010) grounded their survey-based study of online knowledge production in the knowledge gap and political participation literatures. Bélanger and Carter (2010) invoked the digital divide in a study of US attitudes toward Internet voting, finding that younger and more affluent citizens are more favorably disposed toward it. Denardis and Tam (2007) offered a legalistic analysis of global ICT standards based on democratic theory, ultimately recommending open document formats for public institutions. In the sole study grounded in critical theory, Kannabiran and Petersen (2010) presented a Foucauldian reading of Facebook’s interface.

Implicitly theoretical work

Most of the studies reviewed for this chapter did not discuss theory in any substantial way (although some of these cited social science papers to discuss their empirical results). A few lacked literature reviews altogether (Jiang & Argamon, 2008; Jürgens et al., 2011; Ratkiewicz et al., 2011). Those that included them tend to focus on previous studies’ methodological efficiency and range of application, and they generally frame their contributions in those terms as well (Diakopoulos & Shamma, 2010; Diaz-Aviles et al., 2012; Fang et al., 2012; D. Garcia et al., 2012; Michael Kaschesky, Sobkowicz, & Bouchard, 2011; Kim, Valente, & Vinciarelli, 2012; Sarmento et al., 2009; Skoric et al., 2012; Younus et al., 2011; Zhang et al., 2009). In a representative example, Awadallah, Ramanath, and Weikum (2010) presented a new method for classifying political debate arguments as pro or con. Much previous work in the area had at that point been context independent—for example, judging a statement as inherently positive or negative, whereas pro/con judgments depend upon how the debate position is phrased. Further, previous work had also required manually-classified training data, which is time-consuming and expensive. Awadallah’s approach was both context-sensitive and fully automatic, which constitute substantive contributions in the social computing research tradition.

Perhaps the best way to demonstrate the value of implicitly theoretical work is to describe its attempted goals, most of which fall into one or more of three categories: classification, forecasting, and description. Classification, the largest category, consists of studies that aim to fully or partially automate the process of labeling digital content (mostly but not exclusively text). Some of the classification tasks in this sample include labeling political texts as positive or negative (which is also known as sentiment analysis) (Diakopoulos & Shamma, 2010; Diaz-Aviles et al., 2012; D. Garcia et al., 2012; Sarmento et al., 2009), pro or con (Awadallah et al., 2010), subjective or objective (Younus et al., 2011), and liberal or conservative (Conover et al., 2011; Fang et al., 2012; Golbeck & Hansen, 2011; Jiang & Argamon, 2008). Forecasting studies seek to predict patterns or outcomes in the digital realm or offline; examples include elections (Skoric et al., 2012), public opinion polls (Diaz-Aviles et al., 2012; Hong & Nadler, 2011), and the diffusion of political opinions online (M. Kaschesky & Riedl, 2011; Michael Kaschesky et al., 2011). Descriptive studies are similar to their counterparts in social science except that they use very little or no theory (and sometimes no prior research at all) to guide them. As a result, their attempts to discover how platforms such as Twitter were used in particular contexts vary widely in their methodological specifics (Mascaro et al., 2012; Vallina-Rodriguez et al., 2012).

Each of these categories is implicitly theoretical in its own way. Classification studies do not quite go far enough to qualify as social science; their goal is typically to optimize algorithmic performance rather than to contribute to theory. From a social science perspective they resemble extended method sections, full of details on each of the task's steps and the results of various evaluation metrics. This metaphor clarifies the theoretical implications of advanced classification studies to social science: any theory that requires classification could potentially make use of their methodological innovations. For example, the ability to classify political ideology algorithmically could enable theoretically-grounded studies of political polarization and deliberation to analyze sample or population sizes in the millions. Similarly, an automated system for quantifying political sentiment in social media posts could help researchers better theorize how voters react to targeted political messages outside of experimental settings (for more on the uses of sentiment analysis in digital politics research, see Petchler & Gonzalez, this volume). Forecasting is more the province of natural scientists and economists than most of social science, which is more concerned with explanation.[5] That said, we should recall that forecasting encompasses within it correlation and time precedence, which are two of Babbie's (2012) three essential components of causation. The remaining component, the elimination of potential alternative causes, then becomes the task of the social scientist. In the rush to build models that can predict elections based on user-generated data, it is the social scientist rather than the social computing researcher who will be interested in *why* the model works. Finally, most descriptive studies would not pass muster in most social science journals because of their long-standing bias against atheoretical work. Nevertheless, they can still offer the social scientist a sense of the methodological possibilities afforded by new social computing platforms, which could then be incorporated into research questions and/or hypotheses that build theory.

Conclusion and future work

As I have shown, social computing research has produced much of interest to the digital politics researcher. The field has employed computational methods and Big Data since the 1990s, and still conducts much of the cutting-edge research in these areas. In contrast, political science and communication are still very firmly invested in their traditional methods, which are not always optimally suited for analyzing digital data. Engagement with the best social computing research studies has been and will continue to be essential for all social scientists interested in applying computational methods in their home disciplines. The field's theoretical contributions are not always as obvious, but with a bit of work, students of digital politics will be able to profitably draw upon them for inspiration.

I close this chapter with two general recommendations for social scientists who find this sort of work valuable. The first is simply to learn a programming language suitable for manipulating and analyzing large datasets. While researchers can conduct a few descriptive analyses on large datasets without knowing how to program, most research-grade operations require the ability to work directly with code. Collaborating with social computing researchers may work well for some projects, but as we've seen, they have different standards for what constitutes a contribution (and corresponding publication incentives). Moreover, social scientists can recognize theoretically-relevant patterns in data that computer scientists can't—thus it greatly benefits the former to know how to explore large-scale datasets firsthand. (Imagine having to rely on statistician for all your statistics!) For the beginning computational researcher I recommend learning the Python programming language, both because it offers a number of libraries and modules specifically for collecting, preprocessing, and analyzing data; and also because its growing popularity in academic circles offers critical support for new learners. R, the statistical language and programming environment, is more advanced in terms of the variety and complexity of statistical models it supports, but has a steeper learning curve than Python.

As computational research become more accepted in the disciplines in which digital politics research is conducted, graduate faculties should strongly consider how best to teach its methods to their students. At the time of this writing, very few communication departments in the US teach computational methods in any systematic fashion, and I suspect the situation is not substantially different in political science. Few American communication departments have any experts in computational methods on faculty, and fewer still have more than one. Some of these experts, such as Benjamin Mako Hill (as of this writing at the University of Washington) and Sandra Gonzalez-Bailon (currently at the University of Pennsylvania), received their graduate departments in fields other than communication. Others, such as Drew Margolin (now at Cornell) and I, trained in communication departments that do not emphasize computational methods as a core strength. In light of the paramount importance and ubiquity of digital communication data, I submit that computational methods should become one of communication's premier research methods—on par with survey methods, content analysis, experiments, and depth interviews. And just as every doctoral student need not learn how to conduct and analyze surveys, not everyone needs to learn computational methods, but it ought to be one of communication's major areas of methodological specialization. A detailed explanation of how to achieve this outcome lies beyond the scope of this chapter, but at a minimum, committed departments will need to thoroughly revise their hiring practices, tenure guidelines, graduate curricula, and departmental resources (including appropriate hardware, software, and data subscriptions), among other reforms.

My second recommendation pertains to the construct validity of digital traces. Construct validity is the extent to which an operationalized metric actually measures the underlying concept it is intended to measure (Babbie, 2012). As I have documented elsewhere (Freelon, 2014), social computing research studies do not always amply demonstrate the construct validity of the traces they use as metrics. To take an example from the current sample, Ulicny et al. (2010) purport to measure four concepts of academic and practical relevance in the Malaysian blogosphere: relevance, specificity, timeliness, and credibility. Without any reference to prior literature, they define these concepts in terms of manifest digital traces, including use of a real name, network authority, number of comments, and number of unique nouns, among others. Not only are these metrics biased in favor of what can be collected and measured easily, there is no discussion of whether the metrics are comprehensive, and if not, which aspects of the underlying concepts might be omitted. While a lack of attention to construct validity is by no means universal in social computing research, it is common (Fang et al., 2012; A. Garcia et al., 2009; Jürgens et al., 2011; Mascaro et al., 2012; Younus et al., 2011).

Social science research on politics is ultimately concerned with abstract concepts such as power, influence, preference, ideology, and homophily, among many others. Traces such as retweets, Facebook "likes," social media follow relationships, and hyperlink patterns are only interesting inasmuch as they faithfully and consistently indicate such concepts. Yet just as we should avoid studying traces for their own sake, we should also refrain from simply assuming that retweets are endorsements, hyperlinks signify authority, and "likes" imply approval. Credible arguments for these positions should be submitted and substantiated. In some cases, it will be possible to make logical arguments on the basis of a trace's inherent properties, as in the observation that retweets represent peer-to-peer information propagation. But whenever possible, a trace's imputed meaning should draw on empirical observation: close qualitative observation of how traces are used can help fulfill this purpose (Boyd, Golder, & Lotan, 2010).

The rise of computational techniques in social science has barely begun, and digital politics scholars (including me) still have much to learn. Social computing researchers offer some of the most methodologically sophisticated work currently available, and many of them are interested in very familiar subject matter. For these reasons, we would do well to learn what we can from them.

Suggested reading

On social computing:

Parameswaran, Manoj, and Andrew B. Whinston. 2007. "Social Computing: An Overview." *Communications of the Association for Information Systems* 19 (37): 762–780.

Tian, Yonghong, Jaideep Srivastava, Tiejun Huang, and Noshir Contractor. 2010. "Social Multimedia Computing." *Computer* 43(8): 27-36.

Wang, Fei-Yue, Kathleen M. Carley, Daniel Zeng, and Wenji Mao. 2007. "Social Computing: From Social Informatics to Social Intelligence." *Intelligent Systems, IEEE* 22 (2): 79–83.

Learning computational methods:

Cogliati, Josh, Mitchell Aikens, Kiah Morante, Elizabeth Cogliati, James A. Brown, Joe Oppengaard, and Benjamin Hell. (2013). *Non-programmer's tutorial for Python 3.*

Wikibooks. Available at http://en.wikibooks.org/wiki/Non-Programmer's_Tutorial_for_Python_3

Russell, Matthew A. 2013. *Mining the social web*. Sebastopol, CA: O'Reilly Media.

Stanton, Jeffrey. 2013. *Introduction to Data Science.* Available at http://jsresearch.net/

Notes

[1] Consider for example the panels held on the topic of Big Data and/or computational methods at the 2013 annual meetings of the International Communication Association (ICA) and the Association of Educators in Journalism and Mass Communication (AEJMC), as well as the conference theme of the 2014 annual meeting of the American Political Science Association (APSA)—"Politics After the Digital Revolution."

[2] Many social computing papers are published in the archived proceedings of engineering conferences, which carry the cachet of journal publications in social-scientific fields. Aside from CSCW, these include CHI (Conference on Human Factors in Computing Systems), WWW (International World Wide Web Conference), ICWSM (International Conference on Weblogs and Social Media), and HICSS (Hawaii International Conference on System Sciences).

[3] I chose not to conduct a formal content analysis here mainly due to the great diversity of methods comprising the "computational" category, which proved difficult for a non-expert coder to identify consistently.

[4] Readers interested in more in-depth discussions of these methods than I offer here are recommended to consult Graesser, McNamara, and Louwerse (2010) and Petchler and Gonzalez-Bailon (this volume).

[5] For more on the differences between scientific prediction and explanation, see Shmueli and Koppius (2011).

References

Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery* (pp. 36–43).

Awadallah, R., Ramanath, M., & Weikum, G. (2010). Language-model-based Pro/Con Classification of Political Text. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 747–748). New York, NY, USA: ACM. doi:10.1145/1835449.1835596

Babbie, E. (2012). *The Practice of Social Research*. Belmont, CA: Cengage Learning.

Baumer, E. P. S., Sinclair, J., Hubin, D., & Tomlinson, B. (2009). metaViz: Visualizing Computationally Identified Metaphors in Political Blogs. In *International Conference on Computational Science and Engineering, 2009. CSE ’09* (Vol. 4, pp. 389–394). doi:10.1109/CSE.2009.482

Baumer, E. P. S., Sinclair, J., & Tomlinson, B. (2010). America is Like Metamucil: Fostering Critical and Creative Thinking About Metaphor in Political Blogs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1437–1446). New York, NY, USA: ACM. doi:10.1145/1753326.1753541

Bélanger, F., & Carter, L. (2010). The digital divide and internet voting acceptance. In *Digital Society, 2010. ICDS’10. Fourth International Conference on* (pp. 307–310). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5432779

Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *43rd Hawaii International Conference on System Sciences (HICSS)* (pp. 1–10).

Conover, M. D., Goncalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011). Predicting the Political Alignment of Twitter Users. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)* (pp. 192 –199). doi:10.1109/PASSAT/SocialCom.2011.34

CSCW. (n.d.). Call for Participation Papers. ACM. Retrieved from http://cscw.acm.org/participation_papers.html

DeNardis, L., & Tam, E. (2007). Interoperability and democracy: A political basis for open document standards. In *5th International Conference on Standardization and Innovation in Information Technology, 2007. SIIT 2007* (pp. 171–180). doi:10.1109/SIIT.2007.4629327

Diakopoulos, N. A., & Shamma, D. A. (2010). Characterizing Debate Performance via Aggregated Twitter Sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1195–1198). New York, NY, USA: ACM. doi:10.1145/1753326.1753504

Diaz-Aviles, E., Orellana-Rodriguez, C., & Nejdl, W. (2012). Taking the Pulse of Political Emotions in Latin America Based on Social Web Streams. In *Web Congress (LA-WEB), 2012 Eighth Latin American* (pp. 40–47). doi:10.1109/LA-WEB.2012.9

Fang, Y., Si, L., Somasundaram, N., & Yu, Z. (2012). Mining contrastive opinions on political texts using cross-perspective topic model. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 63–72). Retrieved from http://dl.acm.org/citation.cfm?id=2124306

Fink, E. J., & Gantz, W. (1996). A Content Analysis of Three Mass Communication Research Traditions: Social Science, Interpretive Studies, and Critical Analysis. *Journalism & Mass Communication Quarterly*, *73*(1), 114–134. doi:10.1177/107769909607300111

Freelon, D. (2014). On the interpretation of digital trace data in communication and social computing research. *Journal of Broadcasting & Electronic Media*, *58*(1), 59–75.

Garcia, A., Standlee, A., Beckhoff, J., & Cui, Y. (2009). Ethnographic Approaches to the Internet and Computer-Mediated Communication. *Journal of Contemporary Ethnography*, *38*(1), 52–84.

Garcia, D., Mendez, F., Serdült, U., & Schweitzer, F. (2012). Political polarization and popularity in online participatory media: an integrated approach. In *Proceedings of the first edition workshop on Politics, elections and data* (pp. 3–10). Retrieved from http://dl.acm.org/citation.cfm?id=2389665

Golbeck, J., & Hansen, D. (2011). Computing political preference among twitter followers. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (pp. 1105–1108). New York, NY, USA: ACM. doi:10.1145/1978942.1979106

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2010). Methods of Automated Text Analysis. In *Handbook of reading research* (Vol. 4, p. 34).

Hong, S., & Nadler, D. (2011). Does the Early Bird Move the Polls?: The Use of the Social Media Tool "Twitter" by U.S. Politicians and Its Impact on Public Opinion. In *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times* (pp. 182–186). New York, NY, USA: ACM. doi:10.1145/2037556.2037583

Jiang, M., & Argamon, S. (2008). Exploiting Subjectivity Analysis in Blogs to Improve Political Leaning Categorization. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 725–726). New York, NY, USA: ACM. doi:10.1145/1390334.1390472

Jürgens, P., Jungherr, A., & Schoen, H. (2011). Small worlds with a difference: New gatekeepers and the filtering of political information on Twitter. *Proceedings of the ACM WebSci'11*, 14–17.

Kannabiran, G., & Petersen, M. G. (2010). Politics at the Interface: A Foucauldian Power Analysis. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (pp. 695–698). New York, NY, USA: ACM. doi:10.1145/1868914.1869007

Kaschesky, M., & Riedl, R. (2011). Tracing Opinion-Formation on Political Issues on the Internet: A Model and Methodology for Qualitative Analysis and Results. In *2011 44th Hawaii International Conference on System Sciences (HICSS)* (pp. 1–10). doi:10.1109/HICSS.2011.456

Kaschesky, M., Sobkowicz, P., & Bouchard, G. (2011). Opinion Mining in Social Media: Modeling, Simulating, and Visualizing Political Opinion Formation in the Web. In *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times* (pp. 317–326). New York, NY, USA: ACM. doi:10.1145/2037556.2037607

Kim, S., Valente, F., & Vinciarelli, A. (2012). Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates. In *2012 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5089–5092). doi:10.1109/ICASSP.2012.6289065

Leetaru, K. (2012). *Data Mining Methods for the Content Analyst: An Introduction to the Computational Analysis of Content*. Routledge.

Mascaro, C., Black, A., & Goggins, S. (2012). Tweet recall: examining real-time civic discourse on twitter. In *Proceedings of the 17th ACM international conference on Supporting group work* (pp. 307–308). Retrieved from http://dl.acm.org/citation.cfm?id=2389233

Munson, S. A., & Resnick, P. (2010). Presenting diverse political opinions: how and how much. In *Proceedings of the 28th international conference on Human factors in computing systems* (pp. 1457–1466).

Nahon, K., & Hemsley, J. (2011). Democracy.com: A Tale of Political Blogs and Content. In *2011 44th Hawaii International Conference on System Sciences (HICSS)* (pp. 1–11). doi:10.1109/HICSS.2011.140

Parameswaran, M., & Whinston, A. B. (2007). Social computing: An overview. *Communications of the Association for Information Systems*, *19*(37), 762–780.

Park, S., Ko, M., Kim, J., Liu, Y., & Song, J. (2011). The Politics of Comments: Predicting Political Orientation of News Stories with Commenters' Sentiment Patterns. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (pp. 113–122). New York, NY, USA: ACM. doi:10.1145/1958824.1958842

Potter, W. J., Cooper, R., & Dupagne, M. (1993). The Three Paradigms of Mass Media Research In Mainstream Communication Journals. *Communication Theory*, *3*(4), 317–335. doi:10.1111/j.1468-2885.1993.tb00077.x

Ratkiewicz, J., Conover, M., Meiss, M., Goncalves, B., Patil, S., Flammini, A., & Menczer, F. (2011). Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web* (pp. 249–252). New York, NY, USA: ACM. doi:10.1145/1963192.1963301

Russell, M. A. (2013). *Mining the social web*. Stebastopol, Calif.: O'Reilly Media.

Sarmento, L., Carvalho, P., Silva, M. J., & de Oliveira, E. (2009). Automatic Creation of a Reference Corpus for Political Opinion Mining in User-generated Content. In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion* (pp. 29–36). New York, NY, USA: ACM. doi:10.1145/1651461.1651468

Shmueli, G., & Koppius, O. (2011). Predictive Analytics in Information Systems Research. *Management Information Systems Quarterly, 35*(3), 553–572.

Skoric, M., Poor, N., Achananuparp, P., Lim, E.-P., & Jiang, J. (2012). Tweets and Votes: A Study of the 2011 Singapore General Election. In *2012 45th Hawaii International Conference on System Science (HICSS)* (pp. 2583–2591). doi:10.1109/HICSS.2012.607

Stieglitz, S., & Dang-Xuan, L. (2012). Political Communication and Influence through Microblogging–An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior. In *Hawaii International Conference on System Sciences* (Vol. 0, pp. 3500–3509). Los Alamitos, CA, USA: IEEE Computer Society. doi:http://doi.ieeecomputersociety.org/10.1109/HICSS.2012.476

Ulicny, B., Kokar, M. M., & Matheus, C. J. (2010). Metrics For Monitoring A Social-Political Blogosphere: A Malaysian Case Study. *IEEE Internet Computing, 14*(2), 34 –44. doi:10.1109/MIC.2010.22

Vallina-Rodriguez, N., Scellato, S., Haddadi, H., Forsell, C., Crowcroft, J., & Mascolo, C. (2012). Los Twindignados: The Rise of the Indignados Movement on Twitter. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)* (pp. 496–501). doi:10.1109/SocialCom-PASSAT.2012.120

Wang, F.-Y., Carley, K. M., Zeng, D., & Mao, W. (2007). Social computing: From social informatics to social intelligence. *Intelligent Systems, IEEE*, *22*(2), 79–83.

Weber, I., Garimella, V. R. K., & Borra, E. (2012). Mining Web Query Logs to Analyze Political Issues. In *Proceedings of the 3rd Annual ACM Web Science Conference* (pp. 330–334). New York, NY, USA: ACM. doi:10.1145/2380718.2380761

Wei, L., & Yan, Y. (2010). Knowledge production and political participation: Reconsidering the knowledge gap theory in the Web 2.0 environment. In *2010 The 2nd IEEE International Conference on Information Management and Engineering (ICIME)* (pp. 239–243). doi:10.1109/ICIME.2010.5477878

Younus, A., Qureshi, M. A., Asar, F. F., Azam, M., Saeed, M., & Touheed, N. (2011). What Do the Average Twitterers Say: A Twitter Model for Public Opinion Analysis in the Face of Major Political Events. In *2011 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 618–623). doi:10.1109/ASONAM.2011.85

Zhang, Y., Dang, Y., & Chen, H. (2009). Gender difference analysis of political web forums: An experiment on an international islamic women’s forum. In *IEEE International Conference on Intelligence and Security Informatics, 2009. ISI ’09* (pp. 61–64). doi:10.1109/ISI.2009.5137272

Table 1: The methods of 40 highly-cited social computing research papers

Authors	Title	Traditional quantitative methods	Qualitative methods	Computational methods
Adamic & Glance	The political blogosphere and the 2004 US election: divided they blog			x
Awadallah, Ramanath, & Weikum	Language-model-based pro/con classification of political text			x
Awadallah, Ramanath, & Weikum	Harmony and dissonance: organizing the people's voices on political controversies			x
Baumer et al.	metaViz: Visualizing Computationally Identified Metaphors in Political Blogs	x		x
Baumer, Sinclair, & Tomlinson	America is like Metamucil: fostering critical and creative thinking about metaphor in political blogs	x		x
Bélanger & Carter	The digital divide and internet voting acceptance	x		
Conover et al.	Predicting the political alignment of twitter users	x		x
DeNardis & Tam	Interoperability and democracy: A political basis for open document standards		x	
Diakopoulos & Shamma	Characterizing debate performance via aggregated twitter sentiment	x		
Diaz-Aviles et al.	Taking the Pulse of Political Emotions in Latin America Based on Social Web Streams			x
Fang et al.	Mining contrastive opinions on political texts using cross-perspective topic model			x
Fisher, Becker, & Crandall	eGovernment Services Use and Impact through Public Libraries: Preliminary Findings from a National Study of Public Access Computing in Public Libraries	x	x	
Furuholt &Wahid	E-government Challenges and the Role of Political Leadership in Indonesia: the Case		x	

	of Sragen			
Garcia et al.	Political polarization and popularity in online participatory media: An integrated approach			x
Golbeck & Hansen	Computing political preference among twitter followers			x
Gulati, Yates, & Williams	Understanding the impact of political structure, governance and public policy on e-government	x		
Hong & Nadler	Does the early bird move the polls?: the use of the social media tool 'Twitter' by US politicians and its impact on public opinion			x
Jiang & Argamon	Exploiting subjectivity analysis in blogs to improve political leaning categorization	x		x
Jürgens, Jungherr, & Schoen	Small worlds with a difference: New gatekeepers and the filtering of political information on Twitter			x
Kannabiran & Petersen	Politics at the interface: a Foucauldian power analysis		x	
Kaschesky & Riedl	Tracing opinion-formation on political issues on the internet: A model and methodology for qualitative analysis and results	x	x	
Kaschesky, Sobkowicz, & Bouchard	Opinion mining in social media: modeling, simulating, and visualizing political opinion formation in the web			x
Kim, Kavanaugh, & Pérez-Quiñones	Toward a model of political participation among young adults: the role of local groups and ICT use	x		
Kim, Valente, & Vinciarelli	Automatic detection of conflicts in spoken conversations: ratings and analysis of broadcast political debates	x		x
Mascaro, Black, & Goggins	Tweet recall: examining real-time civic discourse on twitter	x		x
Munson &	Presenting diverse political opinions: how	x		x

Resnick	and how much			
Nahon & Hemsley	Democracy. com: A Tale of Political Blogs and Content			x
Park et al.	The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns	x		x
Ratkiewicz et al.	Truthy: mapping the spread of astroturf in microblog streams			x
Sarmento et al.	Automatic creation of a reference corpus for political opinion mining in user-generated content	x		x
Singh, Mahata, & Adhikari	Mining the Blogosphere from a Socio-political Perspective			x
Skoric et al	Tweets and votes: A study of the 2011 singapore general election			x
Stieglitz & Dang-Xuan	Political Communication and Influence through Microblogging--An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior			x
Ulicny, Kokar, & Matheus	Metrics for monitoring a social-political blogosphere: A Malaysian case study			x
Vallina-Rodriguez et al.	Los twindignados: The rise of the indignados movement on twitter	x		x
Wallsten	Beyond Agenda Setting: The Role of Political Blogs as Sources in Newspaper Coverage of Government	x		x
Weber, Garimella, & Borra	Mining web query logs to analyze political issues			x
Wei & Yan	Knowledge production and political participation: reconsidering the knowledge gap theory in the web 2. environment	x		
Younus et al.	What do the average twitterers say: A twitter model for public opinion analysis in the face of major political events	x		
Zhang, Dang,	Gender difference analysis of political web			x

& Chen	forums: An experiment on an international islamic women's forum			
Total	-	19	5	29