On the Interpretation of Digital Trace Data in Communication and Social Computing Research

Deen Freelon

School of Communication

American University

4400 Massachusetts Ave NW

Washington, DC 20016

Freelon@american.edu

The final version of this preprint manuscript will be published in the *Journal of Broadcasting & Electronic Media*.

Abstract

The widespread availability of analytical tools for Big Data offers enormous opportunities and challenges for communication researchers. In contrast to user-generated texts, *digital trace data* (evidence of online user activities such as hyperlinks and retweets) represent a new methodological frontier for the field. However, interpretive strategies remain scattered and ad hoc with few best practices to guide them. To help remedy this situation, this article reviews recent scholarship in both communication and social computing research that has incorporated three common types of trace data: hyperlinks, Twitter followers, and retweets. It finds that while researchers in both fields have interpreted each trace in a variety of ways, they have largely declined to explain the validity of their interpretations.

On the Interpretation of Digital Trace Data in Communication and Social Computing Research

Digital traces surround us. Every link, like, share, follow, and friend request leaves behind a record which can be collected and analyzed by researchers who know how and where to look. Recent advances in collection software and data availability have driven an explosion of research on *digital trace data*, which are defined formally as evidence of human and human-like activity that is logged and stored digitally (Howison, Wiggins, & Crowston, 2011).[1] These promise to reveal interesting patterns about how people communicate online that would be unlikely to emerge by simply asking them. This essay reviews some of the recent literature on digital traces in communication research and social computing, an interdisciplinary research field that relies heavily on such data. Studies from both fields are included here to enable cross-disciplinary triangulation for best practices and recommendations. My main goal is to explore how three types of traces—hyperlinks, Twitter followers, and retweets—have been interpreted and how those interpretations have been justified.

The field of social computing, which overlaps to some extent with human-computer interaction (HCI) and social informatics, shares key research interests with communication. However, the two fields differ in two major respects. First, social computing focuses exclusively on the use of computers, whereas the communication's purview is much broader. Wang, Zeng, Carley, and Mao (2007), synthesizing from multiple sources, define social computing as the "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). Accordingly, studies within this tradition that address concepts and data of relevance to communication scholarship (e.g. political, interpersonal, or organizational communication) nearly always restrict their investigations to behaviors in which computers are centrally involved. In contrast, scholars of,

say, political communication cannot effectively limit their interests to computer-mediated interactions—they also need to address the roles of traditional media and face-to-face communication

The second major difference between trace data's role in communication and social computing concerns the role of theory. In the former, theory provides the rationale for the vast majority of empirical inquiries; indeed, insufficient theoretical relevance is a frequent rationale for rejecting communication journal submissions. But it is largely optional in social computing, whose publication venues are mostly conferences sponsored by broad professional organizations such as the ACM (Association for Computing Machinery), the IEEE (Institute of Electrical and Electronics Engineers), and the AAAI (Association for the Advancement of Artificial Intelligence). These conferences are known primarily by their acronyms and include CHI (Computer-Human Interaction), CSCW (Computer-Supported Collaborative Work), and KDD (Knowledge Discovery and Data Mining), among hundreds of others. The papers published through these venues are roughly as long as standard social science journal articles (6-8,000 words) or shorter, but typically lack substantial literature review and discussion sections. Instead, most of the allotted space is typically devoted to a combination of in-depth methodological description, statistical analysis, and data visualization (though qualitative work is by no means absent, social computing research on trace data is mostly quantitative). Issues such as the theoretical positioning of the project and the broader significance of the results are ordinarily addressed only briefly if at all.

A key question for both social computing and communication concerns the social meanings of trace data. The mere ubiquity of digital traces does not justify studying them—their real value lies in what they tell us about the people who generated them. Some studies explicitly

portray traces as indicators of some higher-level concept of interest (e.g. influence, popularity, credibility, etc.), while in others, their meaning is more implicit. Studies also differ in the extent to which they explain why traces should be interpreted in certain ways: some base their conclusions on their empirical findings, others in example-based argument, and still others simply assert that a given trace has a given meaning without further explanation. As the literature on argumentation and persuasion amply demonstrates, these standards of evidence vary widely in quality (Baesler & Burgoon, 1994; Hoeken, 2001; Reinard, 1988). Understanding the various ways in which trace data are interpreted and the quality of the arguments being marshaled in support of those interpretations is critical in both assessing the quality of trace data research and in establishing best practices for future studies. To these ends, the three main sections of this essay review how prominent studies in both social computing and communication have used and interpreted three types of digital traces: hyperlinks, Twitter followers, and retweets. These particular traces were chosen because all have been studied extensively in both communication and social computing. The two guiding research questions are:

1. What meanings have communication and social computing researchers imputed to trace data?
2. What kinds of arguments have they used to establish the validity of these meanings?

Method

As the number of communication and social computing studies analyzing the three types of trace data number in the hundreds, it is impossible to review them all in a single article. However, the subset examined here is not a convenience sample, since that would have constituted selecting on the dependent variable. Studies were instead chosen using two similar, systematic methods, each tailored to the publication venues of its corresponding field.

Communication

The first task in selecting studies for each discipline was identifying a set of appropriate publication venues. For communication, a master list was created of all journals reviewed in two landmark studies of the theoretical history of communication research (Bryant & Miron, 2004; Graber & Smith, 2005). On Jan 27, 2013, each of the following keywords— "hyperlink," "twitter follower," and "retweet"—was searched in Google Scholar within each journal and the number of results was recorded for each journal/trace combination. Of these, the five journals containing the highest number of results for each trace were identified. The most-cited articles among the search results (as designated by Google Scholar) were downloaded and searched in descending order of number of citations for the presence of the search term for which it was returned. The rationale for selecting among the most-cited articles as opposed to random selection was that the former would be more likely to contain widely-used interpretations of trace data. For retweets and Twitter followers, the top ten most-cited articles containing each keyword at least twice outside the references were included for analysis in this study.[2] Due to the comparatively large number of studies meeting the keyword criteria for hyperlinks, the possibility emerged that one or two journals could dominate the sample and reduce variety in trace data usage. Therefore, for hyperlinks only, the top two most-cited articles within each of the top five hyperlink journals meeting the keyword criteria were selected.

Social computing

Because no central authority was available from which to select leading publication venues, the social computing article selection strategy was slightly different. It began by using the names of the two largest professional organizations in computer engineering research, the ACM and the IEEE, to narrow search results in Google Scholar. Entering each organization's

acronym into Google Scholar's "published in" field returns publications sponsored by that organization. The same keywords used in the communication article selection process were thus reused for the ACM and the IEEE. To ensure that differences between the two organizations did not distort the results, the top five most-cited articles within each organization containing each keyword at least twice outside the references were chosen. When several articles were found to be demonstrations of new software with no empirical analyses of trace data, an additional condition was added requiring the presence of such analyses. This process yielded a total of ten social-computing articles for each type of trace data to match its corresponding articles from communication.

All articles were read carefully to answer the two research questions. Each of the next three sections focuses on one type of trace data and addresses both research questions in turn.

Hyperlinks

The research history of hyperlinks is the longest of the three types of trace data, originating in the late 1990s in communication and earlier in social computing. In the studies reviewed here, scholars of communication and social computing tend to interpret hyperlinks fairly differently. One of the most common interpretations in communication explicitly acknowledges the diversity of social relations links can imply. In an early example, Foot and colleagues note that "hyperlinks are... mediators of a wide range of associative relations between producers of Web materials" (Foot, Schneider, Dougherty, Xenos, & Larsen, 2003, n.p.). Similar sentiments are expressed by Halavais, who likens links to "roads, telephone lines or citations" (Halavais, 2000, p. 12) and Trammell and colleagues ("hyperlinks can manifest a drive to be connected to others on the Internet or to share information") (Trammell, Tarkowski, Hofmokl, & Sapp, 2006, p. 12). Gillan (2009) and Van Aelst and Walgrave (2002) invert this idea,

emphasizing that scholars should refrain from making a priori assumptions what hyperlinks may signify.

While the social diversity argument is difficult to dispute, the communication studies differ in the extent to which they specify which roles their links actually play. Earlier studies claimed that the fact that hyperlinks *can* play multiple social roles was by itself a sufficient justification for studying hyperlink networks, but refrained from investigating empirically what those roles were (Foot et al., 2003; Halavais, 2000). Later studies explore this question more directly, discovering that certain social uses of links predominate over others (Coddington, 2012; Trammell et al., 2006). Some of their findings overlap with studies portraying hyperlinks as fulfilling singular social functions without acknowledging diversity. Among the communication studies reviewed here, the most common functions ascribed to hyperlinks (both by those who acknowledge diversity and those who do not) are credibility, additional information, and self-expression. Credibility was prominent in journalistic contexts, with both Coddington (2012) and Matheson (2004) mentioning hyperlinks' "ability to provide credibility to the linker by giving readers a transparent means of determining for themselves the basis for the author's claims" (Coddington, 2012, p. 216). Relatedly, Matheson (2004) and Dimitrova, Kaid, Williams, and Trammell (2005) note that links can provide additional information to news readers who want to know more about particular aspects of a story. Two studies of blogs and personal web pages emphasize hyperlinks' capacity for self-expression through engagement with individual interests and social identities (Papacharissi, 2002; Trammell & Keshelashvili, 2005; Trammell et al., 2006).

The social computing studies were concerned with a decidedly different set of social roles for hyperlinks. The two most widely-shared interpretations among the sample were as

indicators of influence (Bross, Quasthoff, Berger, Hennig, & Meinel, 2010; Mathioudakis, Koudas, & Marbach, 2010; Ulicny, Kokar, & Matheus, 2010) and of relevance (Bhattarai, Rus, & Dasgupta, 2009; Jamali & Abolhassani, 2006). This general view closely matches that of PageRank, the algorithm that premises the order of Google's search results upon the number of highly-linked pages linking to them (indeed, PageRank is cited or mentioned in several of these articles). Two articles were devoted to spam detection, and neither of these explicitly states the purpose(s) of links. Instead, each team of authors uses links as one of several components of a novel spam-detection algorithm (Chen, Tan, & Jain, 2009; Lee, Caverlee, & Webb, 2010). Another article discusses the role of hyperlinks in helping users search Twitter and the web, but similarly declines to specify any social roles (Teevan, Ramage, & Morris, 2011). The final article likens hyperlinks to references or citations (Tsagkias, de Rijke, & Weerkamp, 2011).

The disciplinary differences between these two groups of articles in their respective views of hyperlinks are difficult to summarize. Communication and social computing each seems to have its own set of overlapping interpretations which resists easy comparison. Clearer differences emerge, however, when we examine closely each group's justifications for what hyperlinks mean. These fall into four basic categories, which also apply to the other trace data types: *citation*, *empirical findings*, *example*, and *no justification*. These categories are frequently cited and discussed in the argumentation literature (Baesler & Burgoon, 1994; Hoeken, 2001; Reinard, 1988) and differ in their comparative value. Empirical (particularly statistical) evidence is usually preferred to analogic examples because it generalizes more validly (Hoeken, 2001), but the latter are still better than nothing. The value of citation evidence depends upon the credibility and relevance of the cited study, which can obviously vary widely (Reinard, 1988). If

a source is cited to support a claim about the meaning of trace data, the source should be based on firm (ideally empirical) epistemological grounds.

Citation is not only the most common justification practice among the hyperlink studies in communication, it is also the easiest: authors either adopt a previous study's interpretation wholesale or incorporate it into their own synthesis. Every such study reviewed here except Van Aelst and Walgrave (2002) cites at least one other study's claims about the social function of links. Some authors explore at length the various ways other scholars have interpreted hyperlinks (Coddington, 2012; Foot et al., 2003) while others touch on this only briefly (Dimitrova et al., 2005; Gillan, 2009; Halavais, 2000; Papacharissi, 2002; Trammell et al., 2006). Besides length, these articles also likely differ in the similarity between the empirical settings of the cited and citing works—not all preexisting interpretations of hyperlinks will necessarily be valid for the research at hand.

A few studies strengthen their interpretational claims by grounding their justifications in empirical findings. In an early example focusing on personal webpages, Papacharissi notes that "[m]ost links were related directly to the content of the page, so that the links would point to similar content" (2002, p. 652). She uses these data to conclude that "[w]ithin… a personal Web site, appearance was asserted with a variety of social status markers, predominantly hyperlinks" (Papacharissi, 2002, p. 654). Applying similar methods to their study of Polish blogging practices, Trammell et al. find that "[r]ather than being motivated by self-promotion, the bloggers' linking habits are consistent with social utility motivations… outweighing informational ones" (2006, p. 716). In contrast to the etic focus of these two studies, Coddington (2012) takes an emic approach, discovering through interviews how journalists view their own linking practices.

The third justification strategy, example, denotes the use of specific examples, similes, analogies, or similar reasoning in interpreting trace data. Halavais' (2000) aforementioned comparison of hyperlinks to roads, telephone lines, and academic citations falls into this category. Van Aelst and Walgrave suggest "territorial competition" (2002, p. 486) as one possible reason social movement organizations did not link to one another very often. And Matheson labels the journalistic use of hyperlinks a "mesh of authority" (2004, p. 457) that is distributed across news articles and source material.

While all the communication articles included at least some justification for their interpretations of hyperlinks, this was not true of the social computing articles. Justification attempts were few and far between; instead, interpretations are typically stated as plainly as any self-evident fact: "a hyperlink is usually an explicit indicator that one Web page author believes that another's page is related or relevant" (Jamali & Abolhassani, 2006, n.p.); "if a page is referred by many other pages, the relevance of this target page increases" (Bhattarai et al., 2009, n.p.); "by following-up links… [t]he representation of the most influential opinion leaders is therefore feasible" (Bross et al., 2010, n.p.). Seven of the ten social computing articles about hyperlinks either lack either justifications for their interpretations or lack interpretations altogether (those not already mentioned are Chen et al., 2009; Lee et al., 2010; Teevan et al., 2011; and Tsagkias et al., 2011). The remaining studies—Mathioudakis et al. (2010), Ulicny et al. (2010), and Varlamis et al. (2010)—rely solely on citations for justification.

The main difference between the two groups of highly-cited hyperlink articles should now be clear: while each has its own distinctive collection of interpretations, the communication articles justify theirs much more often than did the social computing articles. This is likely reflective of the differing purposes of the two disciplines: communication is largely oriented

toward advancing theory, while social computing is concerned more with prediction and software development. In the case of hyperlinks, many social computing studies have likely internalized the hyperlink interpretations popularized by Google, Yahoo, Technorati, and other services because they yield what many end users perceive as “high-quality” results (Page, Brin, Motwani, & Winograd, 1999).

Twitter followers

Unlike Facebook, Twitter’s default functionality allows users to connect to one another without approval. There is broad agreement among both the communication and social computing studies that the articulated connections created through Twitter have some social significance (boyd & Ellison, 2007), but scholars have identified a range of potential meanings for them. Although communication research on Twitter is still in its infancy, the studies in the current sample have begun to converge around a small set of shared meanings. The notion of diversity is in strong evidence, albeit in a much different form than was observed in the hyperlink studies: while Marwick and boyd (2011) explore the differences in follower relationships that different kinds of users perceive, another contingent of researchers emphasizes the potential for follower counts to be misinterpreted: Moe, for example, warns that "follower count[s] should not be seen as measure of impact” (2012, p. 1233). Taking this idea a step further, both Karpf (2012) and Ausserhofer and Maireder (in press) observe that the widespread perception of follower counts as indicators of popularity creates incentives for unscrupulous users to obtain followers by dishonest means. Karpf formalizes his pessimism about the general enterprise of imputing social meaning to traces as “Karpf’s rule,” which holds that “*Any metric of digital influence that becomes financially valuable, or is used to determine newsworthiness, will become increasingly unreliable over time”* (Karpf, 2012, p. 650, italics in original). A less

pessimistic perspective might exhort researchers to consider possible incentives for bad-faith actors to distort different trace metrics for financial or publicity purposes—as Karpf explains, this particular moral hazard applies strongly to Twitter follower counts.

Potential for manipulation notwithstanding, a second interpretation of following behavior is that followers comprise the communities or audiences of highly-followed individuals and organizations. In their study of the 2012 Eurovision contest, Highfield, Harrington, and Bruns (in press) emphasize Twitter's role as a platform for fans/followers to provide running commentary concerning their preferred candidates. Lovejoy and Saxton (2012) suggest that nonprofit organizations can use Twitter as a tool to attract new members and keep existing ones engaged. Some researchers working along similar lines elide the distinction between "followers" on Twitter and people who "follow" a cause or institution in the traditional sense. When Greer and Ferguson (2011) discuss television "audiences" in a study of local TV stations' Twitter accounts, it is unclear when the term refers to viewers, Twitter followers, or both. Boyle's use of the term "follower" in his study of a Mormon newspaper's tweets (2012) is even more ambiguous given its distinct religious and social media connotations. Similar ambiguities obtain in Bennett and Segerberg (2011) and Papacharissi (2012). However, these omissions are less problematic because Twitter followers do not figure prominently in their arguments.

The evidence suggests that communication scholars justify their interpretations of Twitter followers far less often than for hyperlinks. Over half of the studies reviewed here either offer no justification or no interpretation at all (Bennett & Segerberg, 2011; Boyle, 2012; Greer & Ferguson, 2011; Highfield et al., in press; Lovejoy & Saxton, 2012; Papacharissi, 2012). Among those who do justify their portrayals of followers, Karpf (2012) and Ausserhofer and Maireder (in press) cite the example of falsifying one's follower count as a warning for researchers

interested in measuring influence. Moe (2012) arrives at the same conclusion by citing previous studies. But it is Marwick and boyd (2011) who provide strongest contribution in this area, devoting their entire paper to an emic investigation of Twitter following. They document an important distinction between Twitter users with few and many followers: while the former "typically spoke about friends," the latter "commonly described their audience as 'fans'" (Marwick & Boyd, 2011, p. 118). This is the only communication study in the sample to base its conclusions about the social implications of Twitter following on empirical data.

The social computing studies generally interpret Twitter followers less emphatically than the communication studies when they offered any interpretations at all. The most common strategy was simply to give a very narrow technical definition of following in Twitter and either leave it at that or build upon it later. These definitions often stress the unilateral character of following. For example, "Twitter employs a social-networking model called "following", in which each twitterer is allowed to choose who she wants to follow without seeking any permission" (Weng, Lim, Jiang, & He, 2010, p. 261; see also Bakshy, Hofman, Mason, & Watts, 2011; Golder & Yardi, 2010; Java, Song, Finin, & Tseng, 2007; Krishnamurthy, Gill, & Arlitt, 2008; Kwak, Lee, Park, & Moon, 2010; Wang, 2010). Similar to the communication literature, a few studies point out the pitfalls of equating high follower counts with influence: for example, Yang and Leskovec "find that Twitter users who have the most followers are not the most influential in terms of information propagation" (2010, n.p.; see also boyd, Golder, & Lotan, 2010; Weng et al, 2010). While no study defines influence solely in terms of followers, several incorporate follower counts as one of several components of the concept (Bakshy et al., 2011; Weng et al., 2010; Kwak et al., 2010). Two studies discuss the possibility that following can hold multiple context-specific meanings, and both classified users into categories based on the ratio of

followers to followed (Java et al., 2007; Krishnamurthy et al., 2008). These conclusions represent an etic equivalent of Marwick and boyd's (2011) aforementioned research—indeed, Krishnamurthy et al. (2008) reaches similar conclusions as Marwick and boyd, classifying users either as "broadcasters" or "acquaintances."

Unlike the corresponding hyperlink research, most of the social computing studies on Twitter followers offer some basis for their claims about what the traces mean. For this trace only, they outperform the corresponding communication studies. The single most popular means of substantiating these claims, employed in half the studies reviewed, is inductive investigation. Krishnamurthy et al. (2008) and Java et al. (2007) derive their user categories from quantitative analyses of follower patterns, while boyd et al.'s (2010) observations about Twitter users' "imagined audiences" emerge from crowdsourcing (see also Golder & Yardi, 2010; Yang & Leskovec, 2010). After empiricism, the next most prevalent justification strategy was citation (Suh, Hong, Pirolli, & Chi, 2010; A. H. Wang, 2010; Weng et al., 2010). The remaining two studies simply declare what following means without justification: as a network through which influence manifests (Bakshy et al., 2011, p. 66), or as influence itself (Kwak et al., 2010).

Retweets

"Retweets are not endorsements" is a disclaimer commonly included in Twitter biographies. The communication research reviewed here largely heeds this warning, but it showcases a variety of alternative portrayals that demonstrates the retweet's complexity as a social signal. Most studies recognize its core function of spreading content to new audiences (Anstead & O'Loughlin, 2011; Elmer, 2012; Highfield et al., in press; Larsson & Moe, 2012; Meraz & Papacharissi, 2013; Moe, 2012; Small, 2011). This major exception aside, I found only small pockets of agreement on other social implications of retweeting. As with hyperlinks and

followers, several teams note the diversity of potential interpretations of retweets (Highfield et al., in press; Larsson & Moe, 2012; Meraz & Papacharissi, 2013; Small, 2011). Both Small and Highfield et al. distinguish specifically between "informational" and "conversational" retweets:

> …many manual retweets serve a significantly more conversational function than 'button' retweets, because they can be edited before sending… 'Button' retweets, on the other hand, constitute merely a verbatim passing-along of the original message, but do not enable retweeting users to include any additional comments with the retweeted message. (Highfield et al., in press, p. 8)

Meraz and Papacharissi (2013), citing boyd et al. (2010), point out that retweets can start conversations, amplify other users' voices, and signal that one is listening. At one point Larsson and Moe concur that "the actual meanings of a mention or retweet still need interpretation" but later declare that "retweet activity is crucial as a measure of whose views are made important on Twitter" (2012, p. 739). This view is also shared by Ausserhofer and Maireder, who claim that "the more people mention or retweet a specific account, the more authority is attributed to it" (in press, p. 3). The complete absence of any interpretation was rare, occurring only in Chadwick (2011) and Wilson and Dunn (2011).

A plurality of communication studies decline to justify their claims about what retweets mean. Of the nine studies that offer one or more interpretation, four present them as unsupported statements (Anstead & O'Loughlin, 2011; Elmer, in press; Moe, 2012; Zappavigna, 2011). The remainder were supported by citations and examples alone except for Small (2011), who tests the claims of prior studies about the meaning of retweets against her own empirical findings. In doing so, she specifies how often the retweets in her data served the purpose of conversation and how often they merely rebroadcasted previously-posted content (Small, 2011).

Most of the social computing studies contain a basic definition of retweeting in the course of discussing their findings. While two offer no additional interpretation (Achrekar, Gandhe, Lazarus, Yu, & Liu, 2011; Wu, Hofman, Mason, & Watts, 2011), the rest go further. The most widely-cited paper in this collection is boyd et al., (2010) who undertake a mixed-method study that was one of the first to explore the retweet's diverse set of potential social meanings. Suh et al. (2010) rely heavily on boyd et al. in their discussion of the range of meanings retweets may convey. Three specific interpretations predominate among the remaining articles: interest (Mustafaraj, Finn, Whitlock, & Metaxas, 2011; Vieweg, Hughes, Starbird, & Palen, 2010), trust (Adali et al., 2010; Castillo, Mendoza, & Poblete, 2011), and influence (Bakshy et al., 2011; Kwak et al., 2010). These functions overlap to some degree—people are unlikely to allow themselves to be influenced by someone they are uninterested in and do not trust—and yet they are clearly distinct.

Solid theoretical and/or empirical foundations would help us distinguish between competing claims about what retweets mean, but they are unfortunately in short supply among the social computing studies—seven of the ten offer no justifications for their interpretations. Among the remaining papers, boyd et al. (2010) base their conclusions on qualitative research, and Suh et al. (2010) concur by citing that paper. Somewhat uncharacteristically, Bakshy et al. (2011) discuss in depth the relationship between digital traces of redistribution and the expansive concept of influence. The passage is worth quoting at length:

> ...the type of influence we study here is of a rather narrow kind: being influenced to pass along a particular piece of information. As we discuss later, there are many reasons why individuals may choose to pass along information other than the number and identity of the individuals from whom they received it—in particular, the nature of the content itself.

> Moreover, influencing another individual to pass along a piece of information does not necessarily imply any other kind of influence, such as influencing their purchasing behavior, or political opinion. (2011, p. 68)

Such depth of explanation of the limits of digital measurement was not especially common among the communication studies and extremely rare in social computing across all three traces.

Discussion

This study analyzed a sample of highly-cited articles in communication and social computing to identify trends in how they interpret and justify their interpretations of digital trace data. Two broad answers to these research questions can now be presented: first, each form of trace data has been interpreted in multiple ways; and second, scholars use four main strategies to justify their interpretations, not all of which are equally valid. What matters most for the current purposes is not so much what those interpretations are specifically, but rather their number and variety. The fact that there is no consensus about how any of these data should be interpreted among even this relatively small set of studies ought to color our perceptions of studies that hold one interpretation above all others. The broader the data set, the more likely it is that one interpretation will not fit all. At the same time, we must ask whether it is enough simply to observe that a given trace has multiple possible meanings before conducting a study. A logical follow-up question would be: how frequently does each type of use appear in the data? In the absence of empirical research on this question, we can only guess as to the answers.

The second answer concerns the main trends in how the ostensible meanings of digital traces are justified. Scholars of argumentation generally rank empirical evidence above examples in terms of evidence quality, with the value of citation depending on the relevance of the source to the case at hand (Baesler & Burgoon, 1994; Hoeken, 2001; Reinard, 1988). They further

agree, at least implicitly, that any reason of at least minimal validity is preferable to no reason at all. But this study found that substantial proportions of articles from both disciplines failed to justify the social implications they imputed to trace data. This occurred in both the communication and social computing samples but more extensively in the latter. While claims that traces represent influence, trust, credibility, etc. may sound intuitive, they need convincing support. Citation was the most popular justification strategy across the board, and while it certainly surpasses nothing, it is not always ideal—the cited studies may not adequately support their own interpretations. Scholars should thus take care to cite only well-grounded works of clear relevance to their research. This is particularly important given that digital traces can convey so many disparate social signals.

The most encouraging efforts in justifying trace data interpretations were those that included example-based logical arguments and/or adduced empirical data. Karpf (2012) and Bakshy et al. (2011) in particular offer very thoughtful considerations of what Twitter followers and retweets respectively might mean in terms of the abstract, elusive concept of "influence." Though neither of these is based on empirical data, they shine as well-reasoned cautionary notes about the difficulties of imputing social value to digital traces. Hoeken notes that examples can offer high-quality evidence when the fit between the example and the argument is strong (2001, p. 152), and such reasoning is particularly important to help guide early empirical work on new kinds of traces. But the gold standard in establishing valid meanings of trace data remains empirical research, and boyd and colleagues occupy the forefront of this contingent—the two studies singled out as examples in the above sections are both highly cited, and deservedly so (boyd et al., 2010; Marwick & boyd, 2011). These studies take the rare step of actually asking social media users how they interpret retweets and Twitter followers respectively, identifying a

variety of meanings. Similarly, one prong of Coddington's multi-method study on journalists' views of links on Twitter asked them directly what they thought, adding emic depth to claims already firmly established in theory and etic observation (Matheson, 2004; Robinson, 2006; Wall, 2005).

The emic/etic distinction is especially relevant to the interpretation of trace data and warrants further discussion. Aside from the few studies that ask users directly about their perceptions of traces (boyd et al., 2010; Coddington, 2012; Golder & Yardi, 2010; Marwick & boyd, 2011), the empirical research reviewed here mostly takes an etic approach in assessing what traces mean. While such work has merit, it should be balanced by approaches that pursue two different types of emic questions about trace data. The first asks how senders perceive their own messages, or how they would like those messages to be interpreted. This is what Coddington (2012) and boyd et al. (2010) sought to answer in asking their respective participants what their links and retweets meant, respectively. The second emic question asks about recipients or audiences' perceptions of traces, which can obviously differ from those of senders. The only study in the current collection that came close to addressing this second question was Golder and Yardi's (2010), which found that Twitter users tend to view high follower counts as status symbols. This happens to jibe with boyd et al.'s (2010) finding that well-followed Twitter users see their followers as fans, but this will not be the case for all traces. An important goal for future research, then, is to mine the potential gaps between the two questions to discover when the answers coincide and when they diverge.

Conclusion: What communication can learn from social computing and vice versa

Given the distinctions established in this study between communication and social computing, a fitting concluding question concerns lessons each field stands to learn from the

other. This study found the research in its social computing sample to be largely quantitative and light on theory, and as such relatively unlikely to interpret the traces it examines (with the exception of Twitter followers). At the same time many such studies concern themselves with concepts long of interest to communication scholarship such as trust, relevance, and especially influence. Authors interested in these and similar concepts should strongly consider drawing on qualitative communication research methods (particularly interviews and close textual readings) to add validity to their operationalizations. While long-standing theoretical definitions of influence and the like will probably be too broad for most studies, the simple practice of substantiating which roles digital traces are and are not claimed to perform will help clarify the relevance and breadth of the empirical conclusions. Subsequent studies can expand upon the definitions of earlier ones, revising and correcting as appropriate. This would not constitute theory-building for its own sake—business, government, and the nonprofit sector all have interests in effectively linking digital traces with high-value concepts at scale (Manyika et al., 2011).

The primary lessons communication would do well to learn from social computing are also methodological. At the moment, the ability to retrieve and analyze mass quantities of trace and other forms of digital data are sparsely distributed among communication researchers, for whom computer programming is not a traditional research method. Aside from the technical barrier this situation presents to scholars who would otherwise have interesting things to say about digital traces, those who have the required skills have little access to discipline-specific methodological best practices and often must develop their own solutions. On the one hand, such "kludgy" methods can help researchers complete their work faster, but over the long term, methodological standards—even loose ones—will save them time and help them avoid

procedural missteps (Karpf, 2012). Communication scholars should not import the standards of social computing wholesale, but they do provide a valuable starting point.

The study of digital trace data is still in its infancy, and its level of sophistication will undoubtedly increase with time. Communication and social computing will doubtless maintain their own distinct approaches, which is appropriate given the differences in their foundational mandates. But the two fields have much to learn from one another, and each should continue to draw on the other's strengths when called for. Digital traces collectively constitute an enormous wellspring of diverse data types fed by countless platforms and services—no single discipline should attempt to navigate it alone.

References

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., & Liu, B. (2011). Predicting Flu Trends using Twitter data. In *Proceedings of the First International Workshop on Cyber-Physical Networking Systems* (pp. 702 –707). doi:10.1109/INFCOMW.2011.5928903

Adali, S., Escriva, R., Goldberg, M. K., Hayvanovych, M., Magdon-Ismail, M., Szymanski, B. K., … Williams, G. (2010). Measuring behavioral trust in social networks. In *Proceedings of the 2010 IEEE International Conference on Intelligence and Security Informatics (ISI)* (pp. 150 –152). doi:10.1109/ISI.2010.5484757

Anstead, N., & O'Loughlin, B. (2011). The Emerging Viewertariat and BBC Question Time Television Debate and Real-Time Commenting Online. *The international journal of press/politics*, *16*(4), 440–462.

Ausserhofer, J., & Maireder, A. (in press). National Politics on Twitter. *Information, Communication & Society*, *0*(0), 1–24. doi:10.1080/1369118X.2012.756050

Baesler, E. J., & Burgoon, J. K. (1994). The Temporal Effects of Story and Statistical Evidence on Belief Change. *Communication Research*, *21*(5), 582–602. doi:10.1177/009365094021005002

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 65–74). New York, NY, USA: ACM. doi:10.1145/1935826.1935845

Bennett, W. L., & Segerberg, A. (2011). Digital media and the personalization of collective action. *Information, Communication & Society*, *14*(6), 770–799. doi:10.1080/1369118X.2011.579141

Bhattarai, A., Rus, V., & Dasgupta, D. (2009). Characterizing comment spam in the blogosphere through content analysis. In *IEEE Symposium on Computational Intelligence in Cyber Security 2009* (pp. 37 –44). doi:10.1109/CICYBS.2009.4925088

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).

Boyd, D. M., & Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230. doi:10.1111/j.1083-6101.2007.00393.x

Boyle, K. (2012). Latter-Day Tweets: The Mormon Times's Use of Twitter as a Reporting Tool. *Journal of Media and Religion*, *11*(4), 189–199. doi:10.1080/15348423.2012.730320

Bross, J., Quasthoff, M., Berger, P., Hennig, P., & Meinel, C. (2010). Mapping the Blogosphere with RSS-Feeds. In *2010 24th IEEE International Conference on Advanced Information Networking and Applications (AINA)* (pp. 453 –460). doi:10.1109/AINA.2010.95

Bryant, J., & Miron, D. (2004). Theory and Research in Mass Communication. *Journal of Communication, 54*(4), 662–704. doi:10.1111/j.1460-2466.2004.tb02650.x

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web* (pp. 675–684). doi:10.1145/1963405.1963500

Chadwick, A. (2011). The Political Information Cycle in a Hybrid News System: The British Prime Minister and the "Bullygate" Affair. *The International Journal of Press/Politics, 16*(1), 3–29. doi:10.1177/1940161210384730

Chen, F., Tan, P.-N., & Jain, A. K. (2009). A co-classification framework for detecting web spam and spammers in social media web sites. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 1807–1810). doi:10.1145/1645953.1646235

Coddington, M. (2012). Building Frames Link by Link: The Linking Practices of Blogs and News Sites. *International Journal of Communication, 6,* 2007–2026.

Dimitrova, D. V., Kaid, L. L., Williams, A. P., & Trammell, K. D. (2005). War on the Web The Immediate News Framing of Gulf War II. *The Harvard International Journal of Press/Politics, 10*(1), 22–44.

Elmer, G. (in press). Live research: Twittering an election debate. *New Media & Society*. doi:10.1177/1461444812457328

Foot, K., Schneider, S. M., Dougherty, M., Xenos, M., & Larsen, E. (2003). Analyzing linking practices: Candidate sites in the 2002 US electoral Web sphere. *Journal of Computer-Mediated Communication, 8*(4).

Gillan, K. (2009). The UK anti-war movement online. *Information, Communication & Society*, *12*(1), 25–43.

Golder, S. A., & Yardi, S. (2010). Structural Predictors of Tie Formation in Twitter: Transitivity and Mutuality. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on* (pp. 88 –95). doi:10.1109/SocialCom.2010.22

Graber, D. A., & Smith, J. M. (2005). Political Communication Faces the 21st Century. *Journal of Communication*, *55*(3), 479–507. doi:10.1111/j.1460-2466.2005.tb02682.x

Greer, C. F., & Ferguson, D. A. (2011). Using Twitter for Promotion and Branding: A Content Analysis of Local Television Twitter Sites. *Journal of Broadcasting & Electronic Media*, *55*(2), 198–214. doi:10.1080/08838151.2011.570824

Halavais, A. (2000). National Borders on the World Wide Web. *New Media & Society*, *2*(1), 7–28. doi:10.1177/14614440022225689

Highfield, T., Harrington, S., & Bruns, A. (in press). Twitter as a Technology for Audiencing and Fandom. *Information, Communication & Society*. doi:10.1080/1369118X.2012.756053

Hoeken, H. (2001). Convincing citizens: The role of argument quality. In D. Janssen & R. Neutelings (Eds.), *Reading and writing public documents* (pp. 147–169). Philadelphia: Benjamins.

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity Issues in the Use of Social Network Analysis with Digital Trace Data. *Journal of the Association for Information Systems*, *12*(12), 767–797.

Jamali, M., & Abolhassani, H. (2006). Different Aspects of Social Network Analysis. In *IEEE/WIC/ACM International Conference on Web Intelligence, 2006.* (pp. 66 –72). doi:10.1109/WI.2006.61

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD workshop on Web mining and social network analysis* (pp. 56–65). doi:10.1145/1348549.1348556

Karpf, D. (2012). Social science research methods in Internet time. *Information, Communication & Society*, *15*(5), 639–661. doi:10.1080/1369118X.2012.665468

Krishnamurthy, B., Gill, P., & Arlitt, M. (2008). A few chirps about twitter. In *Proceedings of the first workshop on Online social networks* (pp. 19–24). doi:10.1145/1397735.1397741

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web* (pp. 591–600). Retrieved from http://dl.acm.org/citation.cfm?id=1772751

Larsson, A. O., & Moe, H. (2012). Studying political microblogging: Twitter users in the 2010 Swedish election campaign. *New Media & Society*, *14*(5), 729–747. doi:10.1177/1461444811422894

Lee, K., Caverlee, J., & Webb, S. (2010). Uncovering social spammers: social honeypots + machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 435–442). New York, NY, USA: ACM. doi:10.1145/1835449.1835522

Lovejoy, K., & Saxton, G. D. (2012). Information, Community, and Action: How Nonprofit Organizations Use Social Media*. *Journal of Computer-Mediated Communication*, *17*(3), 337–353. doi:10.1111/j.1083-6101.2012.01576.x

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*, 1–137.

Marwick, A., & Boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133. doi:10.1177/1461444810365313

Matheson, D. (2004). Weblogs and the Epistemology of the News: Some Trends in Online Journalism. *New Media & Society*, *6*(4), 443–468. doi:10.1177/146144804044329

Mathioudakis, M., Koudas, N., & Marbach, P. (2010). Early online identification of attention gathering items in social media. In *Proceedings of the third ACM international conference on Web search and data mining* (pp. 301–310). doi:10.1145/1718487.1718525

McMillan, S. J. (2000). The Microscope and the Moving Target: The Challenge of Applying Content Analysis to the World Wide Web. *Journalism and Mass Communication Quarterly*, *77*(1), 80–98.

Meraz, S., & Papacharissi, Z. (2013). Networked Gatekeeping and Networked Framing on #Egypt. *The International Journal of Press/Politics*. doi:10.1177/1940161212474472

Moe, H. (2012). Who Participates and How? Twitter as an Arena for Public Debate about the Data Retention Directive in Norway. *International Journal of Communication*, *6*, 1222–1244.

Mustafaraj, E., Finn, S., Whitlock, C., & Metaxas, P. T. (2011). Vocal Minority Versus Silent Majority: Discovering the Opionions of the Long Tail. In *2011 IEEE third international conference on social computing (socialcom)* (pp. 103 –110). doi:10.1109/PASSAT/SocialCom.2011.188

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: bringing order to the web. Retrieved from http://ilpubs.stanford.edu:8090/422

Papacharissi, Z. (2002). The Presentation of Self in Virtual Life: Characteristics of Personal Home Pages. *Journalism & Mass Communication Quarterly*, *79*(3), 643–660. doi:10.1177/107769900207900307

Papacharissi, Z. (2012). Without You, I'm Nothing: Performances of the Self on Twitter. *International Journal of Communication*, *6*, 1989–2006.

Reinard, J. C. (1988). The Empirical Study of the Persuasive Effects of Evidence The Status After Fifty Years of Research. *Human Communication Research*, *15*(1), 3–59. doi:10.1111/j.1468-2958.1988.tb00170.x

Robinson, S. (2006). The mission of the j-blog Recapturing journalistic authority online. *Journalism*, *7*(1), 65–83.

Small, T. A. (2011). What the Hashtag? *Information, Communication & Society*, *14*(6), 872–895. doi:10.1080/1369118X.2011.554572

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010). Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *2010 IEEE Second International Conference on Social Computing (SocialCom)* (pp. 177 –184). doi:10.1109/SocialCom.2010.33

Teevan, J., Ramage, D., & Morris, M. R. (2011). #TwitterSearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 35–44). doi:10.1145/1935826.1935842

Trammell, K. D., & Keshelashvili, A. (2005). Examining the New Influencers: A Self-Presentation Study of A-List Blogs. *Journalism & Mass Communication Quarterly*, *82*(4), 968–982. doi:10.1177/107769900508200413

Trammell, K. D., Tarkowski, A., Hofmokl, J., & Sapp, A. M. (2006). Rzeczpospolita blogów [Republic of Blog]: Examining Polish Bloggers Through Content Analysis. *Journal of Computer-Mediated Communication*, *11*(3), 702–722. doi:10.1111/j.1083-6101.2006.00032.x

Tsagkias, M., de Rijke, M., & Weerkamp, W. (2011). Linking online news and social media. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 565–574). doi:10.1145/1935826.1935906

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

Van Aelst, P., & Walgrave, S. (2002). New media, new movements? The role of the internet in shaping the “anti-globalization” movement. *Information, Communication & Society*, *5*(4), 465–493. doi:10.1080/13691180208538801

Varlamis, I., Eirinaki, M., & Louta, M. (2010). A Study on Social Network Metrics and Their Application in Trust Networks. In *2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 168 –175). doi:10.1109/ASONAM.2010.40

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1079–1088). doi:10.1145/1753326.1753486

Wall, M. (2005). Blogs of war. *Journalism*, *6*(2), 153–172.

Wang, A. H. (2010). Don't follow me: Spam detection in Twitter. In *Proceedings of the 2010 International Conference on Security and Cryptography (SECRYPT)* (pp. 1 –10).

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.

Weare, C., & Lin, W. Y. (2000). Content analysis of the world wide web. *Social Science Computer Review*, *18*(3), 272–292.

Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010). TwitterRank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining* (pp. 261–270). doi:10.1145/1718487.1718520

Wilson, C., & Dunn, A. (2011). Digital Media in the Egyptian Revolution: Descriptive Analysis from the Tahrir Data Sets. *International Journal of Communication*, *5*.

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web* (pp. 705–714).

Yang, J., & Leskovec, J. (2010). Modeling Information Diffusion in Implicit Networks. In *2010 IEEE 10th International Conference on Data Mining (ICDM)* (pp. 599 –608). doi:10.1109/ICDM.2010.22

Zappavigna, M. (2011). Ambient affiliation: A linguistic perspective on Twitter. *New Media & Society*, *13*(5), 788–806. doi:10.1177/1461444810385097

Footnotes

[1] This paper does not discuss digital text, because while it is technically a form of trace data, many articles and books have already been devoted to its study (see e.g. McMillan, 2000; Weare & Lin, 2000).

[2] This study reviews 11 retweet articles rather than ten because there were exactly 11 articles among the five journals selected for the retweet trace that met the keyword criteria.