

Click and Mortar: Organizations on the Web*

Walter W. Powell, Aaron Horvath, Christof Brandtner
Stanford University

Forthcoming, *Research in Organizational Behavior*

Final draft

Contact info: woodyp@stanford.edu, ahorvath@stanford.edu, cbrandtn@stanford.edu

Abstract: The webpages of organizations are both a form of representation and a type of narrative. They entertain, persuade, express a point of view, and provide a means to organize collective action and economic exchange. Increasingly, webpages are the primary point of access between an organization and its environment. An organization's online presence offers a major new source of rich information about organizations. In this paper, we develop three perspectives on websites from an organizational point of view: as identity projects, tools, and relational maps. We draw on data from the online and offline presences of "brick and mortar" nonprofit organizations in the San Francisco Bay Area to both illustrate how a digital transformation shaped these organizations and identify a host of new methods that can be used to study organizations using webpages. Finally, we reflect on both the strengths of these new sources of data as well as possible limitations and conclude with theoretical implications for organizational scholars.

*We are very grateful to members of the Networks and Organizations workshop at Stanford University for feedback on an early draft. We also thank Diane Burton, Art Brief, and Barry Staw for their comments. Peter Monge at USC assigned the paper to his graduate class and we benefited from their responses. Powell is indebted to Achim Oberg for his introduction to the many uses of websites as data sources. We thank Catherine Gray, Sebastian Schuster, and Jacob Waggoner for excellent research assistance, and the Stanford Center on Philanthropy for research support.

Introduction

Today, websites are a primary point of access and communication between consumers and organizations. This relationship is true not just for Amazon and Alibaba – for whom webpages are the central point of contact – but for charities, social movements, schools, and government agencies as well. Activities as diverse as scheduling an appointment, checking your bank account, offering a donation, applying for a job, or making a purchase are done through organizational websites. Webpages, which we take to be the corpus of an organization’s online presence, represent an important new source of rich information on organizations. They are much more detailed than annual reports, and are read and used much more widely. Due to their interactive nature, webpages are more “live” than most other forms of organizational communication. The use of websites can be verified through analytics data, which permit measurement of how often a website is accessed and from where. Consequently, we think it is incumbent on organizational scholars to develop new methods and concepts to analyze the contents of webpages, and to consider the ways in which organizations are changing due to their online presence.

As with any new idea or technology, as it becomes available and groups adopt it, the meaning and functions of the technology change, and the nature of the problems it was designed to solve expand. Consequently, new uses develop. Webpages are ubiquitous. They are both a form of representation and a type of narrative. Like narratives, they have many purposes (Espeland and Sauder, 2016). They inform, entertain, persuade, and express a point of view. Webpages display a distinctive design, which may tell us a good deal about the structure and functioning of organizations. They contain varying amounts of text and visual content, and this content suggests important differences between, say, a formal bureaucracy and a grassroots social movement. Webpages are also a means to organize collective action, and have been widely used in various forms of political participation (Bennett and Segerberg, 2013). In turn, these new participatory practices and new forms of contribution are, in some cases, creating new types of workplaces, with novel points of contact with the external environment (Contractor, 2013).

The internet is a powerful, transformative technology. Through it, organizations that were once tethered locally become accessible globally. The web also enables individuals to access organizations and participate in them well beyond the boundaries of their local communities. This access has profound implications for the spread of ideas both to and from organizations. To be sure, similar changes accompanied the advent of other major new technologies. The emergence of the postal system, newspapers, radio, telephone, and television had considerable influence over how organizations learned, communicated, and traded. As these technologies progressed, they variously enabled organizations – and their audiences – to cross boundaries of time, space, and knowledge.

The web, however, may be even more dramatic in its ramifications for social science research than these earlier technologies. Science makes progress not only through discovery and ideas, but also through the development of new tools and methods (Kuhn, 1962). Network scientist Duncan Watts (2011: 266) has argued that mobile, web, and internet communications are the equivalent of the telescope for the social sciences as they have “the potential to revolutionize our understanding of ourselves and how we interact.” Economist Shane Greenstein (2015; 419-42) contends the web is different from other technologies because it has fostered “innovation from the edges,” with inputs originating from multiple places through dispersed decision-making. Sociologists Golder and Macy (2014:130) suggest that studies of “digital footprints collected from online communities and networks enable us to understand human behavior and social interaction in ways we could not do before.”

The development of organizational theory also follows a trajectory of boundary crossing. Weber’s ideal-type bureaucracy strove to be hermetically sealed from much external influence. Thompson’s (1967) classic treatise emphasized how organizations tried all manner of strategies to “buffer” themselves from their external environments. As organization theory progressed, it began to engage with the external environment: first, in the Carnegie School tradition as an information processing problem to be solved by the organization, and later as the source of resources that organizations attempt to control and manage (Pfeffer and Salancik, 1978). Ensuing research on inter-organizational networks and

studies of organizational fields and populations further transformed our understanding of organizational boundaries.

Organizational theory, by its very nature, has long been grounded in the material and technical reality of the time period in which ideas were developed. Technological change has played an important role in both the evolution of organizations and organizational theory. Chandler (1977), for example, argued that the large industrial firm came about because an explosion of new faster communication and transportation technologies made it more efficient to organize within firms, thereby giving rise to the salaried manager. But new technologies do not drive the evolution of organizations toward some teleological end-point; they interact with organizations in complex ways. Digital technologies have enabled open communities such as Linux and Wikipedia (Benkler, 2011; Jemielniak, 2014), facilitated cross-departmental and cross-national coordination within organizations (Kellogg, Orlikowski, and Yates, 2006; Hinds, Liu and Lyon 2011), and permitted protestors to challenge impenetrable authoritarian regimes (Castells, 2012; Tufekci and Wilson, 2012). In so doing, some theorists speculate that the internet appears to have induced a flattening of hierarchy in many organizational forms (Castells, 2000; Benkler, 2002; von Hippel, 2005). These democratizing effects – both internally and externally – foster new modes of collaboration and production, and reduce barriers for new organizations to form. At the same time, these tools may well help to reproduce power and inequality through unequal access to them (Hargittai, 2010).

We argue that organizations scholars should pay more attention to the digital economy in general, and to organizational webpages in particular, as both empirical phenomena and data on organizational behavior. To illustrate the potential uses of webpages for researchers, we draw on a longitudinal research project on San Francisco Bay Area nonprofit organizations. We have rich data spanning the years 2000 to the present. In our first round of interviews, approximately 30% of the organizations had webpages and even fewer had interactive ones that could be used for making a donation, scheduling an appointment, or offering opinions about services. Today, more than 85% of the remaining organizations have webpages,

with numerous interactive features. We use our study as a unique opportunity to see how a digital transformation shaped these organizations, and to make the case that webpages offer revealing data to organizations scholars.

We begin with a brief discussion of our sample and the varied methods we employed to study the organizations. We compare different sources of data and methods in terms of their ease of use, the depth of information they offer, and how they may inform organizational research. We then discuss three aspects of webpages that offer promising insights into organizational behavior, using features of the organizations in our sample to illustrate how they react differently to the web. As we proceed, we reflect on how central ideas in organizational analysis, including convergence and divergence, decoupling, and organizational boundaries, are illuminated by analysis of webpages. We discuss additional methods for analyzing web-based data, and the tools that are presently available. We also reflect on limitations of webpage data.

The San Francisco Bay Area Nonprofit Sector

Put simply, the internet “happened” to the organizations in our sample. Roughly two-thirds of them existed prior to 1994, approximately a third were founded between 1994 and 2000 after the internet became more publically available with the advent of Netscape (for a chronology of the internet’s development, see Greenstein, 2015:12-30). Over the course of the study, the organizations in our sample took to the web at an impressive rate – an average of nine new websites a year – bringing the total number of websites from 61 out of 200 in 2002 to 151 out of 172 in 2012.¹ Whereas having a website was the exception early in this century, not having one is the anomaly now. Figure 1 illustrates the adoption timeline for our sample of organizations.

¹ The data were collected by manually inspecting IRS 990 forms for each of the organizations in the sample between 2002 and 2012. Starting in 2001, all versions of the 990 form, there is a field for the organization to list its website URL. We counted the first year a URL appeared as the birth year of the organization’s website. Several organizations – many with websites – listed nothing in this field. We therefore consider this estimate of web growth to be conservative.

(Figure 1 here)

From 2003 to 2005 our team of researchers analyzed a representative sample of 501(c)3 charitable organizations operating in the San Francisco Bay Area region, which includes the urban areas of San Francisco, Oakland, and San Jose, the suburbs of Marin and San Mateo counties, and formerly rural counties of Napa, Solano and Sonoma. Organizations exempt from tax under section 501(c)3 of the Internal Revenue Service code are operating charitable nonprofits and eligible to receive tax- deductible contributions. Using data from the U.S. Internal Revenue Service, digitized by the National Center for Charitable Statistics (NCCS), we identified the population of IRS form 990 filers in the year 2000, which was the most complete recent year available back then. There were 7,106 operating charities in the ten county area, from which we drew a random sample.

We approached these organizations first by mail, then followed up by telephone and email. We eventually contacted 264 viable organizations to develop a sample of 200 operating charities, resulting in a response rate of 76%. The organizations in the sample were typically rather small, with approximately half having annual budgets of less than \$200,000. Nevertheless, some were quite large with budgets in the tens of millions. The sample organizations, as with the full Bay Area population, came primarily from four areas: human services (37%), education (21%), arts and culture (14%), and health (11%).

In the initial study we conducted extensive face-to-face interviews from 2002 to 2004 with the executive directors of staffed organizations or with the board presidents of volunteer based organizations. The interviews were conducted using a standard interview protocol, and lasted approximately 90-120 minutes. The interviews were conducted at the offices of the nonprofits, where interviewers could see the leaders in their work surroundings. We used additional information from IRS form 990s for corroboration, which provided financial data regarding expenses and funding sources. The primary research method was survey research, and the resulting publications were based on quantitative analyses of survey data, and selective use of quotes from the lengthy interviews (Hwang and Powell, 2009; Suarez, 2010; Bromley, Hwang, and Powell, 2014).

A decade later we followed up on these organizations. Only seventeen no longer existed, while another eleven underwent significant changes. Two of these became private foundations, one was absorbed by the public sector, three moved out of state, and five merged with or were absorbed by other organizations. The overall resilience of the sample might seem remarkable, given the intervening economic turmoil, but when times get difficult, nonprofit service agencies get busy. They do not fold up shop when there are people in need. We drew on a variety of data sources to study the subsequent development and evolution of these organizations. We administered an online, follow-up survey in 2014, which had a response rate of 65%. Such a survey would not have been possible in 2004. The respondents, the executive directors or presidents of their respective organizations, were asked questions regarding their education, training, work background, tenure in their current post, and changes in their organization over the previous ten years. Once again, we made use of administrative and financial data from the IRS form 990, accessed digitally via NCCS, for fiscal years 2002-2012.

We also encountered rich sources of public documents that were unavailable for the full sample just a decade earlier. We analyzed webpages and social media profiles. The tax filings of the organizations are now accessible on the web. Interestingly, these tax filings no longer represent only a source of financial and administrative data, but have become more public-facing representations of the organizations, including much more information about their mission and activities, and increasingly referring to performance and impact in quantified terms. We used LinkedIn to access resumes and obtain career histories for many of the executive directors; previously we had a difficult time asking directors for their resumes at the close of interviews. We also conducted in-depth interviews in 2015 with a small subset of executive directors. Because the information available through webpages is now so abundant, we did not feel the need to do full-scale interviews with the entire sample of directors of the organizations. Instead we used the interviews to contextualize the data we obtained from surveys, websites, and tax filings.

The San Francisco Bay Area is a fertile region for nonprofits, with a greater per-capita presence than in other metropolitan areas of California. The Bay Area environment also has numerous private foundations, wealthy community foundations, and a new generation of generous donors. The sample we obtained, however, matches the overall U.S. population of nonprofits in terms of age and activities. But it is widely known that the Bay Area is a leading edge of experimentation with new ideas. The region has a storied history of activism and radical politics, which lead to the creation of rights and environmental organizations early in the 20th century and to AIDS activism in the latter part of the century. Today, San Francisco Bay Area organizations of all stripes tend to be initiators and early adopters of novel ideas and practices. Many of these have spread widely. Thus it is likely that what is established in this region may well be new in other locales. We think, however, that organizational representation on the web is by no means unique to this population of organizations.

Strengths and Weaknesses of Diverse Data Sources

The use of different data sources and methods, including administrative records, ethnography, and surveys, is well-established in the study of organizations. Websites, on the other hand, invite skepticism and claims that online material is “less real” than what may be learned in person or through financial documents. We argue to the contrary. In the course of our research, we used a variety of methods, including quantitative analysis of administrative and fiscal records, interviews with top leaders, online surveys, and website analysis. These different analytical strategies revealed disparate features of organizations that, when put in dialogue, provide novel insights. Indeed, each strategy comes with its own strengths and weaknesses. They vary both with respect to the ease and quality of data collection and how they shed light on organizations. Furthermore, these data sources are not always in agreement. A website can detail what an executive director would never share in person. Instead of treating conflict or congruence among data sources as a flaw, we see it as a means of grappling with the complex character of organizations.

To better understand the role of webpages in the quiver of analytical approaches, we consider their correspondence with other data sources used in the course of our research. We focus on several key dimensions: the type of data provided by the source and its relative thick- or thinness; the ability of independent researchers to arrive at similar conclusions from the same information; the accessibility of data; the initial and marginal costs of data collection; the comparability of data across cases; and the capacity to examine data longitudinally. For the most part, the data sources are complementary. Each has practical advantages and shortcomings that others do not. As we demonstrate below, these data sources are conceptually complementary as well. Each provides insights that others do not. The discussion that follows is based on our experiences collecting data for this sample. Building from this, we generalize beyond our sample to discuss the respective pluses and minuses for each data source. To this end, Table 1 summarizes our arguments.

(Table 1 here)

Administrative data. Our initial sample was drawn from IRS tax records. These data were advantageous for providing the full population of San Francisco Bay Area nonprofit organizations, from which we drew a random sample. Beyond sampling, the specific form 990 tax filings provide somewhat thin data, covering a breadth of topics but in varying detail. The forms list expenses and grants received, but do not specify from whom the grants were obtained. Earned income data are provided, but no details are given specifying the activities that generated the revenues. Comparisons across different organizations are easily done for the same calendar year, but when the submitted information used different tax forms (forms 990 and 990-EZ), this compromised detailed comparative and longitudinal analyses. The tax forms offer limited data on personnel, only occasionally reporting salaries of the five highest paid employees of the organization. On the other hand, because we were able to obtain information for all the organizations in the sample, replication was simple.

Today, public access to digitized tax filings is possible through the National Center for Charitable Statistics at the Urban Institute, which provides aggregate datasets for 1989 through 2013 and a

searchable database of scanned and e-filed 990 tax filings. Longitudinal comparisons can be challenging, however, especially because tax filing rules change over time, and data formats can change markedly. Such digitized data are prone to error. Specific to nonprofits, administrative data can be misleading. What appear to be irregular fluctuations in revenue may actually be normal funding and grant cycles. Thus, the costs of cleaning data and making multi-year comparisons are considerable.

Our use of nonprofit tax filings is similar to other administrative data among organizations scholars. The *Yearbook of International Organizations*, maintained by the Union of International Associations, provides basic information on intergovernmental and international non-governmental organizations (Schofer and Longhofer, 2011). The popular COMPUSTAT database furnished by Standard and Poor's is a veritable gold mine for researchers interested in either the financial performance of global companies or organizational demographics (Zuckerman, 1999). At the same time, researchers looking for thicker data on organizations would be left empty-handed by such data.

Interview data. To build on the administrative data in the early period of our study, we arranged for in-depth interviews with the executive leaders of the 200 organizations. These interviews were highly labor-intensive. Even with a team of four dedicated researchers and the principal investigator, it took over two years to set up, conduct, and transcribe the interviews. The salary costs were steep, but they produced remarkably detailed data. Once we had the interview protocol done through pilot testing and trained the interviewers, comparability across interviews was relatively easy, unless the interviewee proved to be a loose cannon. Clearly, the start-up costs of this kind of interviewing are steep, and, in our case without generous support from Stanford Graduate School of Business, simply not possible. Even though the interview transcripts were invaluable, another team of researchers from another university would find the open-ended aspects of the interviews difficult to replicate.

Replicating this interview effort a decade later would have been difficult. Although nearly half of the directors we spoke with in 2003 and 2004 remain in their positions, many organizations had seen at least one new person in this position. Some organizations, such as PTAs, experienced turnover annually.

Even more problematic, a second round of interviews would have been cost-prohibitive. Thus, we looked for other means to collect follow-up information.

Surveys. In 2014, we conducted a survey with our original sample using a Qualtrics online questionnaire. This was fast and inexpensive, and even though the 65 percent response rate was not as exceptional as with the initial interviews, it was very high by the standards of survey research (Baruch and Holtom 2008). We spent two months with letters, emails, and phone calls to urge executives to take the surveys. Once the data were obtained, it was easy to transform answers to numerical queries, and we used Dedoose (a qualitative coding software) to analyze the open-ended questions. The attention and detail that nonprofit leaders gave to open-ended questions was exceptional, but there is no comparison of responses to online questions to the depth of the spoken word. Interviews provide far richer data. On the other hand, comparisons of open-ended questions on the online survey were easier due to comparable length and control.

Surveys suffer from similar limitations as interviews. Directing our questions at leaders invites executive bias. Responses from employees and staff may have provided very different perspectives. Turnover, as discussed, may compound this problem over time, making comparability a challenge. Asking questions about relatively concrete aspects of the organization, however, mitigates this bias to some degree. For example, we asked a series of questions about professional practices, including whether or not the organization is volunteer run or has a paid, full-time director. Other questions, such as open-ended inquiries about challenges the organization faced, may be subject to greater variability.

Webpages. One line of questioning in our follow-up survey asked how organizations engaged with the internet. Reviewing both closed and open-ended responses, it became apparent that webpages played a different role for organizations than they had just 10 years before. As mentioned, the percent of organizations with websites tripled. More than merely being present online, organizations were actively engaged online. They regularly maintained and updated websites, were active users of social media, and responded to reviews and criticisms on review sites such as Yelp. Significant energy was being put into a

new sphere of organizational activity, and evidence about how nonprofits engaged with their environments and constituents was readily available. To follow up on illuminating responses about web use in our survey, we inspected how organizations presented themselves online. In the following section, the strategies used to analyze these sites, including quantitative content analysis and factor analysis, are detailed.

Compared with the other sources of data, webpages have a different character. Instead of being characterized by a standard format, they offer a plethora of data collection options. They can be used for content analyses, linguistic analyses, and network analyses – all of which reveal different aspects of organizations. Clearly the term “web data” elides the myriad information that may be extracted and the variety of methods used to collect and analyze it. Webpages do not offer a discrete research method as much as they represent a new avenue for research. Whereas ethnography and participant observation emphasize the researcher’s presence, interviews and surveys entail direct communication with representative respondents, and content analyses and administrative data take advantage of artifacts left behind by organizations, websites are a combination of these. The researcher can passively observe action through websites (some have even embraced “netnography” [Kozinets 2002]), engage in or observe direct communications with organizational representatives, or scour the traces left behind by organizations. Researchers can also go beyond a single organization and consider relationships across organizations via their weblinks. Thus, websites are not a single source of data, but a wealth of data sources that serve different purposes.

Across the many ways websites can be used, there are nonetheless commonalities. As Golder and Macy (2014:132) put it, “the web sees everything and forgets nothing.”² Every click and key press resides somewhere. Webpages are time stamped, so they are similar in some respects to newspaper-based event data (Earl et al., 2004). They can be observed unobtrusively, thus no bias is introduced by the ordering or wording of questions. Because researchers can observe websites anonymously, there is little risk that

² At the same time, there appears to be a growing interest among communications scholars and some sociologists about the role of ephemerality, anonymity, and self-destructive communications online (Kotfila 2015).

researcher presence will influence outcomes. Webpages can be observed over time, which affords longitudinal analysis prospectively, although current capacities for retrospective website analyses are limited. Depending on the type of website data being collected, the size of websites, and the size of the sample, data storage can quickly present a significant challenge: scraping website content can produce an overwhelming amount of data. Webpages are easily accessible, so barring data storage issues, the costs for increasing sample size are marginal. Even when they suffer from “link rot,” that is, the website degrades or becomes inactive (Meneses et al, 2016), this can be used as a measure of organizational demise. In terms of thickness, webpages can offer both text and images, however coding the content of webpages is, as with other media, dependent on the sensibilities of the analyst.

These features are an advantage for both researchers and consumers. A researcher can, with relative ease, examine the archive of materials that an organization posts online. For researchers as well as consumers, particularly potential members or donors, webpage archives serve as a track record of organizational decisions and public statements. An archive of press releases, news items, blog posts, and tweets provides a digital history. Consumer reviews and comments about an organization can be used to guide decision-making, and for websites that provide reviews over long periods of time, these can offer traces of an organization’s history. Such transparency, however, may lead some organizations to maintain a relatively quiet presence online, or actively purge information or contentious reviews. We might not expect companies such as Halliburton to be as forthright about their inner workings as some of the parent-teacher associations in our sample.

Consider how several popular concepts in organizational studies, including organizational boundaries, networks, and decoupling, are differently illuminated by particular data sources and research methods. In our interviews with directors, we asked them to list the organizations with whom they interact most frequently and fatefully. Their responses to this network query are subject to a number of biases – selective recall, the inclination to stress what one is currently working on, a desire to present the organization in a positive light. More generally, such name-generator questions routinely lead to an

undercount of network size (Fischer, 2009). Similarly, annual reports and press releases attempt to cast organizations favorably. Administrative and tax data may well have more accuracy, due to legal requirements, but they are quite limited in detail.

To be sure, webpages are subject to many of the biases afflicting other data sources. This can be both an opportunity and a challenge for researchers. Indeed, some pages capture current efforts as fast moving and present organizations as highly engaged. Such fluidity can be a useful measure of organizational effort, and is more accurate than an executive's memory and more detailed than an annual report. Additionally, although websites may reflect organizations to have a tidy public face, they also tend to be collective projects – text, images, comments, links, and other material are often the product of different parties, some of whom may not be on the organization's payroll. This is not to say that the organization does not have a hand in shaping its site, but compared to, say, an annual report, it is an articulated hand with fingers operating with some degree of independence. Often, annual reports are posted on webpages. And webpages involve frequent updates, interactive elements, and web links that are absent with annual reports. The challenge, which we take up below, is how to assess whether representations on the web reflect the everyday routines of an organization or are decoupled from them.

What do websites tell us about organizations?

We explore three approaches to webpages and consider how each sheds light on organizations. First, we consider webpages as an expression of organizational identity, a representation of how an organization presents itself to the world. As many executive directors noted to us, websites are increasingly the first point of contact. How the organization presents itself online can affect whether or not it receives funding or recruits volunteers. Moreover, how an organization presents itself through its webpage reveals a great deal about how it sees itself. Second, we conceive of websites as tools, a venue for getting things accomplished. In this form, websites can both supplant and complement activities that were previously accomplished through other means and they can also represent an entirely new plane for organizational efforts. We observe that much of what happens virtually can have real consequences

offline. Third, by considering the way constituencies connect with organizations and how weblinks connect organizations, we see webpages as relational maps, a picture of the interactions an organization has with its external environment.

Websites as identity projects

Webpages serve expressive functions as well as instrumental ones. They are the public face of an organization. Instead of visiting the organization in person or on the phone, customers and supporters increasingly turn to the organization's webpage. What they are met with – design, aesthetics, and text – reveals a great deal about the organization. The director of a low-income housing organization emphasized that their website was “where people go, it's where funders go...it's where volunteers go, it's where donors go.” Webpages can therefore provide important insights into the ways an organization presents itself to its constituents, similar to annual reports (Bromley and Sharkey forthcoming; Meyer and Höllerer, 2016), press releases (Grimmer, 2013), or testimonies (Suddaby and Greenwood, 2005). The format of webpages, such as the number of drop-down menus, reflects differences in the size and function of departments within organizations. The visual images used on webpages suggest differences in the audiences to which organizations wish to appeal, and represent a form of positioning or branding.

In examining professional management in the nonprofit sector, we used both qualitative and quantitative analyses of website text to learn about the ways organizations express their identity. Almost invariably, organization websites featured an “About Us” page listing the organization's mission, as well as information on its founding, lead staff, and links to annual reports and past newsletters. Such pages seem designed for first time visitors as they detail the who, what, when and why of the organization. Mission statements, presented on almost every organization's webpage, concisely document how the organization sees itself and would like to be seen. In fact, they are often the product of organization-wide collaboration, occasionally aided by consultants and arrived at during staff retreats. What impression do

these statements give and what do they say about organizational identity? To examine this question, we collected the mission statements of all organizations and compared them to those from ten years ago. The changes were at times considerable, as in the case of the Playground Kids Club³:

The Playground Kids Club was founded in 1985 by the PTA in order to provide quality day care to Playground students. (2003)

In a child centered, academically challenging environment, which cultivates character and celebrates learning, the families, community, and staff of Playground Elementary School, as partners, will develop creative, exemplary learners with the skills and enthusiasm to shape a changing world. (2013)

In this example, the mission becomes both more abstract and more expansive over time. We wanted to establish that there was a systematic pattern to the transformation of how nonprofits present themselves in mission statements. A simple quantitative analysis showed that almost all mission statements were revised significantly between the two waves of data collection, as illustrated in the histogram of the Jaccard similarity coefficient presented in Figure 2. Together with two research assistants, we scored the expansiveness and abstraction of all statements following a predefined and theoretically developed three-point coding scheme. The inter-coder reliability was satisfying (Cronbach's alpha of expansiveness: .78; abstraction: .63), and we used these measures to determine whether period and organizational properties predicted properties of the mission statement. We find that contemporary mission statements are on average more abstract than mission statements from 10 years ago ($p < .05$). In contrast, they are not significantly more expansive. Expansiveness, however, is associated with younger organizations ($p < .05$).

(Figure 2 here)

³ All organization names used here are pseudonyms.

Second, to learn more about how nonprofits articulate professionalism, we inductively detected discriminating words that were characteristic of more and less professional organizations in our sample (Monroe, Colaresi and Quinn, 2008). The identification of discriminatory words describes words that are characteristic of a certain type of organization, which can be useful for exploration. The log odds ratio of each observation takes into account each word j 's prevalence π_j on the website of organizations whose staff is above and below the the median level of organizational professionalism.⁴ Those who are above are more professional and those below are less so. As Figure 3 illustrates, the words that are most commonly associated with high levels of professionalism include “president,” “paypal,” and numbers, but also “membership,” “library,” and “research.” In contrast, less professional nonprofits were associated with educational language and “schools,” “students,” and “community.”

$$\log odds ratio_j = \log\left(\frac{\pi_{more,j}}{1 - \pi_{more,j}}\right) - \log\left(\frac{\pi_{less,j}}{1 - \pi_{less,j}}\right)$$

(Figure 3 here)

Third, we applied a topic modeling algorithm to detect latent themes that unify and distinguish websites from one other. Topic modeling relies on the strong assumption that the text on each webpage is an unstructured bag of words (DiMaggio and Blei, 2013). There are various special applications of more sophisticated topic models, such as dynamic topic models (where the content of topics can change over time) and correlational topic models (which show the co-occurrence of various themes). The number of topics is manually specified by the researchers and, as Table 2 shows, determine the meaningfulness of the outcome a great deal.

⁴ Organizational professionalism is an index created from the sum of five binary variables describing each organization's professional practices. These include whether the organization is staffed by paid personnel, uses these personnel (and not volunteers) for service delivery, are led by someone with a managerial title, such as “executive director,” and has a paid executive director who holds a full-time position. Volunteer-based organizations without full-time executive directors score low on this index, whereas staff-run organizations with paid, full-time executive directors score highly. In our sample, 12 nonprofits had a score of zero, including several youth athletic associations, a PTA, and a philatelic society. At the other end, seven organizations had a score of five, including a housing development corporation, an industry association, and a disability services organization.

(Table 2 here)

Column A displays the ten most prevalent topics from a set of fifty, whereas column B shows a topic model with only ten topics. Even though computational linguists have developed measures that help estimating reasonable model specifications (Wallach et al., 2009), the appropriate number of topics depends on data and purpose of the analysis. In this case, a larger number of topics picked up on subtler themes consistent with nonprofit organizing, such as community orientation, volunteering, and a certain bureaucratic tendency toward committees, councils, and conferences, whereas other topics reflected the industries in which the nonprofits from this sample were active, such as K-12 education, housing, and health care. In contrast to the National Taxonomy of Exempt Entities (NTEE) code, which is a classification enacted by the same IRS experts who decide on an organization's exempt status, such topics can be used to inductively detect the activities and foci of an organization as it represents it to the public rather than vis-à-vis the IRS (Barman, 2013). These inductive (unsupervised by the investigator) methods can be complemented by deductive (supervised by the researchers' intuitions) methods, such as estimating the frequency of words in a theoretically guided dictionary (see Weber, 1990; Brandtner, 2016).

In addition to the descriptive insights that researchers can generate from text and images, indicators such as the ones described above can be important outcomes or explanatory variables for research on organizational behavior. All of the above methods can be combined with multivariate statistics, in which the quality of a mission statement or the prevalence of a linguistic topic is estimated based on organizational properties. One type of language analysis, structural topic modeling, is designed to relate latent themes in text to properties of the person or organization that produced the text, such as their political affiliation (Roberts et al., 2014). We provide an illustration for such multivariate analyses. Figure 4 shows a heat map of the strength of correlations between each organization and a specific topic, which allows us to identify common topics as well those that are unique to a specific organization. Each row is one of the 155 organizations whose websites served as data for the topic model, and each column

represents one of 50 topics. The illustration shows that some topics are pervasive and unify the nonprofit organizations in the sample, such as mentions of community, services, and programs, whereas other topics are highly characteristic of a single organization, such as its specific purpose.

(Figure 4 here)

These examples suggest how web pages provide a window into the ways that organizations present themselves to their audiences. We noted a certain trend toward abstraction in the mission statement of organizations and, through the discriminating word analysis, observed a relationship between professionally organized organizations and reference to more formal structures, whereas less professionally directed nonprofits were more associated with community, teams, and clubs as well as K-12 education. Topic modeling also revealed themes such as volunteering, memberships, and community service as frequent threads of communication across a variety of nonprofit websites.

Following Goffman (1959), the way an organization presents itself online shapes how it is perceived; and in turn the way it is perceived shapes the way others interact with it. Such interactions then influence how an organization understands itself and directs its future activity. Within organizations, alignments between online and offline presences may occur. Between organizations, convergence online may lead to further convergence offline. To pursue these intuitions, we turn to a discussion of how self-presentations on the web reflect – and perhaps influence – an organization’s activity offline.

Websites as tools

Websites often reflect what organizations do offline. If an organization engages in feeding the homeless, we would expect this activity to be reflected on its website. Beyond mirroring, however, websites also present new avenues through which organizations can act. At its simplest, the site serves as a sort of community bulletin board, informing visitors about organizational activities. Moreover, the

website may play a direct role in these activities, through advocating for homeless issues, volunteering for shifts at the soup kitchen, or generating revenue through donations and sales.

In these and many related ways, a webpage is a technological device that allows an organization to raise funds, advertise goods, sell products, serve members, and proselytize messages. As websites are used more extensively, organizations are transformed, reshaping their internal processes and increasing their geographic reach. Nor should this relationship be viewed as unidirectional. Even as organizations shape the functionalities of their websites, websites shape the goals and services of organizations.

We witnessed several organizations reinvent their activities around their web presence. Some, for example, used their sites to reach broader, previously inaccessible, audiences. The director of a women's blue-collar trade organization, for example, told us that by simply sharing information on the website she has increased meeting attendance. She proudly said that because of the site, some women now make multi-hour drives to meetings from "further afield than I would expect to show up on a Wednesday night at 6:00." Several cultural organizations, including drama groups, choirs, and theatres used their websites to post information about upcoming shows and allow visitors to buy tickets directly from the site. Though an obvious continuation of offline activities, the extension has the effect of expanding audiences at live shows as well as increasing revenue. Some organizations engaged in entirely new activities. A children's chorus, for example, regularly posts free videos and music online, becoming, in effect, an online entertainment provider. Their site even boasts a testimonial from a satisfied consumer describing how the choral performances kept her occupied during a layover, even drawing a crowd of fellow passengers: "Some folks were moved to tears. A couple of ladies were crying. The performances gave me chills." Such videos may increase attendance at live shows, but they also enable an entirely new way of engaging with the organization. For some, such virtual presence is an ambition. The director of a nonprofit music venue said he "would love [the website] to be a resource of art and culture information and not just a regurgitation of dates of when our performances are." Another organization uses its site to expand book sales, aiming to eliminate the need for expensive storefront space. Webpages may even be a resource

during times of organizational transition. After one organization ceased activities in low-cost home construction, the executive director, still wanting to serve the community, transformed the website into a hub to freely distribute digitized children's books. Rather than being a passive artifact of organizational activity, websites are an active plane on which work occurs.

How do offline organizational characteristics drive website functionalities? To explore the connection between organizations' online and offline activities, we visually analyzed and inductively coded websites following several key principles.⁵ First, the sites were observed and engaged with in the way an unacquainted observer might encounter them for the first time. Indeed, many directors keep this in mind when designing sites, with one telling us that people "still hear about us from word of mouth. But when they do, they go to the website to find out more." Second, the sites were coded based on their functionalities – can you do things *directly* via the site or is the site an html-enhanced community bulletin board where flyers are posted? For example, when coding the ability for a site user to join an organization, the interest was in whether one could actually sign up *via* the site, not just obtain information about joining or contact numbers. Websites were also coded in terms of their aesthetics and imageries. The layouts and images used in site design tend to be reflective of the organizations. Nonprofits focused on children and parents, for example, tend to have many images of kids and families on their websites. Examining the various aesthetics, imageries, and functionalities of the websites, we consider how these features relate to offline organizational characteristics, and discuss what convergence and divergence among websites may say about organizations offline.

To quantify the user experience, we generated three factors using a subset of variables coded from a visual analysis. The factors capture the degree to which the website is an interactive tool; presents empirical documentation of the organization's efficacy; and attempts to convey a friendly "warm glow." These three measures would be particularly difficult to obtain through other data sources. As such, they

⁵ This analysis is necessarily restricted to organizations from whom we received complete survey responses.

demonstrate how an organization's web presence can offer insights not readily available through other sources.

Some web functionalities were unique and germane to specific organizational goals. One organization that focused on helping women and children who are victims of abuse offered an "escape" button on their site that follows the user as she scrolls. If an abuser were to appear unexpectedly, the site visitor can click the button and immediately be sent to a retail shopping website. Such specific tools, however, were unusual. More common was the ability for the user to donate, join, volunteer for an organization, make purchases, sign up to receive a newsletter, and participate in discussions through a blog or social media page. To assess these common functionalities, we combined them into a "functions" factor (Cronbach's $\alpha = .66$).

Some features were so common they did not help to distinguish sites from each other. Having an "About Us" page, for example, was nearly ubiquitous, despite the varying inward and outward orientations of the sites. Other features helped to characterize the aesthetics of the website. We consider this "feel" in two ways: evidence-based and friendliness. The more a site quantified organizational activities, cited evidence from studies and evaluations, offered ratings from external agencies or included a formal organization chart, the higher it was on our "evidence" factor (Cronbach's $\alpha = .60$). The site's organization chart, often found toward the bottom of the home page, provides a listing of links to pages within the organization's website. Functionally, it helps the user to navigate the site, similar to a site map. Aesthetically, by listing the organization's activities and departments, it looks strikingly similar to hierarchical organization charts. By enumerating various activities, this chart indicates some level of bureaucratization and suggests that the organization represents its activities with precision. The "friendly" factor was based on websites that conveyed a sense of warmth, community, and accessibility. Such sites included embedded social media (Facebook and Twitter), the presence of testimonials and stories, pictures of smiling people, and images that scrolled automatically (Cronbach's $\alpha = .68$). Variables used in these factors are summarized in Table 2 below.

(Table 3 here)

How well do offline characteristics predict an organization's web presence? To explore this question, we examine each of the three factors in relation to variables drawn from administrative and survey data. Interestingly, we find a positive relationship between the level of organizational professionalism and the functionality of the website. We also find that increased income from donations and grants positively predicts both the friendly aesthetic of the site and its functionality. Membership organizations that rely on donations are more likely to have pictures of smiling members, and stories of their experiences. Additionally, these organizations are more likely to have the capability for making direct donations through their sites. We also find that larger organizations are more likely to have more evidence-based websites. Strikingly, the founding year of an organization does little to predict web features. There is a slight positive association between organizations founded after 1994 and the use of evidence on the web, but our intuition that youth would predict greater web functionality was not borne out. This suggests that although brick and mortar organizations may bear the imprint of the time they were born, websites invite reinvention and newness.

Despite these notable associations between the character of webpages and distinctive organizational features, there is much similarity among webpages, even among quite disparate organizations. The measures of functionality, friendliness, and evidence should thus not be viewed as singular descriptions of a site but as dimensions along which a webpage can be aligned. Websites are best regarded as a mix of dimensions – both quantified and friendly, for example. To explore these combinations, we plot the websites in a three-dimension space defined by the three factors. In this plot, we hone in on one particular brick and mortar variable of interest, organizational professionalism. The factor plot is shown in Figure 5 below.

(Figure 5 here)

From this plot, we see that increased levels of organizational professionalism are associated with greater levels of website functionality. That is, the more professional the organization, the more a visitor can interact with their website. On one level, this may be unsurprising as greater professionalism is associated with larger budgets. But professionalism also has a direct association with more elaborate websites. And more functionality means more opportunities for interactive features. Indeed, we see that the organizations clustered in the upper right of the plot are not only highly functional, but highly evidence-based and friendly as well. Professionalism is not a perfect predictor of website functionality, however. In the lower left of the plot we see a few organizations with generally less functional, evidence-based, or friendly websites. Another cluster in the data, a ridge toward the center of the plot, is made up of middle to highly professional organizations. Moving from the bottom right to the top middle of the plot, we see that as these organizations' websites become more aesthetically friendly, they also tend to be less quantified and more functional.

Several features are apparent from this analysis. First, among a wide array of organizations – including Yoga ashrams, schools, eldercare organizations, athletic teams, and temporary festivals – there is a good deal of similarity. Second, although there are identifiable connections between the on- and off-line aspects of organizations, the unexpected similarities and dissimilarities suggest that a number of other elements mediate this relationship. Just as the spread of organization charts across for-profit, public, and nonprofit organizations induced conformity in formal organizational structures (DiMaggio and Powell, 1983), webpages create isomorphism as well.

Commonalities across sites may reflect brick and mortar processes: how sites were made, who made them, and the design tools utilized in the process. Designing sites from scratch can be challenging. Websites – and the way organizations are experienced through them – are constrained by technical know-how. Some directors went so far as to pit frustrations with the state of their sites on their struggle to recruit tech savvy youth to work in nonprofits. Others thought it was ridiculous to pay others to develop a site when they could develop a rudimentary one on their own. Not unexpectedly, many organizations

resorted to low cost web hosting services such as Wordpress or Weebly. Some of these platforms directly influence the aesthetics and functions of a webpage, with default settings that allow interactive web features, blogs, and photo galleries. Others invested much more heavily. Approximately 40% contracted web developers to design their pages, another 17% had a web developer on their payroll, and 17% expanded a current employee's role, often a communications or public relations manager, to include webpage development. Web developers might offer greater levels of customization, but even so, their designs and recommendations are likely circumscribed by past projects, current trends, or looking to peer organizations' websites.

What does it mean for an organization to be different from others offline but similar when experienced on the web? Alternatively, what does it mean for an organization to be similar to others offline but different on the web? For the most part, there are discernable relations between the two realms. Nevertheless, we speculate that an inadvertent effect of web similarity is to facilitate the comparison, perhaps commensuration, of organizations with diverse purposes. To the degree that websites and webpage representations create a common language for organizations, this may be a powerful influence. Web developers and website hosting services may dictate the form, aesthetics, and capabilities of a site. Peer influence may operate as new sites borrow from established ones.

These websites mirror – but only imperfectly – the organizations that they represent. Mismatches may suggest managed decoupling, with the organization trying to project or emphasize online features that are less apparent offline. Such differences could be aspirational as well. Constrained in its brick and mortar reality, an organization is freer to project how it would like to see itself online. As with the use of any technological device, these relationships can be recursive. Websites can influence brick and mortar activities. For example, if the organization reaches a broader audience online than offline, does it eventually orient the bulk of its activities to this expanded constituency? We see some evidence of such activity change, both in practice and aspiration. For example, the director of temporary family housing organization lamented that information flow on her organization's site is “all just one way” and not

interactive enough. Another described the site's main use as being a "quick fix" for information that "needs to be more interactive." There was a clear desire for websites to take on active roles in the organization, rather than being merely informational. As websites become more interactive and functional, we expect periods of misalignment and realignment between online and offline worlds. In this way, virtual representations reshape brick and mortar activities.

Webpages as relational maps

Webpages are also reflections of an organization's relationships to a wider community. The layout of, and discourse on, a webpage reveals much about the formal structure of organizations, including which departments are most outward facing and involved in boundary-spanning activities (Phillips and Oswick, 2012). Webpages contain message boards, forums, blogs, and various other online communities that leave an extensive digital footprint, reflecting how different constituencies interact with an organization. In our sample, 50% of the webpages gave direct access to the organization's twitter feed, 73% gave direct access to the organization's Facebook page, and 53% had a public blog. A smaller percentage of the organizations had online forums and webinars.

There are notable differences across websites in terms of possible interaction. Some websites are locations where people not only access information but also produce content. In our sample, these activities ranged from watching a YouTube or Vimeo clip about the organization, or seeing a performance of a choral group, to contributing videos. Fully 40% of the webpages included an online calendar, and 36% had job postings; in both cases, the activities and staffing of the organization are made visible to the public. An art museum had a public sign-up schedule for docents, making the most active volunteers clearly identifiable, and encouraging others to participate. The interactive character of many websites means that they evolve with consumer use and feedback, and vary with regard to which consumers participate in distributed production (von Hippel and Krogh, 2003).

Analysis of how online communities interact with an organization is facilitated by Google Analytics and Alexa; both offer information on how a website is accessed, including how often, from what location, and for how long. They even provide demographic breakdowns of site users, the common links that referred users to the focal site, and the most popular pages within a particular domain. But beyond demographic data on users, webpages also provide affiliation data, such as an organization's membership in various associations and partnerships. Many of the organizations in our sample list their major donors on their webpages, including individual benefactors, foundations support, and corporate sponsors. Finally, most webpages list and include profiles of boards of directors. All of these data offer opportunities for network analysis of communities.

An additional fruitful source of data is the analysis of hyperlinks, the incoming and outgoing references to other webpages. Such data are gathered with a web crawler that starts from one or more identified central websites and then follows and captures the network of links between webpages, in a form similar to snowball sampling. The resulting hyperlinks are a type of reference network, comparable to citations in academic papers or friendship networks as expressed on Facebook. Lists of affiliated organizations with hyperlinks also resemble alliance portfolios, common in the study of inter-organizational networks, or tombstone listings, typical of work on investment banks. Incoming links may represent an organization's status or recognition, whereas outgoing links may reflect an organization's aspirations, i.e. to whom do they wish to be attached, or those to whom it is indebted. The overall portrait of a network of hyperlinks suggests the position of an organization within a particular issue field or domain.

A web crawler allows for the generation of a sample by which members are identified by others on the basis of webpage connectivity. To begin an iterative process of tracing weblinks, one begins with a small group of participants and builds outwards. Using a web crawler to identify relationships means that potential participants are selected on the basis of web affiliations rather than ontological properties. An organization is considered part of a network if it is recognized by and connected to other members; its

form, function, and activities alone do not establish membership. The web crawler thus produces a relational, not a category-based, field, whereby the reference-based procedure permits boundaries to emerge independent of a priori definitions or researcher bias. This self-referencing process, similar to snowball sampling procedures, is especially useful in analyzing communities composed of multiple types of members, referred to as “multimodal networks” (Shumate and Contractor, 2013: 450).

As one illustration from a companion project on global organizations involved in evaluating social impact (Korff, Oberg, and Powell, 2015; Powell et al, 2016), we present a circular connection graph in Figure 6 of the hyperlinks on the webpage of UNICEF, the United Nations Children’s Fund, an intergovernmental organization that offers humanitarian and developmental assistance to children and mothers in developing countries. Although the majority of its programs focus on direct community-level services, UNICEF is also known for its global advocacy and efforts to mobilize expertise and engagement to promote children’s well-being. Evaluation features prominently in these endeavors as a means to achieve “evidence-based decision-making and advocacy, transparency, coherence and effectiveness” (UNICEF website). The weblinks graph shows both the connectivity and diversity of UNICEF’s relations, and its popularity as a partner.⁶ Civil society organizations, business corporations, and public sector agencies, both national and international, establish references to it. These linkages allow UNICEF to draw on a large pool of allies and associates. Figure 6 also depicts UNICEF’s diverse affiliations by representing incoming (dotted lines), outgoing (dashed lines), and bidirectional (solid lines) ties to other organizations. The extensive incoming references draw from across the globe, even including businesses that seldom reference other businesses, indicating UNICEF’s convening ability. UNICEF also receives endorsement from social movements as diverse as the Christian *Bread for the World* and impact investors *Social Capital Markets*. Some of these ties are reciprocated, although overall UNICEF’s bidirectional ties are concentrated in the world of international governmental and nongovernmental organizations.

⁶ Circular connection graphs were developed initially for graphical representation of genomic data (Krzywinski et al, 2009), but have spread to representations of global migration flows (Abel and Sander, 2014), epidemics (Luo et al, 2012) and patterns of music beats (Lamere, 2012).

(Figure 6 here)

Although weblinks require minimal financial resources, they suggest interest in alerting one's audience to another organization's activities. Bidirectional references assert mutual awareness, and a willingness to share attention and show solidarity (for further discussion, see Powell et al, 2016). Unreciprocated outgoing weblinks suggest an aspiration or a building block, as with academic citations. Incoming links reflect prominence. In short, analysis of hyperlinks can provide a portrait of the relational patterns in an organizational field, and capture different degrees of embeddedness with a community.

Discussion

The study of organizations has long involved a dialogue among practice, theory, and sources of data. Ethnographic data allowed organizational theory to move beyond stylized hierarchical charts and Taylorist scientific management to observe the informal relations that undergird organizational activity. More recently, network data enabled scholars to explore relations within and across organizational boundaries. In the back-and-forth interplay between data and theory, data offers ways to develop and challenge theory whereas theory offers guidance for questions to be answered with data (Van Maanen, Sørensen, and Mitchell, 2007). As we have demonstrated, webpages offer opportunity for new understandings of organizational behavior. At an elementary level, they allow researchers to see how organizations present themselves to their external environment and see how the external environment is reflected in these representations. At a more advanced level, webpages enable organizations to offer new services to new and existing audiences, and, in turn, are a pathway through which change can occur. These new understandings raise the question of whether webpages require a re-application of old theories, or if there is something distinct about webpages that demands new theory.

As we have discussed, the internet is one among many transformative communication technologies that have shaped organizational activity. Like newspapers, radio, telephones, and television,

the internet is a broadly accessible communication medium. Broad participation, accessibility, speed, and scale, suggest the internet is an amplified form of earlier communication technologies. Its wide usages in daily life, however, suggests it may be more unique. According to the Pew Internet and American Life Project (2014), over 85 % of Americans use the internet for email (90%), news (78%), health and medical information (71%), and to follow politics or political campaigns (61%), look for information on a hobby or interest (84%), shop (71%), bank (61%), visit a government website (67%), make a donation to charity (25%), rate a product, service, or person (37%), and take a virtual tour of a location (52%). The reach of the internet is so vast that some web evangelists, such as the Mozilla Foundation, consider it a basic human resource along with food and shelter.⁷ We may not go so far, but the fact that the internet serves as such a central platform for people to connect and learn – often through organizations – makes us consider its implications, both online and off, for theory.

Broad participation online comes in two forms. In one form, unlike a telephone call where it may be unclear who else is on the line, a website reveals affiliations and deference through its links to other websites and its references to other organizations. Because webpages are highly accessible, they shorten the distance between organization and audience. To be sure, there are restrictions on the web; paywalls, member logins, and censorship all stand in the way of complete accessibility. The barriers, however, are greatly reduced, leading us, at least to some extent, to reconsider the relationship between the openness of the internet and the inward-orientation of bureaucracy.

In another form, many webpages are highly interactive, allowing “outsiders” to comment, vote, or directly manipulate web content. Broad participation, however, does not necessarily render the internet – or the organizations represented through webpages – democratic. Though people may engage collectively on various platforms, they do not do so unfettered. Many platforms are the result of design choices that algorithmically privilege some material over others, shaping knowledge and behaviors (Benkler, 2016; Tufekci, 2016; for examples, see Kramer, Guillory, and Hancock, 2014, Muchnik et al. 2013, and Bond et

⁷ <https://blog.mozilla.org/blog/2016/02/08/the-internet-is-a-global-public-resource/>

al., 2012). From this view, webpages are sites of power. Such power grows with the centrality of any given webpage. How Google, Baidu, Facebook, and YouTube structure information is consequential, as witnessed by the May 2016 kerfuffle among conservatives about purported bias at Facebook, which triggered demands for affirmative action for conservative news sources. As webpages increasingly become sources or aggregators of news, and provide information about organizations, they wield considerable influence over discourse and the fate of organizations. Behind the patina of objectivity, the algorithms that steer web users toward information are products of human design. Even for less central webpages, selective presentation of information can augment how they are perceived. Nonetheless, many voices can be collectively expressed with impressive speed and consequence, shaping both governmental and organizational activities. Many of the nonprofit leaders in our sample stayed vigilant about who was interacting with them through their sites or social media pages. They were willing to be influenced by their constituents online.

The internet presents both a promise and a challenge in the form of speed and scale. Anyone who has attempted to actively follow a Twitter or Facebook feed knows the experience of drinking from a fire hose. The inundating aspect of communication via the web rapidly brings some ideas and trends into sharper focus while obscuring others. This is yet another boon for researchers, many of whom have gleaned data from rapid fire interactions on massive social networks. For consumers of organization websites, pace matters as well. Webpages present a portal through which organizations may engage with social trends and expectations. As such, it presents another dimension on which organizations may lag behind. Failing to have a vibrant webpage or failing to update it at regular intervals suggests inactivity or unreliability offline – even when the organization is in good health.

The internet enables novel ways of assigning organizational status as well. Some of these are built into systems of how knowledge is found, such as Google's PageRank algorithm (the search engine embodiment of Bonacich's eigenvector centrality measure) which ranks pages by the number and status of webpages that link to it. Other services use more direct human input to rate organizations. For example,

Pitchbook and CB Insights (both venture capital and private equity investment databases) combine financial information, website hits, and social media followers to profile the growth and potential of startups. In the social sector, some nonprofit rating agencies recommend potential donors visit an organization's website assess its quality. This strikes us as a remarkable shift in how organizations are evaluated, especially given the relative newness of their presence on the web.

Limitations

Despite the many opportunities afforded by webpages, researchers face challenges in using organizations' online presences for data analysis. As with all data sources, researchers using web data have to consider the problems of reliability (is my measure accurate and consistent?), validity (did I measure what I intended to?), replicability (can others confirm my findings?), and generalizability (what can the sample tell us about the population?). Some of these problems warrant methodological consideration, others have theoretical implications, and yet others call for scope conditions about the results derived from website data. We briefly introduce the key pitfalls in working with organizational web data and offer remedies and general guidelines. As we consider some concerns, specifically conceptual questions about boundaries, access, and decoupling, we will return to them below when we discuss theoretical implications. Table 4 summarizes six main challenges to working with webpage data, as well as potential remedies.

(Table 4 here)

Fluidity. The foremost challenge of the Internet is that it changes constantly. Webpages are updated frequently and content can appear evanescent. Some online content is actually tailored to the user, a problem that communication scholars refer to as reactivity (Rössler and Wirth 2001). The dynamic nature of the web poses problems for the reliability of measuring and coding web content, for instance when multiple coders are looking at different data listed under the same URL.

Although fluidity is a difficult limitation to overcome, it also provides prospects for new discovery. The rapid pace of web development is an opportunity for researchers to explore organizational developments over relatively short periods of time. Updating also becomes a data point in and of itself. A regularly updated webpage may indicate an organization that seeks to be alive online. Even though some rapid changes are a problem, webpages are an exemplary source of longitudinal data, similar to using yearbooks to study career progression. Some data points are actually more valid if taken from a webpage than from other data bases. For example, positions listed on websites are probably superior to personal resumes, organizational membership in various international associations is easier access than annual directories that often lag several years behind.

To deal with the limitation that fluidity presents, it is necessary to archive data as it is downloaded or scraped.⁸ Screenshots may also be used to engage in visual analyses of sites over time. By doing this, the dynamic aspects are saved rather than overwritten by changes on the site. Repeated snapshots can be used to create longitudinal datasets.

Black box. Webpages are collective projects that, in most cases, do not reflect a single person's preference and efforts at any one point in time, but that are produced jointly and incrementally. The content of webpages may be reflective of functional differentiation within an organization, where the webpage is maintained by a small group of individuals. Similar to other content data, the consumption of webpages depends on the context: some users may use a webpage to decide on a potential donation, whereas another may want to purchase a ticket for an event (Krippendorff, 2013). The black box aspects of organizational webpages push researchers to look for other sources of organizational data. As we have demonstrated, some of what is seen online can be predicted by offline data. Surveys and interviews can

⁸ A primary task for studying organizations on the web is the ability to download the content of webpages. Most solutions, such as Scrapy or BeautifulSoup, are open source programs that can be used for free, but require working knowledge of HTML and a programming language such as Python. There are a number of user-written or commercial programs that are more intuitive for beginners, such as Gehman and Grimes's (2015) *Cultr* toolkit for "scraping culture in action" or the import.io data extraction platform. A special challenge is to scrape webpages longitudinally. The Internet Archive's *Wayback Machine* has provided researchers with an invaluable tool to access archived webpages which, if available, frequently reach back to around the year 2000 (archive.org/web).

reveal the organization's web developer and general approach to putting content online. In our study, interviewees gave us great detail on the webpage production process and who they thought was reading the webpage. This helped to understand the process of content creation as well as allowed insights into the degree to which different messages are decoupled from each other.

Deception. A frequently voiced criticism is that information on webpages is deceiving, because it is not regulated, nor always under the complete control of an organization. Seen in this light, webpages can be the outcome of impression management strategies (Ginzel, Kramer, and Sutton, 1993; Elsbach, 2003). But so too are all organizational communications. Not unlike an annual report, such a representation is useful for learning about the identity of the organization and how it chooses to present itself to its customers and stakeholders. Inconsistency between the talk on webpage discourse and the walk reflected by internal practices and structures is in fact a longstanding theoretical issue in organizational theory. As Brandtner (2016) argues, the inconsistency between different accounts of one and the same organization can be a consequence of competing expectations about how a focal organization should respond to external pressures. We suggest the possibility of deception can be dealt with by triangulating webpage data with other sources and that any observed inconsistencies are rich with theoretical implications.

Context-dependence. One concern about drawing insights from one sample of webpages is that it may not apply to other data. The degree to which different domains of organizational activity are comparable is subject to debate. Similarly, it is not self-evident that webpages should be compared across various industries or domains. Some domains, such as internet retail, have inherently tighter links between their online and offline activities than others, such as a soup kitchen. For the most part, this problem is self-evident and likely to be dealt with through research design or controls. Where it becomes a challenge, however, is when researchers focus their study on a central issue – such as social impact – where organizations from multiple domains are involved in a common issue (Powell et al., 2016). The insights that can be generated by comparing activist blogs, charitable foundations, and management

consultants can be confounded by differences of style and content in how these organizations communicate. Though this issue may not be new to studying organizational webpages, the high-level of inter-domain connectivity revealed by websites may intensify the challenge of studying issue-based fields that attract different types of organizations. On the other hand, weblinks are an excellent new tool for capturing such cross-domain ties.

The issue of context-dependence also touches on generalizability. Findings from a study of online presences in one field are not necessarily applicable to another. This is true for organizational theory in general (e.g., the comparability of hospitals and manufacturing corporations cannot be taken for granted). Though scope conditions are a practical method for dealing with generalizability issues, overly narrow restrictions make theories irrelevant for a larger audience of organizational theorists. Importantly, methods applied to one context (e.g. by developing a specific dictionary for measuring bureaucratic language in the nonprofit sector) should not be blindly applied to another.

Complexity. Webpages may contain a vast amount of information, only some of which can be easily analyzed. Scraping and archiving webpages may require large amounts of hard disk or server space. Computer-aided analysis and principal component analysis can help researchers to reduce the dimensions of information, but this does not remedy the fact that various types of information – such as network data, text, and images – are in reality consumed simultaneously, may interact with each other, and provide different opportunities for measuring the same theoretical concept (Meyer et al., 2013). At the same time, various measures of the same concept provide an opportunity for validation, a crucial component of working scientifically with text data (Grimmer and Stewart, 2013).

Sampling bias. Not all organizations have web presences. For instance, even though nearly 90% of organizations in our sample represented themselves online, the fact that 10% do not have a webpage introduces unwanted bias. The fact that not all organizations are online can lead to undue inferences about the organizational landscape (e.g. when trying to model the restaurant scene of a neighborhood based on Yelp data). Using a list of URLs as a starting point does not guarantee that all relevant organizations are

included, because some might not be present on the web. We suggest sampling theoretically or randomly and subsequently testing for sample selection bias. A catalogue of brick and mortar organizations, such as the NCCS or S&P 500 can be the best starting point for learning about organizational behavior; for studying a specific discourse or issue field, snowball sampling with the help of a web crawler may be the best strategy.

Conclusion

Just as there are many types of organizations, there are many kinds of webpages. The site of a nonprofit service organization differs from an art museum. A museum may have more in common on the web with a retail store, whereas a social service agency may have similarities with a government agency or a for-profit provider of job training services. This diversity of representations, we argue, is a valuable new source of rich data for students of organizations. Webpages are reflections of organizations' interests, services, and identity, and constitute a form of membership in wider communities. But webpages are much more than 'mere' reflections, they also shape the actions of organizations, in both direct (through the provision of functions online) and indirect (through external evaluations, blogs, and weblinks) ways.

Consider what an organization would be like without a webpage. Most remarkably, it is increasingly hard to imagine a 'reputable' or 'legitimate' organization not having a webpage. Consumer decisions are often based on the ease of use and utility of an organization's website. Applying for a job, making a donation, booking a table, and organizing a meeting are now commonly done through websites. Given this recent and quite remarkable transformation in how organizations conduct their affairs, we think it is incumbent on researchers to use these valuable data sources. In our work on the Bay Area nonprofit sector, we found webpages to be an exceptionally rich source of data, for both more rapid access to standard forms of information used in the past and new kinds of content ranging from aesthetic to interactive. Just as scholars of political participation and analysts of open source communities have

discovered, more and more kinds of activities have moved online (O'Mahony and Lakhani, 2011; Ferraro and O'Mahony, 2012; Bennett and Segerberg, 2013). Organizations may have been later entrants in this process, but clearly the nonprofit entities in our random sample are now active digital participants through their webpages.

We have presented a broad toolkit of ways webpages can be used for organizational analysis, and attempted to compare these methods with older, more established ones. Data derived from webpages can be correlated with administrative data; used for qualitative content analysis; scraped at intervals for longitudinal analyses; or weblinks can be collected with a web crawler to examine relational patterns. Images on webpages can be analyzed, or quantitative text analysis can be used to assess large amounts of written materials. In many respects, we have just scratched the surface of possibilities. Other scholars are using webpages to analyze the interactions of citizens and public bureaus, changes in the political positions of politicians, and the development of discursive communities. All of these represent a new frontier in how organizing and organizational life are conducted. We hope we have persuaded others of opportunities afforded by the up until now untapped potential of organizational webpages.

References

- Barman, E. 2013. Classificatory Struggles in the Nonprofit Sector. *Social Science History* 37(1): 103-41.
- Baruch, Y., & Holtom, B. C. 2008. Survey response rate levels and trends in organizational research. *Human Relations* 61(8): 1139-1160.
- Bell, E., Schroeder, J. E., and Warren, S. (Eds.). 2013. *The Routledge Companion to Visual Organization*. New York: Routledge.
- Benkler, Y. 2002. "Coase's Penguin, or, Linux and The Nature of the Firm." *Yale Law Journal* 112(3): 369-446.
- Benkler, Y. 2011. *The Penguin and the Leviathan: How Cooperation Triumphs Over Self-Interest*. New York: Crown Business.
- Benkler, Yochai. 2016. "The Internet: Degrees of Freedom, Dimensions of Power." *Daedalus* 145(1):18-32.
- Bennett, W. L., and A. Segerberg. 2013. *The Logic of Connective Action*. New York: Cambridge University Press.
- Bird, S. 2006. NLTK: the natural language toolkit. In Proceedings of the COLING/ACL on Interactive presentation sessions (July: 69-72). Association for Computational Linguistics.
- Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489 (7415): 295-298.
- Brandtner, Christof. 2016. "Managing the magic: Conditions of decoupling in the U.S. non-profit sector." Working paper at Stanford University.
- Bromley, Patricia, and Amanda J. Sharkey. 2015. "Casting Call: The Expansion of Actorhood in US Firms, 1960-2010." *Academy of Management Proceedings* 2015(1): 15743.
- Bromley, Patricia, Hokyung Hwang, and Walter W. Powell. 2013. "Decoupling Revisited: Common Pressures, Divergent Strategies in the U.S. Nonprofit Sector." *M@n@gement* 15(5): 468-501.
- Castells, M. 2000. *The Information Age: Economy, Society and Culture - Vol. III*. Cambridge, MA: Blackwell.
- Castells, M. 2012. *Networks of Outrage and Hope: Social Movements in the Internet Age*. Cambridge, UK: Polity Press.
- Chandler, A. D. 1977. *The Visible Hand: The Managerial Revolution in American Business*. Cambridge, MA: Harvard University Press.
- Clifton, B. 2012. *Advanced Web Metrics with Google Analytics*. New York: John Wiley & Sons.
- Contractor, N. 2013. "Some Assembly Required: Leveraging Web Science to Understand and Enable Team Assembly." *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 371:1-14.

- DiMaggio, P., Nag, M., & Blei, D. 2013. "Exploiting Affinities Between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of US Government Arts Funding." *Poetics* 41(6): 570-60.
- Earl, J., Martin, A., McCarthy, J. D., & Soule, S. A. 2004. "The Use of Newspaper Data in the Study of Collective Action." *Annual Review of Sociology* 40: 65-80.
- Elsbach, K. 2003. "Organizational Perception Management." *Research in Organizational Behavior*, R.M. Kramer and B.M. Staw, eds. 25: 297-332.
- Espeland, W. and Sauder, M. 2016. *Engines of Anxiety*. New York: Russell Sage Press.
- Ferraro, F. and O'Mahony, S. 2012. "Managing the Boundaries of an Open Project," in *The Emergence of Organizations and Markets*, J.F. Padgett and W.W. Powell, eds., pp. 540-65. Princeton, NJ: Princeton University Press.
- Fischer, C. S. 2009. "The 2004 GSS Finding of Shrunk Social Networks: An Artifact?" *American Sociological Review* 74(4): 657-69.
- Gehman, J. & Grimes, M., 2015. *Cultr: A Toolkit for Scraping Culture in Action*. Available online at <http://cultrtoolkit.com>
- Ginzell, L.E. 1993. "Organizational Impression Management as a Reciprocal Influence Process." *Research in Organizational Behavior* (15): 227-66.
- Goffman, E. 1959. *The Presentation of Self in Everyday Life*. New York: Anchor.
- Golder, S. A., & Macy, M. W. 2014. "Digital Footprints: Opportunities and Challenges for Online Social Research." *Annual Review of Sociology* 40(1): 129-152.
- Greenstein, S. 2015. *How the Internet became Commercial*. Princeton, N.J. Princeton University Press.
- Grimmer, J. 2013. *Representational Style in Congress*. New York: Cambridge University Press.
- Grimmer, J., & Stewart, B. M. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis* 21(3): 267-297.
- Hargittai, E. 2010. "Digital Na(t)ives? Variation in Internet Skills and Uses among Members of The 'net Generation.'" *Sociological Inquiry* 80(1): 92-113.
- Herring, S. C. 2010. "Web Content Analysis: Expanding the Paradigm." Pp. 233-49 in *The International Handbook of Internet Research*, edited by J. Hunsinger, M. Allen, and L. Klastrup. Amsterdam: Springer.
- Hinds, P. L. Liu, and J. Lyon. 2011. "Putting the Global in Global Work." *Academy of Management Annals* 5(1): 1-54.
- Hopkins, D., King, G., Knowles, M., & Melendez, S. 2010. ReadMe: Software for automated content analysis. Institute for Quantitative Social Science.
- Huberman, B. A., and L. A. Adamic. 2004. "Information dynamics in the networked world." In B. A. Huberman and L. A. Adamic, *Lecture Notes in Physics - Complex Networks*, pp. 371-398. Heidelberg: Springer Verlag.

- Hwang, H., and W. W. Powell. 2009. "The rationalization of charity: The influences of professionalism in the nonprofit sector." *Administrative Science Quarterly* 54: 268–298.
- Jemielniak, D. 2014. *Common Knowledge: An Ethnography of Wikipedia*. Stanford, CA: Stanford University Press.
- Jurafsky, D., & Martin, J. H. 2009. *Speech and Language Processing*. Prentice Hall.
- Kellogg, K. C., Orlikowski, W. J., and Yates, J. 2014. "Life in the Trading Zone: Structuring Coordination in Postbureaucratic Organizations Across Boundaries." *Organization Science* 17(1): 22–44.
- Korff, V. P., Oberg, A., and Powell, W. W. 2015. "Interstitial Organizations as Conversational Bridges." *Bulletin of the American Society for Information Science and Technology* 41(2): 34–38.
- Kotfila, C. 2014. "This Message Will Self-Destruct: The Growing Role of Obscurity and Self-Destructing Data in Digital Communication." *Bulletin of the American Society for Information Science and Technology* 40(2): 12–16.
- Kozinets, R. V. 2002. "The Field Behind the Screen: Using Netnography for Marketing Research in Online Communities." *Journal of Marketing Research* 39(1): 61–72.
- Kramer, A. D., Guillory, J. E., and Hancock, J. T. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111(24): 8788–90.
- Krippendorff, K. 2012. *Content Analysis: An Introduction to Its Methodology*. Sage.
- Krzywinski, M. et al. 2009. "Circos: An Information Aesthetic for Comparative Genomics." *Genome Research* 19(9): 1639–45.
- Kuhn, T. S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Lamere, P. 2012. "The Infinite Jukebox." Music Machinery. (Retrieved online on August 5, 2015: <http://musicmachinery.com/2012/11/12/the-infinite-jukebox/>).
- Lewis, K., Gray, K., & Meierhenrich, J. 2014. "The Structure of Online Activism." *Sociological Science* 1: 1–9.
- Luo, Y. et al. 2012. "Comparative Genome Analysis of the High Pathogenicity Salmonella Typhimurium Strain UK-1." *PLoS ONE* 7(7).
- Marres, N., & Weltevrede, E. 2013. "Scraping the Social?" *Journal of Cultural Economy* 6(3): 313–335.
- McMillan, S. J. 2000. "The microscope and the moving target: The challenge of applying content analysis to the World Wide Web." *Journalism & Mass Communication Quarterly* 77(1): 80–98.
- Meneses, L., Jayarathna, S., Furuta, R., and Shipman, F. 2015. Grading Degradation in an Institutionally Managed Repository. In *Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries* (pp. 263–264).
- Meyer, R. E. and Höllerer, M. A. 2016. "Laying a Smoke Screen: Ambiguity and Neutralization as Strategic Responses to Intra-Institutional Complexity." *Strategic Organization*.

- Meyer, R. E., Höllerer, M. A., Jancsary, D., & Van Leeuwen, T. 2013. "The Visual Dimension in Organizing, Organization, and Organization Research: Core Ideas, Current Developments, and Promising Avenues." *Academy of Management Annals* 7(1) 489-555.
- Mohr, J. W., and Bogdanov, P. 2013. "Introduction—Topic Models: What They Are and Why They Matter." *Poetics* 41(6): 545–69.
- Muchnik, L., Aral, S., and Taylor, S. J. 2013. "Social Influence Bias: A Randomized Experiment." *Science* 341(6146): 647–51.
- Neuendorf, K. A. 2002. *The Content Analysis Guidebook*. Thousand Oaks, CA: Sage.
- O'Mahony, S. and Lakhani, K. 2011. "Organizations in the Shadow of Communities," in *Communities and Organizations*. ed. by C. Marquis, M. Lounsbury, and R. Greenwood, pp. 3–36. *Research in the Sociology of Organizations*. Vol.33. Bingley, UK: Emerald Publishing.
- Pew Internet & American Life Project. 2014. Internet use over time. [Data file]. Retrieved from <http://www.pewinternet.org/files/2014/01/Usage-Over-Time- May-2013.xlsx>
- Pfeffer, J., and Salancik, G. R. 1978. *The External Control of Organizations*. New York: Harper & Row.
- Phillips, N., and Oswick, C. 2012. "Organizational Discourse: Domains, Debates, and Directions." *Academy of Management Annals* 6(1): 6–36.
- Powell, W. W., Oberg, A., Korff, V., Oelberger, C., and Kloos, K. 2016. "Institutional Analysis in a Digital Era: Mechanisms and Methods to Understand Emerging Fields." in *New Themes in Institutional Analysis: Topics and Issues from European Research*, edited by G. Krücken, C. Mazza, R. Meyer, and P. Walgenbach. Cheltenham, UK: Edward Elgar.
- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... & Rand, D. G. 2014. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science* 58(4): 1064-1082.
- Rössler, P., & Wirth, W. 2001. Inhaltsanalysen im World Wide Web. *Inhaltsanalyse. Perspektiven, Probleme, Potentiale* 280-302.
- Salganik, M. J., & Watts, D. J. 2009. Web-Based Experiments for the Study of Collective Social Dynamics in Cultural Markets. *Topics in Cognitive Science* 1(3): 439-468.
- Salganik, M. J., Dodds, P. S., & Watts, D. J. 2006. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science* 311(5762): 854-856.
- Schofer, E. and Longhofer, W. 2011. "Structural Sources of Association." *American Journal of Sociology* 117(2): 539-85.
- Shumate, M., and Contractor, N. 2013. "The Emergence of Multidimensional Social Networks." In L. L. Putnam and D. K. Mumby (Eds.) *The SAGE Handbook of Organizational Communication*, pp. 449–474. Thousand Oaks, CA: Sage.
- Suarez, D. F. 2009. "Nonprofit Advocacy and Civic Engagement on the Internet." *Administration & Society* 41(3): 267–289.
- Suarez, D. F. 2010. "Street Credentials and Management Backgrounds: Careers of Nonprofit Executives in an Evolving Sector." *Nonprofit and Voluntary Sector Quarterly* 39(4):696–716.

- Suddaby, R. and Greenwood, R. 2005. "Rhetorical Strategies of Legitimacy." *Administrative Science Quarterly* 50(1): 35–67.
- Thompson, J. D. 1967. *Organizations in Action*. New York: McGraw Hill.
- Tufekci, Z. 2016. "As the Pirates Become CEOs: The Closing of the Open Internet." *Daedalus* 145(1):65–78.
- Tufekci, Z., and Wilson, C. 2012. "Social Media and the Decision to Participate in Political Protest: Observations from Tahrir Square." *Journal of Communication* 62(2): 363–79.
- Van Maanen, J., Sørensen, J. B., and Mitchell, T. R. 2007. "The Interplay between Theory and Method." *Academy of Management Review* 32(4): 1145–54.
- Von Hippel, E, and Krogh, G. V. 2003. "Open source software and the "private-collective" innovation model: Issues for organization science." *Organization Science* 14(2): 209-223.
- Von Hippel, E. 2005. *Democratizing Innovation*. Cambridge, MA: MIT Press.
- Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. 2009. "Evaluation methods for topic models." In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1105-1112.
- Watts, D. J. 2011. *Everything Is Obvious: How Common Sense Fails Us*. New York: Crown Business.
- Weber, R. P. 1990. *Basic Content Analysis*. Thousand Oaks, CA: Sage.
- Zuckerman, E. 1999. "The Categorical Imperative: Securities Analysts and the Illegitimacy Discount." *American Journal of Sociology* 104(5): 1398-1438.

Table 1. Strengths and weaknesses of organizational data sources

	<i>Administrative data</i>	<i>Executive interviews</i>	<i>Original survey data</i>	<i>Website data</i>
<i>Data types</i>	Thin data	Thick data	Various: Closed and open answers	Various: Text, network, images
<i>Replicability</i>	Easy	Very challenging	Challenging	Possible
<i>Public access</i>	Easy	Very hard	Hard	Very easy
<i>Startup costs (1st observation)</i>	Medium, but high cleaning effort	Relatively cheap, given access	High effort to design, implement	High effort to implement
<i>Marginal costs to increase N</i>	Very low	Almost as high as first observation	Medium	Cheap to add observations
<i>Cross-sectional comparability</i>	High, but depends on data source	Low	Good	Good, depends on method of analysis
<i>Difficulty to get longitudinal data</i>	Usually low, but high cleaning effort	High due to attrition and recollection bias	High due to attrition	Publicly accessible (Internet Archive)

Table 2. Most prevalent topics from two topic models (50, 10 topics) based on website text

#		Topics A (10/50)	#		Topics B (10/10)
26	1.32	Information: information time make provide programs part work free state online include support including current required questions meet call additional	1	4.57	Secondary education: school children community people child youth time family parents staff work life group support families skills program education
36	1.05	Services (general): service office members number provided made special system general change issues national department date individual including list future provide	5	2.07	Housing services: housing community services assets center net development program health senior property affordable income family families apartments seniors curry home
11	0.89	Community: place community program area communities project programs site public development management university people building board resources school working work	3	1.93	Higher education: students student university education financial program faculty support college information campus aid educational office courses experience academic school time
23	0.89	General: place time back found find history world end great set home began space working late center major area society	6	1.82	Human services: services information health plan care address member part state date service form school applicant program income medical fund
10	0.87	Family services: community services children family program support volunteers food families volunteer place life people home gift make house provide hope	0	1.61	Bureaucracy: convention connection chapter council information committee call time make office members room hotel contact president conference
17	0.62	Education (general): school students skills program community learning education class family teaching learn social work art schools student classes activities language	8	1.36	Organization: board permission local strategy members committee form training information safety section number conference water program
5	0.45	Work-life: work show people good south today things time ca week great working office make night weekend life coming fun	9	0.89	Arts: music american film world uss arts york art life work artists films festival show area performance works
6	0.30	Sports: club team league game instructor players teams play player age high coach field games horse county sports riding youth	2	0.83	Environment: state trail vehicle area road summit rock water species donation program support canyon system information season free foot call
33	0.28	Human services: children services child county program families care support center family place youth violence health development community education prevention abuse	7	0.62	Education (general): school university college students making waves graduated degree bachelor arts graduate received master public lives state student change high
40	0.20	General: people line party find give award job family august call light items news made hope minutes	4	0.60	Science: brain ny phd press books people body life energy mind time women information md men person males ca studies

Note: Some words deleted to preserve anonymity. Topic labels (bold) chosen based on high-frequency words in each topic and reading of the websites.

Table 3. Summary of web features

Tool	Percent	Can a web user...	Factors*
Sales	28%	<i>buy products?</i>	1
Donations	77%	<i>make donations?</i>	1
Newsletters	53%	<i>sign up for a newsletter?</i>	1
Join	44%	<i>join or volunteer for the organization via website?</i>	1
Calendar	41%	<i>see a calendar of upcoming events?</i>	1
Blog	53%	<i>read or participate in blog?</i>	1
Facebook	73%	<i>connect to the organization's Facebook feed?</i>	1, 3
Twitter	50%	<i>connect to the organization's Twitter feed?</i>	1, 3
Org. chart	14%	<i>see links arranged in visible hierarchy?</i>	2
Quantification	49%	<i>read quantified descriptions of organization activities?</i>	2
Cite evidence	17%	<i>find evidence (reports, studies)?</i>	2
Ratings listed	9%	<i>find external ratings of the organization?</i>	2
Video	65%	<i>watch videos via the website?</i>	3
Testimonials	37%	<i>read testimonials about products or services?</i>	3
Scrolling images	56%	<i>see scrolling or moving images?</i>	3
Smile pictures	67%	<i>see pictures of smiling people or playing children?</i>	3

*Factor 1 = Functional ($\alpha=.66$); Factor 2 = Evidence ($\alpha=.60$); Factor 3 = Warmth ($\alpha=.68$)

Table 4. Challenges of organizational analysis using webpages, and remedies

	Limitations	Solutions
<i>Fluidity</i>	Dynamic updates, removal of web content, content reactive to users	Change is longitudinal variation Archive data
<i>Black box</i>	Unclear who creates, consumes, and updates web content	Triangulate data Theorize (boundaries and access)
<i>Deception</i>	Organizations may choose content strategically, engage in impression management	Theorize (decoupling) Triangulate data
<i>Context-dependence</i>	Limited comparability, organizational fields are inhabited by various organizations	Scope conditions Missing information is data
<i>Complexity</i>	High information density, multiple media, large file sizes	Sample theoretically Validate vigorously Triangulate data
<i>Sampling bias</i>	Not all organizations are online, catalogue of URLs is not reflective of full population	Sample theoretically Missing information is data

Figure 1. Growth in website use among Bay Area nonprofits, 2002-2012

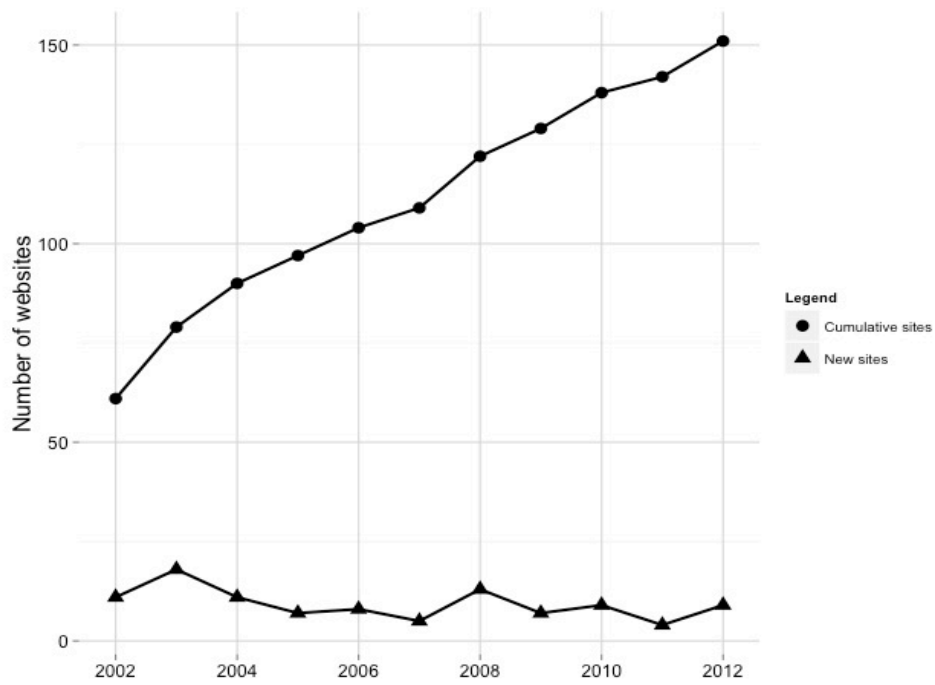


Figure 2. Histogram of Jaccard coefficient indicating change in contents of mission statements over time

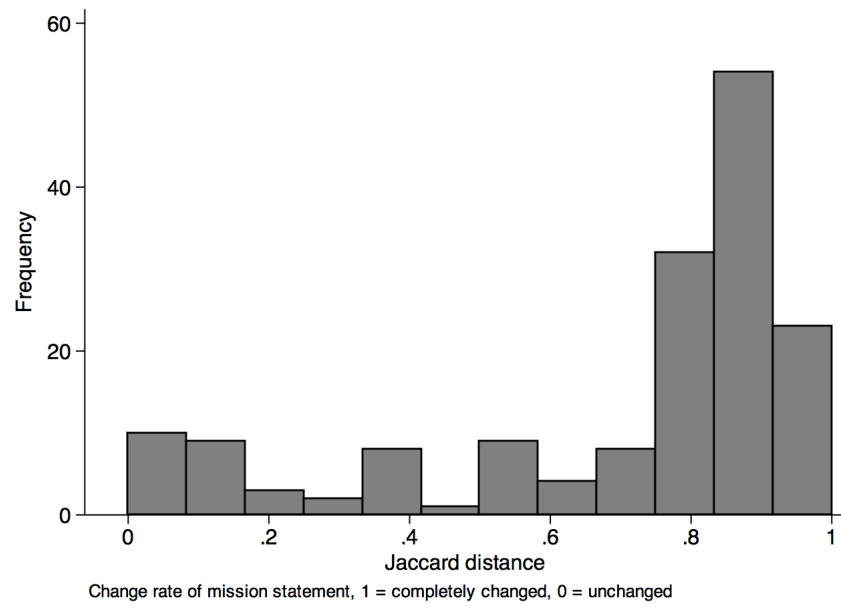
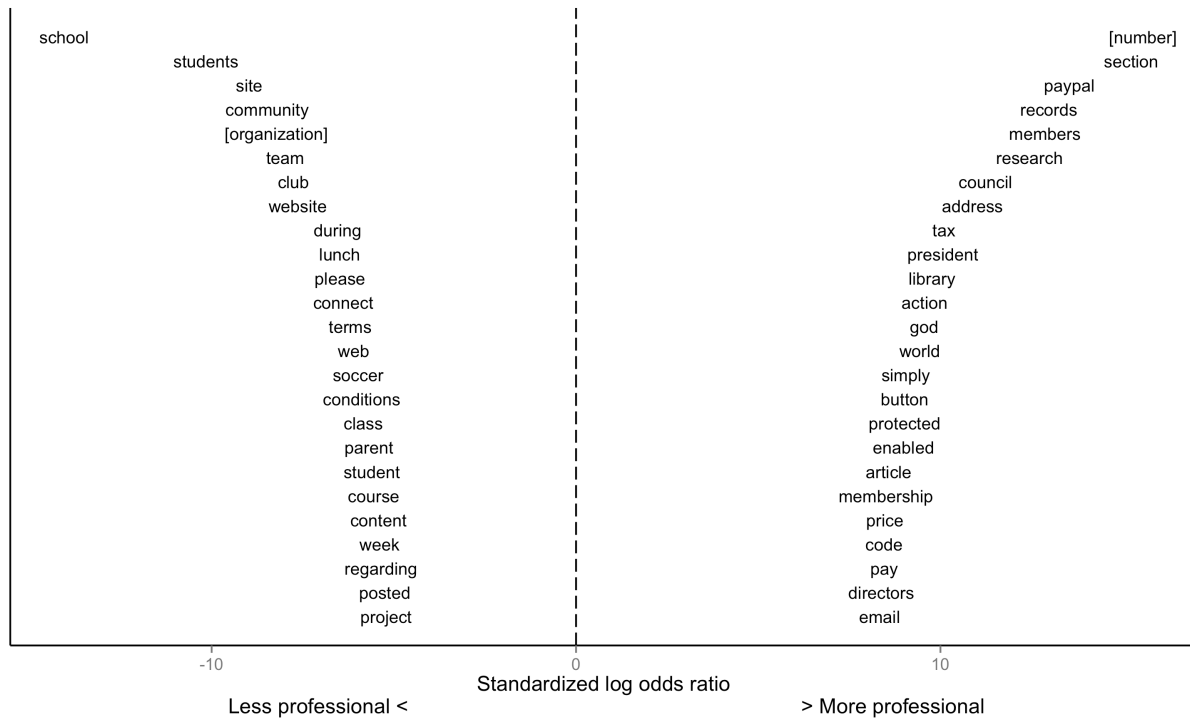


Figure 3. Standardized log odds ratio indicting most discriminating words used on websites



Note: Words ranked by the degree to which they are indicative of (less or more) organizational professionalism. Note that weighted word frequencies such as the standardized log odds ratio may conceal contextual meaning. In this sample, “team” is frequently used in the context of youth sports clubs, which are amateur groups and less professional.