6

INFERRING INDIVIDUAL-LEVEL CHARACTERISTICS FROM DIGITAL TRACE DATA

Issues and Recommendations

Deen Freelon

The digital age has generated innumerable new data sources for scholars of communication. Political communication researchers have probably benefited from this bonanza more than some subfields (interpersonal and organizational, for example) due to the public nature of many of its objects of study. From social media and websites to digitized versions of offline texts, digital data sources allow us to explore political communication research questions in unprecedented ways. Thus far we have only scratched the surface of the methodological possibilities afforded by the many datasets now available to us through a few clicks.

One important category of digital data for our subfield is *digital traces*. These are the records of online activity recorded by the servers that undergird all internet-based communication (Freelon, 2014). Traces can be created manually or generated automatically: user-generated text, hyperlinks, social media follows, "likes" and "favorites," and timestamps are all examples. (Not included are analog media content that is later digitized for preservation; in other words, traces are necessarily native to digital contexts.) These records collectively contain enormous empirical potential to answer all manner of politically relevant questions.

One of the greatest challenges for researchers interested in digital traces is managing the gap between their research's conceptual focus and the set of readily available traces. Not every type of trace will be equally valuable from a particular research standpoint, and not every interesting concept will be measurable using the traces to which we have access. Researchers should never assume without support that a given trace or trace-derived construct indicates a given underlying concept, however intuitive it may seem. Some traces may require only brief explanations of how and why they relate to their theoretical referents. For others,

more elaborate arguments and data transformations may be necessary to sufficiently justify particularly theoretical uses.

The purpose of this chapter is to contribute to the development of a framework for assessing the construct validity of theoretical inferences drawn from digital traces. Most high-quality trace-based empirical research does this to some extent, but what is missing is an abstract set of standards and heuristics by which the quality of its inferences may be assessed. This will help ensure the rigor of such research, which is especially important given that it is still in its infancy. I define four nested, platform-independent domains that researchers should bear in mind when choosing traces for analysis: technical design, terms of service (TOS), social context, and the potential for misrepresentation. I demonstrate the value of this framework in discussions of three general categories of techniques for trace inference: *direct indication, proper names,* and *speech patterns.* I apply the framework to these techniques by drawing examples from three individual-level characteristics of great interest to political communication researchers: gender, race/ethnicity (R/E), and geographical location. Each of these has seen a diverse range of empirical attempts to infer them from traces in the relevant literature.

Four Domains Affecting the Construct Validity of Trace Data

The well-known dictum that "raw data is an oxymoron" (Bowker, 2005, p. 184; Gitelman, 2013) is rarely illustrated more clearly than in the case of digital traces. Their use is so common in contemporary social science research that authors rarely bother pointing out that they were not generated with research in mind (Howison, Wiggins, & Crowston, 2011). Unlike survey or experimental data, which are constructed to optimize the quality of the research based on them, trace data are created incidentally through everyday internet use. Because their fitness as research data is not guaranteed for any particular purpose, researchers should argue convincingly that particular inferences can be drawn from them.

When deciding whether to infer a specific characteristic from a specific trace, four domains warrant close consideration: technical design, terms of service, social context, and the potential for misrepresentation. These are hierarchically nested (see Figure 6.1) in that the outer domains constrain the range of choices available within the domains they enclose. The outermost domain, technical design, defines the absolute limits of what can and cannot be done within a sociotechnical system. Proceeding inward, a platform's terms of service designate which technically possible behaviors may result in official punishment, including account suspension and deletion. Social contexts can only be built upon behaviors that operate within a platform's terms of service, and misrepresentation is a characteristic of certain social contexts. Together, these domains provide a comprehensive foundation for arguments about the relationships between traces and their ostensible referents.

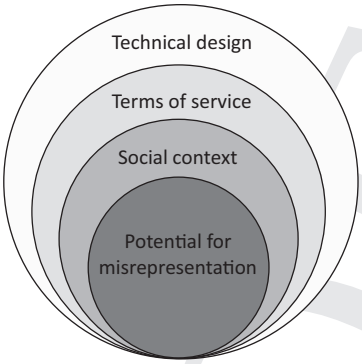

FIGURE 6.1 Four Nested Domains Affecting the Construct Validity of Digital Traces

The following subsections discuss each in turn. They rely heavily on examples drawn from social media, which offer many diverse examples of theoretically valuable traces.

Technical Design

The design of a communication system determines the kinds of communications it can support (Freelon, 2010, 2015a; Sack, 2005; Wright & Street, 2007). Almost none of the 20th century's mass media allowed their audiences to respond, so their channels were dominated by elites. Online media permit such responses, which grants ordinary users an unprecedented menu of communication options. Every platform's menu is unique: for example, Twitter allows its users to unilaterally "follow" one another by default, while Facebook requires its users to approve each "friend" request. Snapchat deletes messages as soon as the recipients have viewed them, and the now-defunct social network This allowed users to share just one hyperlink per day. Design features such as these set the absolute boundaries of digital behavior.

This general point is fairly well understood by most students of communication technologies, so I will not belabor it here. However, before continuing it is worth briefly discussing how the official labels of some design features invite specific inferences. One example of this is "likes" as implemented by Facebook and Twitter, which allow users to mark individual posts with a thumbs-up or heart icon, respectively. A researcher could infer positive sentiment toward "liked" posts on the basis of the feature label alone. But such an argument would not be satisfactory, because "likes" can convey much more than just positive sentiment, which is not a monolithic concept in any case (Gerlitz & Helmond, 2013). This example demonstrates that even in the most seemingly obvious of cases, researchers should not take traces at face value. "Likes" may reliably indicate positive sentiment in some or even most cases, but we need more evidence than the platform creators' intentions to soundly argue as much.

Terms of Service (TOS)

Terms of service are the lengthy documents we all must agree to (without necessarily reading) before registering as users on most social media platforms. They specify the platforms' rules and the consequences of violating them. Unlike the descriptive rules of technical design, which forcibly forbid or require certain behaviors, terms of service are often prescriptive rules that must be voluntarily obeyed. Many TOS can be broken fairly easily in the normal course of using a platform, whereas the restrictions imposed by technical design cannot. For example, notwithstanding a few exceptions I cannot post a tweet of more than 140 characters or befriend more than 5,000 Facebook users. I can, however, create a pseudonymous profile on Facebook or reveal another private user's personal information on Twitter or Facebook, both TOS violations punishable by account suspension. Most platforms' TOS require users to obey the laws of the countries in which they live in addition to whatever other rules they choose to set.

When widely known and enforced, a TOS's provisions can influence user behavior, which in turn affects how traces can be interpreted. Facebook's real-name policy has cultivated a norm of real name use on the platform, which has important implications for inferences of gender and race, as I discuss below. But because Twitter's TOS does not require its users to tweet under their real names, comparable inferences cannot always be made for that platform. In a very different example, researchers interested in TOS violations such as hate speech or advocacy of violence must account for the fact that platforms often remove such content quickly. Such vigilance, while laudable from most users' perspectives, complicates the task of measuring the prevalence of such behavior. However, researchers may be able to effectively track certain TOS violations that are not vigorously enforced by the platform (e.g. Matias et al., 2015).

Social Context

As both boyd et al. (2010) and Gerlitz and Helmond (2013) attest, traces can contain multitudes. It therefore stands to reason that a single trace cannot assume all of its possible meanings in any given instance. Social context can help researchers decide whether a given trace interpretation is plausible for the study at hand. ~~To return to a previous example,~~ reporters often warn that retweets should not be construed as endorsements (Metaxas et al., 2015), which implies a tendency to assume that they are considered as such in at least some contexts. In contrast, several studies have shown that retweets are valid indicators of ideology among communities of users that discuss politics (Aragón, Kappler, Kaltenbrunner, Laniado, & Volkovich, 2013; Conover et al., 2011; Freelon, Lynch, & Aday, 2015; Freelon, McIlwain, & Clark, in press). Thus we have emic evidence that agreement should not be inferred from retweets in one context, and etic evidence that it should in a different context. Such empirical evidence should be adduced whenever possible to support trace-based inferences.

Researchers should strive to understand the social contexts of their research as thoroughly as possible, obvious as that may sound. Unfortunately, trace-based research conducted using computational methods does not always reflect such understanding (Freelon, 2015b). One task for which this is especially important is the inference of certain identity characteristics (e.g. gender and race) from first names. One widely accepted means of inferring gender from first names is to use a dictionary of popular first names keyed to the genders they most often predict. Twitter poses a problem for this method because it allows its users to post under whatever pseudonym they like, as noted above. Most importantly from the standpoint of social context, available evidence suggests that pseudonym use may not be evenly distributed across social groups. Participants in "Black Twitter," for example, have been known to choose sui generis screen names that cleverly allude to pop culture figures and media (see Clayton, 2013). LGBTQ individuals also frequently adopt nontraditional names to express their identities. Such pseudonyms may confound tools for inferring identity characteristics specifically for those social groups who do not use their given names, or whose given names are unique. This in turn may lead to disproportionately high levels of "unknown" categorizations for members of these groups. Understanding such social contexts can help researchers address the methodological challenges they present.

Potential for Misrepresentation

In a perfect world, every digital trace would directly index a specific action committed by a specific human being. Needless to say, they do not, and a key reason for that is willful misrepresentation by duplicitous parties. Thieves and fraudsters have created false traces whenever and wherever they can profit from doing so, with varying levels of success. Machine-generated spam promising wealth, health, beauty, love, and other human desiderata will almost certainly be familiar to anyone with an email account. Some politicians and other would-be notables have purchased non-human followers for themselves on Twitter and other social network sites to cultivate the illusion of popularity (Cresci, Di Pietro, Petrocchi, Spognardi, & Tesconi, 2015; Stringhini et al., 2013). These are just two digital examples of Campbell's law, which holds that valuable social metrics will inevitably be gamed and distorted (Campbell, 1979; cf. Karpf, 2012).

Campbell's law implies that not every trace will be subject to the same degree of pressure toward misrepresentation. The greater the opportunity for tangible benefit, the greater the potential for misrepresentation. Commercial spammers who target social media mostly focus on a particular range of businesses, including finance, dietary and health products, marketing, and consumer electronics (Lee, Eoff, & Caverlee, 2011; Sridharan, Shankar, & Gupta, 2012). Other things being equal, datasets devoted to such commercial topics should exhibit more of a spam problem than those covering other topics. One major exception is episodes of contentious politics in some non-Western countries, which have seen unknown

parties inundating political conversations with machine-generated nonsense (Thomas, Grier, & Paxson, 2012; Verkamp & Gupta, 2013). Bot-detection methods and region-specific expertise can help researchers discern when automated conversation hijacking will be a more or less serious concern.

Individual-Level Characteristics

In the following sections, I will evaluate three general types of traces from which individual-level characteristics are often inferred. These trace types do not have established names, so here I refer to them as *direct indication*, *proper names,* and *speech patterns*. Each of these is common in sociotechnical systems and has been analyzed by a substantial body of work. Further, each is relevant to multiple individual-level characteristics, of which I discuss three: gender, R/E, and geographical location. I chose these particular characteristics for several reasons: first, each harbors clear value for one or more political communication theories. Second, and accordingly, empirical attempts to detect each from text are common in the communication, political science, sociology, and/or social computing literatures. Third, each of these characteristics has an objective answer known to someone somewhere, even if discovering it is prohibitively difficult for researchers. Everyone has a gender identity, an ethnic identity, and a physical location. In contrast, there are no objective ways of judging whether something is (for example) funny, racist, or attractive. An exploration of how best to infer such subjective characteristics from digital traces is beyond the scope of the current chapter.

Direct Indication

One of the simplest methods of inferring individual characteristics from digital traces is simply to take users at their word. One important way platforms allow researchers to do so is through *direct indication*, by which I mean dedicated fields through which users can (or must) declare specific facts about themselves. Facebook, for example, offers users a mandatory open-text box into which users can enter whatever gender label(s) fits them best.[1] Instagram offers direct indication for gender, but the only three options are "Male," "Female," and "Not Specified" (the default). Twitter's design does not permit direct indication of gender, and none of the three permit direct indication of R/E. Because Facebook requires users to indicate a gender while Instagram does not, direct indication is a superior source of gender information for the former than it is for the latter.[2] Facebook enacts its real-identity TOS requirement in part by requiring users to indicate their gender as a condition of account creation, and Instagram enacts its looser identity policy by not requiring it. Social context is likely to be a major issue for platforms in which users can choose to hide their gender. Facebook's terms of service require gender indication uniformly across all social contexts, but on Instagram it may be ~~more~~ customary in some contexts (e.g. politically-charged

ones) than in others. For the same reason, systematic gender misrepresentation is less likely on Facebook than it is on Instagram.

Many sociotechnical systems also permit the direct indication of location information. For social media, this usually means either GPS-generated location data or text strings that may or may not correspond to identifiable physical locations. There are at least four key design considerations here: first, whether the system offers a dedicated location field; second, if such a field exists, whether location indication is opt-in or opt-out; third, whether the field supports GPS; and fourth, whether the field supports auto-complete for locations. An opt-out policy obviously makes a location field much more useful from a research perspective than opt-in, ethical considerations notwithstanding. Studies of Twitter, whose location field is opt-in, have found substantial numbers of users for whom locations cannot be ~~resolved above the country level~~ (Hecht, Hong, Suh, & Chi, 2011; Leetaru, Wang, Padmanabhan, & Shook, 2013; Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011). This is because they either left the field blank or entered a string that did not match a known place name ~~above the country level~~. Where available, GPS is the gold standard for location data, as it can estimate a user's physical location to within several feet. But GPS is rarely enabled by default, and when it is not (as with Facebook, Twitter, and Instagram), very few users bother to turn it on (Leetaru et al., 2013; Morstatter, Pfeffer, Liu, & Carley, 2013). Thus, studies that include only users who have turned on GPS location run a high risk of systematically omitting certain types of users. One study that investigated this question directly on Twitter found that younger, urban, higher-income, Black, and Hispanic users were more likely to opt in to GPS (Malik, Lamba, Nakos, & Pfeffer, 2015). The same is likely true for textual locations: certain groups may be more likely to use official place names, while others may more often use informal location names or nonstandard abbreviations that cannot be resolved by automated location-guessing systems. Users can also misrepresent their locations, either intentionally or unintentionally. Among other reasons, intentional misrepresentation can occur as an act of solidarity, as when Twitter users around the world changed their locations to "Tehran" in an attempt to stymie the Iranian authorities' attempts to surveil local protesters' tweets (Mueller & van Huellen, 2012). Since location fields are usually static, but people move around quite a bit, researchers may find discrepancies between users' listed and actual locations. Though unintentional, these discrepancies may reduce the validity of trace inferences drawn in studies of major protests, concerts, conferences, and other events to which many people travel.

Proper Names

People's names can convey a wide range of relevant facts about them. In English and many other widely-spoken languages, most given names are overwhelmingly used by either males or females (Flowers, 2015). Thus, gender can usually be inferred from a person's given name (e.g. Burger, Henderson, Kim, & Zarrella,

2011; Freelon, Becker, Lannon, & Pendleton, 2016; Liu & Ruths, 2013; Mislove et al., 2011; Sloan et al., 2013). Most platforms have one or more fields through which users can or must name themselves, but as discussed above, not all require the use of official names. Even some sites that do not require real names have strong social norms toward real name usage (Google Plus, for example). For such sites, researchers can use simple dictionary-based gender-guessing programs for some populations such as Genderize (https://genderize.io/), Genderator (https://github.com/bmuller/genderator), and Gender API (https://gender-api.com/). These programs contain large dictionaries of common names, each of which is indexed to the gender it most often indicates: male for "John," female for "Mary," etc. Gender-guessing programs typically assign the corresponding gender for each name present in its dictionary and "Unknown" values to all names not present or that indicate maleness about as often as femaleness. Real-name requirements that are strictly enforced reduce the potential for misrepresentation, since misrepresentation risks negative consequences. So as long as the names in your gender-guessing program largely match those in your population, you would have a strong ~~use~~ case for using it.

But this is not always the case. Gender-guessing programs reflect their creators' social backgrounds, and often skew toward traditional English and American names (e.g. Muller, 2012). Online social contexts that lie beyond the programmers' familiarity may therefore not be the best match for such programs. Services such as Genderize and Gender API attempt to overcome this limitation by using dictionaries of hundreds of thousands or millions of names across dozens of languages. But these particular services' dictionaries are closed and proprietary, preventing users from seeing for themselves how comprehensive they are (see Guo's chapter in this volume for more discussion of this issue). And even the most expansive and open dictionary will not suffice for populations that favor unique names and spellings, such as African Americans. Handling non-Latin character sets effectively is another major issue for inferring gender from names written in those scripts. Of course, in spaces where pseudonym use is the norm, none of these methods may provide acceptable levels of accuracy.

Several major differences in inferring R/E vs. gender from proper names become apparent immediately. First, few if any major social media platforms even permit, let alone require, users to indicate R/E directly. Most of the time it can only be inferred indirectly, with users' family names (rather than given names) being a popular data source (Chang, Rosenn, Backstrom, & Marlow, 2010; Fiscella & Fremont, 2006; Mislove et al., 2011). As with gender, the overall efficacy of this method for inferring R/E depends on the availability of the user's full name, which in turn depends on the four domains. Design-wise, the number and label(s) of the "name" field(s) is a major consideration. Facebook offers "first name" and "last name" fields as well as a field for "other names" including nicknames, maiden names, former names, and the like. Twitter offers two name-related fields: one simply labeled "Name," the other labeled "Username." Both of these can be easily

altered at will, unlike Facebook's primary name fields, which can be changed only once every 60 days and are subject to multiple content restrictions. Thus, Twitter's technical environment is more hospitable to creative pseudonyms, from which gender is more difficult to infer. Some will take more advantage than others of opportunities to create pseudonyms, for example those interested in sensitive content such as drugs, guns, pornography, and racial hatred (Peddinti, Ross, & Cappos, 2014). Social contexts in which impersonation is relatively common, such as politics and entertainment (Freelon & Karpf, 2015; Highfield, 2016), should also receive extra scrutiny. But even given the presence of a credible full name, dictionary-based gender-guessing techniques are not equally effective for all races and ethnicities. It works better for Latinos and Asians than for African Americans, women, and individuals of higher socio-economic status (Fiscella & Fremont, 2006). Generally, it appears that the more heterogeneous the sample, the worse this technique is likely to perform on ~~a particular segment of~~ it.

Speech Patterns

A third trace-based inference technique exploits group-level differences in speech patterns. The idea here is that people who share a certain trait (gender, R/E, location, etc.) will tend to speak in ways that distinguish them from those who do not possess the trait. From a design standpoint, as long as a system allows users some degree of free textual expression (as all social media do), it will always offer the researcher something to analyze. However, the method's viability may depend on the extent to which the design restricts users' expressive latitude, for example through limitations on the number of posts or characters per post permitted (Burger et al., 2011; Peersman, Daelemans, & Van Vaerenbergh, 2011). Terms of service play a similar role: prohibitions on certain forms of expression could potentially affect a researcher's ability to infer certain characteristics from speech. For example, both Twitter and Facebook's terms forbid users from engaging in abusive behavior, which could potentially include words that predict membership in gender, racial, or affinity groups. But social context is probably the most consequential domain for this method, especially as it applies to R/E and gender. The predictive value of speech patterns for identity characteristics relies on the strength of the correlation between group membership and social context. In other words, speech pattern-based inference techniques assume that women will talk in distinctly "female" ways, men will talk in "male" ways, Blacks will talk in "Black" ways, etc. The truer this assumption, the more valid the results will be. It has long been known that men and women tend to speak differently in the aggregate, and speech-based studies of gender detection have exploited this fact to achieve gender classification rates of 70–90% (Bamman, Eisenstein, & Schnoebelen, 2014; Burger et al., 2011; Sap et al., 2014). While these rates are indeed high, they mask the fact that these methods may systematically misclassify certain subsets of individuals—those with heterogeneous social networks,

for example (Bamman et al., 2014). Therefore, researchers should use the empirical record and expert knowledge about the population under study to ascertain whether speech-based classification will perform adequately for any given case. This would of course include any group-level proclivities toward identity tourism (Nakamura, 1995), which could prove misleading.

Several studies have also used the full text of user posts to identify their physical locations (e.g. Cheng, Caverlee, & Lee, 2010; Li, Wang, Deng, Wang, & Chang, 2012; Mahmud, Nichols, & Drews, 2012; Stefanidis, Crooks, & Radzikowski, 2013). In platforms that lack direct indication, this may be the only means of location identification available. Most of the technical design and TOS constraints are the same as for detection R/E and gender, so I will not reprise those here; the most substantial differences lie in the social domains. Simply put, the tendency to talk about one's physical location is probably not evenly distributed across a given platform's user population. The issue here is similar to GPS opt-in, with the main difference being that it is very easy to determine when someone has opted out of GPS. But it's much more difficult to determine when people regularly mention places they have been. A typical approach would process a corpus of social media posts through a dictionary of location names and analyze the matches (see Leetaru et al., 2013). But users will likely differ greatly in the volume of identifiable locations they post. Moreover, while some users' hits may represent places they've been, others might be places they want to go, places they've visited in the past, or places in the news. This could be seen as a form of misrepresentation, albeit one created as an unavoidable side effect of this technique's basic assumptions. The solution is the same: expert knowledge and a full consideration of the extent to which deviations from the assumptions might harm the analysis.

Conclusion

As we have seen, the value of trace data inference techniques depends on differences between cases that can be expressed in terms of the four domains of technical design, TOS, social context, and potential for misrepresentation. The type of pre-research analysis demonstrated here will help researchers judge when particular techniques will be more and less effective. Unfortunately, there is no quantitative threshold to cleanly separate "effective" and "ineffective" research applications. Instead, researchers will have to make their cases based on the specifics of each situation and on prior research practice.

I hope I have made it clear that in many cases inferring individual characteristics from trace data will *not* be a straightforward affair. The staggering multiplicity of ways users can express gender, race/ethnicity, and location presents nontrivial challenges for both manual and computational methods. Moreover, in some research contexts none of the available methods will yield sufficient classification rates or levels of validity. There are no guarantees in research, especially when

using data that was not created for that purpose. Only through shared empirical standards and frameworks will we be able to recognize the differences between higher- and lower-quality trace inferences.

The four dimensions of trace inference offer a useful framework for evaluating the construct validity of trace-based inferences. Each contributes independently to the quality of the argument that a given trace reliably and validly indicates a given concept within a given context. Clearly the technical design must enable the provision of certain traces for researchers to be able to analyze them, and so much the better if the design actively encourages it. A platform's TOS is almost as powerful in this regard, although its relevance depends upon how strictly it is enforced. Social context reminds us that inferential validity depends to a large extent on use: inferences that are valid in one context may not be in another. The same goes for the potential for misrepresentation, which is technically a special case of social context, but recurs frequently enough to warrant its own category.

While this chapter analyzed three types of trace inference techniques individually, I should point out that many studies base their inferences on multiple trace indicators. In studies that use machine-learning classification, multiple indicators are additive: the more of them point toward a particular category, the stronger the confidence that that category is the correct one. For example, using speech patterns and proper names together has been demonstrated to increase the percentage of correct gender classifications (Burger et al., 2011). But the framework introduced in this chapter suggests that inferences based on some traces may be inherently more valid than those based on others. Consider the difference in validity between classifying an individual as a woman, or Black, or in New York City because 1) she stated as much directly, 2) her name is disproportionately common among members of the first two categories, or 3) she tends to speak in ways typical of people in those categories. Per the previous section, it is possible, and likely in some cases, that the subsets of individuals given incorrect and "unknown" judgments will differ systematically across these three techniques. But perhaps even more importantly, we should ask ourselves: are the outcomes of each of these inference techniques really epistemologically equivalent?

This should direct us to think not only about how to maximize classification rates, but also about the reasons behind our misclassifications and the extent to which certain groups may be excluded from analysis. We already know that most social media platforms (aside from perhaps Facebook in some cases) are not representative of any broader offline population. Ignoring possible bias in our classification techniques may skew social media datasets even further. It seems important to know, for example, if a given technique works twice as well for one subpopulation as it does for another. To address this issue, we must understand how it applies to our particular datasets and introduce methods that include all the sub-populations relevant to our study.

The discussions above demonstrate the four dimensions' value for the specific characteristics of gender, R/E, and location, but it is not limited to them.

Articulating the dimensions at such a high level of abstraction allows them to be applied to other characteristics and platforms. They will almost certainly prove useful for the study of subjective concepts, which are not addressed in this chapter. And as platforms inevitably rise and fall in popularity and digital communication technologies continue to advance, these dimensions will retain their relevance because they are not tied to any single platform.

Methods, too, will continue to develop. As computational researchers, we find ourselves in something of an arms race with platform developers: as soon as methodological best practices for widely used traces begin to solidify, new traces emerge from which new inferences might be drawn. For example, in early 2016 Facebook expanded its menu of one-click reaction indicators from one (the "like" button) to six (new buttons for "love," "haha," "wow," "sad," and "angry"). ~~As with the "like" button, the five new reactions are available as data through Facebook's API for users whose profiles are configured for maximum publicity.~~ Facebook executives may well intend and believe that these buttons transparently convey the sentiments on their labels across all users and contexts. But as I hope I have sufficiently argued here, such claims should only be established on a foundation of solid conceptual and empirical substantiation.

Notes

1 When users first sign up for a Facebook account, they must indicate that they are "Male" or "Female." They can access the more inclusive open-text option only after their account has been created.

2 Facebook's API does not grant access to this gender information, although it is visible to a user's friends and can also be accessed by Facebook apps with the user's permission.

References

Aragón, P., Kappler, K. E., Kaltenbrunner, A., Laniado, D., & Volkovich, Y. (2013). Communication dynamics in Twitter during political campaigns: The case of the 2011 Spanish national election. *Policy & Internet, 5*(2), 183-206. https://doi.org/10.1002/1944–2866.POI327

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics, 18*(2), 135-160. https://doi.org/10.1111/josl.12080

Bowker, G. C. (2005). *Memory practices in the sciences*. Cambridge, MA: MIT Press.

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1301-1309). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2145432.2145568

Campbell, D.T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning, 2*(1), 67-90. https://doi.org/10.1016/0149-7189(79)90048-X

Chang, J., Rosenn, I., Backstrom, L., & Marlow, C. (2010). ePluribus: Ethnicity on social networks. *ICWSM, 10*, 18-25.

Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating Twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 759-768). New York, NY: ACM. https://doi.org/10.1145/1871437.1871535

Clayton, T. (2013, July 28). Black Twitter's best screen names. Retrieved from www.theroot.com/blog/the-grapevine/black_twitter_the_best_screen_names/

Conover, M. D., Ratkiewicz, J., Francisco, M., Goncalves, B., Flammini, A., & Menczer, F. (2011). Political polarization on Twitter. In *Proceedings of the 5th International Conference on Weblogs and Social Media* (pp. 89-96). Barcelona, Spain: AAAI.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2015). Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems, 80*, 56-71. https://doi.org/10.1016/j.dss.2015.09.003

Fiscella, K., & Fremont, A. M. (2006). Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research, 41*(4p1), 1482-1500.

Flowers, A. (2015, June 10). The most common unisex names in America: Is yours one of them? Retrieved from http://fivethirtyeight.com/features/there-are-922-unisex-names-in-america-is-yours-one-of-them/

Freelon, D. (2010). Analyzing online political discussion using three models of democratic communication. *New Media & Society, 12*(7), 1172-1190.

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

Freelon, D. (2015a). Discourse architecture, ideology, and democratic norms in online political discussion. *New Media & Society, 17*(5), 772-791. https://doi.org/10.1177/1461444813513259

Freelon, D. (2015b). On the cutting edge of big data: Digital politics research in the social computing literature. In S. Coleman & D. Freelon (Eds.), *Handbook of digital politics* (pp. 451-472). Northampton, MA: Edward Elgar.

Freelon, D., Becker, A. B., Lannon, B., & Pendleton, A. (2016). Narrowing the gap: Gender and mobilization in net neutrality advocacy. *International Journal of Communication, 10*(0), 5908-5930.

Freelon, D., & Karpf, D. (2015). Of big birds and bayonets: Hybrid Twitter interactivity in the 2012 presidential debates. *Information, Communication & Society, 18*(4), 390-406. https://doi.org/10.1080/1369118X.2014.952659

Freelon, D., Lynch, M., & Aday, S. (2015). Online fragmentation in wartime: A longitudinal analysis of tweets about Syria, 2011–2013. *The ANNALS of the American Academy of Political and Social Science, 659*(1), 166-179. https://doi.org/10.1177/0002716214563921

Freelon, D., McIlwain, C. D., & Clark, M. D. (in press). Quantifying the power and consequences of social media protest. *New Media & Society*.

Gerlitz, C., & Helmond, A. (2013). The like economy: Social buttons and the data-intensive web. *New Media & Society, 15*(8), 1348-1365. https://doi.org/10.1177/1461444812472322

Gitelman, L. (2013). *Raw data is an oxymoron*. Cambridge, MA: MIT Press.

Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 237-246). New York, NY: ACM. https://doi.org/10.1145/1978942.1978976

Highfield, T. (2016). News via Voldemort: Parody accounts in topical discussions on Twitter. *New Media & Society, 18*(9), 2028-2045. https://doi.org/10.1177/1461444815576703

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems, 12*(12), 767-797.

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

Lee, K., Eoff, B. D., & Caverlee, J. (2011). Seven months with the devils: A long-term study of content polluters on Twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*. Retrieved from www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2780

Leetaru, K., Wang, S., Padmanabhan, A., & Shook, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday, 18*(5). https://doi.org/10.5210/fm.v18i5.4366

Li, R., Wang, S., Deng, H., Wang, R., & Chang, K. C.-C. (2012). Towards social user profiling: Unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1023-1031). New York, NY: ACM. https://doi.org/10.1145/2339530.2339692

Liu, W., & Ruths, D. (2013). What's in a name? Using first names as features for gender inference in Twitter. In *AAAI Spring Symposium: Analyzing Microtext* (Vol. 13, p. 1). Retrieved from https://pdfs.semanticscholar.org/b60d/04043a60e46670f182b2debb485e9d17ce46.pdf

Mahmud, J., Nichols, J., & Drews, C. (2012). Where is this tweet from? Inferring home locations of Twitter users. In *Sixth International AAAI Conference on Weblogs and Social Media*. Retrieved from www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4605

Malik, M. M., Lamba, H., Nakos, C., & Pfeffer, J. (2015). Population bias in geotagged tweets. In *Ninth International AAAI Conference on Web and Social Media*. Retrieved from www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10662

Matias, J. N., Johnson, A., Boesel, W. E., Keegan, B., Friedman, J., & DeTar, C. (2015). *Reporting, reviewing, and responding to harassment on Twitter* (SSRN Scholarly Paper No. ID 2602018). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2602018

Metaxas, P., Mustafaraj, E., Wong, K., Zeng, L., O'Keefe, M., & Finn, S. (2015). What do retweets indicate? Results from user survey and meta-review of research. In *Ninth International AAAI Conference on Web and Social Media*. Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.713.6936&rep=rep1&type=pdf

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2011). Understanding the demographics of Twitter users. *ICWSM, 11*, 5th.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's streaming api with Twitter's firehose. *Proceedings of ICWSM*. Retrieved from www.public.asu.edu/~fmorstat/paperpdfs/icwsm2013.pdf

Mueller, P. S., & van Huellen, S. (2012). A revolution in 140 characters? Reflecting on the role of social networking technologies in the 2009 Iranian post-election protests. *Policy & Internet, 4*(3-4), 184-205. https://doi.org/10.1002/poi3.16

Muller, B. (2012). *Genderator*. Python. Retrieved from https://github.com/bmuller/genderator

Nakamura, L. (1995). Race in/for cyberspace: Identity tourism and racial passing on the internet. *Works and Days, 25*(26), 13.

Peddinti, S. T., Ross, K. W., & Cappos, J. (2014). "On the internet, nobody knows you're a dog": A Twitter case study of anonymity in social networks. In *Proceedings of the Second ACM Conference on Online Social Networks* (pp. 83-94). New York, NY: ACM. https://doi.org/10.1145/2660460.2660467

Peersman, C., Daelemans, W., & Van Vaerenbergh, L. (2011). Predicting age and gender in online social networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents* (pp. 37-44). New York, NY: ACM. https://doi.org/10.1145/2065023.2065035

Sack, W. (2005). Discourse architecture and very large-scale conversation. In R. Latham & S. Sassen (Eds.), *Digital Formations: IT and New Architectures in the Global Realm,* 242-282.

Sap, M., Park, G., Eichstaedt, J. C., Kern, M. L., Stillwell, D., Kosinski, M., … Schwartz, H. A. (2014). Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1146-1151).

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., & Rana, O. (2013). Knowing the tweeters: Deriving sociologically relevant demographics from Twitter. *Sociological Research Online, 18*(3), 7.

Sridharan, V., Shankar, V., & Gupta, M. (2012). Twitter games: How successful spammers pick targets. In *Proceedings of the 28th Annual Computer Security Applications Conference* (pp. 389-398). New York, NY: ACM. https://doi.org/10.1145/2420950.2421007

Stefanidis, A., Crooks, A., & Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. *GeoJournal, 78*(2), 319-338.

Stringhini, G., Wang, G., Egele, M., Kruegel, C., Vigna, G., Zheng, H., & Zhao, B.Y. (2013). Follow the green: Growth and dynamics in Twitter follower markets. In *Proceedings of the 2013 Conference on Internet Measurement Conference* (pp. 163-176). New York, NY: ACM. https://doi.org/10.1145/2504730.2504731

Thomas, K., Grier, C., & Paxson, V. (2012). Adapting social spam infrastructure for political censorship. Presented as part of the 5th USENIX Workshop on Large-Scale Exploits and Emergent Threats. Retrieved from www.usenix.org/conference/leet12/workshop-program/presentation/thomas

Verkamp, J.-P., & Gupta, M. (2013). Five incidents, one theme: Twitter spam as a weapon to drown voices of protest. Presented as part of the 3rd USENIX Workshop on Free and Open Communications on the Internet. Retrieved from www.usenix.org/conference/foci13/workshop-program/presentation/verkamp

Wright, S., & Street, J. (2007). Democracy, deliberation and design: The case of online discussion forums. *New Media & Society, 9*(5), 849.