BIG DATA: PITFALLS, METHODS AND CONCEPTS FOR AN EMERGENT FIELD

Author: Zeynep Tufekci

Affiliation: Princeton University and University of North Carolina

Email: zst@princeton.edu

PLEASE NOTE THAT THIS IS A DRAFT. FEEDBACK, CRITICISM AND SUGGESTIONS ARE WELCOME

PLEASE EMAIL ME AT ZST@PRINCETON.EDU OR USE TWITTER (@techsoc).

THANK YOU FOR YOUR INTEREST!

Abstract

Big Data, large-scale aggregate databases of imprints of online and social media activity, has captured scientific and policy attention. However, this emergent field is challenged by inadequate attention to methodological and conceptual issues. I review key methodological and conceptual challenges including: 1.Inadequate attention to the implicit and explicit structural biases of the platform(s) most frequently used to generate datasets (the model organism problem). 2.The common practice of selecting on the dependent variable without corresponding attention to the complications of this path. 3.Lack of clarity with regard to sampling, universe and representativeness (the denominator problem). 4.Most big data analyses come from a single platform (hence missing the ecology of information flows).

Conceptual issues include: 1.More research is needed to interpret aggregated mediated interactions. Clicks, status updates, links, retweets, etc. are complex social interactions. 2.Network methods imported from other fields need to be carefully reconsidered to evaluate appropriateness for analyzing human social media imprints. 3.Most big datasets contain information only on "node-to-node" interaction. However, "field" effects--events that affect a society or a group in a wholesale fashion either through shared experience or through broadcast media—are an important part of human socio-cultural experience. 4.Human reflexivity –that humans will alter behaviors around metrics-- needs to be assumed and built into the analysis. 5.Assuming additivity and counting interactions so that each new interaction is seen as (n+1) without regards to the semantics or context can be misleading. 6.The relationship between network structure and other attributes is complex and multi-faceted.

BIG DATA: PITFALLS, METHODS AND CONCEPTS FOR AN EMERGENT FIELD

The dramatic proliferation of technologically mediated human interaction produces online imprints which are increasingly aggregated into large databases. Such large datasets, especially of *social media imprints,* commonly referred to as *big data,* have been analyzed by scholars, corporations, politicians, journalists, and governments (Lazer et al., 2009; boyd & Crawford, 2012). Although big data is being variously touted as the key to rigor in social science and as an important basis for policy, this emergent field suffers from inadequate attention to methodological and conceptual issues.

Methodological issues which will be examined in this paper include the following:

1. Inadequate attention to the implicit and explicit structural biases of the platform(s) most frequently used to generate datasets (the model organism problem);

2. The common practice of selecting on the dependent variable without corresponding attention to the complications of this path. (Most hashtag analyses, for example, involve selecting on the dependent variable.)

3. Lack of clarity with regard to sampling, universe and representativeness (the denominator problem);

4. Most big data analyses come from a single platform (hence missing the ecology and the natural setting of information flows and interaction).

The conceptual issues related to *big data* analysis that are examined in this paper include:

1. More research is needed to interpret aggregated mediated interactions. Clicks, status updates, links, retweets, etc. are complex social interactions with varying meanings, logics and implications.

2. Network methods imported from other fields need to be carefully and thoroughly reconsidered to evaluate appropriateness for analyzing human social media imprints.

3. Most big datasets contain information only on “node-to-node” interaction. However, that is not the only mechanism in human societies. “Field” effects--events that affect a society or a group in a wholesale fashion either through shared experience or through broadcast media—are an important part of human socio-cultural experience.

4. Human reflexivity –that humans will alter behaviors around metrics-- needs to be assumed and built into the nuanced analysis of data.

5. Assuming additivity and counting interactions so that each new interaction is seen as (n+1) without regards to the semantics or context can be misleading.

6- The relationship between network structure and other properties and variables is often complex and cannot always be resolved by *big data* methods.

METHODOLOGICAL CONSIDERATIONS:

1. Model organisms and research: Twitter as the field’s *Drosophila melanogaster:*

While there are many social media platforms, Twitter is used disproportionately in large scale *big data* research, especially those involving millions or billions of data points. This is mostly due to data availability, tools availability and popularity, and ease of analysis. While Facebook is the largest social media platform, truly public (open the Web) data on Facebook less and thus Facebook is less accessible by scraping or via Facebook’s API as many more Facebook users (estimated to be more than 50%) have taken their profiles private compared with Twitter users (estimated to be less than 10%). The Twitter stream has also been relatively easy to access using a variety of widely available and popular methods (the Twitter *firehose*, *the spritzer*, white-listed accounts, etc.) while Facebook’s API is both lesser-known and has fewer ready-made tools (thanks to an anonymous reviewer for pointing out Facebook’s API allows access to public parts of Facebook). While Twitter has been changing and also closing off some of the easier means of access, the key difference remains that bulk of Facebook activity occurs within the privacy of profiles that are largely inaccessible for purposes of research while almost all of Twitter (accept for Direct Messages) is on the publicly visible Internet except.

Twitter also has a simple and clean data structure. In contrast to the finer-grained privacy settings on other social media platforms, Twitter profiles are either "all public" or "all private." Every aspect of non-private profiles –the full social graph, retweets, mentions, lists—are accessible via APIs (Application Programming Interfaces). Thus, with a maximum of 140 characters per tweet, and only a few basic functions (retweet, mention, and hashtags) to map, the datasets generated by Twitter are relatively easy to structure, handle and analyze.

Consequently, Twitter has emerged as a "model organism" of *big data*. In biology, "model organisms" refer to species which have been selected for intensive examination by the research community in order to shed light on key biological processes such as fundamental properties of living cells. Model organism-program of research in biology has been spectacularly successful (Fields & Johnston, 2005). However, this investigative path is not without trade-offs. For example, all dominant biological model organisms –such as the fruit fly *Drosophila melanogaster*, the bacterium *Escherichia coli*, the nematode worm *Caenorhabditis elegans*, the mouse *Mus musculus*-- have been selected because they are easy to breed in artificial settings (lab-friendly), have rapid life cycles (quick results) and small adult size (adding to their lab-friendliness), incorporate "rapid and stereotypical" development (making experimental comparisons easier), and involve "early separation of germ line and soma," (reducing certain kinds of variability) (Bolker, 1995; Jenner & Willis, 2007). However, these very characteristics which make them useful for studying certain biological mechanisms come at the expense of illuminating others (Gilbert, 2001; Jenner & Wills, 2007). For example, being easy to handle in the laboratory, in effect, implies having been selected for "relative insensitivity to environmental influences"

(Bolker, p. 451-2) and thus relative lack of suitability to examining environmental interactions. The rapid development cycle depresses mechanisms present in slower growing species and small adult size can imply “simplification or loss of structures, the evolution of morphological novelties, and increased morphological variability” (Bolker, p. 452).

In other words, model organisms are not necessarily representative of their taxa. Consequently, biologists, especially in subfields such as eco-evo (ecology and evolution) and evo-devo (evolution and development), have turned to expanding their set of model organisms to capture the broader, more diverse processes.

The dominance of Twitter as the “model organism” for social media in big data analyses similarly skews analysis. Each social media platform carries with it certain affordances which structure its social norms and interactions and may not be representative of other social media platforms, or general human social behavior. In the case of Twitter, the key characteristics that structure its interactions are short message length, rapid turnover, public nature, and a directed graph of social network interaction.

Similar to the rapidly reproducing, small model organisms, Twitter is dominated by rapid turnover of small chunks of text, which means that it is more suitable to certain kinds of interactions while discouraging others. It is thus lacking in some of the characteristics that blogs, LiveJournal communities, or Facebook possess such as longer-length texts, norms of lengthier reaction time, stronger integration of visuals with text, and evolution of conversations over longer periods of time.

Twitter’s lightweight application interface, suitable to mobile devices and accessible via texting, means that it is often the platform of choice when posting outdoors, from high-tension events (demonstrations, etc.) or from low-bandwidth environments. The retweet mechanism also engenders its own complex set of status-related behaviors and norms which do not necessarily translate to other online platforms.

Crucially for behavior within the network, Twitter is a directed graph--which means that a person can “follow” someone else without their consent or mutuality. In contrast, Facebook’s backbone is mostly an undirected graph—“friending” requires mutual consent. Similarly, Livejournal’s core interactions tend to occur within “friends lists”. While Facebook’s mutuality is built in, LiveJournal’s depends on strong norms governing interaction. Twitter has neither of these things, which means that Twitter is more likely than other platforms to sustain bridge mechanisms and cross-over communities. As an example, analysis indeed shows that bit.ly shortened links distributed on Twitter using revolutionary hashtags played a key role as an information conduit from within the Arab uprisings to the outside world (Aday et al. 2012). That cannot be interpreted however, to mean that other social media *as a whole* also mostly acted as a bridge mechanism--or even that Twitter was solely a bridge mechanism.

Finally, Twitter is used by about 10% of the US population (Pew Research Center, 2013) which is certainly far, far from a representative sample. While Facebook has a wider diffusion rate, its rates of use are structured by race, gender, class and other factors and are also not representative (Hargittai, 2007). Using these sources as “big data” model organisms raises important questions of representation and visibility as different demographic or social groups may have different behavior –online and offline—and may not be fully represented or even sampled via current methods.

This does not imply that Twitter is an inappropriate social network to study. Research in the model organism paradigm can be quite illuminating, as it allows a large community of scholars to coalesce around similar datasets and problems. The field should not, however, lose sight of specific features of each platform and questions of representativeness.

2. Hashtag Analyses and Selecting on the Dependent Variable:

Many studies extract relevant tweets using hashtags, a twitter convention for marking a specific tweet as part of a particular conversation. For example, the Tunisian uprising was marked by the hashtag #sidibouzid while the initial Egyptian protests, scheduled for January 25, 2011, were associated with the hashtag #jan25. While hashtag studies can be a powerful method for examining information flows, all hashtag analyses, *in effect and by definition*, select on the dependent variable--with all concomitant features and weaknesses of this methodological path. Hashtag datasets should also be seen as self-selected samples with data "*missing not at random*" and interpreted and analyzed accordingly (Allison, 2001; Dunning and Freeman, 2007).

"Selecting on the dependent variable" occurs when inclusion of a case in a sample depends on a particular outcome of the very variable being examined. Such samples have limited analytic power and can have misleading results since the variation on the dependent variable is limited. For example, analysis of the conditions for emergence of revolutions or wars performed only by examining cases with wars or revolutions that have occurred will, necessarily, have limited conceptual power (Geddes, 1990) as it is missing the cases which might have similar characteristics but in which there have been no wars or revolutions. In hashtag datasets, a tweet gets included in the dataset precisely because it has a particular outcome already attached to it. Further, most hashtags used to build big datasets are successful hashtags—ones that got well-known, distributed widely and generated large amount of interest. It is likely that dynamics of such events differ significantly from those of less successful ones. Selecting on the dependent variable can introduce a range of errors specifics of which depend on the characteristics of the uncorrelated sample.

Deciding to use a particular hashtag is an act of self-selection. This is especially true for political hashtags which are used mostly among politically-engaged participants. Samples of self-selected cases are fraught with selection effects—i.e. a self-selected population will not only have different overall characteristics than the general population, they may also exhibit significantly different correlational tendencies, which create thorny issues of confounding variables. Famous errors have this kind include the hormone replacement therapy (HRT) controversy in which researchers had, erroneously, believed that the HRT conferred health benefits to post-menopausal women when, in fact, this was based on observational studies of women who self-selected to take the HRT. In reality, HRT therapy was adopted by healthier women; randomized, properly-sampled and blind studies showed that HRT was, in fact, harmful.

A hashtag is often loaded with assumptions, meaning and cultural or political structure. Thus, hashtag use, besides being an act of self-selection, often involves participation and engagement with the framework that embeds the hashtag. Sometimes, the use of the hashtag is a declaration of particular sympathy. In other cases, there may be warring messages as the hashtag emerges as a contested cultural space. For example, twitter users who use the hashtag "#jan25" are more likely to be sympathetic to the Egyptian revolution. In contrast, "#Bahrain" tends to be used both by supporters and opposition of the uprising in Bahrain. On the other hand, "#cairotraffic" can unite different political factions (but divide by social class). (This observation was based on two years of regular monitoring of activity --checking for an hour once a week -- on both hashtags as well as conversations with people active on those hashtags. Upon the helpful suggestion of a reviewer, I also collected more systematic data on three occasions and downloaded 100 tweets and found that only about 1 in 100 #jan25 tweets were neutral while the rest were all supporting the revolution while about 5 #Bahrain tweets out of 100 were neutral (ads for competition to win Blackberry) while about 15-10 out of 100 tweets accused the protestors of being terrorists, while the rest were supportive of the uprising.)

This is not to argue that hashtag datasets are not useful. In contrast, they can provide a fascinating and illuminating glimpse into specific cultural and socio-political conversations. However, hashtag dataset analysis need to be accompanied by a thorough discussion of the culture of the specific hashtag and analyzed with careful consideration of any methods of their constitution which introduce significant selection and sampling biases.

3. The Missing Denominator: We Know Who Clicked But We Don't Know Who Saw Or Could.

One of the biggest methodological dangers of big data analyses is insufficient understanding of the underlying sample and denominators. It's often not enough to understand how many people have "liked" a Facebook status update, clicked on a link, or "retweeted" a message without having a sense of how many people saw and chose to –or not to– take that option. That kind of normalization is rarely done, or may even be actively decided against because the results start appearing more complex or more trivial (Cha et al., 2010).

While the precise denominator is often not calculable, in many cases, it may be possible to calculate preliminary estimates. One measure might be "potential exposure," corresponding to the maximum number of people who might have seen a message. This also highlights a key issue for in big data research: the data is often proprietary (boyd and Crawford, 2012). However, it might be possible to work with these platforms to get estimates of visibility, click-through and availability. For example, Facebook has allowed its researchers to disclose information about potential audiences for status updates –for example, the mean and median fraction of a user's friends that see the post is about 34-35% of the universe of friends, though the distribution of the variable seems to have a large spread (Bersntein et al, 2013).

With more research and more disclosure from proprietary platforms, it may also be possible to calculate “likely” exposure numbers based on “potential” exposure—similar to the way election polling models likely voters or TV ratings industry tries to capture people watching a TV rather than just being in the room. Steps in this direction are likely to be complex and difficult, but without such efforts, our ability to interpret raw numbers will remain limited as we won’t even have an estimate of the denominators. Many platforms, such as Facebook, Twitter, OKCupid and others, have in-house researchers who participate in academic conferences and publish in academic journals. The academic community should engage these researchers for more disclosure and access in these regards.

It’s also important to normalize underlying populations when comparing “clicks,” “links,” or tweets. For example, Aday et al. (2012) compares number of clicks on bit.ly links in tweets containing hashtags associated with the Arab uprisings and concludes that “new media outlets that that use bit.ly are more likely to spread information outside the region than inside it.” However, it is hard to contextualize this conclusion without a sense of the respective populations of Twitter users inside and outside the countries in question. Egypt’s population is about 80 million, hence about 1 percent of the global population--any topic of global interest about Egypt could very easily generate more *absolute* number of clicks outside the country even as the activity within the country remained much more concentrated in *relative* proportions.

Coupled with the lack of representativeness and (often) lack of random sampling in the sources of big data –as discussed above—the denominator and sampling issues raise many troubling questions about bot the representativeness and fairness of generalizing from many available big data sets.

4. Missing the Ecology for the Platform:

Most existing *big data* analyses of social media are confined to a single platform, often Twitter, as discussed above. However, most of the topics of interest to such studies, such as influence or information flow can rarely be confined to the Internet, let alone to a single platform. Understandable difficulty in obtaining high-quality multi-platform data does not mean that we can treat a single platform as a closed and insular system, as if human information flows were all gases in a chamber. They are not.

The emergent media ecology is an integrated mix of old and new media rather than strictly segregated by platform or even device. Many "viral" videos take off on social media *only* after being featured on broadcast media, and this step itself is often preceded by being highlighted on intermediary sites such as Reddit or Buzzfeed. Political news flowing out of political Arab uprisings to broadcast media was also often curated by sites that had emerged as trusted local information brokers such as Nawaat.org. These and other examples show the object of analysis should be this integrated ecology, that rather than stay at "which site" or "which link."

Further confirming how new and old media are not separable ecologies, link analyses in datasets of hashtags associated with the Arab uprisings show that the most common links from social media are to websites of broadcast media (Aday, 2012; the Archivist #jan25 archive). Even in most political settings, most users likely alternate between Facebook, Twitter, broadcast media, cell-phone conversations, texting, face-to-face and other methods of interaction and sharing information.

For example, to justify drawing broader conclusions from a single-platform big data study, Onnela et al. (2007) argue that analysis from a single platform, cell phone networks, is justified because, as they argue: " although mobile phone data capture just a slice of communication among people, research on media multiplexity suggests that the use of one medium for communication between two people implies communication by other means as well." However, media multiplexity theory

(Haythornthwaite, 2002) suggests something different than that: that the stronger the tie, the more the means of communication employed—not that "one medium of communication implies communication by other means". In other words, there is no assumption that using one medium implies communication by other means as well—it depends on the context, the strength of the tie, the content of the message, the availability of the communication, the suitability of the medium, among other factors.

These challenges do not mean that nothing valuable can be used from single-platform analyses. However, all such analyses must take into account that they are not examining a closed system --And the Onnela et al. (2007) study certainly has interesting results and was published in the prestigious PNAS)—and that there likely isn't justification to draw some of the broader claims.

More research is needed to understand the actual multi-platform connectivity patterns. It's possible that the solution to this "big data" limitation may not be solvable by "big data" alone. Sometimes, the way to study people is … to study people.

CONCEPTUAL CONSIDERATIONS:

1. What's in a Retweet? Understanding our Data.

There needs to be more in-depth conceptual research to deepen our understanding of what exactly social media footprints mean--and what can we legitimately infer from "big data" analyses of those footprints. What's a click? What does a retweet mean? In what context? By whom? How do different communities interpret these interactions?

In many studies, for example, retweets or mentions are used for measuring "influence." (Cha et al, 2010). A social media interaction may be a reasonable proxy for influence under certain conditions and times; however, there are many conceptual steps and implicit assumptions embedded in going from retweeting to being influenced. Most of the time, what is actually being measured is information exposure and/or reaction to information. The meaning of that retweet could range from affirmation to denunciation to sarcasm to approval to disgust.

As an example, take the recent case of the twitter account of fashion store @celebboutique. On July, 2012, the account tweeted with glee that the word "#aurora" was trending and attributed this to the popularity of a dress named #aurora in its shop. The hashtag was trending, however, because it was the site of a movie theatre massacre on that day in which 12 people were killed. (It was later revealed that the @celebboutique account was run by "social media employees" outside the United States.) There was a massive and expansive backlash against @celebboutique. I counted more than 200 mentions and many hundreds of retweets of angry messages per sixty seconds. This went on for about an hour before the company realized its mistake and stepped in. This was followed by more condemnation—the mentions were a few hundred a minute at a minimum, often rolling too fast to analyze (see here (Gilad, 2012) for more analysis). Retweets of condemnatory tweets were also very prolific. Hence, without understanding the context, the spike in @celebboutique mentions could easily be misunderstood in just the way @celebboutique's social media operatives misunderstood why #aurora was trending.

While this is an extreme example, it should be clear that any straight measure of "mentions" and "retweets" as influence is conceptually and methodologically limiting. While clicks, retweets and mentions do indicate variables such as attention and engagement, the manner of this engagement is often complex and varied.

2. Limits of Methodological Analogies: All Networks Don't Operate the Same Way.

Are social media networks similar to networks of airlines? Does information work the way germs do? Such questions are rarely explicitly addressed even though many papers import methods from other fields on the implicit assumption that the answer is a relatively unqualified yes. To step back further, even representing social media interactions as a network requires a whole host of implicit and important assumptions—and these should be considered explicitly rather than assumed without analysis (Butts, 2009).

Epidemiological or contagion-inspired analyses often treat connected edges in social media networks as if they were "neighbors" in physical proximity. In studies of epidemiology, it is reasonable to treat "physical proximity" as a key variable relevant to network structure by assume that neighboring or adjacent nodes are "susceptible" to disease transmission. There is a very good reason for this: the underlying model is a very well-developed, empirically-verified germ-theory of disease in which small microbes travel in actual space (where distance matters) to infect the next person by entering their body. This is a very physical process with well-understood properties, and underlying probabilities can often be calculated with some precision.

Creating an analogy from social media interactions to physical proximity may be a reasonable and justified under certain conditions and assumptions—but this step is rarely subjected to critical discussion of whether it is warranted. There are significant differences between information traveling in social media networks and germs. Adjacency in social media is multi-faceted; it is not always mappable to physical proximity; and human "nodes" are subject to information from a wide range of sources (not just those they are connected to in a particular social media platform). Finally, there is rarely straightforward interaction between information exposure and rate of infection as there can be for diseases.

[Note to readers: I'm compiling examples of such research to be added to this draft as a paragraph here; suggestions are welcome at zst@princeton.edu]

Network methods from other fields may indeed be useful and appropriate. There are clearly similar dynamics in many different types of networks, human and otherwise, and the multiple fields can learn much from each other. However, importation of the method needs to rely on more than "they're both networks;" it is crucial to examine the specific properties of nodes, edges, connectivity, flow, interaction and structure in different networks.

3. Field Effects: Humans Do Not Interact Only in Networks

Another difference between spatial or epidemiological networks and human social networks is that human social information flows do not occur only through node-to-node networks but also through field effects. Field effects can be thought of as large-scale societal events that impact a large group of actors contemporaneously through changes in whole social, cultural and political fields. Big events, national moods, weather occurrences, etc. all have society-wide field effects and do not depend and often do not diffuse through interpersonal interaction (although they also greatly impact interpersonal interaction by affecting the agenda, mood and disposition of individuals). Such events can be experienced collectively in a direct fashion or shared through broadcast media.

For example, most observers agree that Egyptian social media networks were important in the uprising which began in Egypt in January 2011. Indeed, social media appear to have been a key conduit of protest information (Tufekci, 2012) as well as the transfer of the oppositional narrative beyond Egypt's borders (Aday et al., 2012). However, there was almost certainly another important information diffusion dynamic. As many observers, participants and scholars agree that the Tunisian revolution was a major turning point for the possibility of an uprising in Egypt (Ghonim, 2012; Lynch, 2012).

In other words, analysis of social media structures would not have revealed a major difference between the second and third week of January of 2011 but something major had changed. To translate it into epidemiological language, the "nodes" in the network had a greatly different "susceptibility" and receptivity to the information being passed, and a much different political calculation of what was possible. The downfall of the Tunisian president, which showed that even an enduring autocracy in the Middle East was susceptible to street protests, energized the opposition and changed the political calculation in Egypt. This information was diffused through multiple methods and broadcast media, such as Al Jazeera, played a key role. Thus, the communication of the Tunisia effect to the Egyptian network was not dependent on network structure.

Such field effects are very important in human socio-political and cultural dynamics and the turn to networks as a key metaphor in social sciences, while fruitful and productive in many dimensions, should not diminish our attention to the multi-faceted nature of human socio-cultural and political interaction.

4. You Name It, Humans Will Game it: Reflexivity of Nodes and the 'Human Planck Constant'

Unlike disease vectors or gases in a chamber, humans understand, evaluate and respond to the same metrics that big data researchers are measuring. In all social media platforms where there are explicit or implicit effects of metrics, analytics and algorithms, it is safe to assume that the *"human Planck constant"* has been tripped and that the existence of the metric will alter behavior.

The "Human Planck constant" can be defined as the implicit or explicit presence of metrics or algorithms that form the basis for discernible differential rewards to which at least some actors will respond by tailoring behavior towards desired outcomes. The metric need not be present explicitly in order for the reflexivity to be triggered; it only needs to be discernible in its effects. It is hard to untangle effects of such "gaming" behaviors by looking solely at social media footprints using as the human reflexive effort is directed precisely at thwarting or gaming those known methods of observation and measurement.

For example, political activists, especially in countries where the unrest and repression have received less mainstream global media attention, such as Bahrain, often undertake deliberate attempts to make a hashtag "trend." Indeed, "to trend" is increasingly used as a causative verb among these users, as in: “let’s trend #Hungry4BH”. These efforts are generally not mere blind stabs at massive tweeting; they often display a fine-tuned understanding of Twitter’s (undocumented) “term frequency–inverse document frequency” algorithm for selecting trending topics. (While Twitter’s trending algorithm is not public, there have been cases of attempts reverse-engineering it (See (Lotan, 2011) – thanks to a reviewer for suggestion I add this information--) hence people attempting to game it are not merely groping in the dark though they are probably also operating through trial-and-error]

I’ve observed multiple hashtags trend worldwide as a result of such carefully-concerted campaigns. (Those from Bahrain include #100strikedays, #KillingKhawaja, #StopTearGasBahrain, #Bloodyf1, and #F1DontRaceBahrain, among others).

Another example of users consciously altering their social media practices in order not to be visible in a particular (measurable way) is the practice of “subtweeting” –talking about a topic or a person with a deliberate misspelling or the omission of the @ sign before the person’s name so that the tweet does not show up in that person’s timeline—or, consequently on social media analytics. (Subtweeting can be done on a continuum which range from dropping the @ sign but using the correct name to deliberate misspelling to implied conversations each with different implications for big data analytics.)

For example, on Egyptian social media and blogs, there was a large amount of discussion of an article about gender oppression in the Middle East written by Egyptian-American writer and speaker Mona El Tahawy. Sociologist Alex Hanna and Marc Smith extracted the tweets which mentioned her or linked to the original article. The consequent network analysis showed great polarization in the discussion, with two distinctly clustered groups. This was valuable and timely analysis and it contributed to our understanding of the discussion. However, while watching this discussion explode online, I noticed that many of the high-profile bloggers and young activists who were discussing the article –and greatly influencing the conversation-- were “sub-tweeting”. Later discussions with this community quickly revealed that this was a deliberate and informed choice that the community had made exactly because they did not want to appear to engage her directly, or appear to be give her attention. Sub-tweeters are probably a much smaller and more discriminating community than tweeters as a whole and it may appear that ignoring this phenomenon won’t significantly impact results. This may be true to a degree--but overall, this example highlights the broader problem that big-data analysis is not always good at differentiating people who are important, engaged or influential on a given topic.

Such behaviors –aimed at avoiding detection, amplifying a signal or attaining other goals by deliberate gaming of algorithms and metrics-- should be expected in all analysis of human social media imprints. The “gaming” motivations may range from getting attention for a political cause to simply avoiding a parent on social media.

5. Additivity as a Property: N+1 is Not Always N+1:

Most big data analyses assume that social media imprints are cumulative and that they can be added as if any one imprint were identical to any other imprint. However, this assumption is rarely justified either methodologically or conceptually (for example, thorough semantic analyses). As discussed above in the case of @celebboutique, social media imprints can be positive and a lot of activity around a topic may not necessarily by interpretable simply by counting the frequency of any metric. One of the most cited papers using Twitter Big Data sets titled the “Million Follower Fallacy” (Cha, Haddadi, Benevenuto, & Gummadi, 2010) measures influence by counting mentions and retweets—in such an analysis, would @celebboutique appear to be quite influential? Addition, while appearing simple, harbors a set of epistemological and ontological assumptions which should be made explicit and justified. *[Draft section considering expanding or dropping.]*

6. Network Structure Does Not Reveal All

Many *big data* analyses of social media footprints often attempt to use structural network properties to infer other properties of the links between alters. That path, however, needs to proceed with careful awareness of the limits of information contained solely in network structure.

For example, one of the most common efforts has been to connect strength of the tie between alters (an individual's network consists of "ego," the individual, and ego's "alters," the others to whom ego is connected) to properties of the structure of the network. This is often based on the hypothesis developed by Mark Granovetter in his article "The Strength of Weak Ties" (1973), where he proposed that, under certain strict conditions, bridge ties between network clusters would be more likely to be weak ties. However, many subsequent efforts to use this hypothesis conflate all weak ties with bridge ties (which is not warranted), and also ignore the strict assumptions (such as that of triadic closure between strong-tie alters) that are necessary for Granovetter's hypothesis to hold. Many dense networks are composed of weak ties (like a workplace where most people know each other) and under many conditions, strong-ties can be bridges between otherwise unconnected network clusters.

Another common method of moving from social media imprints to other attributes has been to attempt calculate tie-strength from frequency of interaction. For example, in Onnela et al. (2007), tie-strength is computed as a function of frequency of contact among mobile phone users. However, research shows that the relationship between tie-strength and network structure (and online imprints) is quite complex and multi-dimensional (Gilbert and Karahalios, 2009). Further, frequency of interaction is not a good proxy for tie-strength (Marsden and Campbell, 1984) as it is a highly-confounded variable. This is because, of all the potential dimensions of tie-strength –described by Granovetter (1973) as: "the amount of time, the emotional intensity, the intimacy (or mutual confiding), and the reciprocal services"—situational factors interact most with frequency of contact (Marsden and Campbell, 10984, Petroczi, Bazsó, and Nepusz 2006). Indeed, many people have very frequent contact with their weak ties (workplace colleagues, coordinating events with acquaintances, arranging certain practical tasks, etc. can all involve frequent cell contact).

Research shows that the best, unconfounded indicator of tie-strength is subjective: a feeling of intimacy and closeness (Marsden and Campbell, 1984). And that variable cannot be inferred merely from network structure or frequency of interaction; rather, uncovering it requires other methods such as more extensive content analysis of the communication between the alters or interpersonal research techniques such as surveys or interviews.

CONCLUSION:

Big data is a powerful addition to social science toolkit. However, this emergent field needs to be placed on firmer methodological and conceptual footing. *Big data* cannot answer all questions: concepts like tie-strength, meaning of social media imprints, context of human communications, and nature of interactions are multi-faceted and complex. Our methods of sampling and analyzing social media imprints have largely been driven by availability of data but this recent flood of data has not been matched by conceptual and methodological depth for understanding both the promise and the limitations of *big data*.

Besides more discussion of the methodological and conceptual areas enumerated above, there needs to be broad steps taken to clarify and consolidate the *big data* emergent methods. Primarily:

1- Analysis of *big data* should be geared towards substantive questions and away from being driven by data availability;

2- Big data analysts need to become more explicit in discussing the implicit assumptions and steps in their analytic methods;

3- There needs to be a better understanding of the limits of *big data* analyses. *All* methods have limits and trade-offs and better awareness can lead to better studies and understanding.

3- The grounding and deepening of *big data* methods and conceptual toolkit needs to be an interdisciplinary effort and carried out with the collaboration of scientists with expertise in substantive areas under examination.

Overall, this is an exciting time to be studying many fundamental social questions. Progress will be more sound and comprehensive if the excitement at the opportunity is strengthened by methodological and conceptual rigor.

REFERENCES

Aday, S., Farrell, H., Lynch, M., Sides, J., & Freelon, D. (n.d.). New Media and Conflict after the Arab Spring. Blogs and Bullets II. United States Institute for Peace. Retrieved from http://www.usip.org/files/resources/PW80.pdf

Allison, P. D. (2001) Missing Data Thousand Oaks, CA: Sage Publications

Bernstein, M., Bakshy, E., Burke, M., and Karrer, B. (2013). Quantifying the Invisible Audience in Social Networks. ACM CHI 2013: *Conference on Human Factors in Computing Systems*

Bolker, J. A. (1995). Model systems in developmental biology. BioEssays, 17(5), 451–455. doi:10.1002/bies.950170513

boyd, danah, & Crawford, K. (2012). Critical Questions for Big Data. Information, Communication & Society, 15(5), 662–679. doi:10.1080/1369118X.2012.678878

Butts, C. T. (2009). Revisiting the Foundations of Network Analysis. Science, 325(5939), 414–416. doi:10.1126/science.1171022

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. (2010). Measuring user influence in twitter: The million follower fallacy. Measuring user influence in twitter: The million follower fallacy (pp. 10–17). Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1538/1826

Dunning, T., & Freedman, D.A. (2008) Modeling section iffects. in Outhwaite, W. & Turner, S. (eds) Handbook of Social Science Methodology. London: Sage. Return

Fields, S., & Johnston, M. (2005). Whither Model Organism Research? Science, 307(5717), 1885–1886. doi:10.1126/science.1108872

Hargittai, Eszter. (2007). Whose space? Differences among users and non-users of social network sites. *Journal of Computer-Mediated Communication*. 13 (1), 276-297

Haythornthwaite, Caroline. 2002. “Strong, Weak, and Latent Ties and the Impact of New Media.” The Information Society 18:385-401.

Jenner, R. A., & Wills, M. A. (2007). The choice of model organisms in evo–devo. Nature Reviews Genetics, 8(4), 311–314. doi:10.1038/nrg2062

Geddes, B. (1990). How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics. Political Analysis, 2(1), 131–150. doi:10.1093/pan/2.1.131

Gilbert, S. F. (2001). Ecological Developmental Biology: Developmental Biology Meets the Real World. Developmental Biology, 233(1), 1–12. doi:10.1006/dbio.2001.0210

Gilbert, E., and K. Karahalios. (2009). “Predicting tie strength with social media.” Pp. 211–220 in Proceedings of the 27th international conference on Human factors in computing systems.

Granovetter, Mark. 1973. “The Strength of Weak Ties.” The American Journal of Sociology 78:1360-1380.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., et al. (2009). Computational Social Science. Science, 323(5915), 721–723. doi:10.1126/science.1167742

Lotan, Gilad, 2011. Data Reveals That “Occupying” Twitter Trending Topics is Harder Than it Looks! http://blog.socialflow.com/post/7120244374/data-reveals-that-occupying-twitter-trending-topics-is-harder-than-it-looks

Marsden, P. V., & Campbell, K. E. (1984). Measuring Tie Strength. Social Forces, 63(2), 482–501.

Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., et al. (2007). Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences, 104(18), 7332–7336. doi:10.1073/pnas.0610245104

Petroczi, A., F. Bazsó, and T. Nepusz. 2006. “Measuring tie-strength in virtual social networks.” Connections 27:39–52.

Pew Research Center. (2013). *Twitter Reaction to Events Often at Odds with Overall Public Opinion*. Retrieved from http://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/

PLEASE NOTE THAT THIS IS A DRAFT. FEEDBACK, CRITICISM AND SUGGESTIONS ARE WELCOME

PLEASE EMAIL ME AT ZST@PRINCETON.EDU OR USE TWITTER (@techsoc).

THANK YOU FOR YOUR INTEREST!