'Critically Engaging with Social Media Research Methods'

Dhiraj Murthy, Goldsmiths College

Final version to appear in Trends and Challenges in Social Science Research (Sage)

Abstract: As social media technologies such as Twitter, Instagram, and YouTube have become highly ubiquitous, social life itself has become reconfigured. Though early notions of an offline/online binary remain in some quarters of social research, there is a realization amongst most that this binary is reified. As such, the study of social interactions within social media is a fundamental sociological question. This chapter argues that social researchers need to engage with the study of social media in order to comprehensively understand modern social life. This chapter also provides insights into how we, as social researchers, can critically collect and discern social formations via social media. Twitter is specifically used in this chapter to provide an example of how the medium provides opportunities for mixed qualitative and quantitative social analysis. Ultimately, this chapter also argues that the understanding of large social questions is increasingly contingent on us deciphering and understanding social knowledge formed and maintained within social media platforms.

Introduction

C Wright Mills (1954) in his barbed essay 'IBM Plus Reality Plus Humanism=Sociology' attacks the sociological establishment, arguing that the discipline has been divided into 3 camps: 'The Scientists' ("who would love to wear white coats with an IBM label"), 'The Grand Theorists' (who, through "turgid prose" proffer unintelligible, overly complicated theories of society, and the 'Third Camp' (who are critical and who seek to understand important, socially relevant macro-social

questions in the micro and the macro). Mills argues that the first two camps feed upon their own intellectual narcissism and egocentricity, building manufactured boundaries between their camp and the other. In many ways, Mills was forecasting some of the artificially binaristic divides between quantitative and qualitative sociology and encouraging us to join the Third Camp.

In this chapter, I would like to do the same in the context of social media and the Big Data revolution it is part of. Our social lives are increasingly mediated by social media technologies. Social scientists need to be technically, theoretically, and empirically knowledgeable of these ubiquitous media. And, most of all, they need to have the ability to be critical. Mills' IBM should be substituted for Facebook, Twitter, or Google and you have the 'The Scientists' of digital, computational sociology. Like The Scientists of Mills' time, contemporary computational social media research is sadly often divorced from theory and generally completely removed from the larger social questions that Mills is imploring us to not lose sight of.

Major sociological conferences, including the American Sociological Association (ASA) annual meeting and the International Sociological Association (ISA) meeting have sessions focused on hyper-empirical readings of social media just doors down from macro-theoretical discussions of social media. This camp mentality puts sociology behind, rather than ahead of the curve. It also encourages some digital, computational sociologists to put on white lab coats (with the Facebook, Microsoft or Google logo on it).[1] Indeed, immediately prior to the 2014 ASA annual meeting, Facebook held an invite only workshop for sociologists whom they felt understood demographic social media analytics. Facebook bused sociologists from San Francisco down to their headquarters in Silicon Valley to introduce them to the company's

internal data, methods, and to build research relationships with them. We know Facebook is working with academics across a range of disciplines (as the infamous 'Facebook experiment' (Kramer, et al. 2014) highlighted). So why not sociologists?

The point of this chapter is not to argue against Facebook or those who construct grand theories of digital society. Rather, it is to take up Mills' charge of a more unified, critical discipline of sociology, which seeks to understand the bigger picture at the level of macro-social forces and formations. However, unlike in Mills' time, many of these macro-social questions now need specialist computing expertise. And like Mills' time, theoretical insights remain needed. Because social media and emergent digital technology continues to be increasingly important to our social worlds, the lab coats and theoreticians need to work together and support a critical, digitally cognizant, sociology and not fall prey to a camp mentality.

The purpose of this chapter is to make this case and provide accessible examples of mixed methods to encourage sociologists to collect and interpret social media data and not leave that work solely to the white lab coats at technology companies (or I should say T-shirt and jeans wearing). Big data methods are moving quickly and unless the social sciences as a whole make a significant step change in terms of trying to understand social media and other big data, our complex online interactions and footprints will be left dependent on corporate interpretations (or computer science methods and interpretations at best).

Social media and the empirical crisis of sociology

In the discipline of sociology, Savage and Burrows (2009) compellingly argue that the discipline is at a crisis point in terms of not paying enough attention to studying big,

social transactional data that is stored in everything from social networking sites to large corporate databases. Five years after Savage's and Burrow's polemic, we remain in this crisis. The rising popularity of social media produces tremendous volumes of social data. However, the field of sociology itself has had difficulty in accepting interdisciplinary methods (such as natural language processing) and the actual utility of such data in understanding social behavior. Because so much of our social life exists in or is dependent on social media, it is critical that we use a wide variety of interdisciplinary methods to try and discern the extremely complex and nuanced social processes that are being performed, enacted, and articulated in these technosocial spaces. Sloan et al. (2013) argue that using the Twitter spritzer stream (1% of all tweets) is an important data source for social science. They argue that understanding the demographics of Twitter users via analysis of these data[2] will provide very large sample sizes for social scientific analytic methods. As it stands now, there is much to be understood about demographics as well as general sociolinguistic behavior on Twitter to get to this point.

The popularity of social media sites has grown enormously and much of social research has been fundamentally altered by our daily engagement in technologically mediated communication. For example, quantitative sociology has been traditionally driven by manageable, structured data sets. Digital sociology—the sociology of online networks, communities, and social media— is now quickly emerging as a major field due to the rise of social networking sites like Facebook and Twitter (Orton-Johnson 2013). Data from these and other social media sites has been regularly used to study social behavior online (Gold 2012; Marres 2012). As a result of the increased availability and user-friendliness of analytic techniques and exponential jumps in

processing power and storage capabilities, a variety of disciplines including, but not limited to, the digital humanities, social sciences, and information systems are becoming increasingly interested in capturing, storing, and analyzing large data sets that were previously inaccessible to most. Inexpensive and accessible social media cloud services such as HootSuite archives require very little technical expertise. However, sociology as a discipline has lagged behind many other social science and humanities disciplines. Media studies and communications, for example, have surmounted initial learning curves and have embraced social media data as evidenced by recent social media and big data panels at the International Communications Association (ICA) annual meeting.

Social Media, Big Data, and Research Methods

Over the last decade, there has been an exponential increase in the amount of quantitative social media trace data –our 'digital footprint' - available to researchers across the globe. Facebook boasts over a billion users (Vishwanath 2015) while Twitter, the increasingly pervasive micro-blogging service, has grown to over 600 million users generating over 500 million tweets a day (Statistic Brain cited in Ottoni, et al. 2014). Other technology companies are part of a rush to bring a wide variety of broad-based and niche social media services, products, and ecosystems into the global online marketplace. For example, Instagram, a social media site for sharing photos that debuted in 2010 has over 150 million users (Bakhshi, et al. 2014). As a result of this rapid growth, there has been an increasing demand for research methods that allow the collection, storage, and analysis of these vast troves of social trace data. Big data typically refers to data sets so large that they challenge the abilities of more traditional software tools and systems typically used in data collection, storage, and

analysis (Manovich 2011). Though, as Ruppert argues in this edited collection, Big Data is not just about volume, but new forms of organizing data as well. As the desire and need to study these diverse, complex, and often large data sets has grown, individual social researchers have often found it difficult to be sufficiently resourced. They have either clustered into labs or closely collaborated with technology companies, especially Microsoft and Facebook (hence the mention of Mills' lab coats). Because of the changing research relationship this has brought, the ways in which social data is often interpreted are generally subject to external, non-university influencers. This marks a shift from early social media work in which social media data was quasi-open data with unlimited requests to the Twitter API allowed. Early work such as Kwak et al. (2010) mapped the entire Twitterverse, collecting 41.7 million user profiles and 1.47 billion social relations and 106 million tweets. Though these numbers do not carry the same shock value they did in 2010, the important thing is that this work was able to collect complete social media data due to a lack of stringent Terms of Service (TOS), which all major social media platforms have today. In other words, Kwak et al. were studying the entire Twitterverse rather than partial, biased samples of the Twitter population.

Though many Big Data methods, such as that used by Kwak et al. (2010) will not be accessible to individuals sociologists, it is important for us to engage with and experiment with Big Data so that we can maintain critical, reflexive perspectives of social forces and their impact on policy and society (Ruppert, this volume). For individual researchers, there may be a variety of pros and cons associated with their decision to work with or not to work with social media data. But, regardless, researchers should be able to crack open the black box and use accessible tools and software to further the discipline in innovative, but critical ways. The utility of basic

social media analytics will only become increasingly important to the *craft* of social research.

Though news reports (and academic conferences) abound with social media's promise, the speed at which it is changing represents a significant challenge. The ubiquity of social interactions in social media has the potential for developing a richer understanding of online social formations. However, it can be difficult, expensive, and time consuming to store and process 'messy', unstructured social media data. Acquisition of social media data presents particular challenges. Complete social media data collection is technically complex and often cost prohibitive (e.g. The Twitter firehose, which is the full Twitter stream, has an undisclosed access cost which is thought to be approximately $100,000 per month excluding bandwidth costs to collect the tweets).

A large segment of social media literature uses machine learning and other highly computational methods, which discern sentiment, topics, frequency and network composition (Aggarwal 2011; Özyer, et al. 2013; Shu-Heng 2014). Because of the volume of social media data, it is difficult for most qualitative researchers to know where to start without some background in computational social media methods. I have successfully used these techniques in several studies to help form research questions and to assist with hypothesis generation. What is often surprising to colleagues is that these data intensive, computational approaches often help facilitate a grounded theory (Glaser and Strauss 2009) approach by providing some initial context (such as what topics people are discussing on Twitter and Facebook). Specifically, in my work on the use of social media by cancer patients (Murthy, et al. 2011), I used machine learning to provide a bird's eye view of tens of thousands of

cancer-related tweets. From the topics derived from this analysis (see Table 2), I discovered at what stages of cancer people were tweeting (usually the most frequently at diagnosis or in the first stages of chemotherapy). This enabled me to better articulate my research questions and the remit of my study.

The topic clustering approach I used was an artificial intelligence machine learning approach called Latent Dirichlet Allocation (LDA) (Blei, et al. 2003) to discern topic clusters within Twitter data as well as within online scientific social networks (Gross and Murthy 2014). Though this level of analysis requires some levels of technical expertise, it is free and is being used across a variety of disciplines. PHP scripts can be used with minimum modification and programmers can be hired for more advanced corpora. How-to guides on using LDA with social media data are available and, in the case of MALLET, a popular LDA implementation for text, videos and PowerPoint tutorial slides are available.[3]

Visual Methods

An integral aspect of many social media research methods is visualization. Not only do good visualizations clearly present complex findings derived from social media data, but they demystify the black box and serve to open dialogue, which can produce critical interrogations of these data. Although Figure 1 looks straight out of Mills' depiction of The Scientists with log scales, this visual representation makes a significant macro-social point. What it measures is the interval between tweets (on average) for users of different American cities. Data points under the fit line indicates a faster than average tweet rate. What these data reveal is that, surprisingly, cities such as San Francisco, Los Angeles and New York have large populations and below

average tweet frequency rates for their population size. What is particularly sociologically interesting is that almost all the cities below the fit line in the quadrant in the bottom right are predominantly black cities. For example, Atlanta, Washington and Detroit have majority black populations and Cleveland is close to majority black (Massey 2001). On the other hand, Boulder is nearly 87% white and Rochester is nearly 82% white (Sperling and Sander 2004). Generally, cities above the fit line (with slower rates of tweeting) had higher average household incomes than cities below the fit line according to 2012 US Census data (Noss 2013). Rather than reducing this analysis to a purely computational exercise like Mills' critique of The Scientists, such a critical approach provides a clear remit for how mixed methods work could provide more detail of why poorer, black cities tend to tweet at a significantly faster rate than their richer, more white counterparts. Ethnographic accounts could and should investigate race, age, gender, employment and education further. Sociological theory also benefits from such analyses as it has additional contextual data to consider in theoretical accounts.

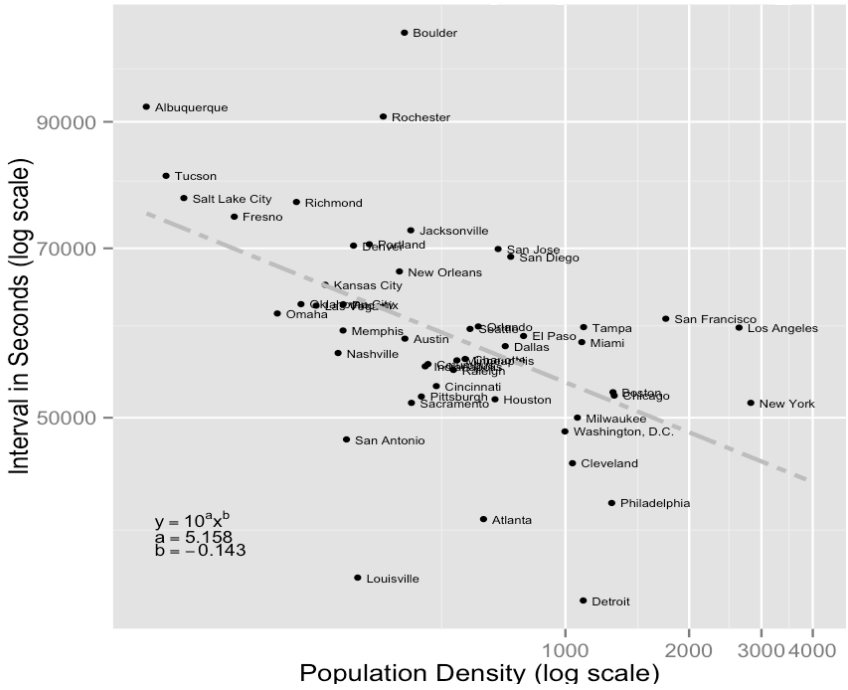

Figure 1: Mean Tweet Interval by Population Density [4]

I have also used very basic matrix-style visualizations (see Figure 2) to help understand complex online communities which use social media (Murthy, et al. 2013). Specifically, I studied scientific virtual organizations in the life sciences, which used various types of proprietary social media. Scientists on these platforms discussed a wide variety of professional and social things on the social media, but it was difficult to discern how these conversations intersected. Because the project also involves ethnographic observation and interviewing, I used these basic visualizations as a research method to glean macro-social insights before going into the virtual ethnographic field. Importantly, visualization can break down barriers between quantitative and qualitative methods/camps and facilitate the asking and answering of large social questions. Also, simple rather than overly complex visualizations help break down quantitative and qualitative boundaries.

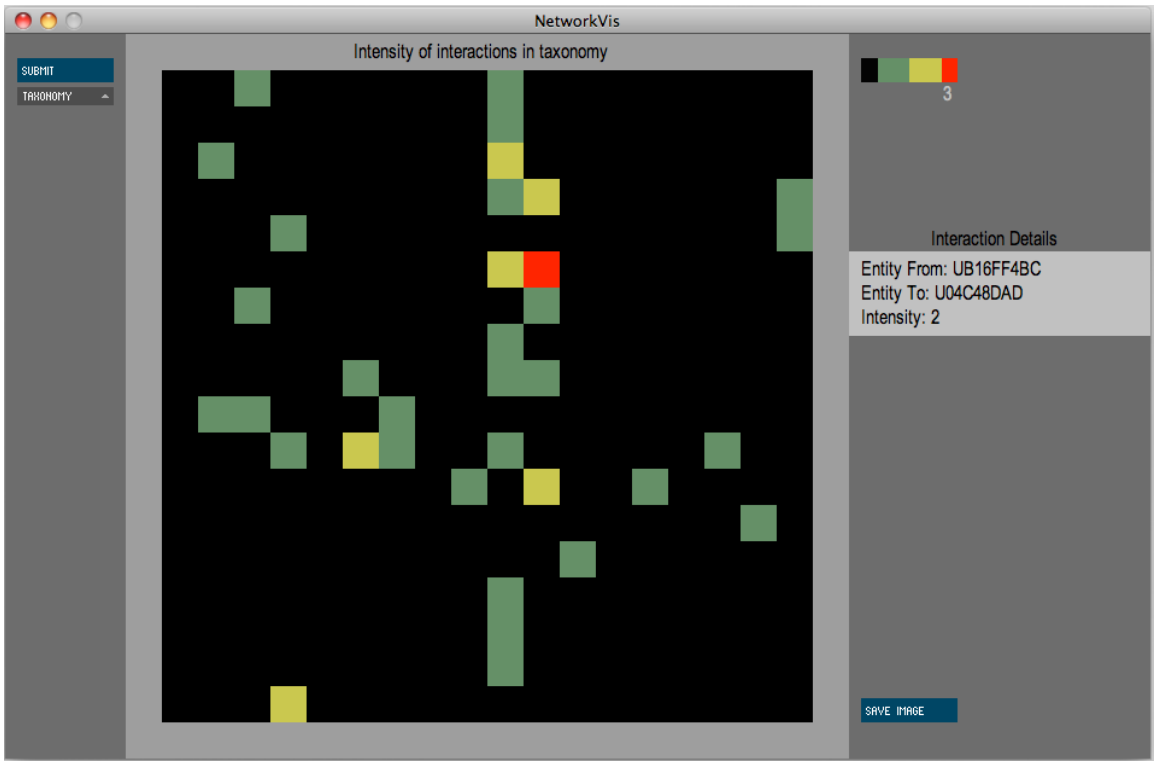

Figure 2: Matrix-style visualization of social media interactions [5]

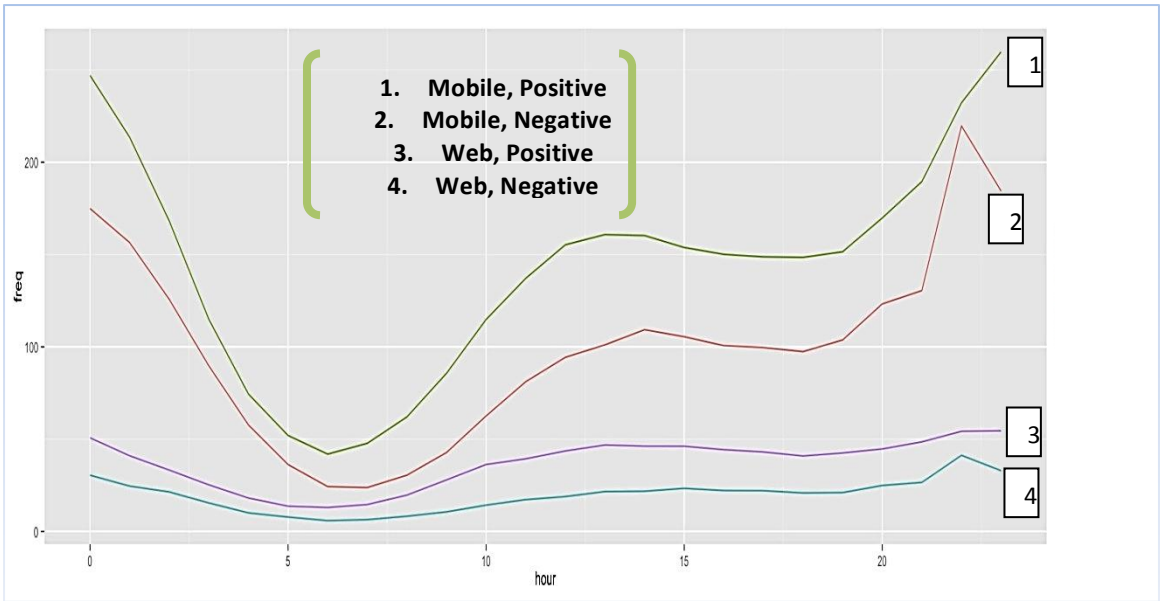

Figure 3: Frequency of tweets by time and by medium (mobile vs. web) [6]

Figure 3 visualizes the frequency of positive versus negative tweets for mobile versus web users by time of day. Drawn from 235 million tweets, this simple visualization reveals several things of value to mixed methods approaches. First, the distillation of hundreds of millions of tweets reveals diurnal social patterns on Twitter. Specifically, we tend to begin our posting on Twitter in the morning (with a steady increase from waking hours to just after lunch). Tweet rates stay steady until after normal working hours and then sharply increase. Tweets from mobile devices are much more frequent than web-based interactions, though both mediums have a higher frequency of positive content. Interestingly, though negative mobile tweets rise more sharply than positive ones in the evening, they also drop off at a faster rate. This simple visual approach to synthesizing vast amounts of social media data using Big Data social media methods is very useful to mixed methods research. In this case, it is able to provide very macro trends of the sentiment of tweets by time of day and how that is potentially linked to mobile versus web use. However, major challenge to this type of approach is how much we can really discern from matching words in tweets to sentiment dictionaries. Indeed, given the millions of tweets being studied, we need to

be aware that what are missing from such methods are rich, micro-level understandings. Integrating other methods is an important way of addressing these limitations. For example, mixed methods could help us understand why negative content spikes so sharply in the evening. Is this due to venting about work, issues around family life, or commute stresses? Big data methods are ill-equipped to answer such questions in much detail. Focused survey or ethnographic work would not only reveal granular detail about these questions, but, importantly, would provide rich case studies of specific groups/locales. Importantly, time- and device-based research methods provide more nuanced detail into how we socially communicate within social media platforms.

N-gram methods

The use of n-grams (with n signifying how many adjacent words should be extracted) is common in computer science and quantitative content analysis in the social sciences. Because the volume of contemporary social media data is so large, methods such as automated machine learning, give one the ability to sort through vast amounts of data and discern general trends/patterns. These results can be used to spur focused qualitative research in particular domains from relationships to mobile sociability. N-grams provide very high-level frequency counts and can provide important macro-level views of very large social media corpora (especially when one compares unigrams, bigrams, and trigrams). Hundreds of thousands of tweets are beyond the scale for possible digital ethnographic methods. Therefore, n-grams provide one type of quantitative observation that can guide the development of qualitative research questions and provide a way to gain some perspective of general trends in social media textual data.

A common method of simple n-gram analysis is rank ordering. Table 1 provides an example of the rank-ordering of commonly used words and phrases[7] in 235 million tweets. Though not without its drawbacks and limitations, this n-gram analysis provides some useful modes to understand themes, trends, and topics within tweets. In my example in Table 1, 'I' is used more than 'you' and 'good' is used more than 'bad'. This type of n-gram analysis can help develop research questions ranging from cultural taste to temporality to egocentricity. Again, raw frequency data such as this is best used to give a bird's eye view of a large social media text corpus, which otherwise would be impossible to discern general trends from.

'I' is ranked 1	'you' is ranked 5	'he' is ranked 69	'she' is ranked 88
'I love' is ranked 5	'love you' is ranked 11	'I love you' is ranked 4	'I hate when' is ranked 84
'Love' is ranked 34	'good' is ranked 64	'bad' is ranked 183	
'buy' is ranked 431	bought is ranked 1093		
'birthday' is ranked 204	'today' is ranked 69		
'my iphone' is ranked 5075	'the blackberry' is ranked 3855	'the android' is ranked 3855	
'I want to' is ranked 5	'I need to' is ranked 12	'I feel like' is ranked 27	
'want to' is ranked 17	'going to' is ranked 19	'I was' is ranked 21	
'posted a new' is ranked 36	'photo to facebook' is ranked 40		'follow me' is ranked 14

Table 1: Selected Unigram, Bigram and Trigrams by use ranking

Indeed, the ability to get a general sense of topics/themes emerging within large corpora is an often overlooked utility of quantitative data analysis methods and a real resource for mixed methods research. In my work, I have used various forms of Big Data methods to cluster topics in a range of social media. In the context of Twitter, I have done this with cancer-related tweets and have discovered topics, which have

helped me better understand the context of how individuals are using Twitter from diagnosis to recovery or the death of a family member.

Topic 005	Topic 006	Topic 007
my	Good	is
mom	Found	year
got	Start	there
through	Side	god
hope	Effects	scan
hospital	Work	hear
really	Said	heart
dad	While	praying
prayers	Feel	clear
during	morning	continue
keep	Bad	glad
friends	Body	low
happy	Started	beauty
oh	Sick	january
strong	Feeling	bless
put	Pain	checked
grandma	Luck	scheduled
sunday	oncology	recovering
#beatcancer	Yeah	yours
thoughts	Job	breasts
both	Hours	mri
congrats	Once	donenext
prayer	Diet	praise
helped	Later	ultrasound
cousin	Pretty	lord
dose	Nurse	ct
journey	Room	makeup
mean	Gone	
	thinking	

Table 2: Cancer topic clusters derived from machine learning

For example, Table 2 illustrates three topic clusters derived from a large data set of 90,986 cancer-related tweets (based around selected keywords including melanoma, lymphoma, cancer, mammogram, and chemo). Using a machine learning statistical technique called Latent Dirichlet Allocation (LDA), tweets were sorted into 50 topic clusters. Topic 5 particularly emphasizes family, prayers, and the overall journey of cancer patients. Topic 6 more clearly emphasizes the earlier stages of cancer patients

and specific emotions, side effects, and detail. Topic 7 involves breast cancer patients at all stages and involves diagnosis, treatment, family, and religion. All three topics provide insights into what cancer patients, family members and friends are posting on Twitter regarding cancer and what specific words they were using in their tweets and what similarities there are in tweet content for patients with similar cancers. Additionally, the topics provide important data for mixed methods research, which could involve focus groups, interviews, or participant observation with cancer patients, family, friends, service providers, and health professionals. Most importantly, machine learning methods, such as LDA provide social researchers with preliminary data before ethnographic or survey work, potentially saving time, effort, and money in the research process. Without knowing what is circulating on social media, we often have a partial portrait of complex social forces. And following Mills, our ability to have a critical 'sociological imagination' (Mills 2000b) can be affected.

Collecting social media data 'for dummies'[8]

I often hear colleagues mentioning that they would like to integrate social media research methods into their research toolkit, but feel that the learning curve is far too steep. They are scared off by technical challenges, Big Data jargon, or the fear of change. That being said, many in social research do understand how online social interactions are increasingly important to much of our social lives. When I conduct training on social media research or teach on the subject to students, I usually get a lot of confessions of math phobia or techno-illiteracy. Though very large-scale work using social media data (e.g. millions of tweets or posts) is very technically challenging to study, small samples (i.e. small n's) are usually not. Indeed, easy to use and free/cheap software tools[9] can collect, and visualize social media data off-the-

shelf. These tools include NodeXL, HootSuite Archives, DiscoverText, netvizz, and Gephi.

When I first started social media research with blogs and early social networking sites in 2004, all of my data collection had to be custom designed and analysis was often messy as data had to be laboriously 'cleaned' for statistical analysis, content analysis, and other mixed methods. Much has changed even in the last 5 years since I first wrote about 'digital ethnography' in *Sociology* (Murthy 2008). In that article, I discussed the fears of sociologists towards social networking, social media, and even digital fieldwork diaries. I now often see iPhones and iPads with microphones used to record interviews and to take field notes. These are exactly the devices I encouraged digital ethnographers to use (Murthy 2011), but was honestly not sure about their future uptake in the discipline. However, the use of tools to collect and interpret social media data has not taken off in the same way. Much of this is due to the same perceived fears that sociologists had in the previous decade about digital field notes and virtual ethnography. For example, social research methods textbooks now generally cover digital ethnography well, but social media research methods still remain minimally covered or not covered at all.

Social media data collection and basic analytics has become a part of many off-the-shelf packages. For basic social media research, one can easily and cheaply use cloud-based managed services. For example, Hoot Suite, the social media dashboard, offers a 'Pro' package that has an inexpensive monthly fee (and free trial) which includes the ability to visualize social demographics with geolocation, measure sentiment metrics, and archive up to 100,000 posts. DiscoverText is more tailored towards academic research and offers cloud-based access to live social media feeds, including

Facebook and Twitter. The service is more expensive (at $99/month at the time of writing), but includes basic analytics and reporting tools. However, both of these services are heavily limited by the fact that they do not allow one to easily export data to use in NVivo, SPSS, Excel or Word. This is not inherently problematic for very small scale social research, but poses challenges as projects scale up.

One solution to this is to use the open source software package NodeXL, which is a completely free plug-in to Microsoft Excel. Because it uses the Microsoft Excel interface, the learning curve for it is not steep. Additionally, the authors of the NodeXL software package have published an easy-to-follow textbook (Hansen, et al. 2010) which clearly describes how to import data from Facebook, Twitter, YouTube, and other services and how to extract basic analytics. For example, one can easily discern the top hashtags, tweets, and users within a Twitter keyword search. A major advantage of NodeXL is its integration of a Social Network Analysis (SNA)-based visualization package. As Miller and Dinan argue in their chapter in this book, SNA has been an important Big Data research method.

Social network analysis allows one to visualize the ways in which individuals are interacting and specifically connected within a network. In the case of Twitter data, it also enables one to see the ways in which individual tweets spawn conversations and who is involved. Because NodeXL uses SNA, more advanced users can tailor visualizations using the standard SNA measures such as degree and centrality, NodeXL can be easily used to visualize networks of YouTube commenters and Facebook friend networks, for example. The major limitation of NodeXL is that it is only compatible with Excel running within Microsoft Windows. However, Mac users have successfully installed Windows as a virtual or dual operating system and

installed NodeXL without incident.[10]

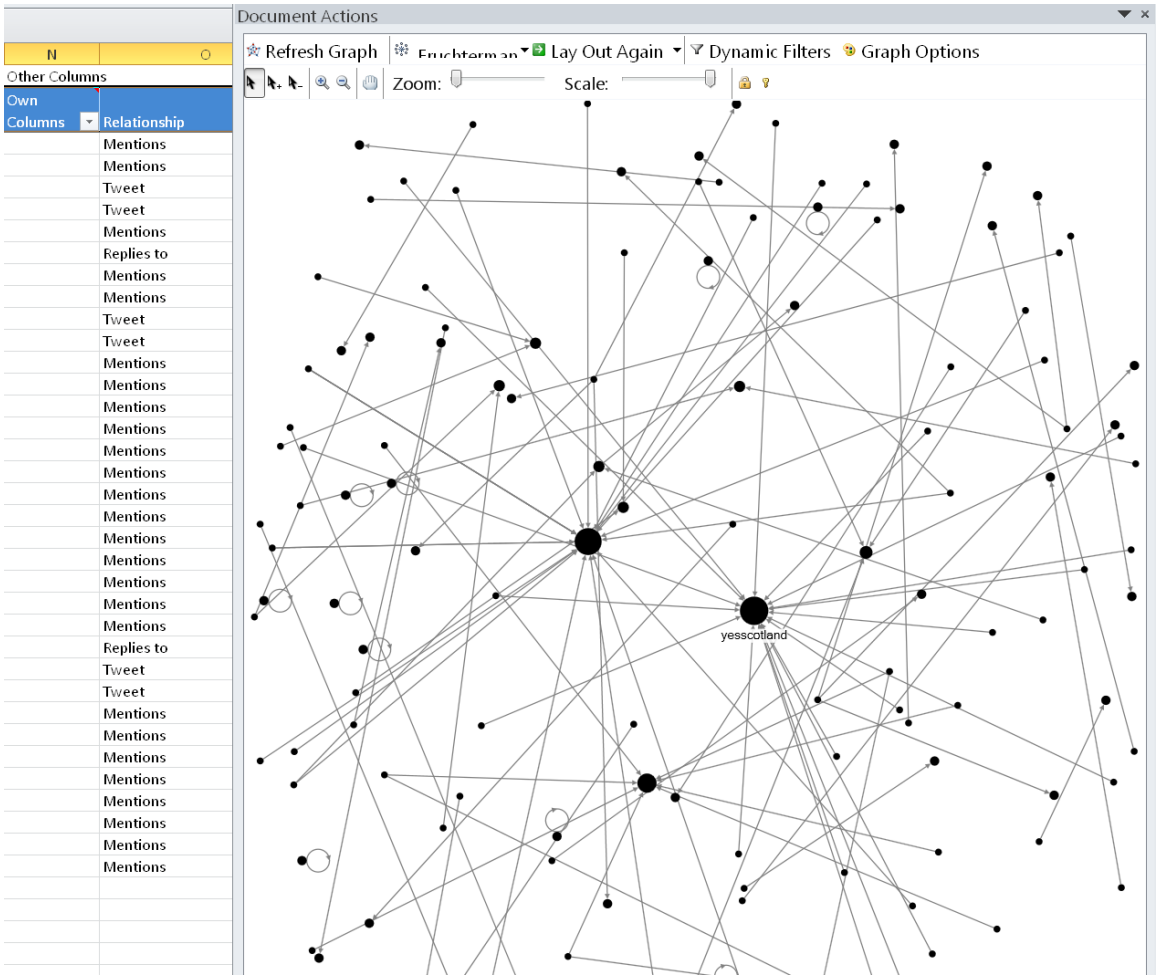

Figure 4: NodeXL network visualization of pro-independence tweets during the 2014 Scottish Independence Referendum

Figure 4 visualizes 100 tweets collected from the #voteyes hashtag, which was used to promote voting yes in the 2014 Scottish referendum for independence. The group yesscotland can be seen to be quite central in receiving mentions on Twitter. Interestingly, when I expanded the data collection to include thousands of tweets, yesscotland remained the most mentioned. Mixed methods such as digital ethnography could be used to conduct online observations of yesscotland as well as other central Twitter users and groups to better understand the Yes campaign's use of Twitter. As every major political campaign includes uses of social media, it is important for sociologists to be able to access these tools. What should surprise most

is that the collection of these tweets and the visualization of them were done in under 10 minutes, which is not only testament to NodeXL's basic ease of use, but also of a certain level of democratization of these technologies. Of course, more nuanced visualizations analytics take much longer. However, much can be done with a basic understanding of the package.

Collecting Facebook data for dummies

Though Twitter is used as the case study in this chapter, Facebook remains a popular medium for social communication and is fundamentally important to social research. NodeXL can collect egocentric networks from Facebook and this is useful for understanding an individual's social graph. However, unlike Twitter, Facebook has a much more closed API and data collection system. Generally, the data one is able to collect from Facebook is public. For example, data can be collected regarding comments on a public Facebook page. One can also collect data relatively easily about oneself or a user who is logged onto their Facebook profile and gives access to collect data. Apps can also be designed to gather data about a user and their friend network. However, when doing this, there are complex ethical considerations that need to be taken into account including the fact that a user is giving access to their friend network and sufficient time and effort needs to be put into putting research through Institutional Review Boards (IRBs) or other ethical research review processes (see Zimmer (2010) for a discussion about Facebook research and ethics). That being said, apps have been successfully used in research projects (similar to plug-ins within browsers). As a first port of call, many social researchers start collecting Facebook data through an already existing Facebook app called netvizz.[11] This app was designed to allow researchers to either collect small-relatively large amounts of data from Facebook. This includes public scraping of comments of public Facebook pages

and can be used to discern which users are particularly active within those networks. Additionally, commenters within a Facebook page can be extracted and a network map using social network analysis can be drawn using various visualization tools (including NodeXL). Additionally, for some forms of research, it may be useful for researchers to get consent from respondents and have them log into their Facebook accounts from the computer of the researcher. When doing this, netvizz can collect their data, which can include what Facebook pages they have liked. Again, this can be visualized using SNA and, again, there are complex ethical considerations (Rieder 2013). Netvizz itself does not do any of the visualization or analysis. Rather, it is a Facebook app that is used to collect the data. The software that is most commonly used to visualize collected Facebook data via netvizz is Gephi,[12] an open source visualization software package. It is easy to use and there exists a substantial support community online for Gephi users.[13] This includes YouTube videos, how to's, and PowerPoint slides. Gephi has full tutorials on its website that include specific Facebook data tutorials and even tutorials in French, Spanish, and Chinese.[14]

Conclusion

Reflections on the work of C. Wright Mills and the need to critically think about the sociological imagination in the digital age prompted much of how this chapter was conceived. At a critical turning point in his career, Mills took up a post at the Bureau for Applied Social Research (BASR) and was invited to progress his career through a role in administration. He writes in an autobiographical letter to Tovarich that he declined that offer as he prized his role as an 'independent craftsman' (Mills 2000a: 252). What Mills could not have imagined, however, is that the power of ubiquitous

technology, ironically the progeny of lab-coated IBM'ers, has enabled individual sociologists to have a critical, independent craft of studying new forms of computer mediated data. Rather than having to be embedded within large research organizations, new research methods allow sociologists to be independent with their craft. This is important in light of recent abuses of big data (e.g. PRISM and the infamous 'Facebook experiment' (Kramer, et al. 2014)). This chapter has also sought to demystify social media research methods as a matter of principle towards this agenda. If we leave social media research to larger social research bodies and corporate entities, we risk losing a vibrant sociological imagination, which is especially needed in our digital age. Importantly, the independent craft needs hacking of methods (which mixed methods does well) and the retention of a critical outlook, rather than the trend of being required to partner with corporate bodies to conduct this type of research. A key argument I have tried to make is that each of us can learn some tools of the trade which are either free or nearly free and begin our own explorations of social media data rather than relying on corporate or sanitized accounts. And for those not interested in learning how to use these tools, having sufficient social media literacy is important to the social sciences and humanities.

I made a case for the use of digital ethnographic methods in *Sociology* (Murthy 2008). In this article, I outlined the need for digital ethnographic methods including digital field notes, digital observation, and digital ethnography. The same need now exists for methods surrounding social media. It is tempting to simply look at social media data and make quick observations. However, having methods to systematically and rigorously study social media data is fundamental to advancing empirical and theoretical knowledge in diverse social, scientific, and humanities fields. Additionally, the ethics of using social media data are not being sufficiently taught.

This has much to do with the fact that sociologists themselves are not well-versed in social media ethics. There are major ethical implications in using social media data in social research and the discipline needs to take this literacy gap seriously.

This chapter emphasizes accessible, cost-effective (and usually free) solutions that can leverage the power and potential of social media data and research methods. While large-scale social media research is most prominent in the literature, there is little information available about scaling these solutions to fit the needs of individual social researchers or small projects. There is a perception that studying social media requires a large research budget or one needs solutions such as custom programmed tweet collectors. However, many off-the-shelf solutions exist that have similar tools/capabilities to process social media - but on a smaller scale, smaller budget and have the ability to scale up. This potentially empowers individual researchers or researchers in non-computational fields to pursue research questions that were previously unfeasible due to technical and budget challenges. If we in the social sciences do not gain some level of literacy in social media data research, we will be inadvertently propping up barriers between quantitative and qualitative research. As Mills cautions, it is dangerous to sociology to have camps that mystify their work, whether under theoretical jargon or the cover of white lab coats. We are at a critical juncture when we could potentially take on the camp mentality Mills warned us about. And if we do, our ability to understand social world is likely to be significantly hampered.

References

Aggarwal, C. C. 2011 *Social network data analytics*, New York: Springer.
Bakhshi, S., Shamma, D. A. and Gilbert, E. 2014 'Faces engage us: photos with faces attract more likes and comments on Instagram' *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*: ACM.
Blei, D. M., Ng, A. Y. and Jordan, M. I. 2003 'Latent dirichlet allocation', *the Journal of machine Learning research* 3: 993-1022.
Glaser, B. G. and Strauss, A. L. 2009 *The discovery of grounded theory: Strategies for qualitative research*: Transaction Publishers.
Gold, M. K. 2012 *Debates in the digital humanities*: U of Minnesota Press.
Gross, A. and Murthy, D. 2014 'Modeling virtual organizations with Latent Dirichlet Allocation: A case for natural language processing', *Neural Networks* 58(0): 38-49.
Hansen, D., Shneiderman, B. and Smith, M. A. 2010 *Analyzing social media networks with NodeXL: Insights from a connected world*: Morgan Kaufmann.
Kramer, A. D. I., 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): 8788-8790.
Kwak, H., Lee, C., Park, H. and Moon, S. 2010 'What is Twitter, a social network or a news media?' *Proceedings of the 19th international conference on World wide web*: ACM.
Manovich, L. 2011 'Trending: the promises and the challenges of big social data', *Debates in the digital humanities*: 460-75.
Marres, N. 2012 'The redistribution of methods: on intervention in digital social research, broadly conceived', *The Sociological Review* 60(S1): 139-165.
Massey, D. S. 2001 'Residential Segregation and Neighborhood Conditions in U.S. Metropolitan Areas', in N. J. Smelser, W. J. Wilson and F. Mitchell (eds) *America Becoming:: Racial Trends and Their Consequences*, Vol. 1, Washington DC: National Research Council.
Mills, C. W. 1954 'IBM plus reality plus humanism= sociology', *Saturday Review* 37(18): 22-23.
Mills, C. W. 2000a 'C. Wright Mills Letters and Autobiographical Writings', Berkeley: University of California Press.
Mills, C. W. 2000b *The sociological imagination*: Oxford University Press.
Murthy, D. 2008 'Digital ethnography an examination of the use of new technologies for social research', *Sociology* 42(5): 837-855.
— 2011 'Emergent digital ethnographic methods for social research', *Handbook of Emergent Technologies in Social Research, Oxford University Press, Oxford*: 158-179.
Murthy, D., Gross, A. and Bond, S. 2012 'Visualizing Collective Discursive User Interactions in Online Life Science Communities', *arXiv preprint arXiv:1204.3598*.
Murthy, D., Gross, A. and Oliveira, D. 2011 'Understanding Cancer-Based Networks in Twitter Using Social Network Analysis' *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*.
Murthy, D., Gross, A., Takata, A. and Bond, S. 2013 'Evaluation and Development of Data Mining Tools for Social Network Analysis' *Mining Social Networks and Security Informatics*: Springer.
Noss, A. 2013 'Household income: 2012', *US Census Bureau*.

Orton-Johnson, K. P. N. 2013 'Digital sociology : critical perspectives'.
Ottoni, R., Las Casas, D., Pesce, J. P., Meira Jr, W., Wilson, C. and Mislove, A. 2014 'Of Pins and Tweets: Investigating how users behave across image-and text-based social networks' *Association for the Advancement of Artificial Intelligence*
Özyer, T., Erdem, Z., Rokne, J. and Khoury, S. (eds) 2013 *Mining social networks and security informatics*, Dordrecht, NL: Springer.
Rieder, B. 2013 'Studying Facebook via data extraction: the Netvizz application' *Proceedings of the 5th Annual ACM Web Science Conference*: ACM.
Savage, M. and Burrows, R. 2009 'Some further reflections on the coming crisis of empirical sociology', *Sociology* 43(4): 762-772.
Shu-Heng, C. (ed) 2014 *Advances in computational social science : the fourth world congress*, Tokyo: Springer.
Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P. and Rana, O. 2013 'Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter', *Sociological Research Online* 18(3): 7.
Sperling, B. and Sander, P. 2004 *Cities Ranked& Rated: More than 400 Metropolitan Areas Evaluated in the US and Canada*: John Wiley & Sons.
Vishwanath, A. 2015 'Habitual Facebook Use and its Impact on Getting Deceived on Social Media', *Journal of Computer Mediated Communication* 20(1): 83-98.
Zimmer, M. 2010 '"But the data is already public": on the ethics of research in Facebook', *Ethics and information technology* 12(4): 313-325.

[1] Rather than being judgmental towards those that choose to go this route, my argument is that the discipline should do more to try to be welcoming of computational methods and social media data collection and analysis.
[2] Which includes understanding gender and other demographic information.
[3] See http://mallet.cs.umass.edu/
[4] Alexander Gross assisted in the production of this figure.
[5] Further examples can be found in Murthy et al. (2012).
[6] This figure was produced with Alexander Gross and Sawyer Bowman.
[7] Methods-wise, unigrams are rank-ordered in comparison to other unigrams, bigrams with other bigrams, and trigrams with other trigrams.
[8] The popular 'for dummies' book series published by Wiley does have several social media-related titles, but they are mostly marketing related.
[9] Ranging from inexpensive or even free to fully managed by a 3rd party.
[10] For example, by using VMware Fusion (see http://www.vmware.com/products/fusion).
[11] Netvizz can be accessed at: https://apps.facebook.com/netvizz
[12] Gephi can be accessed at: http://gephi.github.io/
[13] See the Gephi forum at https://forum.gephi.org/
[14] See: http://gephi.github.io/users/