10-2014

Sifting the Sand on the River Bank: Social Media as a Source for Research Data

P. Takis Metaxas
pmetaxas@wellesley.edu

Eni Mustafaraj
Wellesley College, emustafa@wellesley.edu

Recommended Citation

Panagiotis T. Metaxas, Eni Mustafaraj (2014). "Sifting the Sand on the River Bank: Social Media as a Source for Research Data." it - Information Technology. Volume 56, Issue 5, Pages 230–239, ISSN (Online) 2196-7032, ISSN (Print) 1611-2776, DOI: 10.1515/itit-2014-1047, September 2014

Sifting the Sand on the River Bank: Social Media as a Source for Research Data

Panagiotis Metaxas, Eni Mustafaraj
Department of Computer Science
Wellesley College, Wellesley, MA, USA

ABSTRACT

Computational social science has been described as a new field at the intersection of computer science and social sciences, aiming to study the ways that society evolves, interacts, and reacts. Like prospectors sifting the sand in a river bed for gold, computational social science researchers are looking into the streams of social media for insight on our social interactions. Enabled by the availability of and easy accessibility to vast amounts of data generated by social entities, as well as by powerful computing hardware and algorithms, its researchers conduct observations of social interaction and experiments testing social theories in scales not realizable before. In this paper, after a short review of the characteristics of this new area, we discuss issues related to the types of data sought and used, and some of the challenges in collecting and interpreting the data. Throughout the paper we also examine some of the pitfalls awaiting and the standards that need to be observed.

1. INTRODUCTION

One of the major technological developments in the early 2000's was the expansion and prominence of the so-called social Web. As Aristotle famously wrote in Politics[1], humans are by nature social animals, and the Web technologies developed today have enhanced the means of their communication. They can do so at every moment, with increasingly fewer space boundaries, and with a far greater variety of people than before – many of whom they may never meet in person. The information technologies that have made this possible, have also provided for the recording and storing of these interactions.

Setting aside the serious implications that such recording has for privacy and security, in this article we focus on the unprecedented potential that these data have for researchers who study social interactions. In fact, while the study of social interactions in the past meant spending considerable amounts of time gathering data from surveys, interviews, handwritten notes and library archives, today's computational social scientists [21] and web scientists [16] can gather large amounts of data in relatively short time. The time saved can potentially be put into the development, analysis and testing of relevant theories.

The ease in collecting data may be misleading if suggesting that the analytical part of the research has become any simpler. In fact, in some instances, it may have made analysis harder, as the collections of so-called "Big-Data" makes it both overwhelming to analyze and potentially obscuring important stories in the data [6]. *Overwhelming*, because the fact that the vast majority of all data in the world have been generated in the last two years[2] requires often the use of powerful computers, algorithms and storage that many researchers may not be prepared to handle. And *obscuring* because focusing on the big picture may diminish the opportunities of discovering important threads that do not make it to the top of the data analysis (e.g., [24]).

In this paper, we discuss issues related to the types of *data sought* and used and some of the challenges in *collecting and interpreting* the data. The remaining of the paper is organized as follows: The next section 2 presents the two different veins of research informed by social media. Section 3 discusses the challenges encountered when collecting social media and section 4 addresses the challenges in interpreting the results of their analysis. The last sections has our conclusions. Throughout the paper we also examine some of the pitfalls awaiting, and the standards that need to be established.

2. VEINS OF RESEARCH USING SOCIAL MEDIA DATA

Today's technologically informed research in social sciences is still interested in some of the same themes that traditional research is: analysis of social observations and design of research experiments. But while this new technology has not resulted yet in major changes in the types of questions asked, many research components have changed. In particular, what has changed is the ease with which *observations* can be gathered, and the scale of *experiments* being conducted. We discuss them in the next two subsections.

2.1 Observational

[1]Wikipedia page: http://en.wikipedia.org/wiki/Politics_(Aristotle)

[2]According to [41], "Every day, we create 2.5 quintillion bytes of data – so much that 90% of the data in the world today has been created in the last two years alone."

Probably the greatest use of social data today is in terms of conducting observational studies of the *digital traces* that people leave while communicating or searching online. Social media companies and search engines collect – and in some cases make partially available – vast amounts of data generated by individuals' and groups' actions while they perform their everyday routines mediated by online technology. This new direction has started to be known under the names of "computational social science" (term coined by [21]) and as a component of "web science" [16].

As an example of this direction, [15] studied mood variations over time across 84 countries by collecting and analyzing public Twitter messages of 2.4 million people. They found that, "around the world, the day dawns full of promise. But moods go downhill over the course of the day, rebounding again in the evening." And while some may argue that "we already knew that" the difference is that we have moved from anecdotal small-scale evidence to a detailed, across-boundaries and cultures, evidence [39].

One of the more surprising examples of observational research comes from data collected by search engines, such as Google and Yahoo, of user searches over time. In [13] it was shown that the information collected by users searching for symptoms and remedies of flu can be used to forecast with high accuracy flu epidemics several days in advance of the official announcement by the U.S. Center for Disease Control (CDC). There is reasonable causality inference for this connection: When people search for flu symptoms, they have a serious need for such information, and this need is recorded by search queries. Metadata also make it possible to pinpoint the geographical area and the intensity of the epidemic and has led to the development of a website that reports regularly on flu epidemics on a global scale[3]. Additional research indicates that flu epidemics could also be detected from other social media data such as Twitter data (see, e.g., [1]).

The possibility that one could forecast important social events based on social media data has given birth to a large collection of research papers trying to predict a variety of important events. We counted that, in 2011, about a fifth of all papers presented in the major social media conferences contained the word "predict" in their title [33]. Typical applications of predictions include forecast of product sales, such as movie box office revenues on the opening week [2]. However, some of the more visible publications involve stock market fluctuations and electoral results. The success of such predictions is still debatable. While there were early reports indicating that stock market fluctuations could be predicted [5], the monetary fund that was established making use of social media data to predict the stock market closed after about only one month of operation.[4] Even more disappointing was the early excitement that followed the announcement that social media data could easily predict the outcome of German elections. In [18] it was pointed out that the data were inadvertently filtered in ways that made the prediction possible. Ironically, not all social media researchers have noticed this development. At the time of this writing, the original, flawed paper has been more widely cited (as example of successful prediction) than its correction by [18]. Multiple attempts to show the predictability of elections using simple, non-representative samples have consistently failed to deliver.

The latter example highlights an important aspect in the kinds of observational research that is conducted today involving replication of prior studies. In this spirit one can see, for example, the work by [11] aiming to replicate some of the methods predicting electoral results in the US or the work to replicate the prediction of box office revenue of movies by [2]. It has been argued persuasively (e.g., by [17]) that *replication* of results is essential for the health of a growing young observational field. Unfortunately, in computational social science venues, replication has not yet caught up[5]. Responsibility for such an omission is shared by authors and editors of research venues: certain authors who see the opportunities in a new topic, rarely returning to it. And conference review committees who see replication as a low-hanging fruit not worth promoting. Since, in general, negative results in science are considered as less valuable [10], inexperienced reviewers do not appreciate that negative results of replication studies are quite essential in verifying the validity of research. In that respect, computational social science is more like medicine, rather than a science field[6]. An important related issue is the lack of sharing research data for meta-analysis, which can be attributed to the privacy issues that such sharing may raise, on the restrictions placed on sharing data imposed by companies that assert ownership on them, and in the failed attempts to safely anonymize data for sharing [36].

2.2 Experimental

In addition to observational studies, there is a comparatively smaller, but currently growing use of Internet-based technologies to enable the recruitment of participants from all over the world into organized experiments. For example, through crowdsourcing sites[7], researchers have answered physiological questions such as the age at which cognitive abilities have peaked, the ability to recognize faces or to estimate the size of a collection [40]. This line of research is also addressing problems related to diversity of subjects participating in the experiments. In particular, one commonly cited objection to experiments conducted in a lab environment rises from the fact that these experiments are using college undergraduate students as study subjects. But there are encouraging findings here: in [12] it was shown that Web experiments compare well with those conducted in the lab.

Probably the greatest gain of online crowdsourced experiments is the tremendous scaling in the number of participants that can participate. However, careful design of the experiments can add other dimensions such as simultaneous experimentation in parallel worlds. A prime example of that is the "Music Lab" experiment that tested the influence of popularity in the perceived quality of products.[8] [32] describes how it was possible to set up eight parallel statistically equivalent worlds to test this influence. Similar

[3]Google Flu Trends site: http://www.google.org/flutrends/

[4]See "Last tweet for Derwent's Absolute Return" in Financial Times, May 24, 2012.

[5]At the time of this writing [4] reports a major effort in replicating recent well-known results in psychology. We expect this effort to expand soon in other social sciences.

[6]See blog post "Research Replication in Social Computing" by P. Metaxas, May 1, 2012, http://hvrd.me/IiXXvv

[7]See, for example, sites Test My Brain: http://testmybrain.org/ and Lab In The Wild: http://labinthewild.org/about.php

[8]Music Lab web site: http://www.princeton.edu/m̃js3/musiclab.shtml

web-based experiments are being conducted within so-called "virtual labs" [27].

One of the better known sociological experiments was conducted by psychologist Stanley Milgram in the 1960's [26]. With his "small-world" experiment, Milgram challenged what we knew about our connectivity to our world and suggested that our society is a network characterized by short-path lengths. His original experiment involved a very small number of participants (296 in the first version and 160 in another). Yet, it was still possible to calculate an average connectivity path (of length six, hence, "six-degrees-of-separation" is the name by which this result is known.[9])

The small number of participants and the high non responsive rate in Milgram's experiment were typical of the times and its technologies. Thirty-five years later, researchers were able to repeat Milgram's small-world experiment using email communication. While the idea was the same, the scale of the experiment changed dramatically. [9] sent more than 60,000 emails to reach 18 persons in 13 countries. The results were shockingly similar to Milgram's (average path lengths between five and seven), but allowed for better understanding of how these paths were being established. Since then, researchers have repeated this experiment using LinkedIn[10] contacts as well as Facebook[11] and Twitter[12] relationships.

Experimentation can involve the study of users' behavior under conditions designed by the researchers that run in live online settings. But manipulating users' behavior can raise significant ethical issues, as the recent "emotional contagion" experiment by Facebook researchers has made clear[19]. While this particular experiment created an outcry due to its potential to manipulate secretly voter's opinions[13], it also showed the limitations of the use of Institutional Review Boards (IRBs) in guaranteeing the ethical use of humans as research subjects.

3. PROCESS AND CHALLENGES IN OBTAINING SOCIAL MEDIA DATA

Like the old prospectors sifting sand on river banks hoping to find gold, computational social scientists have to decide when and where to collect their data in order to examine a hypothesis or test a theory. And like rivers, social sites are usually streaming their data in a variety of speeds. At one end of the stream-speed spectrum, Web sites and blog contents are changing at low speeds making it easier to collect all of their contents for a relatively long amount of time. At the other end, online social sites update content with amazing speeds and pose extra challenges with respect to data collection.

Most of the available research with social media data relies on data generated within the platforms while users are performing certain recordable actions, such as posting, commenting, approving or broadcasting. Researchers generally rely in one of the three following methods to get access to this data: a) ready-to-use datasets provided by data owners; b) Application Programming Interfaces (APIs) made available by the platform owners; and c) crawling and scraping of websites.

Each method has its advantages and disadvantages as we will discuss in the following subsections.

3.1 Ready-to-use datasets provided by data owners

The easiest way to get access to large amounts of data is when the data owners make the data available to the research community for certain research purposes. Usually, this move comes with potential benefits to the companies. The most prominent example was the Netflix Prize[14], that started in 2006 and run until 2009, when a winner team was declared and awarded one million dollars. The purpose of this dataset was to help researchers develop algorithms that predict movie ratings by Netflix customers, which would enable the company to improve the recommendation experience of its users. The example set by Netflix has been followed by many organizations that nowadays use a dedicated platform, Kaggle, which runs world-wide competitions with data provided by the interested organizations [14]. Similar efforts, but with more open-ended research goals in mind, were initiated by datasets provided by Yahoo[15], Google, Microsoft, etc. Other datasets that provide access to billions of blogs posts is the spinn3r datasets[16] of 2009 and 2011. For email communications, the Enron email dataset[17] has been studied extensively due to its role in the trial and conviction of the senior personnel of the company. A slightly different collection of datasets that has attracted considerable attention, is the one that can be downloaded from Wikipedia's websites. As part of its efforts to be transparent, Wikipedia keeps detailed records of every change performed on its pages. Such databases of user activities can be studied for varied research purposes. A recent example was the automatic discovery of sockpuppet accounts [35].

Advantages: Data owner-provided datasets provide ease in data collection and allow comparative research as well as replication studies, since research groups will be using the same dataset.

Disadvantages: Such datasets might be limited to what the owners want researchers to study. In some cases, one still has to write custom code to extract, clean, and format the data before using it for analysis. Additionally, there are issues of privacy, ownership, and consent described in subsection 3.4.

3.2 APIs made available by the platform owners

Many owners of online services make their data available through APIs. When this is done well, it provides the best tool for researchers. However, the technologies of data col-

[9]Wikipedia page: http://en.wikipedia.org/wiki/Six_degrees_of_separation

[10]See"Seeing who you know and how you know them just got easier with Linkedin" http://blog.linkedin.com/2014/01/29/seeing-who-you-know-and-how-you-know-them-just-got-easier-with-linkedin/

[11]See "The Anatomy of the Facebook Social Graph" http://arxiv.org/abs/1111.4503

[12]See "Six Degrees of Separation, Twitter Style" http://www.sysomos.com/insidetwitter/sixdegrees/

[13]See "Facebook Could Decide an Election Without Anyone Ever Finding Out" by J. Zittrain, New Republic, June 1, 2014

[14]Wikipedia page: http://en.wikipedia.org/wiki/Netflix_Prize

[15]Yahoo Labs dataset: http://webscope.sandbox.yahoo.com/catalog.php

[16]Spinn3r dataset at the ICWSM site: http://www.icwsm.org/data/

[17]Enron Email dataset at CMU: https://www.cs.cmu.edu/ enron/

lection are also changing all the time. It used to be the case, for example, that the Google Search API[18] would provide a wealth of information including web page links, backlink collections, and consistency between its reported results and the browser search results. Gradually this changed. Due to the challenges posed by "black hat" search engine optimization (SEO) hackers and web spammers [23], the back link collections were greatly restricted and the results reported by the API greatly diverted from those returned by a browser. An additional reason for this diversion is due to the personalization of results and usability experiments that search engines are continuously conducting. On the other hand, new APIs were made available such as the Google Trends and book n-grams[19] providing opportunity for new types of research, such as "culturonomics" [25].

Additionally, the primary goal of many web platforms for offering APIs to their data is to allow an ecosystem of third-party applications that will provide additional features to users. Twitter was one of the companies that, at least in the first years of its existence, allowed generous use of its data through the API to thousands of developers and researchers. However, in 2011, Twitter stopped this practice[20] and has been constantly changing the way its API works, by making it very difficult to access large amounts of data. For example, the Twitter Streaming API restricts collections once the volume of retrieved tweets with the desired search terms make up for more than 1% of the Twitter volume. This, of course, happens especially in the most interesting situations, such as political debates, natural disasters, and other world-wide captivating events. A recent study [28] that compares this sample with the complete data returned by the so-called Twitter Firehose has discovered disparities that need to be carefully addressed by researchers. However, while Twitter still allows researchers to collect their own datasets, it prohibits the sharing of such datasets with other researchers, beyond the level of the list of tweet IDs in the dataset. This prevents comparative research on the same dataset. Currently, Twitter directs researchers to its pay-for-content partners such as Topsy[21].

Advantages: APIs provide high-quality data. Programs for collecting data through APIs are often available in multiple programming languages. Often, it's the only allowed way to collect data from a website.

Disadvantages: Websites and companies can change or discontinue their API services at any moment. Often the APIs have restrictions on what can be done with the data (e.g., Facebook requires that researchers request specific permission from every user). APIs give only a partial view of the data collected by the service operators.

3.3 Crawling and scraping of websites

Collecting data through crawling and scraping is the oldest technique used on the web and is guaranteed to work when there is no access through owner-datasets or APIs. However, it's technically the most challenging and cannot be used by researchers who lack training in computer science. The crawling process refers to the automatic visiting of webpages starting from a base URL. It is used by search engines to index the Web and it can be repurposed to download the entire content of a website for offline processing. Software programs such as cURL or GNU wget[22] are commonly used for crawling websites. Once the content of HTML pages is retrieved from the web, one uses the process of web scraping to extract the useful information from the HTML file. For this, dedicated programs based on existing libraries such as Java's HtmlCleaner or Python's BeautifulSoup need to be written. Recently, dedicated professional tools have started to emerge, e.g, https://www.kimonolabs.com/.

This combination of crawling and scraping works well for websites where content is retrievable through permanent URLs, so that the crawler can access all pages. However, many websites nowadays have moved toward dynamic loading of social content. For example, in the past it was possible to crawl and scrape the comments from NY Times articles. This is not possible with the above-mentioned techniques anymore, because loading comments on the page requires user interaction. Workarounds for scraping dynamic content include the use of tools such as Selenium[23] that can be used to generate scripts that simulate user interaction with a page.

Advantages: With crawling and scraping researchers can collect data in large amounts (millions of posts by users), that are not accessible by other means, see [34], [3].

Disadvantages: Many websites have terms of service that explicitly prohibit crawling. Additionally, researchers will need to write custom code to extract data from HTML pages. Some issues of consent and replicability are described in subsection 3.4.

3.4 Issues with data access

Data Privacy - Data privacy is a major issue with all the described collection methods, but has received attention mostly in regard with the company-provided datasets (e.g the Netflix dataset [30], or the NYC taxi drivers dataset[24]). While companies try to anonymize the data before making them available, the concern is that computing power and correlations between anonymous and non-anonymous data has made it possible to de-anonymize a portion of users in such datasets [36]. In some cases this information may even be inside the dataset, as it was the case for the AOL dataset that was de-anonymized by journalists.[25] Yet, in other cases de-anonymization can be obtained by crawling public websites, such as Flickr [31].

Data Ownership - The data collected through the three above mentioned methods mostly represent only the "tip of the iceberg" of what the companies themselves possess in terms of user-generated activity. In fact, researchers receive access to the portion of data that is already public through websites' interfaces, while the most interesting information (the digital traces of usage), conducive to detecting unconditioned patterns or understanding user behavior, is visible

[18]Google APIs Explorer: https://developers.google.com/apis-explorer/

[19]Google Trends: http://www.google.com/trends/ Book grams: https://books.google.com/ngrams

[20]Twitter Kills the API Whitelist http://readwr.it/h0Vo

[21]Topsy: http://topsy.com/

[22]cURL: http://curl.haxx.se/ GNU wget: http://www.gnu.org/software/wget/

[23]Selenium: http://docs.seleniumhq.org/

[24]The Guardian: "New York taxi details can be extracted from anonymized data". By Alex Hern. Published: June 27, 2014.

[25]NYTimes: "A Face Is Exposed for AOL Searcher No. 4417749" By Michael Barbaro and Tom Zeller Jr. Published: August 9, 2006.

neither to the users themselves, nor to the public. Private data is used by the companies to improve their service, as well as to better target advertisement to their customers. The power imbalance created by this one-sided data ownership generated sufficient pushback and eventually all big platforms, Facebook, Twitter, and Google added tools to allow users to download their own data when desired[26].

User Consent - Within the field of CSS, computer scientists have been at the forefront of research due to their ability to collect data automatically. Because they tend to regard everything published on the Web as "public", there has been a growing chasm between them and social scientists when it comes to the topic of "user consent". Efforts to bridge the gap in ethical handling of social media data were the focus of two early workshops in 2009 [37] and 2010 [38], presciently titled "Research Ethics in the Facebook Era". The recent controversy surrounding Facebook's "emotional contagion" study [19], uncovered the fact that no agreement has been reached within the research community, and the two fronts remain divided (a list of researchers' and media reactions is available here[27]).

Research Replicability As we discussed in subsection 2.1, replicability is not an established practice within the CSS community. Adding to the lack of good will are the difficulties that the previous three issues present. Due to data ownership and in name of protecting data privacy, most big companies have created their own research departments to perform observations and experiments. Some companies, such as Facebook[28], have programs that allow academicians to spent a sabbatical in their research labs to access their data. Recently, Twitter introduced a program titled "Data Grants"[29] to give access to its entirety of data. The fact that it chose only six projects out of 1300 proposals[30], reemphasizes the power imbalance generated by data ownership. Such situations have led to increasing concern on fairness and correctness of results based on privately held data [17], leading some publishing venues to introduce requirements of dataset openness as a criterion for publication.

4. CHALLENGES IN INTERPRETING THE RESULTS

In the previous section we described some of the technical and ethical issues arising from the collection of data from social media. In this section we will discuss the challenges a researcher may face once the data have been collected. These challenges are related to the *completeness* of the data, to the *computational ability* in processing and analyzing them, and to issues affecting the *interpretation* of the results.

First we will examine in some detail challenges arising from applying the so-called *closed world assumption* of the analysis, and challenges related to the *filtering* streams of data. The *storing and processing* needs generated by trillions of data items every day and the challenges posed by the variety and level of maturity of *algorithms* used for analysis (e.g., algorithms developed in different theoretical environments such as in Artificial Intelligence, Graph Theory, Physics, Applied Statistics, etc) are two important technical challenges outside the realm of this paper.

People and other social entities interact in a variety of ways and in a non-continuous fashion. Some online discussion that emerged on blogs might migrate in one or more Facebook pages, and may emerge as a trending topic on Twitter featuring a variety of hashtags [7]. At the same time, important input may come through traditional media channels such as radio and TV. It would be impossible to collect every bit of information related to a viral discussion. Part of the due diligence for a researcher is to check all potentially relevant contributions and collect data in a comprehensive way that, arguably, capture the discussion.

It is important to observe that it may not be essential to collect all relevant information to answer a question. Instead, a diligent researcher hopes to collect the most informative and/or influential pieces of data. Thankfully, this may be possible. However, the fact that the data collected do not represent all of the relevant information should be in the minds of the researchers who should present a convincing argument of how the missing data may, or may not affect their analysis.

As an example of how the closed world assumption may affect data, consider the inclusion of real-time results in the top-10 of search engine search results that was implemented by Bing in October, and by Google in December, 2009.[31] While users of the two major search engines were expecting that the top-10 search results represented the outcome of a sophisticated ranking by search engines[23], the inclusion of real-time postings on Twitter, Facebook and others broke this assumption. Postings from social sites would rise to the third position of search results independently of quality. Political activists noted the change and started repeating tweets related to the crucial 2010 Special Senatorial Elections in Massachusetts to fill the seat of the deceased Senator Ted Kennedy. Repetition of the same tweet by a user was allowed by Twitter at that time, as it was probably expected not to occur intentionally: repeating a tweet would likely annoy the followers of the original sender. However, political activists repeated one-third of all related tweets in an effort to spam search engine users who were looking for news regarding the candidates of the elections. Their actions aimed at launching a Google bomb[32] through Twitter [22], violating the closed world assumption. A researcher wanting to understand the political discourse at the time should better be careful to filter those repeated tweets as they were not part of the discourse, but part of a scheme to influence voter perception in a different medium. Not surprisingly, online Social Media companies pay attention to research developments, and moved to diffuse future Google bombs through Twitter. A few months after the publication of [22], search engines removed real-time results from search results and placed them on a sidebar, while Twitter disabled verbatim repetition of a tweet by the same user.

[26]Wired: "How to Download and Archive Your Social Media Memories", July 15, 2014.

[27]Archive: http://laboratorium.net/archive/2014/06/30/the_facebook_emotional_manipulation_study_source

[28]Research Publications at Facebook: https://www.facebook.com/publications

[29]https://blog.twitter.com/2014/introducing-twitter-data-grants

[30]https://blog.twitter.com/2014/twitter-datagrants-selections

[31]Search Engine Land blog: "Google Launches Real Time Search Results", Dec 7, 2009 by Danny Sullivan.

[32]A Google bomb is created when spammers are able to circumvent the ranking algorithms of search engines, promoting their own content to the top-10 search results. See [23] for more information.

On the other hand, consider the efforts to evaluate the predictive power that Twitter volume may have on prediction electoral results. Assuming a closed world, researchers collected data for several months in advance of the election and found that counting the tweets containing text references to German political parties was enough to predict the electoral outcome with impressive accuracy. However, [18] has shown this prediction to be wrong, as it ignored relevant tweets mentioning parties which the original researchers were not looking for. We would argue that this embarrassment can be avoided when researchers are aware of the possibility of "apophenia" or seeing patterns where none actually exist [6]. One can avoid apophenic effects by searching for theoretical explanations on why correlations may occur. Without them, correlations achieve causation status by default.

In another example described by [20], the closed world assumption was accidentally violated from within. It was observed that the Google Flu Trends site was overestimating the reported cases of flu, in some cases by almost 130% [8]. One of the reasons that affected the estimates was the inclusion of "suggested searches" in the search engine's interface. People searching for remedies to flu symptoms were offered the opportunity of doing more searches by clicking on searches that others had recently done. As a result, not only did the volume of searches increase (something that was addressable through normalization), but the volume of irrelevant symptoms also affected the generation of real data.

On the other hand, careful filtering of information can discover important stories obscured by Big Data. We present two examples about it. The first example comes from the study of the 2010 Special Elections in MA, USA. As a result of their study to understand the political discourse about the elections on Twitter, [22] found that the Twittersphere was extremely polarized and measured the extent of polarization. A closer examination of the collected data revealed the first political Twitter bomb,[33] an effort to spread lies from a set of 9 fake accounts that were set up for this reason, eventually reaching about 60,000 accounts within hours. As a result of this accidental finding, [22] influenced the study of Twitter bots and the way we think today about social media data.

The second example comes from the study of user interactions in drug war-torn Mexico in 2011, conducted by [29]. For a number of years, citizens of central and north Mexico have faced dangerous daily situations between drug cartels fighting for area control, and between drug cartels and various Mexican authorities. The murderous attacks of the cartels on professional reporters have resulted in making the latter fearful of reporting, leaving the ordinary citizens uninformed and confused. As a result, ordinary citizens have turned to Twitter as a means of collecting and disseminating information of risk situations by anonymous citizens[34]. While the [29] study analyzed the overall picture of anonymous interaction between citizens worried for their safety, the Big Data methods employed failed to see an important story within the community of anonymous reporters. A follow up paper [24] discovered the attack on a prominent citizen reporter by trolls. Going beyond Big Data wide-view methods, [24] used a variety of in-depth methods, including changes in the patterns of interaction by the audience, interviews with anonymous reporters, and blog exchanges[35].

[33]A Tweeter bomb refers to posting a large number of tweets targeting unsuspecting individuals from multiple accounts with the goal of advertising a certain message. See http://en.wikipedia.org/wiki/Twitter_bomb

[34]Trusting Anonymous Twitter Users http://hvrd.me/M5KDBw

5. CONCLUSIONS

Like prospectors sifting the sand in a river bed for gold, computational social science researchers are looking in the streams of social media for insight on our social interactions. In this paper we discussed issues related to this relatively new interdisciplinary research area. First, we gave an overview of the observational and experimental veins of research conducted in the last decade, and we then focused on some of the issues related to collecting, analyzing and interpreting the data. Throughout the paper we discussed possible pitfalls of the process, some resembling the discovery of fool's gold in the eyes of the prospectors.

As the ancient philosopher Heraclitus of Ephesus[36] once said, "You never step into the same river twice". This is certainly true for the recording and analyzing streams of social media data. The means by which data are collected, through datasets, APIs or crawling can have implications on the analysis that will be performed. Practical issues, such as storing space and processing power may limit our abilities to analyze them. Filtering, interventions by social entities and ownership rules can further challenge our interpretation of the analysis. And of course all of these challenges have significant implications for the replicability of results. Several successful or controversial examples are presented to substantiate our description of how social media data are used for research today.

Even though this is a new field, it already has had a major impact on interdisciplinary collaboration, bringing together researchers from the sciences, the social sciences, and humanities. We expect that this trend will continue and will expand quickly, providing new insights in the way society works.

6. ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their constructive comments. This research was supported by NSF grant CNS-1117693.

7. REFERENCES

[1] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *EMNLP*, pages 1568–1576. ACL, 2011.

[2] S. Asur and B. A. Huberman. Predicting the future with social media. In *IEEE/WIC/ACM Intl Conf on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 492–499, Washington, DC, USA, 2010. IEEE Computer Soc.

[3] M. S. Bernstein, A. Monroy-Hernández, D. Harry, P. André, K. Panovich, and G. G. Vargas. 4chan and/b: An analysis of anonymity and ephemerality in a large online community. In *ICWSM*, 2011.

[4] J. Bohannon. Replication effort provokes praise–and 'bullying' charges. *Science*, 344:788–789, 2014.

[35]Looking beyond ÒBig DataÓ analysis to discover those who make a difference http://bit.ly/1ADL0dt

[36]Wikipedia article: http://en.wikipedia.org/wiki/Heraclitus

[5] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2011.

[6] D. Boyd and K. Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5):662–679, 2012.

[7] A. Bruns and S. Stieglitz. Towards more systematic twitter analysis: metrics for tweeting activities. *International Journal of Social Research Methodology*, 16(2):91–108, 2013.

[8] D. Butler. When Google got flu wrong. *Nature*, 494(7436):155–156, Feb. 2013.

[9] P. S. Dodds, R. Muhamad, and D. J. Watts. An experimental study of search in global social networks. *Science*, 301:827–829, 2003.

[10] D. Fanelli. Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3):891–904, 2012.

[11] D. Gayo-Avello, P. T. Metaxas, and E. Mustafaraj. Limits of electoral predictions using twitter. In L. A. Adamic, R. A. Baeza-Yates, and S. Counts, editors, *ICWSM'11*. The AAAI Press, 2011.

[12] L. Germine, K. Nakayama, B. Duchaine, C. Chabris, G. Chatterjee, and J. Wilmer. Is the web as good as the lab? comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, 19(5):847–857, 2012.

[13] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009. 10.1038/nature07634.

[14] A. Goldbloom. Data prediction competitions – far more than just a bit of fun. In W. Fan, W. Hsu, G. I. Webb, B. L. 0001, C. Zhang, D. Gunopoulos, and X. Wu, editors, *ICDM Workshops*, pages 1385–1386. IEEE Computer Society, 2010.

[15] S. Golder and M. Macy. Diurnal and seasonal mood vary with work, sleep, and day length across diverse cultures. *Science (New York, N.Y.)*, 333(6051):1878–1881, 2011.

[16] J. Hendler, N. Shadbolt, W. Hall, T. Berners-Lee, and D. Weitzner. Web science: An interdisciplinary approach to understanding the web. *Communication of the ACM*, 51(7):60–69, 2008.

[17] B. A. Huberman. Sociology of science: Big data deserve a bigger audience. *Nature*, 482(7385):308ï£¡308, 2012.

[18] A. Jungherr, P. Jurgens, and H. Schoen. Why the pirate party won the german election of 2009 or the trouble with predictions: A response to Tumasjan, a., Sprenger, t. o., Sander, p. g., & Welpe, i. m. ï£¡predicting elections with twitter: What 140 characters reveal about political sentimentï£¡. *Social Science Computer Review*, 30(2):229–234, 2012.

[19] A. D. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 11(24), 2014.

[20] D. Lazer, R. Kennedy, G. King, and A. Vespignani. Big data. the parable of google flu: traps in big data analysis. *Science Magazine*, 343(6976):1203–1205, March 2014.

[21] D. Lazer et al. Social science: Computational social science. *Science*, 323(5915):721–723, February 2009.

[22] P. Metaxas and E. Mustafaraj. From Obscurity to Prominence in Minutes: Political Speech and Real-Time Search. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, Apr. 2010.

[23] P. T. Metaxas. Web spam, social propaganda and the evolution of search engine rankings. In J. Cordeiro and J. Filipe, editors, *WEBIST (Selected Papers)*, volume 45 of *Lecture Notes in Business Information Processing*, pages 170–182. Springer, 2009.

[24] P. T. Metaxas and E. Mustafaraj. The rise and the fall of a citizen reporter. In H. C. Davis, H. Halpin, A. Pentland, M. Bernstein, and L. A. Adamic, editors, *WebSci*, pages 248–257. ACM, 2013.

[25] J.-B. Michel et al. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[26] S. Milgram. The small world problem. *Psychology Today*, 1(1):61–67, 1967.

[27] D. Watts. Computational Social Science: Exciting Progress and Future Directions *The Bridge on Frontiers of Engineering*, 43(4), 2013.

[28] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley. Is the sample good enough? comparing data from twitterï£¡s streaming api with twitterï£¡s firehose. *Proceedings of ICWSM*, 2013.

[29] E. Mustafaraj, P. T. Metaxas, S. Finn, and A. Monroy-Hernández. Hiding in plain sight: A tale of trust and mistrust inside a community of citizen reporters. In *ICWSM*, 2012.

[30] A. Narayanan, V. Shmatikov. How to break anonymity of the Netflix prize dataset. arXiv preprint cs/0610105, 2006.

[31] A. Narayanan, E. Shi, and B. I. P. Rubinstein. Link prediction by de-anonymization: How we won the kaggle social network challenge. *CoRR*, abs/1102.4374, 2011.

[32] M. Salganik and D. Watts. Web-based experiments for the study of collective social dynamics in cultural markets. *Topics in Cognitive Science*, 1(3):439–468, 2009.

[33] H. Schoen, D. Gayo-Avello, P. T. Metaxas, E. Mustafaraj, M. Strohmaier, and P. Gloor. The power of prediction with social media. *Internet Research*, 23(5):528–543, 2013.

[34] S. Y. Schoenebeck. The secret life of online moms: Anonymity and disinhibition on youbemom.com. In *ICWSM*, 2013.

[35] T. Solorio, R. Hasan, and M. Mizan. A case study of sockpuppet detection in wikipedia. In *Workshop at NAACL-HLT 2013*. ACL, 2013.

[36] L. Sweeney. Simple demographics often identify people uniquely. Working paper, Carnegie Mellon University, Data Privacy, 2000.

[37] N. Bos, K. Karahalios, M. Musgrove-Chavez, E. Poole, J. Thomas, S. Yardi. Research ethics in the Facebook era: privacy, anonymity, and oversight Workshop Proceedings of CHI'09. ACM, 2009.

[38] A. Bruckman, K. Karahalios, R. Kraut, E. Poole, J.

Thomas, S. Yardi. Revisiting Research Ethics in the Facebook Era: Challenges in Emerging CSCW Research Workshop Proceedings of CSCW'10, ACM, 2010.

[39] D. Watts. *Everything is Obvious: Once You Know the Answer*. Crown Business, 2011.

[40] J. B. Wilmer, L. Germine, C. F. Chabris, G. Chatterjee, M. Williams, E. Loken, K. Nakayama, and B. Duchaine. Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences*, 107(11):5238–5241, 2010.

[41] P. Zikopoulos, D. deRoos, K. Parasuraman, T. Deutsch, J. Giles, and D. Corrigan. *Harness the Power of Big Data – The IBM Big Data Platform*. Mcgraw-Hill, 2012.