Big Data in Online Social Networks: User Interaction Analysis to Model User Behavior in Social Networks

Divyakant Agrawal, Ceren Budak, Amr El Abbadi,
Theodore Georgiou, and Xifeng Yan

Department of Computer Science,
University of California, Santa Barbara
{agrawal,cbudak,amr,teogeorgiou,xyan}@cs.ucsb.edu

Abstract. With hundreds of millions of users worldwide, social networks provide incredible opportunities for social connection, learning, political and social change, and individual entertainment and enhancement in a multiple contexts. Because many social interactions currently take place in online networks, social scientists have access to unprecedented amounts of information about social interaction. Prior to the advent of such online networks, these investigations required resource-intensive activities such as random trials, surveys, and manual data collection to gather even small data sets. Now, massive amounts of information about social networks and social interactions are recorded. This wealth of big data can allow social scientists to study social interactions on a scale and at a level of detail that has never before been possible. Our goal is to evaluate the value of big data in various social applications and build a framework that models the cost/utility of data. By considering important problems such as Trend Analysis, Opinion Change and User Behavior Analysis during major events in online social networks, we demonstrate the significance of this problem. Furthermore, in each case we present scalable techniques and algorithms that can be used in an online manner. Finally, we propose the big data value evaluation framework that weighs in the cost as well as the value of data to determine capacity modeling in the context of data acquisition.

Keywords: Social Networks, Big Data, Social Analytics, Data Streams, Complex Networks.

1 Introduction

One of the main challenges confronting researchers in many diverse fields is the analysis and understanding of very large data sets. Not only do physical scientists face this challenge when observing natural phenomena or studying experimental results, but social scientists are also being exposed to ever increasing and diverse data sets. In spite of the challenges associated with big data, this phenomenon is enabling scientists approach traditional problems from new perspectives. In the

A. Madaan, S. Kikuchi, and S. Bhalla (Eds.): DNIS 2014, LNCS 8381, pp. 1–16, 2014.

context of social sciences prior to the advent of online networks, various investigations required resource-intensive activities such as random trials, surveys, and manual data collection to generate even small data sets. Now, many social interactions take place in an online environment, and as a result, massive amounts of data about social networks and social interactions are recorded. This wealth of data, presenting an almost-natural yet not easily controllable laboratory for social experiments, can allow social scientists to study social interactions at a scale and at a level of detail that has never been possible before. In fact, it has been argued that online social networks present social scientists with a unique opportunity to observe and analyze interactions in social networks. The right set of *data summarization* tools can help scientists extract *knowledge* out of these ever increasing and diverse data sets.

Modern on-line social networks, such as Facebook, Twitter, and Renren contain a wealth of public information regarding the interactions, likes and dislikes, interests of hundreds of millions of individuals who form large segments of the global society. Facebook and Twitter each claim about 800 million users, and at any given moment millions of interactions occur among the users of each of these social networks. Communication exchanges are occurring on a continuous basis, with about 500 million tweets per day on Twitter and a peak of 143199 tweets per second observed during the airing of a movie in Japan on August 2013 [1]. Both the topic as well as the pattern of such communications can provide deep insights in diverse and sometimes critical contexts. For example, it was reported that during the 2008 Santa Barbara fires, on-line social networks were considered more reliable and up to date with fire locations and evacuation information than the traditional media outlets; or that tweets often spread faster than the tremors of an earthquake [46]. The benefits of online social networks during emergency events extends to natural disasters such as hurricanes and earthquakes [21,51]. In general, in emergency situations both the content as well as the spread of information in social networks can provide valuable *knowledge* that is critical in life saving situations.

The utility of online social networks is not limited to emergency events. Recent evidence indicates that 45% of users in the U.S. say that the Internet played a crucial or important role in at least one major decision in their lives in the last two years, such as attaining additional career training, helping themselves or someone else with a major illness or medical condition, or making a major investment or financial decision [19]. In fact, basic human activities have changed in the context of the Internet and social networks, and new possibilities have emerged. For instance, the process by which people locate, organize, and coordinate groups of individuals with shared interests, the number and nature of information and news sources available, and the ability to solicit and share opinions and ideas across myriad topics have all undergone dramatic change as a result of interconnected digital media. Furthermore, increasing reliance on the "wisdom of crowds" has been demonstrated to both solve and effectively predicate diverse human behavior. Aggregating the efforts of anonymous crowds has been demonstrated to help address complex issues [20,36]. There is also growing

evidence that communities and the exchange of information among connected individuals result in increasing the overall knowledge of the community. Capturing this knowledge is a major challenge [31], and if accurately captured, channeled and articulated, it can help solve many of humanities challenges, such as harnessing human capacity to overcome such endemic challenges as world hunger and illiteracy. This wealth of information, though present in social networks, is buried under a large amount of noise in big data. Even in the cases where the vastness of data is not necessarily a disadvantage, the advantage, or rather the amount of it, needs to be questioned. Lately, there has been growing rhetoric that argues that the information you can extract from any big data asymptotically diminishes as your data volume increases [58].

So, is more data always better? As counter intuitive as it sounds, the answer to this question is not simply "yes". Instead, the answer, while being less satisfactory, is "depends". For instance, the value of *big* data in identifying correlations between two measures x and y in a data set is questionable [14]. It's not hard, even with a data set that includes just 1,000 items, to get into a situation in which we are dealing with many, many millions of correlations. This means that out of all these correlations, a few will be extremely high just by chance: if you use such a correlation for predictive modeling, you will lose [14]. We explore different Social Behavior problems through an analysis of large datasets. We study the problem of Trend Analysis (trending topics) in various levels; from simple trend detection to multi-dimensional trend analysis. We analyze how Opinion changes in a social context and how sentiment varies as global and local opinions change. Finally we explore the area of Event Detection and Summarization and how users behave and break news during large scale and real life events. What is common between all these social behavior problems is that Big Data plays a critical and not always beneficial role. Our long term goal is to further study the impact and implications of Big Data and introduce a framework that can analyze a social problem and attempt to answer the following critical questions: (1) Is a dataset appropriate (utility of the data) and (2) do we need more or less data (amount of data)?

2 Analytic Approaches for User Behavior Modeling

In this section we present related and established work on two of the social behavior applications we study, Trend Analysis and Opinion Change.

2.1 Trend Analysis

Online social networks contain a large variety of information and identifying specific information items of importance has been of interest since their inception. A simple metric for identifying the importance of specific informational topics can be evaluated by the accumulated interest such topics receive from users over time. Such measures are already in use as in the case of YouTube (view counts of videos) or Digg.com (number of *diggs* a story receives). Considering the usefulness and possible impact of this measure, we first start with a

formal *count-based trends* in which topics accumulate value over time according to the number of times they have been mentioned. Assume users of a network can choose to (or not to) broadcast their opinions about various topics at any point in time. Assume further that we can abstract away what the topic is from what a user broadcasts. In this setting, we model a *mention* by node n_i on a specific topic T_x as a tuple $\langle n_i, T_x \rangle$. We refer to the history of such tuples as *stream* and denote it using S. Under this model, *count-based* trendiness of T_x can be defined as:

$$f(T_x) = \sum_{n_i \in N} C_{i,x} \tag{1}$$

where $C_{i,x}$ represents the number of mentions of the form $\langle n_i, T_x \rangle$ in S, i.e. the number of mentions by node n_i of topic T_x and N is the set of users in the social network.

The top-k topic detection problem, when the score is defined in this way, is simply to find the frequent items in a stream of data, also referred to as *heavy hitters*. This problem can be easily solved by keeping track of the accumulated count for each topic discussed in the social network. However, for large and dynamic data sets, it is desirable to look for approximate solutions. Given the large scale of online social networks today, both in terms of number of users and volume of activity, even the simple count-based trends detection calls for such approximate solutions. The *frequent elements problem* has been well studied and several scalable, online solutions have been proposed [8,13,43,40]. The algorithms for answering frequent elements queries are broadly divided into two categories: *sketch-based* and *counter-based*. In the sketch-based techniques [8,13], the entire data stream is represented as a summary "sketch" which is updated as the elements are processed. On the other hand, counter-based techniques [43,40] monitor a subset of the stream elements and maintain an approximate frequency count. The Space Saving algorithm, a counter-based algorithm [43], has been identified to have the best throughput amongst its class of frequency counting algorithms [12]. We therefore plan to use it as a building block for discovering count-based trends.

In recent years, there has been a great increase in research relating to online social networks. While early works focused on static social networks analysis [32,33], more recent research evolved to study more complex and dynamical notions such as information diffusion [27]. As importance of information trends in social networks increased, there has been a number of studies that focused on information trends from various perspectives [4]. For instance, Kwak et al. [33] study and compare trending topics in Twitter reported by Twitter [56] with those in other media. The results show that the majority of topics are headline news or persistent news in nature. In [35] Leskovec et al. study temporal properties of information shared in social networks by tracking "memes" across the blogosphere.

Recently, a number of works have studied structural properties of graphs in a streaming or semi-streaming fashion. The computation of network indices based

on counting the number of certain small subgraphs is a basic tool in the analysis of the structure of large networks. A type of problem that is significantly related to the problem studied here is counting triangles in a graph stream. There are three types of solutions to this problem: exact counting [5], streaming [6] and semi-streaming algorithms [7]. Detecting trends that are not oblivious to the underlying structure of an online social network requires online solutions and therefore these techniques are not directly applicable.

Another important characteristic of news or discussions in social networks is the spatial properties of the agents that are involved in the discussion or the source of the news. A recent work by Teitler et al. [53] collects, analyzes, and displays news stories on a map interface, thus leveraging their implicit geographic context. A follow-up study performs similar techniques to identify geographical information in news in Twitter. Although these works that focus on temporal and spatial characteristics of trends are important for a better understanding of the notion of trends, they are orthogonal to the approaches introduced in this study, as they focus on identifying tweet clusters based on locations and not trend detection. Recently, there has been more effort in online analysis of geo-trends in social networks [38,18]. Hong et al. [18] focus on user profiling from a geographical perspective by modeling topical interests through geo-tagged messages in Twitter. This problem is orthogonal to the problem studied in here as it focuses on user-centric modeling in an offline manner while our approach aims at detecting trends in an online fashion. Similar to our work, MacEachren et al. [38] study the problem of identifying significant events in different localities for crisis management. However, this work provides a high level framework while we provide efficient algorithmic tools with accuracy guarantees.

2.2 Opinion Change

The proliferation of social media, forums, and networks has witnessed the power of networks that propagate news, opinions, and stances on a scale and speed that have never been seen before. Unfortunately, due to the lack of appropriate metrics and models, we are not able to characterize, quantify, and predict persuasion that is occurring everywhere in the social-cyber space. At the core of analyzing persuasion over networks, there is a fundamental problem: how to measure, model and simulate the opinions and shifts of opinion of users in a network, with reasonable accuracy. While it is difficult to derive an accurate model on the individual level, we have demonstrated in our work this year that it is possible to build a collective model over groups of people that share similar opinions over a set of specific topics. By studying and comparing the position and the dynamics of position, persuasion patterns and knowledge that are hidden in complex social and information networks may be revealed.

In order to detect reasons for public opinion change/persuasion, one can first track sentiment variation towards the interested target. If a significant change in crowd sentiment is observed on Twitter, one can analyze tweets during the corresponding period to discover the reasons. There are three challenges for this task: (1) Tweets are very noisy and cover many general topics/events which

do not really contribute to sentiment change. How to filter out these unrelated topics/events is a serious issue. Text summarization techniques are not appropriate for this task since text summarization aims at covering all topics/events in the text collection. Similarly, extracting the most frequently mentioned words during the change to represent the reasons is not a good idea, as these words may actually come from the background or general topics/events which have been discussed for a long time. (2) Events sometimes are complex and are composed of a number of small events. The change of opinion may be caused by only one subevent but not the whole event. How to find these fine-grained reasons is generally very challenging. (3) The third challenge is how to properly represent the reasons. Keywords or topics output by Topic Modeling methods [55] can describe the underlying topic to some extent; but this is not as intuitive as natural language sentences.

Topic-based User Sentiment Analysis and Classification. In order to study and analyze the change in users topic sentiments across time, we first must discover their sentiment from communication data. There have been numerous prior sentiment classification methods introduced which focus upon the determination of sentiment (classifying as positive or negative) within messages sent between users on Twitter [41,57,23,16,9] and other social media websites [45,42,50,39,30,49]. We will later describe our existing approach that classifies tweets from a large, real-world Twitter dataset combining Tan et al.'s technique [57], and Mudhakar et al.'s Multinomial Bayes classifier [45].

Modeling of Sentiment Change in Social Networks. Most current models for the spread of ideas and influence in social networks are based on diffusion of the idea from a node to its (directed) neighbors. For example, if the status x(t) at time t of node i is either active (an adopter of the idea) or inactive, then such a model might postulate that the status at the next time interval is given by $x_i(t+1) = \sum w_{ij} x_j(t)$, where the weighted sum is taken over node i and its immediate neighbors. This type of model, which has its roots in social network theory, was developed to explain small group dynamics [15], and has also been used to simulate dynamics of fish schools [25]. Variants of this basic diffusion model are the Voter Model, and Independent Cascade and Linear Threshold Models [26,29,10]. These types of models have been very successful in explaining information propagation, and they lend themselves well to theoretical analysis. On the other hand, they have known shortcomings. For example, according to social network theory, a given node will keep updating its status even if the status of its neighbors is unchanging. This seems unrealistic, particularly in the context of social networks. Our plan is to determine the extent to which diffusion models can be validated on real social network opinion data, and then to consider potential improvements and extensions to the models.

Controlling Opinion and Detecting & Countering Control. A large number of works relating to influence maximization and opinion control on social networks have been previously introduced. Many of these have taken a threshold

approach to modeling decision and opinion change of users, theorizing that the current number of neighbors with an opinion decides the current opinion of a node [26,17,47]. This type of threshold approach is therefore timing and order-independent, as the timing of neighboring users' opinion changes does not affect the resulting opinion decisions. It is only recently that works have begun to take into account the effect of ordering within influence and opinion cascades, acknowledging that timing and sequence play a vital role in the spread of opinions. In [11], Chierichetti et al. looked at the effect of sequence among neighbors opinion changes, (determining optimal orderings of sentiment changes) to analyze this effect on product adoption cascades. However, while the sequence of neighboring opinions are studied in [11], the actual timing and rates of these opinion adoptions how closely clustered they are in time are ignored.

3 New Approaches for User Behavior Analysis and Modeling

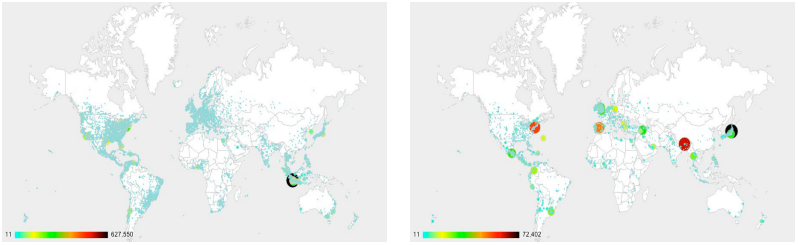
In this section we describe future research on User Behavior in Social media and the impact that Big Data has in each case.

3.1 Semantic-Based Information Trends

Information that is shared in a social network may have certain semantic properties such as the location and time. For instance, one might be interested to know the trends in California alone or short/long term trends . Such queries cannot be answered using trends analysis at the scale of the entire network. Therefore we believe there is a need for trend definitions that explore such dimensions. Our belief is also supported by the growing body of research in this field [54,52]. In this section, we first discuss trends that explore the spatial and temporal characteristics of data.

Spatial trends can be defined in various ways. For instance, the goal can simply be to detect heavy hitters for each location. However, such a technique fails at identifying topics of true geographical nature since a topic of global importance incidentally also has a high frequency of occurrence in various localities without really being related to such locations. Distinguishing such a topic from ones that are trending in only certain localities is not possible without considering the *correlations* between places and topics. Therefore, we plan to focus on the problem of identifying the correlation of information items with different geographical places. We propose *GeoWatch*: an algorithmic tool for detecting geo-trends in online social networks by reporting *trending* and *correlated* location-topic pairs. *GeoWatch* also captures the temporality of trends by detecting geo-trends along a sliding window. With the use of different window sizes, trends of different time granularity can be detected. Our analysis on a Twitter data set shows that such geo-trend detection can be very important in detecting significant events ranging from emergency situations such as earthquakes to locally popular flash

crowd events such as political demonstrations or simply local events such as concerts or sports events. The fast detection of emergency events such as the March 11 Japan earthquake indicates the possible value of *GeoWatch* in crisis management. In Figure 1, we present a heat map of tweets for a period of approximately 2 months of tweets (March 9 to May 8, 2011). More particularly, we capture the volume of tweets originating from various cities in Figure 1(a) and tweets about cities in Figure 1(b). In these plots, every city associated with more than 10 tweets is marked– color and size is proportional to the number of tweets. Our approach helps identify various characteristics of the social network usage. The two figures resemble each other but there are certain interesting distinctions. It is worthwhile to note that the part of the map corresponding to Japan is denser in Figure 1(b). This is mostly due to the Japan Earthquakes that took place within the time period captured in our data set. This important event spanned a long time period due to the after effects and was an important headliner, making it a trending topic in Twitter. On the contrary, a drop in significance can be observed for countries such as Indonesia when comparing the tweets *in* cities to tweets *about* cities. This big difference originates from the fact that Indonesia is a highly active country in Twitter [22], while there are no important events taking place in its cities that would result in people mentioning them. As part of our proposed work, we aim to provide further insights into the data by making it clear which localities (in terms of geography and time) are similar in behavior or which localities play a critical role in a given topic trending. This way, we not only help users focus on a given localities (or time period) and observe the trends there, but we also allow users to focus on a given set of topics and look at them from the perspective of geo-spatial characteristics.



(a) Tweets *in* Cities (b) Tweets *about* Cities

Fig. 1. Heat Map for # of tweets in/about cities of the world

Problem Definition. Given a stream S of location-topic pairs of the form (l_i, t_j), a window size of N, and three user defined frequency thresholds θ, ϕ, and ψ in the interval [0, 1]; our goal is to keep track of all locations l_i s.t. $F(l_i) > \lceil \theta N \rceil$ alongside their frequencies as well as all topics t_x and their frequencies $F(t_x)$. In addition, in order to detect the correlations, we aim to find all pairs (l_i, t_x) s.t. $F(l_i) > \lceil \theta N \rceil$, $F(l_i, t_x) > \lceil \phi F(l_i) \rceil$, and $F(l_i, t_x) > \lceil \psi F(t_x) \rceil$; where $F(l_i, t_x)$ is the number of information items on topic t_x from location l_i in the most recent N items in S; $F(l_i)$ is the aggregate number of occurrences of all the items from

l_i in the current time window; and $F(t_x)$ is the aggregate number of items on t_x. The window size can be set in terms of maximum number of elements or an actual time window such as an hour or a day. In the latter case, the number of elements N in the current window is variable.

Methodology and Data Structures. We now explore a sketch-based structure for *GeoWatch* to detect correlations between locations and topics. The problem of detecting correlations in multi-dimensional datastream has been studied to detect advertising fraud in clickstream data [44]. However, the solution is counting-based and hence only supports insert operations and cannot deal with information deletion. As can be seen from Figure 2, *GeoWatch* consists of two main components. *Location-StreamSummary-Table* contains a $StreamSummary_{l_i}$ structure for each location l_i that has a current estimated relative-frequency of at least θ. In order to provide a solution in a sliding window where deletions as well as insertions of elements need to be supported, *Location-StreamSummary-Table* also needs to include a sketch structure. This sketch structure is maintained to keep track of frequencies of locations in a sliding window by allowing both insertion and deletion operations [24]. In general *GeoWatch* uses sketches to keep track of the frequencies of tracked elements. The second component is the *Topic-StreamSummary-Table*, a hash table that monitors the topics that are potentially correlated with at least one location and a sketch structure to keep track of the topic frequencies. For each tracked topic this structure also keeps track of the number of locations the topic is trendy for. Once this value reaches 0, the topic is removed from *Topic-StreamSummary-Table*.

Even though the development of *GeoWatch* also captures the notion of temporality through the use of sliding windows, its main focus is the spatial characteristics of data. As part of proposed work, we will investigate analyzing information trends at different temporal granularities such as by the minute, hour, days, and so on and doing so in an efficient manner. Furthermore, we aim to identify topics that suddenly become popular, i.e., a topic that is not necessarily a heavy-hitter in the traditional sense but exhibits a sharp increase in frequency over a short period of time. In order to discover such trends, it is necessary to consider both the frequency and the temporal order of elements in a data stream. While many of the data stream algorithms ignore temporal order, there have been several works that have incorporated some notion of the temporal aspect [28,3,2,34].

Big Data Implications. The whole approach we want to take is purely a way to deal with the Big Data nature of the problem. While an exact solution would be 100% accurate, counting and storing all the pairs makes it impossible. Dealing with an information stream that produces thousands of updates per second and being able to report trending pairs in real time dictates approximate counting, space efficient data structures and sliding windows. Now a good question would be if we can sample the data in the stream and get equally good results. Our initial experiments for the proposed approach show that the quality should be nearly perfect but further studying how reducing the data volume can affect the quality or even the proposed algorithms is a very important direction.

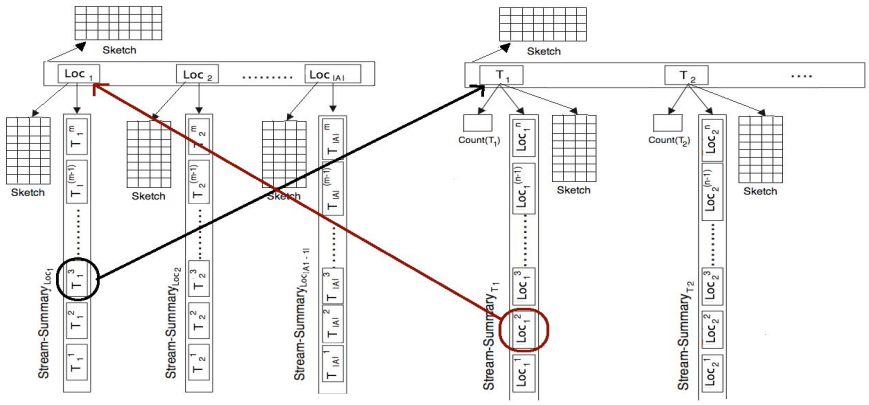


Fig. 2. Overview of Data Structure: The two main sub-components are *Location-StreamSummary-Table* (on the left) and *Topic-StreamSummary-Table* (on the right). *Location-StreamSummary-Table* keeps track of ϕ-frequent topics for each of the θ-frequent locations. *Topic-StreamSummary-Table* keeps track of ψ-frequent locations for each topic that is ϕ-frequent for at least one location. Here the third most important topic for Loc_1 is T_2 and the second most important location for T_2 is Loc_1.

3.2 Multi-dimensional Trend Analysis

A natural extension of the spatio-temporal trend analysis would be to extract trends that focus on multiple dimensions; location and topic being just two of them. There are no limitations on the nature of the dimensions: it can be demographics like age or gender, it can be a location hierarchy, it can be a user's characteristic like opinion, political support or product preference. By analyzing data in a highly dimensional space we can discover trends like "an unusual number of people in the age interval of 18-25, that owns an iPhone, and live in Louisiana mention the topic #CES2013". This information can be extremely valuable to companies, advertisers, political parties and others that need to understand their audience and how they behave, how they are distributed on the map, what topics they are interested in and many other aspects depending on the monitored dimensions. While this problem is the generalization of the spatio-temporal trend analysis described in the previous section, the introduced challenges are not straight forward to solve. Even the exact solution of counting all observed tuples can be very expensive in both computational time and space. And it gets trickier with an online solution where both time and space have to be at most sub-linear.

Big Data Implications. In this particular problem, having Big Data is both beneficial and problematic. On one hand, the curse of having many dimensions suggests that having more data will result in more dense trends but on the other hand, efficiently counting "interesting" frequent tuples (and not all of them as we do in the Database field of frequent item counting), in an online manner, is very challenging.

3.3 Opinion Change

Opinion change consists a two-fold problem: First, the actual opinion or sentiment has to be identified and then, a change must be observed. Opinions change around us all the time and studying the behavior of users and how they make up (or not) their minds can be very useful. As a huge amount of people use social media and express themselves freely, the mining of opinions and how these change should be a rather easy task but depending on the definition of Opinion it can be quite the opposite.

While the general definition of Opinion describes it as a viewpoint or statement about a subjective matter, in many research problems we assume more specific and simpler definitions. For example, sentiment analysis is considered to be a type of opinion mining even if it's only focused on extracting the sentimental score from a given text. So assuming a more simplistic definition of Opinion, we can view people's preferences of political parties or products as opinions. Therefore, there is a wealth of signals to mine from social data like posts on Twitter of Facebook and extract opinions, identify if they change while time passes and analyze what events or other factors contribute in these changes. However, different types of opinions require different types of analysis. For example, Twitter users express their musical preference much more frequently and easily than they do with politics. Also, the reasons why someone might change their opinion on Apple products can be very different and less deep than why they would change they political lean.

Preference in Mobile Devices. The first experiment we conducted while studying the correlation between opinion change and the contribution the social network has on it, focused on Twitter users and their preference on mobile devices. Utilizing the tweet's "source" field that indicates the software client where a tweet was sent from, we were able to build temporal profiles for every user that had a non trivial amount of tweets. These profiles contain the type of the mobile device the user is using (iPhone, iPad, Blackberry, Android, Windows Phone) for every day they tweeted. We assumed this feature as the opinion of a user at any given time, for their mobile preference. Note that the dataset was quite accurate since it reflected the actual device a person was using. Using these timeseries we were able to tell when people switched to a different type of device (e.g. from Android to iPhone).

Having identified the "change" of opinion we then studied if there was a network effect in the process of the decision making. We know when people switched devices and we can also find the account they followed at that time. By joining the social graph with the opinion signal we were able to compute the distribution of mobile devices for the neighbors of every user that changed device. While the hypothesis and generally the literature suggests that people in one's network have an impact on that person, it was not validated in our experiment. There wasn't any statistical significance in the observed results and the fact that we had the complete dataset didn't make a difference. So what we learned from this experience, is that having a lot of data, no matter how Big

or complicated, can't always make up for information that simply is not there. If people trust more what their real life friends say about phones, then just observing their follow graph can be more misleading than beneficial. There has been some recent work [37] on how it is better to focus on specific groups of users when studying behavior rather than the whole population which is very noisy and can also be quite biased. This further underlines the fact that truncating a dataset in a smart and correct way and reducing Big Data to just Data can be sometimes mandatory for specific social applications. We would like to further explore this direction in a framework that can automatically identify such cases.

3.4 User Behavior during Events

The last application we want to study in this context has to do with real life and real time Events. It has become a second nature these days to talk in social media about things that happen in real life. When something happens there is an information rally from users that try to break the news first, write updates and consume content. Common people at the right place and the right moment can give away information on something that happens, before any news agencies. This gives the chance to literally anyone to have their 10 minutes of fame and also highlights the importance of non power-users in social networks. With the recent events at Boston's Marathon on April 15, 2013 (Boston Marathon bombings), we observed a unique situation where people were live reporting from the scene of crime. In the context of analyzing what is happening and shaping information as in a news feed, it is extremely important to be able to capture such cases as soon as possible. Building an application that can report breaking news requires minimizing the reporting latency while maximizing the recall and accuracy of the reported content. It is easy to wait for a story to appear on major news channels but this compromises the latency in about 50% of the times [48]. On the other hand, extracting breaking news from users that are not priorly known to generate such content may lead to unpredictable and questionable quality.

We are proposing the study of a method that can discover in real time, and as soon as possible, the unique users that for a short time span have a very large reporting value (could be even larger than from a news reporting site). This problem belongs in the area of information diffusion and we can view these people as one-time only innovators where their discovery is a time sensitive task. Being able to assign a breaking score to users based on how other users are consuming their content is expected to be an approach to the right direction. One could view this problem as a trend detection problem where instead of topics, as discussed in the previous sections, we have users. A trending user is a user that trends in terms of consumption of their produced content; we count how many times other people are sharing or just consume what they say (e.g. tweets) in a time-window. A time window approach sounds reasonable since as with most trend analysis applications, trendiness is temporal. Users that are interesting to follow during a specific event, are not so likely to generate interesting content for other types of events, therefore we want their trendiness to decade.

Big Data Implications. While we believe that this problem can be viewed as a trend detection problem, it is unclear if it shares the same properties of the other trend analysis tasks we described. Would the same datasets behave equally good? Do we need extra features to get better quality? We believe that this question is very important and worth study.

4 Research Vision: A Scoring Framework for Big Data

As we discussed in the previous sections, Big Data in the context of social behavior analysis can result to largely different benefits or challenges depending on the nature of the studied problem. While each problem is important on its own, all share the same denominator, the fact that Big Data is used and that it's not clear how we should use it. We are proposing the development of a framework that can score a dataset when given a research problem. While we wouldn't be able to score an unknown dataset for a new problem we can score new datasets on well studied problems (like trend analysis) or give a confidence score on how well an existing dataset may work for a new problem. In both cases we need to identify the features and characteristics that make a dataset suitable for a specific application and then we should be able to extrapolate. We also need to take into consideration the actual cost of obtaining a data set. Data can be extremely large while some problem might require a very specific subsets. Or in other cases, equally good quality can be achieved after sampling. Therefore, a cost model must capture the following salient points:

- Cost of acquisition: Can have different values associated with each column (phone number can be more costly then country of residence, user profile can be more costly then a tweet)
- Cost of storage: Traditional methods could suffice. Can we bring cloud into the picture here?
- Cost of processing: Same here, can we bring the cloud?
- SLA: The nature of social problems usually dictate a fast response. Processing large amounts of data can result in long processing time which might be unacceptable.
- Value: While the first 4 parameters can simply be input to the cost model, the value is harder for a user to pinpoint. Especially working in a probabilistic and highly unpredictable space such as computational social science, it is hard to pinpoint the real value of a solution or data. We propose a data-centric approach to identify the value function in a per-application basis. In particular, we learn from a training data the relationship between value and characteristics such as amount, time and location of data.
- Meta-data on data: We believe that the value of the data is not simply embedded in its amount. The behavior of data changes through time and space.

Acknowledgments. This work is partially supported by NSF Grant IIS-1135389 and a gift from the Bill and Melinda Gates Foundation.

References

1. New tweets per second record, and how! https://blog.twitter.com/2013/new-tweets-per-second-record-and-how
2. Aggarwal, C.C., Han, J., Wang, J., Yu, P.S.: A framework for clustering evolving data streams. In: Proc. 29th Int. Conf. on Very Large Data Bases, pp. 81–92. VLDB Endowment (2003)
3. Aggarwal, C.C., Yu, P.S.: Online analysis of community evolution in data streams. In: Proc. SIAM International Data Mining Conference (2005)
4. Allan, J. (ed.): Topic detection and tracking: event-based information organization. Kluwer Academic Publishers, Norwell (2002)
5. Alon, N., Yuster, R., Zwick, U.: Finding and counting given length cycles. Algorithmica 1717, 209–223 (1997)
6. Bar-Yossef, Z., Kumar, R., Sivakumar, D.: Reductions in streaming algorithms, with an application to counting triangles in graphs. In: SODA 2002, pp. 623–632 (2002)
7. Becchetti, L., Boldi, P., Castillo, C., Gionis, A.: Efficient semi-streaming algorithms for local triangle counting in massive graphs. In: KDD 2008, pp. 16–24 (2008)
8. Charikar, M., Chen, K., Farach-Colton, M.: Finding frequent items in data streams. In: Widmayer, P., Triguero, F., Morales, R., Hennessy, M., Eidenbenz, S., Conejo, R. (eds.) ICALP 2002. LNCS, vol. 2380, pp. 693–703. Springer, Heidelberg (2002)
9. Chen, B.: Topic oriented evolution and sentiment analysis. Ph.D. Dissertation, Penn State University (2011)
10. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: KDD 2009, pp. 199–208 (2009)
11. Chierichetti, F., Kleinberg, J., Panconesi, A.: How to schedule a cascade in an arbitrary graph. In: EC 2012, pp. 355–368 (2012)
12. Cormode, G., Hadjieleftheriou, M.: Finding frequent items in data streams. Proc. VLDB Endow. 1(2), 1530–1541 (2008)
13. Cormode, G., Muthukrishnan, S.: What's Hot and What's Not: Tracking Most Frequent Items Dynamically. TODS 2005 30(1), 249–278 (2005)
14. The curse of big data, http://www.analyticbridge.com/profiles/blogs/the-curse-of-big-data
15. Friedkin, N.E.: The attitude-behavior linkage in behavioral cascades. Social Psychology Quarterly, 73–196 (2010)
16. Glorot, X., Bordes, A., Bengio, Y.: Domain adaptation for large-scale sentiment classification: A deep learning approach. In: ICML 2011 (2011)
17. Hartline, J., Mirrokni, V., Sundararajan, M.: Optimal marketing strategies over social networks. In: WWW 2008, pp. 189–198 (2008)
18. Hong, L., Ahmed, A., Gurumurthy, S., Smola, A.J., Tsioutsiouliklis, K.: Discovering geographical topics in the twitter stream. In: WWW 2012, pp. 769–778 (2012)
19. Horrigan, J., Rainie, L.: When facing a tough decision, 60 million americans now seek the internet's help: The internet's growing role in life's major moments (2006), http://pewresearch.org/obdeck/?ObDeckID=19 (retrieved October 13, 2006)
20. Howe, J.: The rise of crowdsourcing. North 14(14), 1–5 (2006)
21. Hughes, A.L., Palen, L.: Twitter adoption and use in mass convergence and emergency events. International Journal of Emergency Management 6(3/4), 248 (2009)
22. Indonesia, brazil and venezuela lead global surge in twitter usage, http://www.comscore.com/Press_Events/Press_Releases/2010/8/Indonesia_Brazil_and_Venezuela_Lead_Global_Surge_in_Twitter_Usage

23. Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T.: Target-dependent twitter sentiment classification. In: HLT 2011, pp. 151–160 (2011)
24. Jin, C., Qian, W., Sha, C., Yu, J.X., Zhou, A.: Dynamically maintaining frequent items over a data stream. In: CIKM 2003, pp. 287–294. ACM (2003)
25. Katz, I., Tunstrom, K., Ioannou, C., Huepe, C., Couzin, I.: Inferring the structure and dynamics of interactions in schooling fish. In: PNAS 2011, pp. 18720–18725 (2011)
26. Kempe, D., Kleinber, J., Tardos, E.: Maximizing the spread of influence through a social network. In: KDD 2003, pp. 137–146 (2003)
27. Kempe, D., Kleinberg, J.M., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM International Conference on Knowledge Discovery and Data Mining, pp. 137–146 (2003)
28. Kifer, D., Ben-David, S., Gehrke, J.: Detecting change in data streams. In: Proc. 30th Int. Conf. on Very Large Data Bases, pp. 180–191. VLDB Endowment (2004)
29. Kimura, M., Saito, K., Nakano, R., Motoda, H.: Extracting influential nodes on a social network for information diffusion. Data Mining and Knowledge Discovery 20, 70–97 (2010)
30. Kimura, M., Saito, K., Ohara, K., Motoda, H.: Learning to predict opinion share in social networks. In: AAAI 2010, pp. 1364–1370 (2010)
31. Kittur, A., Kraut, R.E.: Harnessing the wisdom of crowds in wikipedia: quality through coordination. In: Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work, CSCW 2008, pp. 37–46. ACM, New York (2008)
32. Krishnamurthy, B., Gill, P., Arlitt, M.: A few chirps about twitter. In: Proceedings of the First Workshop on Online Social Networks, WOSN 2008, pp. 19–24. ACM (2008)
33. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media. In: WWW 2010, pp. 591–600 (2010)
34. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proc. 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 497–506 (2009)
35. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: KDD 2009, pp. 497–506 (2009)
36. Libert, B., Spector, J.: We are smarter than me: how to unleash the power of crowds in your business, 1st edn. Wharton School Publishing (2007)
37. Lin, Y.R., Margolin, D., Keegan, B., Lazer, D.: Voices of Victory: A Computational Focus Group Framework for Tracking Opinion Shift in Real Time. In: WWW 2013, pp. 737–747 (2013)
38. MacEachren, A.M., Robinson, A.C., Jaiswal, A., Pezanov, S., Savelyev, A., Blanford, J., Mitra, P.: Geo-Twitter analytics: Application in crisis management. In: 25th International Cartographic Conference (July 2011)
39. Macropol, K., Singh, A.K.: Content-based modeling and prediction of information dissemination. In: ASONAM 2011, pp. 21–28 (2011)
40. Manku, G.S., Motwani, R.: Approximate frequency counts over data streams. In: VLDB 2002, pp. 346–357 (2002)
41. Mehta, R., Mehta, D., Chheda, D., Shah, C., Chawan, P.: Sentiment analysis and influence tracking using twitter. International Journal of Advanced Research in Computer Science and Electronics Engineering 1, 72–79 (2012)
42. Melville, W.G.P., Lawrence, R.D.: Sentiment analysis of blogs by combining lexical knowledge with text classification. In: KDD 2009, pp. 1275–1284 (2009)

43. Metwally, A., Agrawal, D., El Abbadi, A.: An integrated efficient solution for computing frequent and top-k elements in data streams. ACM Trans. Database Syst. 31(3), 1095–1133 (2006)
44. Metwally, A., Emekçi, F., Agrawal, D., El Abbadi, A.: Sleuth: Single-publisher attack detection using correlation hunting. Proc. VLDB Endow. 1(2), 1217–1228 (2008)
45. Mudhakar, S., Srivatsa, L., Abdelzaher, T.: Mining diverse opinions. In: MILCOM 2012, pp. 1–7 (2012)
46. Palen, L.: Online social media in crisis events. Educause Quarterly (3), 76–78 (2008)
47. Patterson, S., Bamieh, B.: Interaction-driven opinion dynamics in online social networks. In: SOMA 2010, pp. 98–105 (2010)
48. Petrovic, S., Osborne, M., McCreadie, R., Macdonald, C., Ounis, I., Shrimpton, L.: Can Twitter replace Newswire for breaking news? In: ICWSM 2013, pp. 713–716 (2013)
49. Rosenfeld, A., Hummel, R.A., Zucker, S.W.: Scene labeling by relaxation operations. IEEE Transactions on Systems Man and Cybernetics 6, 420–433 (1976)
50. Sachan, M., Contractor, D., Faruquie, T.A., Subramaniam, L.V.: Using content and interactions for discovering communities in social networks. In: WWW 2012, pp. 331–340 (2012)
51. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, pp. 851–860. ACM, New York (2010)
52. Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., Sperling, J.: Twitterstand: news in tweets. In: GIS 2009: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 42–51. ACM, New York (2009)
53. Teitler, B.E., Lieberman, M.D., Panozzo, D., Sankaranarayanan, J., Samet, H., Sperling, J.: Newsstand: a new view on news. In: GIS 2008, pp. 1–10 (2008)
54. Teitler, B.E., Lieberman, M.D., Panozzo, D., Sankaranarayanan, J., Samet, H., Sperling, J.: Newsstand: a new view on news. In: GIS 2008: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 1–10. ACM, New York (2008)
55. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology 61, 2544–2558 (2010)
56. Twitter, `http://www.twitter.com`
57. Wang, X., Wei, F., Liu, X., Zhou, M., Zhang, M.: Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In: CIKM 2011, pp. 1031–1040 (2011)
58. Wu, M.: The big data fallacy and why we need to collect even bigger data. TechCrunch (2012)