On Diversifying Source Selection in Social Sensing

Md Yusuf S Uddin, Md Tanvir Al Amin, Hieu Le, Tarek Abdelzaher
Department of Computer Science
University of Illinois at Urbana Champaign
Urbana, IL 61801
{mduddin2, maamin2, hieu2, zaher}@illinois.edu

Boleslaw Szymanski, Tommy Nguyen
Department of Computer Science
Rensselaer Polytechnic Institute
Troy NY 12180
{szymab, nguyet11}@rpi.edu

***Abstract*—This paper develops algorithms for improved source selection in social sensing applications that exploit social networks (such as Twitter, Flickr, or other mass dissemination networks) for reporting. The collection point in these applications would simply be authorized to view relevant information from participating clients (either by explicit client-side action or by default such as on Twitter). Social networks, therefore, create unprecedented opportunities for the development of sensing applications, where humans act as sensors or sensor operators, simply by posting their observations or measurements on the shared medium. Resulting social sensing applications, for example, can report traffic speed based on GPS data shared by drivers, or determine damage in the aftermath of a natural disaster based on eye-witness reports. A key problem, when dealing with human sources on social media, is the difficulty in ensuring independence of measurements, making it harder to distinguish fact from rumor. This is because observations posted by one source are available to its neighbors in the social network, who may, in-turn, propagate those observations without verifying their correctness, thus creating correlations and bias. A corner-stone of successful social sensing is therefore to ensure an *unbiased sampling* of sources that minimizes dependence between them. This paper explores the merits of such diversification. It shows that a diversified sampling is advantageous not only in terms of reducing the number of samples but also in improving our ability to correctly estimate the accuracy of data in social sensing.**

I. Introduction

This paper investigates algorithms for diversifying source selection in social sensing applications. We interpret social sensing broadly to mean the set of applications, where humans act as the sensors or sensor operators. An example application might be a participatory sensing campaign to report locations of offensive graffiti on campus walls, or to identify parking lots that become free of charge after 5pm. Another example might be a damage assessment effort in the aftermath of a natural or man-made disaster, where a group of volunteers (or survivors) survey the damaged area and report problems they see that are in need of attention. Social sensing benefits from the fact that *humans* are the most versatile sensor. This genre of sensing is popularized by the ubiquity of network connectivity offered by cell-phones, and the growing means of information dissemination, thanks to Twitter, Flickr, Facebook, and other social networks.

Compared to applications that exploit well-placed physical sensors, social sensing is prone to a new type of inaccuracy; namely, unknown dependence between sources, which affects data credibility assessment. This dependence arises from the fact that information shared by some sources (say via a social network such as Twitter) can be broadly seen by others, who may in turn report the same information later. Hence, it becomes harder to tell whether information received is independently observed and validated by the source or not. When individual data items are inherently unreliable, one would like to use the degree of corroboration (i.e., how many sources report the same data) as an indication of trustworthiness. For example, one would like to believe an event reported by 100 individuals more than an event reported by a single source. However, if those individuals are simply relaying what they heard from others, then the actual degree of corroboration cannot be readily computed, and sensing becomes prone to rumors and mis-information.

Our paper investigates the effect of diversifying the sources of information on the resulting credibility assessment. We use Twitter as our social network, and collect tweets representing events reported during Egypt unrest (demonstrations in February 2011 that led the resignation of the Egyptian president) and hurricane Irene (one of the few hurricanes that made landfall near New York City in 2011). For credibility assessment, we use a tool developed earlier by the authors that computes a maximum-likelihood estimate of correctness of each tweet based on its degree of corroboration and other factors [1]. In our dataset, some of the tweets relay events that are independently observed by their sources. Others are simply relayed tweets. Note that, while Twitter offers an automatic relay function called "re-tweet", there is nothing to force individuals to use it when repeating information they heard from others. It is perfectly possible to originate tweets with similar content to ones received without using the re-tweet function. In this case, information is lost on whether content is independent or not.

While it is generally impossible to tell whether or not content of two similar tweets was independently observed, our premise is that by analyzing the social network of sources, we can identify those that are "close" and those that are "not close". By using more diversified sources, we can increase the odds that the chosen sources offer independent observations, and thus lower our susceptibility to rumors and bad information.

The paper explores several simple *distance metrics* between sources, derived from their social network. Distance may depend on factors such as whether one source is directly

connected to another (e.g., one *follows* the other in Twitter lingo), whether both are connected to a common ancestor (e.g., both follow a common source), or whether both are followed by the same people. By choosing the most dis-similar sources, according to these metrics, we show that we can indeed suppress more rumors and chain-tweets. The impact of different distance metrics on improving credibility assessment of reported social sensing data is compared.

The rest of this paper is organized as follows. Section II describes earlier work done in field of source selection and fact-finding. Section III formulates our source selection problem and proposes a set of source selection schemes that diversify the sources admitted for purposes of data collection. Evaluation results demonstrating the effect of source selection on credibility assessment of collected data are presented in Section V followed by conclusions in Section VI.

II. Related Work

Social sensing has received much attention in recent years [2]. This is due to the large proliferation of devices with sensing and communication capabilities in the possession of average individuals, as well as the availability of ubiquitous and real-time data sharing opportunities via mobile phones with network connection and via social networking sites (i.e., Twitter). A few early applications include CarTel [3], a vehicular data collection and sharing system, BikeNet [4], an application allowing bikers to share their biking experiences on different trails, PhotoNet [5], a data collection service for pictures from disaster scenes, CenWits [6], a search and rescue scheme for hikers, CabSense [7], a participatory sensing application using taxi car fleets, and ImageScape [8], an application for sharing diet experiences.

Social sensing involves humans as sensors or sensor operators. One consequent problem lies in the decreased quality of collected data, since humans are not as reliable as well-calibrated sensors. A significant amount of literature therefore deals with extracting useful information from a vast pool of unreliable data. Prior to the emergence of social sensing, much of that work was done in machine learning and data mining. For example, following techniques inspired by generalizations of Google's PageRank [9], techniques were proposed that represent information by a source-claim network [10], [11], [12] that tells who said what. An iterative algorithm then tries to reason on this graph to extract the most trustworthy information given the degree of corroboration and inferred source reliability. Generally these techniques are called *fact-finders*, a class of iterative algorithms that jointly infer credibility of claims as well as trustworthiness of sources. Hubs and Authorities [10] is a simple fact-finder, where belief in correctness of a claim is computed as the sum of trustworthiness of sources who made that claim, and the trustworthiness of a source is, in turn, obtained from the beliefs in correctness of the claims it makes. Notable fact-finding schemes include TruthFinder [13], 3-Estimates [14], and AccuVote [15], [16].

Several extensions were developed to improve fact-finding results, such as incorporating prior knowledge [17], [18], and accounting for the source's expertise in different topics [19]. Most recently, a maximum likelihood estimation approach was developed that is the first to compute an *optimal solution* to the credibility assessment problem [1]. The solution is optimal in the sense that the resulting assignment of correctness values to claims and sources is the one of maximum likelihood. A confidence interval was also computed to describe the quality of the maximum-likelihood hypothesis [20]. Our paper is related but orthogonal to fact-finding literature. We attempt to improve quality of fact-finding results not by improving the fact-finding algorithm itself, but by improving its input through increasing the odds of independence between the selected sources.

The problem of information source selection has been discussed in web data retrieval [21], [22], [23] and in query sampling [24], [25], [26]. These efforts reason on the attributes of sources as well as the content that those sources generate. In contrast, ours is a content-agnostic approach that relies only on relationships among sources.

In this paper, we use Apollo [27], a generic fact-finding framework that can incorporate different suitable fact-finding algorithms as plug-ins for a versatile set of applications. We use the aforementioned maximum-likelihood estimator [1] as the fact-finding algorithm in Apollo. We demonstrate that the performance of fact-finding can be significantly improved by using simple heuristics for diversifying sources so that it uses sources that are less dependent on one another.

While our work on diversifying sources would not be needed if one could accurately account for dependence between them in data credibility assessment, we argue that, in general, estimating the degree of dependence between sources is very hard. For example, if one source follows another on Twitter and both report the same observation, it is hard to tell whether the second report is simply a relay of the first, or is an independent measurement. Given the ambiguity regarding the originality (versus dependence) of observations, we suggest that diversifying the sources is a useful technique whether or not credibility assessment can take dependence into account.

We implemented our source selection scheme as an online admission controller that is included as an upfront plug-in to the Apollo execution pipeline. Results show that our admission control can both speed up data processing (by reducing the amount of data to be processed) and improve credibility estimates (by removing dependent and correlated sources).

III. Source Selection in Social Sensing

Data in social sensing applications that exploit social networks (e.g., Twitter) can be polluted by users who report events that are not experienced or verified by themselves. This is because individuals are able to reproduce claims that they heard from others. We argue that if information can be collected from a diverse set of sources who have a weak "social" connection between them, there is a higher chance that the information collected thereby would be more independent, allowing a more informed judgment to be made regarding its reliability. In the following, we use the terms

users, sources and nodes as well as the terms tweets, feeds, claims and observations interchangeably.

A. Online User Social Graph and Source Dependence

In an online community platform or online social network, each user maintains a virtual relationship with a set of other users. This relationship entails some degree of information sharing. For example, on YouTube, a user may subscribe for videos posted by another user so that the former gets a notification when the later uploads a new video. In Facebook, there is an explicit friend relationship and a membership of a fan-page of another well-known user. Google$^+$ has more granularity like friends, family members, acquaintances, and other groups, called circles. In this paper, we consider a Twitter-based social sensing application, which allows a *follower-followee* relation. A user following another user means that the former intends to receive the posts made by the latter. We say that if user i follows user j, i is the follower and j is the followee. In Twitter, a user can arbitrarily choose which other users to follow, although the converse is not true. That is, a person can not make another user follow them (a person can, however, block another user from following).

We leverage this relationship in Twitter to form a *social graph* among users. We represent each user by a vertex in the graph. A directed edge from one vertex to another denotes that the latter follows the former. We use the notation $i \rightarrow j$ to denote an edge in the graph (meaning that user i follows user j). Sometimes, a user may not directly follow another, but can follow transitively via a set of intermediate followees. We refer to this as a follow chain. We use $i \rightarrow^k j$ to denote such a chain with k edges in between. Obviously, $i \rightarrow j = i \rightarrow^1 j$. If i follows j via more than one path, $i \rightarrow^k j$ denotes the one with the least number of hops. We also use $F(i)$ to denote the set of users that a node i follows, that is, the set of followees of node i.

It is reasonable to argue that if source i directly follows source j, reports posted by j would be visible to i, making the information posted by i potentially not original. Another possibility could be that both source i and j have another source in common that both of them follow (i.e., they have a common followee). In that case, the common followee may impact both of them, making their observations mutually dependent. In order to extract reliable information from user-generated tweets, our intention is to gather tweets from *independent* sources to maximize the odds of originality of the information (or equivalently minimize the chance that these users influenced one another). The question is how to reduce potential dependence among users as a given the follower-followee relationships between them. In the following, we formulate this source selection problem.

B. Source Selection Problem Formulation

We construct a *dependence graph* consisting of sources as vertices and directed edges between vertices as an indication whether or not a source is potentially dependent on another source (e.g., receives their tweets). Weights assigned to edges reflect the degree to which such influence can happen. These weights depend on the characteristics of the social network and the underlying relationship among sources in the social graph. In the context of Twitter, we simply use the follow relationship between sources. If we consider the follow relationship to be the only way sources could be dependent, the proposed dependence graph is identical to the Twitter social graph itself. In general, it is reasonable to assume that other forms of dependence may also exist.

Let $\mathcal{G} = (V, E)$ be the dependence graph, where an edge ij indicates source i is potentially dependent on j. Each edge ij is assigned a *dependence score*, f_{ij}, that estimates the probability of such dependence. That is, with probability f_{ij}, source i could make the same or similar claims as source j. Many factors affect these dependence scores. For example, when a source directly follows another source, it is more dependent on its followee than a source that follows the same followee via a longer follow chain. The number of common followees between a pair sources can also be an indication of dependence between them. If a given pair of nodes have a large number of common followees, they are prone to be more dependent than a pair that have fewer common followees or no followees at all. Whatever the cause of dependence between sources is—that we describe in the subsequent subsection in more detail—we aim to choose a subset of sources that have the least amount of dependence among them.

In the rest of the paper, we re-draw the dependence graph, $\mathcal{G}$, as a complete graph with transitive dependencies collapsed into a single edge. Hence, f_{ij} exists for every pair of sources i and j (f_{ij}, and is zero only if no influence exists between them. We are interested in estimating the probability that a source makes an *independent* claim, when its claims can be potentially influenced by those made by others. We define an overall *independence score* for each source that gives the probability that it is *not* influenced by other sources in making a claim. This score, denoted by $\beta(i)$ for source i, can be approximated as:

$$\begin{aligned} \beta(i) &= P[i \text{ is independent in making claims}] \\ &= \prod_{j=1}^{n} P[i \text{ is not dependent on } j] \\ &= \prod_{j=1}^{n} (1 - f_{ij}) \end{aligned} \tag{1}$$

One important property of the independence score (that we shall henceforth refer to as the β-score) is that a source cannot have this score in isolation. It is rather a functional form of dependence on other sources. From the definition, we observe that $\beta(i) = 1$ means that source i is absolutely independent (not dependent on any other sources in consideration). We also notice that the β-score declines for a source if the source is influenced by more other sources. To diversify the collection of sources, we consider only a subset of sources whose sum of independence scores is maximum subject to the constraint that no individual source has an independence score below a certain threshold. Let this threshold be τ. That is, we want to

compute the subset of selected sources $S \subseteq V$ that maximizes the sum of β-scores. Therefore, we have:

$$\max \quad \sum_{i \in S} \prod_{j \in S} (1 - f_{ij}) \tag{2}$$

$$\text{s.t.} \quad \prod_{j \in S} (1 - f_{ij}) \geq \tau, \forall i \in S \tag{3}$$

Note that, individual sources can also have an *influence* factor associated with them that can be inferred from the number of followers. If a source has many followers, it may mean that this source produces observations that other users find reliable. This is a source ranking problem and has been addressed in prior work. In this paper, we do not address source ranking. Instead, we verify the promise that *diversifying* the sources can improve the performance of a subsequent ranking algorithm.

The optimization problem stated by Equation (2) can be shown to be an IP (Integer Programming) problem, and is therefore NP-Hard. We can use a greedy approximation by building the solution incrementally. The greedy algorithm assumes that all candidate sources are available apriori so that the source selection can pick a subset of them. Sometimes the set of sources is not known beforehand. Rather, new sources are discovered as they arrive incrementally. In that case, an *online* algorithm seems more appropriate.

In this paper, we consider a system where a stream of tweets arrives at a processing station. Our source selection scheme acts as an admission controller that needs to make an online assessment regarding whether or not a new source is to be selected based on the relationships it has with respect to other sources selected earlier. If the source is selected, all tweets that originate from that source are admitted, and will be passed to the actual processing engine as they arrive. Otherwise, the source is not admitted and all tweets from that source will be dropped on arrival. Hence, our online admission controller is a simple gate that admits tweets based on which source they are coming from. An advantage of admission control as described above is that it is fast and easy. In particular, it is based on sources and not on the content of tweets. In principle, better admission controllers can consider content as well, but they will be significantly slower. Hence, in this paper, we restrict our notion of data sampling to the granularity of entire sources, making it a source selection scheme. In the following, we compare performance of different source selection schemes.

IV. Online Admission Control

The online admission controller makes a decision regarding each tweet upon its arrival to the system. If the source associated with the tweet is already admitted, the tweet is passed to the next step. If not, the candidacy of the source is evaluated in terms of how independent this source is with respect to the earlier admitted sources. The admission controller computes the β-score of the incoming source and then accepts it only if its β-score remains above an admission threshold, τ. Otherwise, it is denied. Let S be the set of sources that have been admitted so far. The source denial rule, as per Equation (3), is:

$$\textit{Denial rule for source } i\textit{:} \quad \prod_{j \in S} (1 - f_{ij}) < \tau \tag{4}$$

For a certain definition of f_{ij} and the associated admission threshold, τ, we can formulate a set of different admission controllers as we describe in the following. In all admission control schemes, if not otherwise stated, admission decisions are final: once admitted, a source is not revoked from the admitted set. In the following discussion, let i be the source who is seeking admission.

1. No direct follower:

$$f_{ij} = \begin{cases} 1 & \text{if } i \text{ follows } j \\ 0 & \text{otherwise} \end{cases}$$

$$\tau = 1$$

Deny, if the source is a direct follower of another admitted source. Recall that if source i follows any of the earlier admitted sources in S, that is, for some $j \in S$, $f_{ij} = 1$, it leads to $\beta(i) = 0$, thus violating the admission condition.

2. No direct follower as well as no common followee:

$$f_{ij} = \begin{cases} 1 & \text{if } i \to j \vee F(i) \cap F(j) \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

$$\tau = 1$$

Deny, if the source directly follows someone in the set or has at least one followee in common with another admitted source.

3. No descendants:

$$f_{ij} = \begin{cases} p^k & \text{if } i \to^k j, 0 < p < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\tau = 1$$

Deny, if the source is a follower of another admitted source possibly via a set of intermediate followees.

4. β-controller: This controller selects sources that progressively improve the sum of β-scores as per Equation (2), while satisfying the constraint (3) for each individual admitted source. This controller considers transitive *follower-followee* relationships among sources and defines the following dependence function:

$$f_{ij} = \begin{cases} p^k & \text{if } i \to^k j \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

for some constant $p < 1$. We used, $p = \frac{1}{2}$.

Let $B(S)$ be the sum of β-scores of admitted sources, i.e, $B(S) = \sum_{j \in S} \beta(j)$. Let i be the new source. The scheme computes:

$$\beta'(i) = \prod_{j \in S \cup \{i\}} (1 - f_{ij}), \forall i \in S \cup \{i\} \tag{6}$$

$$B(S) = \sum_{j \in S} \beta(j) \tag{7}$$

$$B'(S) = \sum_{j \in S \cup \{i\}} \beta'(j) \tag{8}$$

The scheme then admits i only if $\beta'(i) \geq \tau$ and $B'(S) > B(S)$. Note that, when a new source is admitted, the scores of some earlier admitted sources may decrease (this is because they may be followers of this newly admitted source). Upon admittance of the new source, those scores are updated. Among possible choices, we consider two versions of β-controllers, with $\tau = 0, 1$. The one with $\tau = 0$ does not check individual β-scores but admits sources as long as they improve $B(S)$, whereas $\tau = 1$ denies a new source if it has any link with any of the earlier admitted sources (i.e., $\beta < 1$) and also fails to improve $B(S)$.

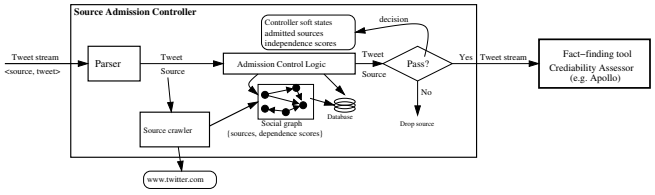


Fig. 1. Schematic model of the admission controller with Apollo's pipeline.

A. Complexity of Admission Controllers

Once accepted, a source is not rejected later, and vice versa. So the decision about a particular source can be stored in a hash table. Once a source arrives, whether that source had already been explored or not, can be checked in $O(1)$ time and the stored decision can be used. If the incoming node is previously unexplored, the admission controller needs to decide about it. For the first three controllers, this decision requires $O(out(i))$ computations, where $out(i)$ is the outdegree of i in the dependence graph. The method is simply to check whether any of those outdegree vertices belong to the set of already decided sources. β-controllers consider ingoing edges also, so they take $O(out(i) + in(i))$ computation steps per admission decision. In short, the admission cost of a new source is at worst in the order of its degree in the dependency graph. But it is O(1) lookup for all the tweets that come from it thereafter. Moreover, social graphs tend to have a power law degree distribution, so very few nodes will require a high computation time for the decision.

B. System Design and Implementation

Our admission controller is used in association with a fact-finding tool called Apollo [27]. It receives a stream of tweets from which it derives credibility scores of sources and claims (i.e., tweets) using an expectation-maximization (EM) technique [1]. Once the iterations converge, Apollo outputs the top credible sources and top credible tweets made by those sources.

Apollo assumes that all sources are independent. Our admission controller filters out tweets before they are fed into the Apollo engine such that the surviving ones are more likely to be independent indeed. Figure 1 shows the design of the whole pipeline.

The pipeline is implemented as a set of stages processing a stream of tweets in JSON format. A parser extracts various information components from each tweet entry. There are two main components to extract: user information, usually a unique Id and screen name of the source who tweeted the current tweet, and the tweet string itself. The admission controller maintains a source information base that is updated as it encounters new sources. Upon encountering a new user, the "source crawler" contacts to the twitter server and collects the Twitter record of that particular user, which includes additional information such as the user's screen name, location, profile url, the number of followers and the number and identities of followees this user has. If not otherwise restricted by any privacy setting for this user, the crawler also collects the complete list of followees (i.e., the other users that this user follows in twitter's user space). As more and more sources are encountered, a social graph among users is constructed. This social graph is stored in a database and is an essential element for source admission control.

An admission controller logic unit implements the admission control rules described in Section IV. It computes dependence scores between pairs of sources and admits new sources as permitted by the corresponding admission rules. When an incoming source is admitted, the associated tweet entry is passed to the next processing stage within Apollo.

V. Evaluation

We evaluated our source selection schemes using two twitter datasets. One is for Egypt unrest, collected in February 2011, during a massive public uprising in Cairo. Another dataset is from hurricane Irene, one of the costliest hurricanes on record in the Northeastern United States, collected in August 2011, when it made landfall near New York City. In both cases, we collected hundreds of thousands of tweets posted by users as the events unfolded during those times. The datasets are summarized in Table I. We were interested in extracting a smaller subset of high quality reports on the progress of these events as computed by the find-finder engine, Apollo. The question is whether a significant improvement occurs in distilling the most important tweets due to the source diversification process described earlier in this paper.

TABLE I
Statistics of two datasets

Dataset	Egypt unrest	Hurricane Irene
Time duration	18 days	$\approx$ 7 days
# of tweets	1,873,613	387,827
# of users crawled	5,285,160	2,510,316
# of users actually twitted	305,240	261,482
# of follower-followee links	10,490,098	3,902,713

In Twitter, both the number of followers and followees per user observe a power law distribution (i.e., heavy tail distribution). More precisely, there exists a very large number of users who have only a few followers, whereas a few sources may have an extremely large number of followers. The same is true for the number of followees. Figure 2(a) plots the complementary cumulative distribution (CCDF) of the number of followers and followees per source across all users recorded

in the Egypt dataset and Irene dataset. The CCDF depicts what fraction of users have the number of followers or followees greater than the corresponding value on the x-axis.

In Figure 2(a), we observe that the number of followers per user, in both datasets, is larger than the number of followees per user. Hence, the followee curve in the plot lies beneath the follower curve. Clearly, when the entire social network is considered, the totals will be the same. However, in our data collection, we see only those who tweet. Hence, we invariably sample the subset of more active users, creating the imbalance between follower and followee counts. We plot the ratio of follower count to followee count (*ff-ratio*) in Figure 2(b). We see that in both datasets only a very small fraction of users have non-zero follower and followee count (1.7% for Egypt dataset and 2.4% for Irene dataset). More than half of these have more followers than followees (ff-ratio > 1). Very few users have an order of magnitude more followers than followees. These are mostly popular entities, such as celebrities, international organizations, and news media.

The goal of the evaluation was to answer two related questions: First, what is the impact of source diversification on data credibility assessment when the social network is well-connected? Second, what is the impact if the social network is very sparse? Since both of our datasets were sparse, to answer the first question, we artificially removed from one of the datasets (namely, the Egypt dataset) all users who did not have any links (together with their tweets). Tweets from the remaining sources were considered. The Irene dataset was kept as is, and used to answer the second question (i.e., demonstrate the impact of our admission controllers in the case when the underlying social network is sparse). Conceptually, our admission controllers, by their very design, exploit links between sources for diversification. Hence, in the absence of many links, their effect should not be pronounced.

Next, we present results from various admission controllers that we described in Section IV. We compare no admission control to several admission control schemes; namely, *no follower* (No FLWR), *no common followee* (No CF) and *no descendant* (No DT), and β*-controller* (Beta). We evaluate the improvement, attained by these admission controllers, in Apollo's ability to rank tweets. Performance was assessed by the fraction of top-ranked tweets that were "good" in that they reported "relevant and true facts". To identify relevant and true facts, we asked volunteers to grade the top-ranked tweets by placing them in one of the following two categories:

- *Fact*: A claim that describes a physical event that is generally observable by many individuals *independently* and can be corroborated by sources external to the experiment (e.g., news media).
- *Other:* An expression of one's personal feeling, experiences, or sentiments. Remarks that cannot be corroborated. Unrelated random text and less meaningful tweets.

Apollo was run with each of the admission control options on consecutive windows of data, called *epochs*, and used to return the top 5 tweets from each epoch. For the Egypt dataset, we divided the timeline into 18 epochs, and collected the top 5 tweets from each, resulting in a total of 90 tweets graded per experiment (i.e., per admission control option). For the Irene dataset, we choose 150 tweets (top 5 tweets from each of 30 epochs). We built a web interface, where volunteers could grade these tweets without revealing which ones were selected in which experiment (i.e., with which admission controller). Once tweets were graded, a *quality score* for each experiment was computed denoting the fraction of tweets that have been identified as *fact*. If more than one volunteer graded the same results and differed in classifying a tweet, we used the average score.

Figure 3 presents the *relative* quality scores of various admission control schemes with respect to the "no admission control" scheme. We present results with two Apollo options, i) with retweets and ii) without retweets. The former option has no effect on the dataset. The latter option discards all tweets that are explicitly tagged by their sources as "retweets" (i.e., a repeat of tweets posted earlier). This discarding is in addition to tweets already dropped by admission control. We observe that, in both datasets, experiments with no-retweet option produce higher quality scores. This is because they eliminate "chain-tweeting", where users relay sentiments and opinions of others. In the absence of such re-tweets, highly corroborated tweets (that percolate to the top) more often reflect situations that independently prompted the respective individuals to report. Such a synchronized reaction typically reflects a higher importance of the reported situation.

In our plots, "Beta 1.0" stands for β-controller with threshold, $\tau = 1.0$. We observe that in general β-controllers result in better quality scores. This observation supports our hypothesis that diversifying sources does indeed improve the quality of information distillation. In contrast, the performance of the other admission controllers is mixed. For the Egypt dataset, simple admission heuristics such as 'no follower', 'no common followee' and 'no descendant' generally offer slightly lower quality scores compared to no admission control. For the Irene dataset, they produce lower scores when retweets are included but higher scores in the no-retweets case.

Note also that, since the Irene dataset has limited connectivity, β-controllers have a more limited impact. They performs similarly to the no admission control case for the with-retweets option, and slightly better for the no-retweets option. This is expected, since sparse social networks offer little opportunities for further diversification.

Figure 4 and Figure 5 show the percentage of sources and tweets that each admission controller admits for the two datasets. It is apparent that some admission schemes are more pessimistic in the sense that they admit fewer sources (and tweets thereby) than others. For the Egypt dataset, on an average, 15–20% tweets are pruned by the admission controllers. For the Irene dataset, however, admission rates across various admission controllers are much higher because of the disconnected nature of the underlying social network.

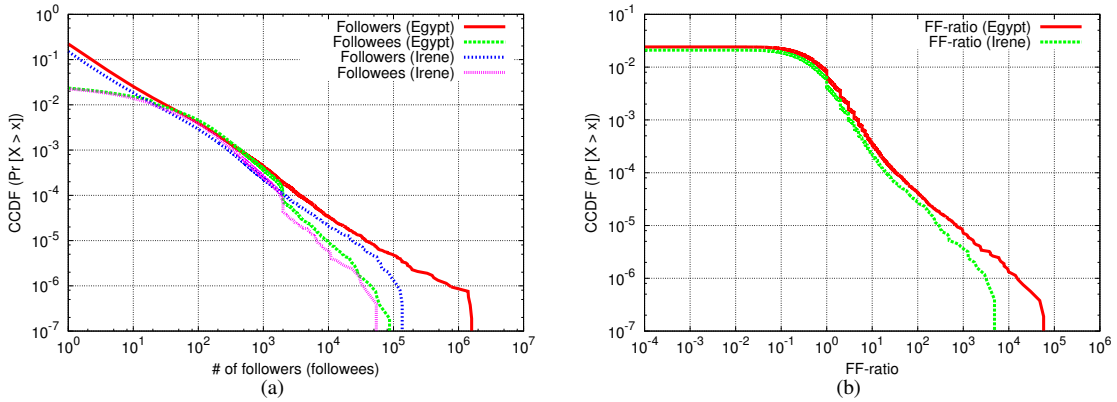


Fig. 2. (a) Complementary distribution (CCDF) of follower and followee count per user, (b) CCDF of ff-ratio per user, in Egypt dataset.

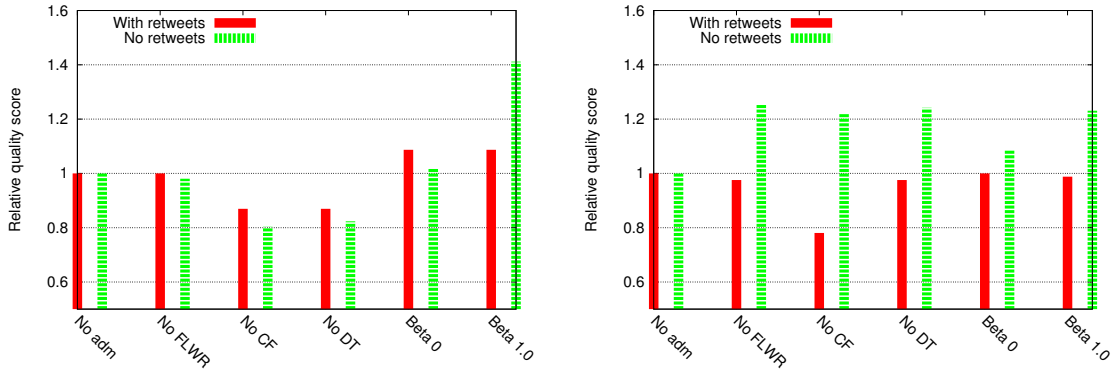


Fig. 3. Relative quality scores across different admission control schemes.

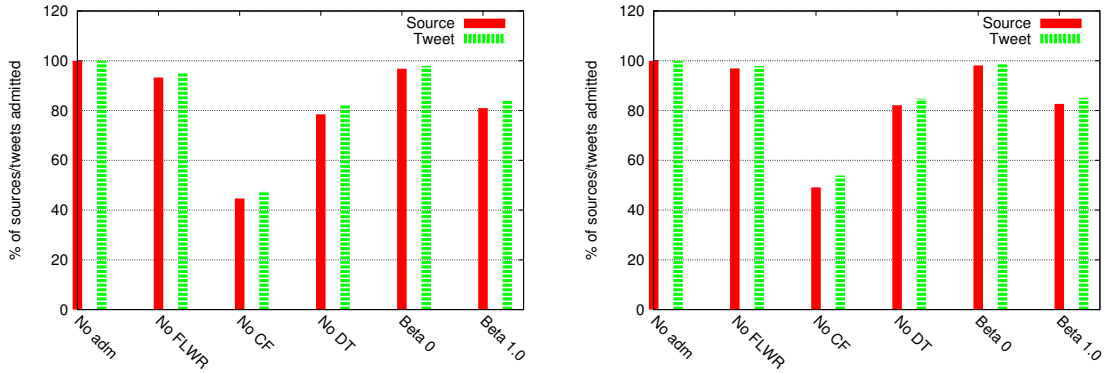
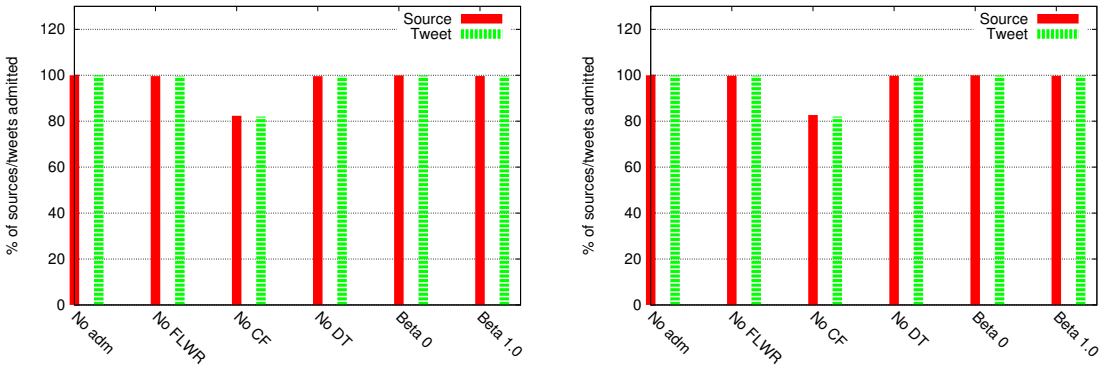


Fig. 4. Admission controller statistics for different admission schemes (Egypt dataset).

VI. Conclusion

In this paper, we considered a fact extraction problem from a large collection of user-generated tweets during two recent events, namely the Egypt unrest and hurricane Irene. We demonstrated that diversifying the sources can improve the results of extracting high quality information (i.e., facts or credible claims) from human-generated content. Human sources on social networks may describe events that do not constitute their own independent observations. This lack of independent corroboration may affect the accuracy of extracting information. We built different online admission controllers that filter tweets based on their sources and feed them into the fact-finding engine, Apollo. We observed that those admission controllers that used local social graph features such as the direct neighborhood of the source in question had inconsistent performance, whereas admission controllers that used more global features tended to perform better. In the current implementation, as a proof-of-concept, we leveraged the "follow" relationship between online users in twitter as an indication of dependence between them. Other attributes that might potentially make sources dependent, such as geographic locations or communities to which users belong, will be investigated in the future.



(a) With retweets (b) Without retweets

Fig. 5. Admission controller statistics for different admission schemes (Irene dataset).

ACKNOWLEDGEMENTS

This research was sponsored in part by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] D. Wang, H. Le, T. Abdelzaher, and L. Kaplan, "On truth discovery in social sensing: A maximum likelihood estimation approach," in *Proc of IPSN*, 2012.

[2] M. Srivastava, T. Abdelzaher, and B. Szymanski, "Human-centric sensing," *Philosophical Transactions of the Royal Society, A*, vol. 370, no. 1958, pp. 176–197, 2012.

[3] B. Hull, V. Bychkovsky, Y. Zhang, K. Chen, M. Goraczko, A. Miu, E. Shih, H. Balakrishnan, and S. Madden, "CarTel: a distributed mobile sensor computing system," in *Proc of SenSys*, 2006.

[4] S. B. Eisenman, E. Miluzzo, N. D. Lane, R. A. Peterson, G.-S. Ahn, and A. T. Campbell, "The bikenet mobile sensing system for cyclist experience mapping," in *Prof of Sensys*, 2007.

[5] M. Y. Uddin, H. Wang, F. Saremi, G.-J. Qi, T. Abdelzaher, and T. Huang, "Photonet: a similarity-aware picture delivery service for situation awareness," in *Proc. of RTSS*, 2011.

[6] J.-H. Huang, S. Amjad, and S. Mishra, "Cenwits: a sensor-based loosely coupled search and rescue system using witnesses," in *Proc. of SenSys*, 2005, pp. 180–191.

[7] Sense Networks, "Cab sense," http://www.cabsense.com/.

[8] S. Reddy, A. Parker, J. Hyman, J. Burke, D. Estrin, and M. Hansen, "Image browsing, processing, and clustering for participatory sensing: Lessons from a dietsense prototype," in *Proc. of EmNets*, 2007.

[9] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[10] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[11] L. Berti-Equille, A. D. Sarma, X. Dong, A. Marian, and D. Srivastava, "Sailing the information ocean with awareness of currents: Discovery and application of source dependence," in *CIDR*, 2009.

[12] L. Blanco, V. Crescenzi, P. Merialdo, and P. Papotti, "Probabilistic models to reconcile complex data from inaccurate data sources," in *CAiSE*, 2010, pp. 83–97.

[13] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," *IEEE Trans. on Knowl. and Data Eng.*, vol. 20, pp. 796–808, June 2008. [Online]. Available: http://portal.acm.org/citation.cfm?id=1399100.1399392

[14] A. Galland, S. Abiteboul, A. Marian, and P. Senellart, "Corroborating information from disagreeing views," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 131–140. [Online]. Available: http://portal.acm.org/citation.cfm?id=1718504

[15] X. L. Dong, L. Berti-Equille, and D. Srivastava, "Integrating conflicting data: The role of source dependence," *PVLDB*, vol. 2, no. 1, pp. 550–561, 2009.

[16] X. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava, "Global detection of complex copying relationships between sources," *PVLDB*, vol. 3, no. 1, pp. 1358–1369, 2010.

[17] J. Pasternack and D. Roth, "Knowing what to believe (when you already know something)," in *Proc. the International Conference on Computational Linguistics (COLING)*, Beijing, China, August 2010. [Online]. Available: http://l2r.cs.uiuc.edu/ danr/Papers/PasternackRo10.pdf

[18] ——, "Generalized fact-finding (poster paper)," in *World Wide Web Conference (WWW)*, Hyderabad, India, March 2011.

[19] M. Gupta, Y. Sun, and J. Han, "Trust analysis with clustering," in *WWW*, ser. WWW '11. New York, NY, USA: ACM, 2011.

[20] D. Wang, T. Abdelzaher, L. Kaplan, and C. C. Aggarwal, "On quantifying the accuracy of maximum likelihood participant reliability estimation in social sensing," in *8th International Workshop on Data Management for Sensor Networks (DMSN 2011)*, 2011.

[21] J. P. Callan, Z. Lu, and W. B. Croft, "Searching distributed collections with inference networks," in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '95, New York, NY, USA, 1995, pp. 21–28.

[22] L. Gravano, H. Garcia-Molina, and A. Tomasic, "Gloss: text-source discovery over the internet," *ACM Transactions on Information Systems (TOIS)*, vol. 24, pp. 229–264, 1999.

[23] B. Yuwono and D. Lee, "Server ranking for distributed text retrieval systems on the internet," in *Proc of Database Systems for Advanced Applications*, 1997, pp. 41 – 49.

[24] F. Abbaci, J. Savoy, and M. Beigbeder, "A methodology for collection selection in heterogeneous contexts," in *Proc of Information Technology: Coding and Computing (ITCC)*, 2002.

[25] L. Si, R. Jin, J. Callan, and P. Ogilvie, "Language modeling framework for resource selection and results merging," in *Proc of Information and Knowledge Management (CIKM)*, 2002.

[26] D. Aksoy, "Information source selection for resource constrained environments," *SIGMOD Rec.*, vol. 34, no. 4, pp. 15–20, 2005.

[27] H. Le, D. Wang, H. Ahmadi, M. Y. S. Uddin, Y. H. Ko, T. Abdelzaher, O. Fatemieh, J. Pasternack, D. Roth, J. Han, H. Wang, L. Kaplan, B. Szymanski, S. Adali, C. Aggarwal, and R. Ganti, "Apollo: A data distillation service for social sensing," University of Illinois Urbana-Champaign, Tech. Rep., 2012.