Crowd-sensing with Polarized Sources

Md Tanvir Al Amin, Tarek Abdelzaher, Dong Wang
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801
Email: {maamin2, zaher, dwang24}@illinois.edu

Boleslaw Szymanski
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, New York 12180
Email: szymab@rpi.edu

Abstract—The paper presents a new model for crowd-sensing applications, where humans are used as the sensing sources to report information regarding the physical world. In contrast to previous work on the topic, we consider a model where the sources in question are polarized. Such might be the case, for example, in political disputes and in situations involving different communities with largely dissimilar beliefs that color their interpretation and reporting of physical world events. Reconstructing accurate ground truth is more complicated when sources are polarized. The paper describes an algorithm that significantly improves the quality of reconstruction results in the presence of polarized sources. For evaluation, we recorded human observations from Twitter for four months during a recent Egyptian uprising against the former president. We then used our algorithm to reconstruct a version of events and compared it to other versions produced by state of the art algorithms. Our analysis of the data set shows the presence of two clearly defined camps in the social network that tend of propagate largely disjoint sets of claims (which is indicative of polarization), as well as third population whose claims overlap subsets of the former two. Experiments show that, in the presence of polarization, our reconstruction tends to align more closely with ground truth in the physical world than the existing algorithms.

I. INTRODUCTION

The paper addresses the problem of reconstructing accurate ground truth from unreliable human observations. It extends recent crowd-sensing literature [1] by investigating reliable information collection from polarized sources. By polarization, we refer to a situation where different groups of sources hold largely different beliefs that color their interpretation, and hence representation, of events they observe. Hence, multiple competing versions of such events are reported. The goal of our algorithm is to identify versions that are more likely to be consistent with ground truth.

We apply our solution to extracting information from Twitter. We view Twitter as a participatory sensing system, where participants voluntarily report events they observe. The view of social networks acting as sensor networks was proposed in a recent survey on human-centric sensing [2]. We do not perform natural language processing on tweets (such a contribution would fall into another venue). Rather, in this paper, we explore the merits of making statistical credibility assessments solely based on propagation patterns of different observations, as well as their degree of corroboration, regardless of their semantics.

There are two different schools of thought in information credibility assessment on Twitter. The first uses a machine learning approach that attempts to model human judgement of credibility. In this approach, classifiers are trained to recognize credible tweets as would be judged by a person (e.g., by a mechanical turk worker). Several recent papers proposed classification features of increasing degrees of sophistication that lead to increasingly good matches between human and machine credibility annotations [3], [4].

The second school of thought comes from sensing literature and adopts an estimation-theoretic perspective. It assumes a unique ground truth that is realized in the physical world, and views humans as unreliable sensors who report such ground truth with possible errors and omissions. Statistical (estimation-theoretic) techniques are then used to determine the likelihood that these sensors are correct, given the correlations between them (e.g., that arise from social ties and retweets). An example of this approach in a recent expectation maximization algorithm that jointly estimates the unknown source reliability as well as the statistical tweet credibility [1]. The work was extended to account for non-independent sources [5] and non-independent claims [6].

The paper adopts the latter school of thought. In this work, we are more interested in understanding the physical world (i.e., in sensing) as opposed to understanding what humans perceive as credible. Following this model, we abstract human observers as binary sensors [5] in that each reported observation is either true or false. The novelty of this paper lies in considering sources that are polarized. Intuitively, polarization affects our model of correlations in (human) sensor outputs: when sources (viewed as unreliable binary sensors) share a more significant bias towards a topic, their observation (bit) errors on that topic are more correlated. On the other hand, when they do not share a bias, their errors are independent. Note that, when sources are correlated, corroboration among them carries less statistical weight than when they are independent. Hence, when statistically assessing the likelihood of error in an observation reported by multiple sources, it is important to know whether the topic of that observation matches the bias of the sources or not. The answer determines whether such sources should be regarded as correlated or not, leading to a topic-dependent source correlation model. Later in the paper, we explore the above intuition more formally to arrive at a polarity-informed maximum-likelihood estimate of statistical credibility for each reported observation.

Another advantage of the estimation-theoretic approach adopted for credibility assessment in this paper is that the resulting estimator has a known error bound. This bound was computed in prior work [7], and remains applicable to ours. Hence, not only do we compute truth estimates but also arrive at confidence intervals in source reliability.
We evaluate our solutions using real-world traces collected from Twitter. We recorded observations from Twitter for four months during a recent uprising against the former Egyptian president. We manually annotated a fraction of tweets depending on their degree of support to the deposed president as pro, anti, or neutral. We henceforth call these tweets claims, with no implication as to their degree of credibility. We then studied the propagation patterns of these different groups of claims and adapted our previous fact-finder to recognize polarization. The fact that different topics propagate on different dissemination trees is intuitive and has already been pointed out in prior literature [8]. The paper is novel in its investigation of the specific case of polarized sources and in accounting for polarization in maximum-likelihood credibility assessment.

The investigation of our particular data set revealed the presence of two clearly defined camps in the social network that tend to propagate only one group of claims, as well as a population that tends to propagate selected claims with less correlation with their polarity. We estimated their respective polarity-dependent propagation networks. Each network was then used to compute correlations among sources for the purposes of computing their error-independence properties. For comparison, we also estimated the propagation network constructed when content polarity is not taken into account, as done in previous estimation-theoretic work on truth estimation [5]. We observed that the latter network matches the respective polarity-dependent propagation networks when describing the graph neighborhood of strongly polarized sources, but diverges when describing the neighborhoods of sources that are more neutral. This causes the previous approach to infer incorrect correlations for neutral sources. The current paper shows that these false correlations lead to degradation in truth estimation in favor of polarized information. Our new approach avoids this pitfall.

The rest of this paper is organized as follows. In Section II, we present a case study for this work that shows how polarized certain situations can be. In Section III, we propose a model for polarized sources, claims, and bias-aware social networks. In Section IV, we present a formulation of the problem and derive algorithms to solve it. Experimental evaluation is presented in Section V. Related work is reviewed in Section VI. Finally, we present conclusions and future work in Section VII.

II. THE CASE OF A POLARIZED NETWORK

We analyzed traces obtained from Twitter during a recent uprising in Egypt that resulted in deposing the president. The collected tweets expressed either a positive or negative sentiment towards the deposed president. These tweets were first clustered such that tweets making the same observation (typically the same sentence or very similar sentences) were put in the same cluster. Each such cluster was viewed as a single claim. By observing the time at which different sources contributed their tweet to a given cluster, it was possible to identify a propagation cascade of the corresponding claim through the social network. Table I presents statistics of the tweets collected.

The overall complementary distribution of cascade sizes is illustrated in Figure 1. Note that, the distribution is approximately heavy tailed. The top cascades account for a large fraction of sources. We manually annotated the largest 1000 cascades as pro, anti, or neutral. Collectively, these cascades accounted for roughly 44K sources and 95K tweets.

To assess polarization, Figure 2 plots the distribution of the probability of a source to tweet pro in the top 1000 cascades. The figure illustrates a very interesting property of these cascades. Namely, it is evident that there are three visibly different types of sources. The first type, accounting for 50% of the sources, has a near 1 probability of tweeting pro. The second type, accounting for more than 20% has a near zero probability of tweeting pro (i.e., mostly tweets anti). The rest of the sources tweet both polarities. They are located in the middle of the plot. We call them “neutral” sources. The figure suggests that the community is clearly polarized. This observation motivates us to ask the question: Does this polarization affect the accuracy of reconstruction of physical world events via social sensing? How reliable are previous data cleaning approaches in the presence of polarized sources? How to circumvent their shortcomings?

We show in our evaluation that, in general, community polarization is strong enough to confuse previous algorithms, and therefore polarity-aware credibility analysis algorithms are necessary.

III. A MODEL FOR POLARIZED SOCIAL NETWORKS

This section presents a model of polarized social networks as sensor networks. In the following subsections, the models for claims, (polarized) sources, and their dependencies are described.

A. Modeling Polarized Claims and Sources

Consider m sources who collectively make n claims (i.e., generate n cascades). The relation between the claims and their sources can be represented by a source-claim network, $SC$, which is a bipartite graph. We conveniently represent it using

<table>
<thead>
<tr>
<th>Query</th>
<th>Egypt OR Morsi OR Cairo OR Location: 100 miles around Cairo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of tweets</td>
<td>4.3M</td>
</tr>
<tr>
<td>Total size</td>
<td>17 GB</td>
</tr>
<tr>
<td>Tweets containing “Morsi”</td>
<td>900K</td>
</tr>
<tr>
<td>English Tweets containing “Morsi”</td>
<td>600K</td>
</tr>
<tr>
<td>Number of cascades</td>
<td>193K</td>
</tr>
</tbody>
</table>

Fig. 1. Complementary Cumulative Distribution of cascade sizes

Table I. Summary of the Tweets Collected
a $m \times n$ matrix, such that $SC_{i,j} = 1$ if source $S_i$ makes claim $C_j$ (i.e., contributes to $j^{th}$ cascade), and 0 otherwise.

We consider a binary model, where each claim can be True or False. This categorization is orthogonal to polarity. To model polarized claims, we introduce a topic indicator, $y_j$, for each claim $C_j$, that takes one of the values from topic set $T = \{pro, anti, neutral\}$. This topic represents the polarity of claim $C_j$. A vector $y$ is defined as the polarity vector for all claims.

In general, a source may make claims of different polarity. We define the reliability of a source as the probability of making correct claims. Note, however, that when making claims that agree with the source’s own bias, the source might become less selective and have a higher probability of making false claims. In contrast, when making claims that are orthogonal to the source’s bias, the source might get more conservative and the probability of correctness increases. This suggests that source reliability is a vector, with one entry per topic. Hence, we model a source $S_i$ by a vector $r_i$ of dimension $|T|$, where $r_{i,t}$ denotes the reliability of the source when making claims of polarity $T_t$.

B. Modeling Polarity-aware Source Dependencies

Prior work on credibility assessment in social sensing [1] developed an algorithm that takes a source-claim network, $SC$, as input, and jointly estimates both reliability of sources and statistical credibility of claims. The algorithm was then adapted to take into account dependencies between sources [5]. As mentioned earlier, such dependencies imply correlated errors that need to be accounted for in statistical analysis.

A dependency between two sources is a directional quantity. It is estimated by observing the probability that one source propagates information obtained from the other (i.e., joins a cascade given that the other source joined it earlier). Representing such correlations by directional links between the respective source nodes, a propagation graph is constructed that constitutes the inherent social (influence) network. Netrapalli and Sanghavi [9] formulate the problem uncovering the latent influence network (or information propagation graph), given a sufficient number of cascades. We use their algorithm to generate social networks given the set of sources, tweets, and their timestamps.

An alternative method of finding the latent network can be to take the Twitter-provided follower-followee graph. However, the follower-followee graph is not always a good representation of actual information propagation paths exercised by users. For example, as most of the tweets are public, when an event of significance transpires in the physical world, interested individuals may search for top tweets and act on those. This method does not require following any particular person and therefore the follower-followee relationship is an incomplete proxy for the underlying information propagation network.

Another possibility is to construct the propagation graph directly from retweets. For example, if source A retweets source B, $k$ times, insert a weighted directed link $(A, B, k)$ in the network. The problem with this approach is that in large cascades it is not clear who exactly (of those who retweeted the same claim earlier) a source was influenced by. Hence, the retweet relation does not necessarily reflect the correct influence topology. The influence network estimation approach proposed by Netrapalli and Sanghavi [9] avoids this problem, which is why we adopt it in this paper.

A further advantage of using the approach of Netrapalli and Sanghavi [9] for estimating the influence propagation network is that we no longer care whether something is a retweet, or a separately authored tweet of similar content. All that matters for this algorithm are the clusters of tweets (of similar content), each forming a cascade, and the timestamp of each tweet in each cascade. Hence, the approach is not restricted to uncovering influence propagation via the Twitter medium itself. A source may influence another externally (e.g., via a different communication medium). The external link can still be uncovered as long as both sources make tweets of similar content.

To model polarity-aware source dependencies, we generate $|T|$ different influence propagation networks, using the aforementioned algorithm [9], by observing claims of a single polarity at a time to infer a single network. The set of these networks is collectively referred to as $SD_B$, where element $SD_B^{t}$ is the network generated by considering only the claims of polarity $T_t$. We call the corresponding networks pro, anti, and neutral networks. For comparison, we also construct a generic network, $SD$, by considering all claims regardless of their polarity. In Section V, we empirically evaluate the quantitative differences between $SD_B$ and $SD$.

Please note that the pro (anti, neutral) network is not a network of only the pro (anti, neutral) sources, rather it is a network created using only the pro (anti, neutral) claims. As a result, these networks may contain overlapping sources if such sources make claims of different polarities. The terms pro source, anti source, and neutral source, when used, therefore refer to the predominant disposition of a source as opposed to exclusive membership of one of the networks.

IV. GROUND-TRUTH ESTIMATION IN POLARIZED NETWORKS

This section formulates the problem of ground truth estimation in polarized networks and describes the algorithm we use to solve it.

A. Problem Formulation

Based on the model described in Section III, the problem is to estimate the statistical credibility of each claim given the
source claim network, $SC$, the polarity of each claim, specified in the vector, $y$ (where $y_j$ is the polarity of claim $C_j$), and the inferred set of influence propagation networks, $SD^B$, one per polarity. Let $z_j$ be the unknown ground truth value of claim $C_j$ (stating whether it is true or false). Formally, we want to compute:

$$
\forall j, 1 \leq j \leq n : \Pr(z_j = \text{True}|SC, y, SD^B)
$$

(1)

B. Solution

As discussed earlier, the bias of a source may cause it to be less selective in making claims of one polarity compared to another. For example, the source might indiscriminately propagate claims that agree with its bias, while being selective in making other claims. Hence, source reliability (probability of making claims that are true) may depend on claim polarity. Let the reliability of source, $S_i$, when making claims of polarity, $T_i$, be denoted $r_{i,T_i}$. For simplicity, in this paper, we assume that the source reliability values for different polarities are independent. The polarities of interest are $T = \{\text{pro, anti, neutral}\}$. Hence, we can break down Expression (1) into three independent subproblems; namely, computing the credibility of $\text{pro, anti, and neutral}$ claims, respectively. This is formally expressed as finding the probabilities below:

$$
\forall j, y_j = \text{pro} : \Pr(z_j = \text{True}|SC_{y_j=\text{pro}}, SD^B_{\text{pro}})
$$

(2)

$$
\forall j, y_j = \text{anti} : \Pr(z_j = \text{True}|SC_{y_j=\text{anti}}, SD^B_{\text{anti}})
$$

(3)

$$
\forall j, y_j = \text{neu} : \Pr(z_j = \text{True}|SC_{y_j=\text{neu}}, SD^B_{\text{neu}})
$$

(4)

where $SC_{y_j=\text{pro}}, SC_{y_j=\text{anti}},$ and $SC_{y_j=\text{neu}}$ are the subgraphs of the source claim network, $SC$, with claims of only the specified polarity present (or equivalently, the array $SC$ with claim columns of other polarities removed).

The independence assumption between source reliability parameters $r_{i,\text{pro}}, r_{i,\text{anti}},$ and $r_{i,\text{neutral}}$ makes it possible to solve for variables (2), (3), and (4) separately, essentially breaking the original problem into three independent subproblems, one for each polarity. In the subproblem corresponding to polarity, $T_i$, we consider the source claim subnetwork $SC_{y_j=T_i}$, and the inferred influence propagation network $SD^B_{T_i}$, then solve jointly for source reliability $r_{i,T_i}$, and statistical claim credibility, $z_j$, where $y_j = T_i$.

Figure 3 illustrates the formation of the subproblems. Here $S_1$ to $S_4$ are the sources, and $C_1$ to $C_5$ are the claims. There is an edge in $(S_i, C_j)$ in the bipartite network if source $S_i$ authored claim $C_j$. The $\text{pro}$ claims are shown in red, the $\text{anti}$ claims are shown in green, and the $\text{neutral}$ claims are shown in white. The proposed polarity-aware algorithm identifies each ‘class’ of claims, and considers the independent subproblems that contain all the claims of that particular class and the sources that make them. The solution to each subproblem results in credibility scores for the claims in that particular class, as well as one element of the polarity-aware reliability vector of the sources.

More specifically, each subproblem is solved using the expectation maximization algorithm presented in [5]. Starting with an initial guess of source reliability parameters, expressed as the vector $\theta_0$, the algorithm performs the iterations:

$$
\theta_{n+1} = \arg \max_\theta \{E_z[SC_{y_j=T_i} \theta_n \{ \ln \Pr(SC_{y_j=T_i}, z^i|SD^B_{T_i}, \theta) \}] \}
$$

(5)

where $z^i$ is the vector of latent variables $z_j$ (claim credibility), for all claims, where $y_j = T_i$. The above breaks down into three steps:

- Compute the log likelihood function
  $$\ln \Pr(SC_{y_j=T_i}, z^i|SD^B_{T_i}, \theta)$$

- The expectation step
  $$Q_\theta = E_{z^i|SC_{y_j=T_i} \theta_n \{ \ln \Pr(SC_{y_j=T_i}, z^i|SD^B_{T_i}, \theta) \}}$$

- The maximization step
  $$\theta_{n+1} = \arg \max_\theta \{ Q_\theta \}$$

where the last two steps are solved iteratively until they converge, yielding updated source reliability estimates and claim credibility, $z_t$ (for claims of polarity $T_i$).

C. Polarity Classification of Claims

Our polarity aware model assumes that there exists a mapping $y$ from claims to polarities. This mapping is required to divide the set of tweets into $|T|$ parts. We manually annotated the top 1000 largest cascades (most propagated claims). However, to use our polarity aware credibility estimation algorithm as a crowd-sensing tool, it is important to include all the claims in the analysis. Therefore, an algorithm to classify each incoming tweet into a particular polarity is required.

We attempted to use readily available learning-based sentiment analysis tools for this purpose that look at the content of the tweets and classify them into positive and negative sentiments. It was not sufficient because the polarity of a tweet is not necessarily correlated with its sense or sentiment being positive or negative. For example, “The government is working for the people”, and “The opposition is working against the people” have positive and negative sentiments respectively; but polarity of both of these claims is likely to be pro-government.
It is possible to design an advanced classifier for this purpose that uses learning techniques or natural language processing methods to classify the tweets into pro, anti, and neutral classes. However, such a classifier requires extensive domain-specific knowledge and its design depends on the choice of polarity classes and their context. Moreover, simple learning-based tools often suffer from low quality and require extensive training. A domain-specific classifier that looks at the content and determines the polarity is therefore hard to generalize.

Instead, given our seed of manual annotations, we used an iterative algorithm that propagates tweet annotations to source annotations, and then from source annotations back to tweet annotations, repeatedly. Hence, sources that predominantly make tweets of a given polarity are identified from the manually annotated tweets and other tweets of the same sources are given the same polarity. This algorithm is clearly an approximation. Nevertheless, even this approximate polarity annotation can lead to an improvement in fact-finding, compared to polarity-unaware analysis.

V. EXPERIMENTS

In this section, we describe the experiments performed to determine how community polarization affects statistical credibility estimation in social sensing. Our experiments use the traces obtained from Twitter during the recent uprising in Egypt resulting in deposing the president (summarized in Table I). The crawling started in July, 2013 and continued for four months.

A. Polarization Analysis

A key hypothesis of our work is that a better solution to the credibility estimation problem is obtained by breaking all tweets by polarity and solving independently for credibility of tweets in each polarity class, $T_t$, given the polarity-specific source-claim matrix, $SC_{ji}=T_t$, and the polarity-specific influence propagation network, $SD^B_{pro}$. This is as opposed to amalgamating all tweets regardless of polarity into one source claim matrix, $SC$, and using a single influence propagation network, $SD$, as inputs to the credibility estimation.

To appreciate the difference between the two solutions, some analysis of the resulting networks is needed. For this analysis, we read the text of the largest 1000 claims and manually annotated them as pro, anti, or neutral. The annotation revealed that there are 199 pro cascades and 109 anti cascades in the top 1000 largest cascades. By utilizing the timestamps of when each source forwarded a given claim, we estimated the inherent social propagation network for each type of claims using the algorithm proposed by Netrapalli and Sanghavi [9].

This resulted in 15,714 edges in the pro network $SD^B_{pro}$, 8,460 edges in the anti network $SD^B_{anti}$, and 33,946 edges in the neutral network $SD^B_{neutral}$. We also estimated the generic network $SD$ using all 1000 cascades together. There are 55,329 edges in that network.

Figure 4 shows the pro network, $SD^B_{pro}$, in red, and the anti network, $SD^B_{anti}$, in green, overlaid together. The neutral network is not shown to maintain visual clarity.\(^1\) This plot suggests that two polarized groups exist with their own different propagation links.

With this preparation, we are ready to answer the question: is considering one amalgamated influence propagation network the same as considering a polarity-specific network, when estimating the credibility of tweets?

The answer is no. It turns out that the neighborhood of neutral sources is not correctly represented in the amalgamated network. This results in improper modeling of source dependencies, which affects credibility estimation when such sources propagate pro or anti tweets. To see the difference in source dependency estimation when neutral sources propagate pro or anti tweets, consider Figure 5, which compares the neighborhood of neutral nodes in the amalgamated influence propagation network, $SD$, versus that in the pro or anti network ($SD^B_{pro}$ or $SD^B_{anti}$). The degree of similarity is measured by the jaccard similarity coefficient between the two sets of neighborhoods. The similarity distribution between $SD^B_{pro}$ and $SD$ is shown in Figure 5(a). The similarity distribution between $SD^B_{anti}$ and $SD$ is shown in Figure 5(b). It is seen that more than 98% of the sources have different neighborhoods in the amalgamated SD network compared to the $SD^B_{pro}$ and $SD^B_{anti}$ networks. This means that the amalgamated network does not properly capture their dependencies. Further inspection suggests that it exaggerates them, leading the statistical estimation algorithm to rely less on such sources (to avoid correlated errors).

The same cannot be said of polarized sources. Figure 6 shows that the generic network $SD$ does not confuse the neighborhood of the strongly polarized sources. Figure 6(a) shows the distribution of neighborhood similarity between $SD^B_{pro}$ and $SD$, and Figure 6(b) shows the distribution of neighborhood similarity between $SD^B_{anti}$ and $SD$. The generic network $SD$ correctly determines the neighborhood for around 80% of the polarized sources. This is expected. Those sources forward mostly one polarity of claims. Hence, the estimation of influence propagation returns the same results whether all or only those claims are considered.

\(^1\)Source are further restricted to only the top 400 cascades for clarity.

Fig. 4. An overlay of two polarized social networks. pro shown in red and anti shown in green
The above figures offer an intuition into the shortcomings of the amalgamated approach from the perspective of credibility estimation: the approach tends to “disenfranchise” neutral sources.

**B. Fact Finding**

We compare the accuracy of our polarity-aware credibility estimation algorithm to its predecessor [5] that does not consider the polarity of tweets. We identify our algorithm by the word ‘Polarized’ and the other algorithm by the word ‘Combined’.

To evaluate the fact-finding performance, we executed three experiments by selecting the largest $n$ cascades, for $n \in \{400, 1000, 5000\}$. Summaries of the datasets used in each experiment are presented in Table II. In each experiment, we classified the claims into the three polarity classes and ran polarity-aware and polarity-unaware estimators. In each case, the fact-finder computed the credibility of input claims $\in [0, 1]$ and the reliability of their sources $\in [0, 1]$.

Figure 7 shows the relation between the output of different algorithms in different experiments. The circle and triangle pointed curves show the fraction of claims that are believed as facts by the combined and the polarized algorithm, respectively. We find that the combined algorithm is less judgmental and believes more claims to be true. The square pointed curve shows the agreement between two schemes. The agreement is computed as the jaccard similarity between the two sets of claims believed as facts by the two algorithms. It is evident that the two algorithms converge more as the number of claims increase. We conjecture that this is because polarized claims were retweeted more and had larger cascade sizes. Hence, the smaller experiments had more polarized claims, offering a larger difference in results between the two approaches.

From Table II, the probability of an arbitrary claim to be polarized is nearly 39% in the 400 claims experiment, while its nearly 31% in the 1000 claims experiment, and only 15% in the 5000 claims experiment. We also classified the tweets for a 10,000 claims and a 25,000 claims experiment, where

---

**Table II. Summary of the Dataset of the Experiments**

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Cascades</th>
<th>Experiment 2</th>
<th>Experiment 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pro Claims</td>
<td>105</td>
<td>199</td>
<td>379</td>
</tr>
<tr>
<td>Anti Claims</td>
<td>50</td>
<td>199</td>
<td>371</td>
</tr>
<tr>
<td>Neutral Claims</td>
<td>245</td>
<td>692</td>
<td>4,250</td>
</tr>
<tr>
<td>Number of Sources</td>
<td>31,480</td>
<td>43,605</td>
<td>68,206</td>
</tr>
<tr>
<td>Number of Tweets</td>
<td>62,864</td>
<td>94,871</td>
<td>184,452</td>
</tr>
<tr>
<td>Pro Tweets</td>
<td>17,603</td>
<td>22,750</td>
<td>27,114</td>
</tr>
<tr>
<td>Anti Tweets</td>
<td>8,509</td>
<td>11,691</td>
<td>19,411</td>
</tr>
<tr>
<td>Neutral Tweets</td>
<td>36,752</td>
<td>60,430</td>
<td>137,927</td>
</tr>
<tr>
<td>Source-Claim Edges (Pro)</td>
<td>6,770</td>
<td>9,302</td>
<td>16,380</td>
</tr>
<tr>
<td>Source-Claim Edges (Neutral)</td>
<td>23,197</td>
<td>41,638</td>
<td>101,017</td>
</tr>
<tr>
<td>Pro Network Edges</td>
<td>12,160</td>
<td>15,714</td>
<td>23,942</td>
</tr>
<tr>
<td>Anti Network Edges</td>
<td>6,292</td>
<td>8,460</td>
<td>19,037</td>
</tr>
<tr>
<td>Neutral Network Edges</td>
<td>19,735</td>
<td>33,946</td>
<td>92,068</td>
</tr>
<tr>
<td>Combined Network Edges</td>
<td>36,472</td>
<td>55,329</td>
<td>130,092</td>
</tr>
</tbody>
</table>
the probability of a claim to be polarized went further down to 11% and 7%, respectively.

Finally, we evaluated the quality of information obtained by the polarized algorithm and the combined algorithm. Here, we present the comparison for the 1000 claims experiment. In this experiment, the polarized algorithm selected 128 pro, 76 anti, and 498 neutral claims as true (a total of 700 claims). The combined algorithm selected 147 pro, 88 anti, and 543 neutral claims (a total of 778 claims). Of the two sets, 662 claims were common in both cases, resulting in a degree of agreement of 81.13%.

The interesting cases, however, are those where the algorithms disagreed. We considered two sets of claims on which there was disagreement. Set A contained the claims that the polarized algorithm believed to be true with a probability 1, but the combined algorithm did not. There were 38 such claims. Conversely, set B contained the claims that the combined algorithm believed to be true with probability 1, but the polarized algorithm does not. There were 116 such claims.

The two sets were merged and presented to a human grader without the information on which claim came from which set. The grader was instructed to carefully research and verify each claim using historic data. Verified facts received a score of 1. Fabricated claims and lies received score of -1. Non-factual claims such as expressions of emotion, slogans, and sentiments were discarded (received a score of 0). After grading was done, we separated the sets again and calculated the scores for each algorithm. The results are presented in Table III.

If we count non-factual claims (i.e., expressions of emotion, etc) then, when the algorithms disagree, 66% of the claims believed by the polarized algorithm are true, compared to 62% for the combined algorithm. More interestingly, the polarized algorithm believes only 2.6% false claims (that received a -1 score), while the combined algorithm believes 8.6% false claims. If we discard non-factual claims from the total (after all, they do not refer to binary facts), then when the algorithms disagree, 96% of the claims believed by the polarized algorithm are true, compared to only 88% for the combined algorithm. Equivalently, the probability of error is reduced (in our case) from 12% to 4%, or by a factor of three!

Finally, combining all scores to get a single overall quality indicator, our bias-aware crowd-sensing algorithm improves the quality by more than 18%.

The results shown above are a step forward. They demonstrate that when sources are polarized, we should consider separately the pro, anti, and neutral claims in performing credibility analysis. Such separation prevents estimation of false dependencies between neutral sources, based on amalgamated retweet patterns. By separating the content and considering only polarity-specific dependencies, errors are reduced.

## VI. Related Work

Crowd-sensing is an increasingly popular area of research, where humans are acting as the sensors generating observations. It extends more traditional participatory sensing models where humans carry the sensor devices that collect data. Human-generated observations have a different error model than traditional sensors, which introduces many interesting questions and challenges.

Wang et al. [1] addressed the question of reliable sensing in the context of human observers. He proposed a model where human observations are treated as binary claims that can be either true or false. The question of estimating credibility of a particular claim can be trivially addressed by voting (i.e., a claim with a larger propagation is deemed more credible). However, this simple approach is highly suboptimal when sources have different degrees of reliability. Wang’s approach [1] jointly estimated source reliability and claim credibility for independent sources. When source are generally not independent, source diversification heuristics were studied that select tweets from only a subset of sources to maximize a measure of mutual independence [10]. A more principled solution that models source dependencies directly and accounts for them in the maximum likelihood framework was described in [5]. Our paper builds on this foundation, while accounting for the polarized sources.

Information propagation through social or other complex networks has been studied extensively [11]–[14]. Netrapalli and Sanghavi [9], Myers and Leskovec [15], and Rodriguez et al. [16] model the propagation of information through social networks as epidemic cascades and use different ways to estimate the propagation graph from multiple cascades. This work nicely complements ours, since the latent influence propagation network is one of the inputs to our maximum likelihood (credibility) estimator. A related problem is community detection. Several efforts addressed the issue of detecting different communities in social networks [17], [18]. These methods can be used to confirm that influence cascades indeed propagate largely within corresponding community boundaries.

Topic-based models to infer user influence and information propagation have been studied in different contexts. Lin et al. [19] proposed a probabilistic model to infer the diffusion of topics through social networks. Pal and Counts [20], and Eytan et al. [21] propose methods to infer topic-based authorities and influential nodes in the context of online social platforms and microblogs. The concept of social media genotype to
model and predict user activity in social media platforms was proposed by Bogdanov et al. [8]. The genotype is a set of features that defines user behavior in a topic-specific fashion. Like us, they argue that a single static network is not a good indicator of user activity. Instead, they derive topic-aware influence backbones based on user genotypes, which we exploit in understanding how different polarities (topics) of information follow different paths in the social network. They focus on predicting user activity, while we are interested in improving the quality of fact-finding.

Finally, our work is related to the more general genre of crowd-sourcing; using the crowd to perform useful tasks [22], [23]. Unlike our paper, where participants are unaware of their participation, this genre of research considers a more controlled and structured environment, where people are generally paid to participate in advertised tasks.

VII. CONCLUSIONS

The paper addressed truth recovery from tweets in the case of a polarized network. It was shown that polarization impairs credibility estimation. The problem was solved by developing a new polarity-aware estimation methodology that improves quality of results by 18%. Several extensions of the current framework are possible. For example, we assume that polarities are already known. Advanced classifiers that aggregate both content and provenance information may prove useful to reduce the need for manual polarity annotation. The idea of polarities can be extended to topics with arbitrary relations and overlap. Also, while this work considered sources that are polarized, it did not regard them malicious. An intent to deceive by an intelligent adversary presents a harder challenge. These extensions are delegated for future work.

ACKNOWLEDGMENT

Research reported in this paper was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement W911NF-09-2-0053, DTRA grant HDTRA1-10-1-0120, and NSF grant CNS 13-29886. 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