Hardness-aware Truth Discovery in Social Sensing Applications

Jermaine Marshall, Munira Syed, Dong Wang
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556
jmarsha5@nd.edu, msyed2@nd.edu, dwang5@nd.edu

Abstract—**This paper develops a new principled framework to solve a hardness-aware truth discovery problem in social sensing applications. Social sensing has emerged as a new application paradigm where a large crowd of social sensors (humans or devices on their behalf) are recruited to or voluntarily report observations about the physical environment at scale. These observations may be either true or false, and hence are viewed as binary claims. A fundamental problem in social sensing applications lies in ascertaining the correctness of claims and the reliability of data sources. We refer to this problem as *truth discovery*. Significant efforts were made to address the truth discovery problem, but an important dimension of the problem has not been fully exploited: *hardness of claims* (how challenging a claim is to be made). A common assumption made in the previous work is that they assumed all claims are of *the same degree of hardness*. However, in real world social sensing applications, simply ignoring the hardness differences between claims could easily lead to suboptimal truth discovery results. In this paper, we develop a new hardness-aware truth discovery scheme that explicitly considers *different hardness degrees of claims* into a rigorous analytical framework. The new truth discovery scheme solves a maximum likelihood estimation problem to determine both the claim correctness and the source reliability. We compare our hardness-aware scheme with the state-of-the-art baselines through three real world case studies (Baltimore Riots, Paris Attack and Oregon Shootings, all in 2015) using Twitter data feeds. The evaluation results showed that our new scheme outperforms all compared baselines and significantly improves the truth discovery accuracy in social sensing applications.**

Keywords—*Social Sensing, Hardness-aware, Truth Discovery, Maximum Likelihood Estimation, Twitter*

I. Introduction

This paper develops a new principled framework to solve a hardness-aware truth discovery problem in social sensing applications. Social sensing has emerged as a new application paradigm where a large crowd of social sensors (humans or devices on their behalf) are recruited to or voluntarily report observations about the physical environment at scale [2]. These observations may be either true or false, and hence are viewed as binary claims. Examples of social sensing applications include crowdsensing/crowdsourcing tasks using different sensors in smartphones [6], obtaining real-time situation awareness for disaster response and crisis management using online social media [33], geo-tagging applications for smart cities using data contributed by common citizens [15]. This paradigm has a few clear advantages over the traditional infrastructure-based sensor networks: (i) social sensing is infrastructure free and it is inexpensive to deploy applications at a large scale; (ii) social sensors are more versatile than physical sensors and they can report a broad category of phenomena (e.g., disasters, traffic congestion, power outage, riots, etc); (iii) social sensing normally has a better coverage than traditional sensing paradigm as social sensors are mobile and naturally scattered around the world. However, the data collection in social sensing is usually open to all and it is impossible to screen all participants (data sources) beforehand. Therefore, a fundamental problem in social sensing applications lies in accurately ascertaining the correctness of claims and the reliability of data sources. We refer to this problem as *truth discovery*.

Significant progress has been made to address the truth discovery problem in social sensing from the sensor network [29], [30], information fusion [12], [25] and data mining [35], [38] communities. A common assumption made in the previous work is: the claims are assumed to be of *the same degree of hardness* (i.e., it is equally challenging for a source to report all of its claims). However, such assumption may not hold in real world social sensing applications, where claims could have different degrees of hardness depending on various factors of the event associated with the claim such as abnormality, time, location, and scale. For example, Table I shows claims reported to Twitter in the aftermath of the Oregon Umpqua Community College Shooting event in October 2015. The first two claims are regarded as hard claims as they requires people to be physically at the prime locations of the events and explicitly report concrete and informative observations. The latter two claims are regarded as easy claims as they are in the form of personal sentiments and repeated information (i.e., Retweets) that can be made by anyone anywhere.

Tweet	Hardness Degree
“There’s a shooter! Run! Run! Get out of there!” –#Oregon students during #OregonShooting. Our latest:#UCCShooting	Hard
The shooter in a massacre at Umpqua Community College in Oregon has been identified.	Hard
My heart goes out to all those who lost loved ones today.	Easy
RT @BanCollectivism: And yet these shooting don’t happen much in ”progressive” countries, you idiot.	Easy

Table I. Claims of Different Hardness Degrees in Oregon Umpqua Community College Shooting Event (2015)

Important challenges exist when we develop a hardness-aware solution to improve the truth discovery accuracy in social sensing. First, social sensing is designed as an open data collection paradigm where the reliability of sources and the correctness/hardness degree of claims are often *unknown a priori*. Second, it is very challenging to find an effective method to automatically and accurately identify the hardness degrees of all claims considering the rich and unstructured data reported by human sensors in social sensing, especially with no prior knowledge of a particular event. Third, sources may have different reliability in reporting claims of different degrees of hardness and such difference cannot be directly identified from the social sensing data.

To address the above challenges, we develop a hardness-aware truth discovery scheme that explicitly incorporates *different hardness degrees of claims* into a maximum likelihood estimation framework. In particular, a Hardness-Aware Expectation Maximization (HA-EM) algorithm is developed to assign true values to claims and reliability to sources more accurately by exploiting the hardness degree of claims. We evaluate our HA-EM scheme through three real world case studies (Baltimore Riots, Paris Attack and Oregon Shootings, all in 2015) based on Twitter data feeds. The evaluation results show that our new scheme outperforms the state-of-the-art baselines and significantly improves the truth discovery accuracy. The results of this paper are important because they allow social sensing applications to accurately estimate the correctness of claims and the reliability of sources by explicitly incorporating hardness degree of claims into a principled framework. To summarize, our contributions are as follows:

- To the best of our knowledge, this study is the first to explicitly consider the *hardness degree of claims* in the truth discovery problem of social sensing using a principled approach.
- We develop an analytical framework that allows us to derive an *optimal solution* (in the sense of maximum likelihood estimation) for the hardness-aware truth discovery problem.
- We show non-trivial performance gains achieved by our hardness-aware truth discovery scheme through three real world case studies in social sensing applications.

The rest of this paper is organized as follows: we discuss the related work in Section II. In Section III, we present the new hardness-aware truth discovery model for social sensing applications. The proposed maximum likelihood estimation framework and the expectation maximization solution is presented in Section IV. Evaluation results are presented in Section V. We discuss the limitations and future work in Section VI. Finally, we conclude the paper in Section VII.

II. Related Work

Social sensing has emerged as a new act of collecting sensory measurements about the physical world from human sources or devices on their behalf [2]. Some early applications include CenWits [10], CabSense [22], and BikeNet [8]. More recent applications in social sensing start to address challenges such as preserving privacy of participants [5], improving energy efficiency of sensing devices [16] and building general models in sparse and multi-dimensional social sensing spaces [3]. An emerging and critical question about data reliability arises when the data in social sensing applications are collected by humans whose "reliability" is not known [1]. Some truth discovery techniques have been developed to address this problem but they did not fully exploit the time dimension of the problem in their solutions [29], [30]. In this paper, we develop a hardness-aware truth discovery scheme that explicitly exploits the claim hardness in social sensing and significantly improves the truth discovery accuracy.

In data mining and machine learning literature, there exists a good amount of work on the topics of *fact-finding* that jointly compute the source reliability and claim credibility [9]. *Hubs and Authorities* [13] established a basic fact-finding model based on linear assumptions to compute scores for sources and claims they asserted. Yin et al. introduced *TruthFinder* as an unsupervised fact-finder for trust analysis on a providers-facts network [34]. Other fact-finders enhanced these basic frameworks by incorporating analysis on properties or dependencies within claims and sources [21], [25]. More recently, new fact-finding algorithms have been designed to address the background knowledge [18], multi-valued facts [37], and multi-dimensional aspects of the problem [36]. In this paper, we use the insights from the above work and develop a new estimation scheme to solve the hardness-aware truth discovery problem in social sensing applications.

Maximum likelihood estimation (MLE) technique has been widely used in sensor network community to solve estimation and information fusion problems [14], [19], [32]. For example, Wang et al. proposed a MLE based target tracking approach to solve the instability problem and offer superior tracking performance in wireless sensor networks [32]. Pereira et al. presented a maximum likelihood estimation algorithm to solve a distributed parameter estimation problem in unreliable sensor networks [19]. Leng et al. built a maximum likelihood estimator to jointly estimate the clock offset, clock skew and fixed delay in sensor networks [14]. However, the estimation variables in the above work are mostly continuous and the sensors are physical sensors. In contrast, we focus on estimating a set of *binary variables* that represent either true or false statements from human sensors. The MLE problem we studied is actually more challenging due to the discrete nature of the estimated variables and the non-trivial complexity of modeling *humans as sensors* in social sensing.

Finally, our work is also related with a type of information filtering system called recommendation systems [11]. Expectation Maximization (EM) has been used as an optimization approach for both collaborative filtering [24] and content based recommendation systems [20]. For example, Wang et al. developed a collaborative filtering based system using the EM approach to recommend scientific articles to users of an online community [24]. Pomerantz et al. proposed a content-based system using EM to explore the contextual information to recommend movies [20]. However, the truth discovery in social sensing studies a different problem. Our goal is to estimate the correctness of observations from a large crowd of unvetted sources with unknown reliability and various degrees of claim hardness rather than predict users' ratings or preferences of an item. Moreover, recommendation systems commonly assume a reasonable amount of good data is available to train their models while little is known about the data quality and the source reliability a priori in social sensing applications.

III. Hardness-Aware Truth Discovery Problem in Social Sensing

In this section, we formulate the hardness-aware truth discovery problem in social sensing as a maximum likelihood estimation problem. We borrowed a social sensing model introduced in [30]. In particular, consider a scenario where a group of M sources, namely, $S_1, S_2, ..., S_M$, who report a set of N observations about the physical environment, namely, $C_1, C_2, ..., C_N$. Those observations may be true or false, and hence are viewed as *binary claims*. For example, in an application that reports the litter locations on city streets, each location may be associated with a claim that is true if the litter is present and false otherwise. We assume, without loss of generality, that the default state of each claim is negative (e.g., no litter on city streets). Hence, sources only report when the positive state of the claim is encountered. Let S_i represent the i^{th} source and C_j represent the j^{th} claim. $C_j = 1$ if it is true and $C_j = 0$ otherwise. We define a *Sensing Matrix* SC, where $S_iC_j = 1$ when source S_i reports that claim C_j is true, and $S_iC_j = 0$ otherwise.

Furthermore, we need to incorporate the hardness degree of claims into our model. To capture the claim hardness, we define a *Hardness Vector* H, where the element h_j represents the hardness degree of claim C_j. Specifically, h_j is a discrete variable with K different values representing K different degrees of claim hardness (e.g., easy, medium, hard).

We formulate the hardness-aware truth discovery problem in social sensing as follows. First, let us define a few important terms that will be used in the problem formulation. We denote the *reliability* of source S_i by r_i, which is the probability that a claim is correct given that source S_i reported it. Formally, r_i is given by:

$$r_i = \Pr(C_j = 1 | S_iC_j = 1) \tag{1}$$

Considering the claims may have different degrees of hardness, we define r_i^k as the reliability of S_i when it reports a claim with a hardness degree of k, where $k = 1, ..., K$. Formally, $t_i^{k,l}$ is given by:

$$r_i^k = \Pr(C_j = 1, h_j = k | S_iC_j = 1) \tag{2}$$

Hence,

$$r_i = \sum_{k=1}^{K} r_i^k \times \frac{s_i^k}{s_i} \quad k = 1, ..., K \tag{3}$$

where s_i^k is the probability that S_i reports C_j with a hardness degree of k. Formally, $s_i^k = \Pr(S_iC_j = 1, h_j = k)$. Note that the probability that S_i reports a claim is: $s_i = \Sigma_{k=1}^{K} s_i^k$.

Let us further define T_i^k to be the (unknown) probability that S_i reports C_j (of hardness degree k), given that the claim is indeed true. Similarly, let F_i^k denote the (unknown) probability that S_i reports C_j (of hardness degree k), given that the claim is false. Formally, T_i^k and F_i^k are defined as follows:

$$\begin{aligned} T_i^k &= \Pr(S_iC_j = 1 | C_j = 1, h_j = k) \\ F_i^k &= \Pr(S_iC_j = 1 | C_j = 0, h_j = k) \end{aligned} \tag{4}$$

Using the Bayes theorem, we can establish the relationship between T_i^k, F_i^k and r_i^k, s_i^k as follows:

Table II. THE SUMMARY OF NOTATIONS

Description	Notation
Set of Sources	S
Set of Claims	C
Sensing Matrix	SC
Hardness Vector	H
Report Probability	$s_i^k = \Pr(S_iC_j = 1, h_j = k)$
Source Reliability	$r_i^k = \Pr(C_j = 1, h_j = k \| S_iC_j = 1)$
Correctness Probability	$T_i^k = \Pr(S_iC_j = 1 \| C_j = 1, h_j = k)$
Error Probability	$F_i^k = \Pr(S_iC_j = 1 \| C_j = 0, h_j = k)$

$$\begin{aligned} T_i^k &= \frac{r_i^k \times s_i^k}{d_k} \\ F_i^k &= \frac{(1 - r_i^k) \times s_i^k}{(1 - d_k)} \end{aligned} \tag{5}$$

where d_k is the prior probability that a randomly chosen claim with a hardness degree of k is true (i.e., $d_k = \Pr(C_j = 1, h_j = k)$). The introduced notations are summarized in Table II.

Therefore, the hardness-aware truth discovery problem studied in this paper can be formulated as a maximum likelihood estimation (MLE) problem: given the Sensing Matrix SC and Hardness Vector H, we aim at estimating the likelihood of the correctness of each claim and reliability of each source. Formally, we compute:

$$\begin{aligned} &\forall j, 1 \le j \le N : \Pr(C_j = 1 | SC, H) \\ &\forall i, 1 \le i \le M : \Pr(C_j = 1 | S_iC_j = 1) \end{aligned} \tag{6}$$

IV. A HARDNESS-AWARE MAXIMUM LIKELIHOOD ESTIMATION APPROACH

In this section, we solve the hardness-aware truth discovery problem formulated in Section III by developing a Hardness-Aware Expectation-Maximization (HA-EM) algorithm.

A. Building The Likelihood Function

EM is an optimization scheme that is commonly used to solve the MLE problem where unobserved latent variables exist in the model [7]. Specifically, it iterates between two key steps: expectation step (E-Step) and maximization step (M-step). In E-step, it computes the expectation of the log-likelihood function based on the current estimates of the model parameters. In M-step, it computes the new estimates of the model parameters that maximize the expected log-likelihood function in E-step. The two steps of EM are shown as follows:

$$\text{E-step: } Q(\theta | \theta^{(n)}) = E_{Z|x,\theta^{(n)}}[logL(\theta; x, Z)] \tag{7}$$

$$\text{M-step: } \theta^{(n+1)} = \arg\max_{\theta} Q(\theta | \theta^{(n)}) \tag{8}$$

where $L(\theta; X, Z) = \Pr(X, Z | \theta)$ is the likelihood function, θ is the estimation parameter of the model, X is the observed data and Z is a set of latent variables.

Now let us consider how to solve the MLE problem we formulated in the previous section by developing a hardness-aware EM scheme. First, we need to define the likelihood function of the MLE problem. In particular, the observed data

X in our problem is the Sensing Matrix SC and the Hardness Vector H. The estimation parameter vector is defined as $\theta = (T_1^k, T_2^k, ..., T_M^k; F_1^k, F_2^k, ..., F_M^k; d_k)$ where $k = 1, ..., K$ and T_i^k, F_i^k and d_k are defined in Equation (4) and (5). Furthermore, we need to define a vector of latent variables Z to indicate whether a claim is true or false. More specially, we have a corresponding variable z_j^k for claim C_j (whose hardness degree is k) such that $z_j^k = 1$ if C_j is true and $z_j^k = 0$ otherwise. Additionally, we define a set of binary indication variables h_j^k such that $h_j^k = 1$ if $h_j = k$ in Hardness Vector H and $h_j^k = 0$ otherwise. Hence, the likelihood function of hardness aware truth discovery problem can be written given as:

$$
\begin{aligned}
&L(\theta; X, Z) = \Pr(X, Z|\theta) \\
&= \prod_{j=1}^{N} \Bigg\{ \prod_{k=1}^{K} \Bigg[\prod_{i=1}^{M} (T_i^k)^{S_iC_j \ \&\& \ h_j^k} \\
&\times (1 - T_i^k)^{(1-S_iC_j) \ \&\& \ h_j^k} \times d_k \times z_j^k \Bigg] \\
&+ \prod_{l=1}^{L} \Bigg[\prod_{i=1}^{M} \prod_{k=1}^{K} (F_i^k)^{S_iC_j \ \&\& \ h_j^k} \\
&\times (1 - F_i^k)^{(1-S_iC_j) \ \&\& \ h_j^k} \times (1 - d_k) \times (1 - z_j^k) \Bigg] \Bigg\}
\end{aligned}
\tag{9}
$$

where $S_iC_j = 1$ when source S_i reports C_j to be true and 0 otherwise. The "&&" represents the "AND" logic for binary variables. The likelihood function represents the likelihood of the observed data (i.e., SC and H) and the values of hidden variables (i.e., Z) given the estimation parameters (i.e., θ).

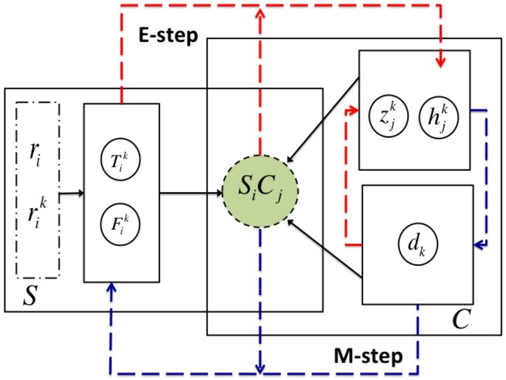


Figure 1. The E and M steps of HA-EM Scheme

We can then derive the E and M steps of HA-EM scheme using EM algorithm based on Equation (8). The E and M steps of HA-EM are shown in Figure 1. The detailed derivations are presented in Section VIII (Appendix). The final solutions of the estimation parameters are:

$$
\begin{aligned}
(T_i^k)^{(n+1)} &= \frac{\Sigma_{j \in SC_i^k} \Pr(z_j^k = 1 | X_j^k, \theta^{(n)})}{\Sigma_{j \in C^k} \Pr(z_j^k = 1 | X_j^k, \theta^{(n)})} \\
(F_i^k)^{(n+1)} &= \frac{\Sigma_{j \in SC_i^k} (1 - \Pr(z_j^k = 1 | X_j^k, \theta^{(n)}))}{\Sigma_{j \in C^k} (1 - \Pr(z_j^k = 1 | X_j^k, \theta^{(n)}))} \\
(d_k)^{(n+1)} &= \frac{\Sigma_{j \in C^k} \Pr(z_j^l = 1 | X_j^l, \theta^{(n)})}{|C^k|}
\end{aligned}
\tag{10}
$$

where SC_i^k is the set of claims (with hardness degree k) that source S_i reports. We also define C^k as the set of claims whose hardness degree is k.

B. Summary of The Hardness-Aware EM Algorithm

Algorithm 1 Hardness-Aware EM Algorithm

```
Input: Sensing Matrix $SC$, Hardness Vector $H$
Output: Estimations of Source Reliability and Claim Correctness
Initialize $\theta$ ($T_i^k = s_i^k, F_i^k = 0.5 \times s_i^k$, $d^l$ =Random number in $(0, 1)$)
$n = 0$
repeat
    $n = n + 1$
    for Each $k \in \{1, 2, ..., K\}$ do
        for Each $j \in C$ do
            compute $\Pr(z_j^k = 1 | X_j^k, \theta^{(n)})$
        end for
        for Each $i \in S$ do
            compute $(T_i^k)^{(n)}, (F_i^k)^{(n)}, (d_k)^{(n)}$
        end for
    end for
until $\theta^{(n)}$ and $\theta^{(n-1)}$ converge
Let $(Z_j^k)^c$ = converged value of $\Pr(z_j^k = 1 | X_j^k, \theta^{(n)})$
for Each $k \in \{1, 2, ..., L\}$ do
    for Each $j \in C$ do
        if $(Z_j^k)^c \geq 0.5$ then
            claim $C_j^l$ is true
        else
            claim $C_j^l$ is false
        end if
    end for
    for Each $i \in S$ do
        calculate $(r_i^k)^*$ from converge values of $(T_i^k)$, $(F_i^k)$ and $(d_k)$
        based on Equation (5)
        calculate $r_i{}^*$ form $(r_i^k)^*$ based on Equation (3)
    end for
end for
```

In summary, the input of the HA-EM algorithm is the Sensing Matrix SC and Hardness Vector H obtained from the social sensing data. The output is the maximum likelihood estimation of estimation parameters and latent variables. The estimation results can be used to compute both source reliability and claim correctness. We summarize the HA-EM scheme in Algorithm 1.

V. Evaluation

In this section, we evaluate the HA-EM scheme using three real world case studies based on Twitter. We choose Twitter as our social sensing application example because it creates an ideal scenario where unreliable content with rich information are collected from unvetted data sources (e.g., people report observations of different hardness degrees on Twitter) [2]. In our evaluation, we compare *HA-EM* to five representative baselines from current literature. The first baseline is *Voting*, which computes the data credibility simply by counting the number of times the same tweet is repeated on Twitter. The second baseline is the *Sums*, which explicitly considers the difference in source reliability when it computes the data credibility scores [13]. The third baseline is *Average_Log*, which explicitly considers both source reliability and the number of claims the source report [17]. The fourth baseline is TruthFinder which used a pseudo-probabilistic model to represent the interdependence between source reliability and claim correctness [34]. The fifth baseline is the *Regular EM*, which was shown to outperform four current truth discovery schemes in social sensing [30].

We have implemented the HA-EM scheme and other baselines in Apollo system, a social sensing platform that we have developed to collect tweets from Twitter and track the unfolding of real world events based on the collected tweets [4]. Examples of such events include terrorist attack, hurricane, earthquake, civil unrest and other natural and man-made disasters. Specifically, Apollo has: (i) a data collection front-end that allows users to collect tweets by specifying a set of keywords and/or geo-locations and log the collected tweets; (ii) a data pre-processing component that efficiently clusters similar tweets into the same cluster by using micro-blog data clustering methods [23].

Using the meta-data output by the data pre-processing component of Apollo, we first generated the Sensing Matrix SC by taking the Twitter users as the data sources and the clusters of tweets as the the statements of user's observations, hence representing the *claims* in our model. We then initialized the values of claims using a simple *domain classifier* that can classify the claims into easy and hard categories based on the content of the tweets. In particular, the domain classifier was built using URL identification such as "http" or "https" commonly found in tweets on Twitter. Each cluster of claims was checked to see if it contained more than one URL and if so it was classified as easy. Otherwise, the claim was classified as hard. The rationale is the claims with URLs are more likely to be the repeated information from other external sources (e.g., news websites), hence are easy to make while claims without URLs are more likely to be made by the users themselves.

One important note is that the above classifier is far from being perfect due to its heuristic nature: it may mis-classify easy claims as hard and vice versa. One goal of our evaluation is to show that our HA-EM scheme can actually achieve a significant performance improvement in truth discovery compared to the state-of-the-art solutions even given this rough and noisy estimation on claim hardness degrees.

For the purposes of evaluation, we selected three real world Twitter data traces, which were collected during the events that happened in 2015. The first trace was collected by Apollo during the *Oregon Shooting* event that happened on October 1, 2015, which caused 10 death including the gunman. It is the deadest event in Oregon's history. The second was collected during *Baltimore Riots* event that happened on April 14, 2015, which were a series of riots that followed the suspicious death of an African American male, Freddie Gray while in police custody. The riots caused several important events to be canceled and a state of emergency declared in the city of Baltimore. The third trace was during *Paris Attacks* event that happened on November 13, 2015, which were a series of terrorist attacks that left 130 dead including at the Bataclan theatre where many were taken hostage. It is the worst terrorist attack to occur in Europe in 11 years. The three data traces are summarized in Table III.

Trace	Oregon Shooting	Baltimore Riots	Paris Attacks
Start Date	October 1, 2015	April 14, 2015	November 13, 2015
Time duration	6 days	17 days	11 days
Physical Location	Umpqua Community College, OR	Baltimore, MD	Paris, France
# of tweets	210,028	952,442	873,760
# of users tweeted	122,069	425,552	496,753

Table III. DATA STATISTICS OF THREE TRACES

We randomly sampled 2020, 1850, and 1620 tweets from the Oregon Shooting, Paris Attacks, and Baltimore Riots data trace respectively for our evaluation [1]. We fed the sampled tweets to the Apollo tool and ran all compared truth discovery schemes on the sampled data. We manually graded all claims using the following rubric:

- *True claims:* Claims that are statements of a physical or social event, which is generally observable by multiple independent observers and corroborated by credible sources external to Twitter (e.g., mainstream news media).
- *False claims:* Claims that do not satisfy the requirement of true claims.

We note that the false claims may include some possibly true claims that cannot be independently verified by external sources. Hence, our evaluation provides *pessimistic* performance bounds on the estimates. Specifically, we focus on the estimation performance of different schemes in terms of correctly identifying the *true claims* because they consist of the actually useful information for social sensing applications. In particular, we used the following evaluation metric to evaluate the performance of all schemes in terms of identifying the correct true claims: Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$, Precision $= \frac{TP}{TP+FP}$, Recall $= \frac{TP}{TP+FN}$ and F1-measure $= \frac{2\times Precison\times Recall}{Precison+Recall}$. The TP, TN, FP and FN represent true positives, true negatives, false positives and false negatives of the classification results.

Figure 2 shows the result for the Baltimore Riots trace. We observe that the HA-EM scheme outperforms difficulty-ignorant truth discovery schemes in identifying more truthful claims and keeping the falsely reported claims least. This is achieved by explicitly incorporating the emotional into the maximum likelihood estimation framework. The performance gain of HA-EM scheme compared to the best performed baseline is significant: *31%* in accuracy, *9%* in precision, *49%* in recall, and *29%* in F1. The high performance gain in recall is achieved by correctly identifying many hard truthful claims that were misidentified as false by other hardness ignorant schemes.

We carried out further experiments to evaluate how newsworthy and important the true claims identified by different algorithms are. Specifically, we independently collected 10 important events reported by media during the Baltimore Riots event to see if they are captured in our true claims. We then

[1] This is because of the limited human power to label the ground truth for all tweets.

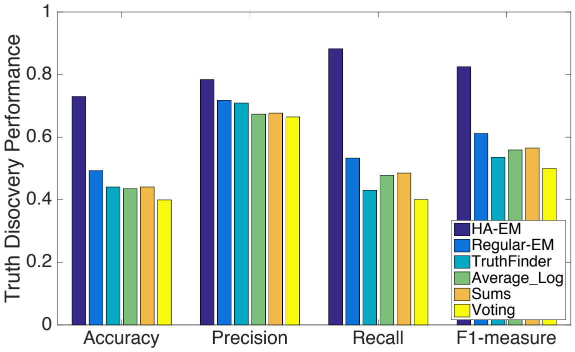


Figure 2. Evaluation on Baltimore Riots Trace

scanned through the true claims identified by each of the algorithms compared to find these events. Results are shown in Table IV (due to space limit, we only showed the comparison between the HA-EM scheme and the best performed baseline: Regular-EM). We observed that all ten events were covered by the true claims from the HA-EM scheme while two of them were missing from the true claims returned from the Regular-EM scheme. This result shows that the truthful claims identified by the HA-EM scheme are more newsworthy and potentially have higher impacts.

We repeated the above experiments on the Oregon Shooting and Paris Attack traces. The results are shown in Figure 3 and Figure 4. We observe that the HA-EM scheme continues to achieve the best performance compared to all baselines in terms of correctly identifying truthful claims. In the Oregon Shooting, the performance gain of HA-EM scheme compared to the best performed baseline is significant: *22%* in accuracy, *9%* in precision, *42%* in recall, and *20%* in F1-measure. The results on Paris Attack trace is similar: *27%* in accuracy, *8%* in precision, *35%* in recall, and *21%*. For the newsworthy events coverage, collecting 10 media events that happened during the Oregon Shooting and Paris Attack events respectively, we observed that the HA-EM found 9 and 10 of them, compared to 6 and 7 found by the best performed baseline. Due to the space limit, we do not show the detailed results here.

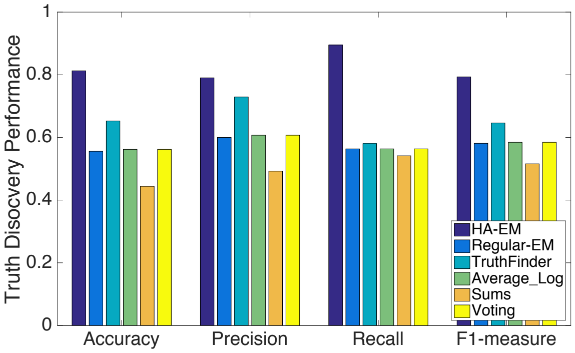


Figure 3. Evaluation on Oregon Shooting Trace

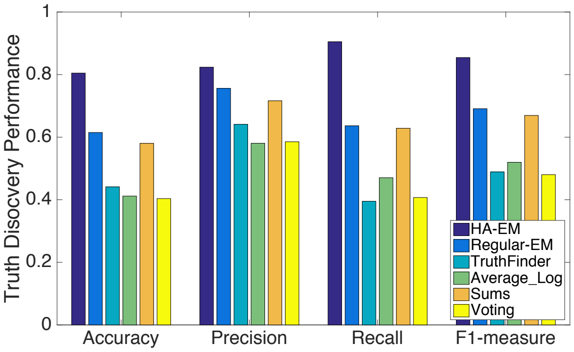
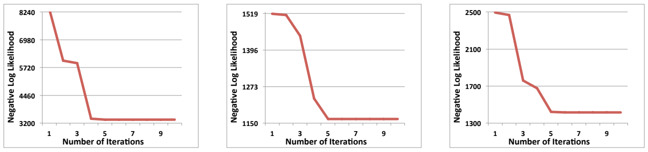


Figure 4. Evaluation on Paris Terrorist Attacks Trace

Finally, we also perform the convergence analysis of the HA-EM scheme. In particular, we studied how the value of negative log-likelihood function (defined in Equation (11)) changes w.r.t to the number of iterations. The results are presented in Figure 5. We observe the HA-EM scheme converges within a few iterations on both data traces. The encouraging results from real world data traces validate the effectiveness of using the HA-EM scheme to obtain more truthful information in a real world social sensing application by explicitly exploiting the emotional information of claims.



(a) Baltimore Riots Trace (b) Oregon Shooting Trace (c) Paris Attacks Trace

Figure 5. Convergence Analysis of HA-EM

VI. Discussions and Future Work

First, we assumed sources to be independent in our current framework. However, sources may be dependent in some social sensing applications, especially when they are connected through social networks. Several analytical models have been recently developed to address non-independent sources in social sensing using estimation theory [28] and machine learning techniques [21]. It is reasonable to integrate these techniques with our HA-EM scheme to explicitly model source dependency in the MLE framework. Furthermore, sources may have different expertise and report claims with different reliability. For example, a civil engineer might be very reliable in reporting the damage of buildings but might not be equally reliable in reporting the habitat of birds. In our current model, source reliability is represented by a scalar variable, which is limited to representing the source reliability on a single dimension. One possible solution is to generalize the source reliability definition from a scalar to a vector, where each dimension of the vector represents the reliability in a particular knowledge domain.

Second, we did not assume dependency between claims. However, reports on different claims might be inherently

#	Media	Tweet found by Hardness-EM	Tweet found by the Best Baseline
1	The post-funeral demonstrations became more tumultuous as the afternoon wore on, with a police car and van being torched and several storefront windows broken.	RT @BrianToddCNN: A police car and van burned on the streets of Baltimore. #BaltimoreRiots http://t.co/Q4k6W9oQLK	RT @BrianToddCNN: A police car and van burned on the streets of Baltimore. #BaltimoreRiots http://t.co/Q4k6W9oQLK
2	A mother in Baltimore caught her son, whom she suspected of rioting, and hit him. She sent him home on live TV.	VIDEO: Mother seen dragging son away from #BaltimoreRiots. http://t.co/gKkZj6sg2j http://t.co/7V0b26SUgj:	RT @Independent: Furious mother marches her son home from Baltimore riots live on TV http://t.co/OiSbX4m4uy http://t.co/cpOSoFC3h6
3	Mayor Stephanie Rawlings-Blake defended her handling of the recent rioting in Baltimore, arguing Tuesday that more aggressive police or military tactics could have escalated the violence	Mayor Stephanie Rawlings-Blake defended her handling of the recent rioting in Baltimore, arguing Tuesday that more aggressive police or military tactics could have escalated the violence	**MISSING**
4	A CVS pharmacy, which had been looted after its windows were smashed, was then set ablaze	cnni: The CVS destroyed in #BaltimoreRiots – complete devastation. More photos as we get them here: http://t.co/YtmVvz53Nm	**MISSING**
5	“”There’s no excuse for the kind of violence that we saw yesterday. It is counterproductive,” Obama said at a press conference from the White House.	CNN: .@BarackObama on Baltimore unrest: ”No excuse for the kind of violence we saw yesterday.” http://t.co/tmH7Kq2otl	CNN: .@BarackObama on Baltimore unrest: ”No excuse for the kind of violence we saw yesterday.” http://t.co/tmH7Kq2otl
6	The American Heart Association announced last night the cancellation of a medical conference in Baltimore due to the unrest in the city.	American Heart Association cancels Baltimore conference: http://t.co/DjkbXJ7P6b by @cardiobrief:	American Heart Association cancels Baltimore conference: http://t.co/DjkbXJ7P6b by @cardiobrief:
7	The Baltimore Orioles postponed a second straight game against the Chicago White Sox on Tuesday after a night of rioting near Camden Yards.	Orioles postpone game vs. White Sox amid riots in Baltimore http://t.co/wNjKaGP025:	Orioles postpone game vs. White Sox amid riots in Baltimore http://t.co/wNjKaGP025:
8	The Baltimore mayor’s office said earlier Tuesday there were 144 vehicle fires, 15 structure fires and nearly 200 arrests in the unrest Monday.	200 arrests, 144 car fires, 15 buildings burned...” http://t.co/61mCjmEMws:	“200 arrests, 144 car fires, 15 buildings burned...” http://t.co/61mCjmEMws:
9	The remarks about giving space to ”those who wished to destroy” generated swift, strong criticism amid more than two dozen arrests, at least 15 police officers injured, and looting and arson in the city.	Baltimore Mayor Stephanie Rawlings-Blake Under Fire For ’Space’ to Destroy Comment	**MISSING**
10	Volunteers and business owners clean up after an evening of riots following the funeral of Freddie Gray on Tuesday.	nytimes: Volunteers pick up broken glass after a night of riots in Baltimore http://t.co/Sx7d2YZIIw (Photo: A.J. Chavar/NYT)	nytimes: Volunteers pick up broken glass after a night of riots in Baltimore http://t.co/Sx7d2YZIIw (Photo: A.J. Chavar/NYT)

Table IV. Ground truth events and related claims found by Hardness-EM vs The Best Performed Baseline (Regular-EM) in Baltimore Riots

correlated. For example, the average speed of segments on the same road normally have similar distributions. The hurricane risk predictions of communities in the same neighborhood are usually highly correlated. Hence it is important to understand how to appropriately incorporate the claim dependency into our MLE framework. Several recent techniques have been developed to model the dependency between claims and take such dependency as prior knowledge in their solutions [18], [26]. Inspired by these results, we will further extend the HA-EM scheme to incorporate the claim dependency (represented by the joint distribution between correlated claims) into the likelihood function and derive a claim-dependency-aware solution. The general guideline of derivation should be similar as the one presented in Section IV.

Third, the ground truth value of a claim was assumed to be time-invariant in our current framework. This assumption holds in the social sensing applications where the states of the claim variables do not change in the observation period [26], [30]. However, in systems where the state of the environment may change quickly over time, it is important to investigate the dynamics of the claim variables as well. Recently, we have developed an extended MLE framework to explicitly handle time variant claims in social sensing [31]. Such extension can be easily integrated with the HA-EM scheme since they use the same underlying MLE framework. Moreover, we focus on binary claims in this paper. This assumption is sufficient in many social sensing applications where the states of the reported event can be represented by a Boolean variables (e.g., litter exists in a given location or not). However, our model can also be easily extended to handle the case where claims have arbitrary discrete values. The authors have recently made some progress in this direction [27]. The key idea is to extend the estimation parameter of our MLE model to cover all possible states of the claim. The general outline of the HA-EM derivation still holds.

VII. Conclusion

This paper develops a hardness-aware maximum likelihood estimation framework to solve the truth discovery problem in social sensing applications. The proposed HA-EM scheme explicitly incorporates the claim hardness into a rigorous analytical framework. The proposed approach jointly estimates both source reliability and claim correctness using an expectation maximization algorithm. We evaluated the HA-

EM scheme through three real world case studies in social sensing applications. The results showed HA-EM achieved non-trivial performance gains in improving the truth discovery accuracy compared to the Regular-EM and other state-of-the-art techniques that ignored the claim hardness in their solutions. The results of the paper is important because it lays out an analytical foundation to exploit different degrees of claim hardness in social sensing using a principled approach.

References

[1] T. Abdelzaher and D. Wang. Analytic challenges in social sensing. In *The Art of Wireless Sensor Networks*, pages 609–638. Springer, 2014.

[2] C. C. Aggarwal and T. Abdelzaher. Social sensing. In *Managing and Mining Sensor Data*, pages 237–297. Springer, 2013.

[3] H. Ahmadi, T. Abdelzaher, J. Han, N. Pham, and R. Ganti. The sparse regression cube: A reliable modeling technique for open cyber-physical systems. In *Proc. 2nd International Conference on Cyber-Physical Systems (ICCPS'11)*, 2011.

[4] Apollo-Toward Fact-finding for Social Sensing. http://apollo.cse.nd.edu/.

[5] I. Boutsis and V. Kalogeraki. Privacy preservation for participatory sensing data. In *IEEE International Conference on Pervasive Computing and Communications (PerCom)*, volume 18, page 22, 2013.

[6] Y. Chon, N. D. Lane, F. Li, H. Cha, and F. Zhao. Automatically characterizing places with opportunistic crowdsensing using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 481–490. ACM, 2012.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

[8] S. B. Eisenman et al. The bikenet mobile sensing system for cyclist experience mapping. In *SenSys'07*, November 2007.

[9] M. Gupta and J. Han. Heterogeneous network-based trust analysis: a survey. *ACM SIGKDD Explorations Newsletter*, 13(1):54–71, 2011.

[10] J.-H. Huang, S. Amjad, and S. Mishra. CenWits: a sensor-based loosely coupled search and rescue system using witnesses. In *SenSys'05*, pages 180–191, 2005.

[11] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender systems: an introduction*. Cambridge University Press, 2010.

[12] L. Kaplan, M. Scensoy, and G. de Mel. Trust estimation and fusion of uncertain information by exploiting consistency. In *Information Fusion (FUSION), 2014 17th International Conference on*, pages 1–8. IEEE, 2014.

[13] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[14] M. Leng and Y.-C. Wu. Low-complexity maximum-likelihood estimator for clock synchronization of wireless sensor nodes under exponential delays. *Signal Processing, IEEE Transactions on*, 59(10):4860–4870, 2011.

[15] N. Maisonneuve, M. Stevens, M. E. Niessen, P. Hanappe, and L. Steels. Citizen noise pollution monitoring. In *Proceedings of the 10th Annual International Conference on Digital Government Research: Social Networks: Making Connections between Citizens, Data and Government*, pages 96–103. Digital Government Society of North America, 2009.

[16] S. Nath. Ace: Exploiting correlation for energy-efficient and continuous context sensing. In *Proceedings of the tenth international conference on Mobile systems, applications, and services (MobiSys'12)*, 2012.

[17] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *International Conference on Computational Linguistics (COLING)*, 2010.

[18] J. Pasternack and D. Roth. Generalized fact-finding (poster paper). In *World Wide Web Conference (WWW'11)*, 2011.

[19] S. S. Pereira, R. Lopez-Valcarce, et al. A diffusion-based em algorithm for distributed estimation in unreliable sensor networks. *Signal Processing Letters, IEEE*, 20(6):595–598, 2013.

[20] D. Pomerantz and G. Dudek. Context dependent movie recommendations using a hierarchical bayesian model. In *Advances in Artificial Intelligence*, pages 98–109. Springer, 2009.

[21] G.-J. Qi, C. C. Aggarwal, J. Han, and T. Huang. Mining collective intelligence in diverse groups. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1041–1052. International World Wide Web Conferences Steering Committee, 2013.

[22] Sense Networks. Cab Sense. http://www.cabsense.com.

[23] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. 2005.

[24] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448–456. ACM, 2011.

[25] D. Wang, T. Abdelzaher, H. Ahmadi, J. Pasternack, D. Roth, M. Gupta, J. Han, O. Fatemieh, and H. Le. On bayesian interpretation of fact-finding in information networks. In *14th International Conference on Information Fusion (Fusion 2011)*, 2011.

[26] D. Wang, T. Abdelzaher, L. Kaplan, R. Ganti, S. Hu, and H. Liu. Exploitation of physical constraints for reliable social sensing. In *The IEEE 34th Real-Time Systems Symposium (RTSS'13)*, 2013.

[27] D. Wang, M. Amin, T. Abedlzaher, D. Roth, C. Voss, L. Kaplan, S. Tratz, J. Laoudi, and D. Briesch. Provenance-assisted classification in social networks. *IEEE Journal of Selected Topics in Signal Processing (J-STSP)*, 2014.

[28] D. Wang, M. T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. C. Aggarwal, R. Ganti, et al. Using humans as sensors: an estimation-theoretic perspective. In *Proceedings of the 13th international symposium on Information processing in sensor networks*, pages 35–46. IEEE Press, 2014.

[29] D. Wang, L. Kaplan, T. Abdelzaher, and C. C. Aggarwal. On credibility tradeoffs in assured social sensing. *IEEE Journal On Selected Areas in Communication (JSAC)*, 2013.

[30] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher. On truth discovery in social sensing: A maximum likelihood estimation approach. In *The 11th ACM/IEEE Conference on Information Processing in Sensor Networks (IPSN 12)*, April 2012.

[31] S. Wang, D. Wang, L. Su, T. Abdelzaher, and L. Kaplan. Towards cyber-physical systems in social spaces: The data reliability challenge. In *The IEEE 35th Real-Time Systems Symposium (RTSS'14)*, 2014.

[32] X. Wang, M. Fu, and H. Zhang. Target tracking in wireless sensor networks based on the combination of kf and mle using distance measurements. *Mobile Computing, IEEE Transactions on*, 11(4):567–576, 2012.

[33] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, (6):52–59, 2012.

[34] 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.*, 20:796–808, June 2008.

[35] X. Yin and W. Tan. Semi-supervised truth discovery. In *WWW*, New York, NY, USA, 2011. ACM.

[36] D. Yu, H. Huang, T. Cassidy, H. Ji, C. Wang, S. Zhi, J. Han, C. Voss, and M. . Magdon-Ismail. The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding. In *The 25th International Conference on Computational Linguistics*, 2014.

[37] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. *Proc. VLDB Endow.*, 5(6):550–561, Feb. 2012.

[38] Z. Zhao, J. Cheng, and W. Ng. Truth discovery in data streams: A single-pass probabilistic approach. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1589–1598. ACM, 2014.

VIII. Appendix

In appendix, we present the details of derivation of E and M steps of HA-EM scheme. First, we derive the Q function for the E-step given by Equation (7) as follows:

$$
\begin{aligned}
&Q(\theta|\theta^{(n)}) = E_{Z|X,\theta^{(n)}}[logL(\theta; X, Z)] \\
&= \sum_{j=1}^{N} \Bigg\{ \sum_{k=1}^{K} \Pr(z_j^k = 1|X_j^k, \theta^{(n)}) \\
&\times \Bigg[\sum_{i=1}^{M} (S_iC_j \&\& \ h_j^k) \times logT_i^k \\
&+ \Big((1 - S_iC_j) \ \&\& \ h_j^k \Big) \times log(1 - T_i^k) + logd_k \Bigg] \\
&+ \sum_{k=1}^{K} \Pr(z_j^k = 0|X_j^k, \theta^{(n)}) \\
&\times \Bigg[\sum_{i=1}^{M} (S_iC_j \&\& \ h_j^k) \times logF_i^k \\
&+ \Big((1 - S_iC_j) \ \&\& \ h_j^k \Big) \times log(1 - F_i^k) + log(1 - d_k) \Bigg] \Bigg\}
\end{aligned} \tag{11}
$$

where X_j^k represents the observed data that is about claim C_j with hardness degree of k. $\Pr(z_j^k = 1|X_j^k, \theta^{(n)}$ represents the probability of claim C_j to be true given X_j^k and current estimation of the parameter θ.

We define $Z(n, j, k) = p(z_j^k = 1|X_j^k, \theta^{(n)})$. It is the conditional probability of the claim C_j (whose hardness degree k) to be true given the observed data X_j^k and current estimate of θ. $Z(n, j, k)$ can be further expressed as:

$$
\begin{aligned}
Z(n, j, k) &= p(z_j^k = 1|X_j^k, \theta^{(n)}) \\
&= \frac{p(z_j^k = 1; X_j^k, \theta^{(n)})}{p(X_j^k, \theta^{(n)})} \\
&= \frac{A(n, j, k) \times (d^k)^{(n)}}{A(n, j, k) \times (d^k)^{(n)} + B(n, j, k) \times (1 - (d^k)^{(n)})}
\end{aligned} \tag{12}
$$

where $A(n, j, k)$ and $B(n, j, k)$ are defined as follows:

$$
\begin{aligned}
A(n, j, k) &= p(X_j^k, \theta^{(n)}|z_j^k = 1) \\
&= \prod_{i=1}^{M} \Big\{ \prod_{k=1}^{K} (T_i^{k,k})^{S_iC_j \ \&\& \ h_j^k} \\
&\times (1 - \sum_{k=1}^{K} T_i^{k,k})^{(1 - S_iC_j) \ \&\& \ w_j^k} \Big\} \\
B(n, j, k) &= p(X_j^k, \theta^{(n)}|z_j^k = 0) \\
&= \prod_{i=1}^{M} \Big\{ \prod_{k=1}^{K} (F_i^{k,k})^{S_iC_j \ \&\& \ h_j^k} \\
&\times (1 - \sum_{k=1}^{K} F_i^{k,k})^{(1 - S_iC_j) \ \&\& \ h_j^k} \Big\}
\end{aligned} \tag{13}
$$

Next we simplify Equation (11) by replacing the conditional probability of $p(z_j^k = 1|X_j^k, \theta^{(n)})$ with $Z(n, j, k)$.

$$
\begin{aligned}
&Q(\theta|\theta^{(n)}) \\
&= \sum_{j=1}^{N} \Bigg\{ \sum_{k=1}^{k} Z(n, j, k) \times \Bigg[\sum_{i=1}^{M} \sum_{k=1}^{K} (S_iC_j \ \&\& \ h_j^k) \times logT_i^k \\
&+ \Big((1 - S_iC_j) \ \&\& \ h_j^k \Big) \times log(1 - \sum_{k=1}^{K} T_i^k) + logd^k \Bigg] \\
&+ \sum_{k=1}^{k} \Big(1 - Z(n, j, k) \Big) \times \Bigg[\sum_{i=1}^{M} \sum_{k=1}^{K} (S_iC_j \ \&\& \ h_j^k) \times logF_i^k \\
&+ \Big((1 - S_iC_j) \ \&\& \ h_j^k \Big) \times log(1 - \sum_{k=1}^{K} F_i^k) + log(1 - d^k) \Bigg] \Bigg\}
\end{aligned} \tag{14}
$$

For the M-step, as we discussed earlier in Section IV, we set partial derivatives of $Q(\theta|\theta^{(n)})$ with respect to θ to 0 in order to get optimal $(T_i^k)^*$, $(F_i^k)^*$ and $(d^k)^*$:

$$
\begin{aligned}
&\sum_{j=1}^{N} \Bigg[Z(n, j, k) \times \Big((S_iC_j \ \&\& \ h_j^k \Big) \times \frac{1}{(T_i^k)^*} \\
&- \Big((1 - S_iC_j) \ \&\& \ h_j^k \Big) \times \frac{1}{1 - \sum_{k=1}^{K} (T_i^k)^*} \Bigg] = 0 \\
&\sum_{j=1}^{N} \Bigg[\Big(1 - Z(n, j, k) \Big) \times \Big((S_iC_j \ \&\& \ r_{ij}^k \ \&\& \ w_j^k \Big) \times \frac{1}{(F_i^k)^*} \\
&- \Big((1 - S_iC_j) \ \&\& \ w_j^k \Big) \times \frac{1}{1 - \sum_{k=1}^{K} (F_i^k)^*} \Bigg] = 0 \\
&\sum_{j=1}^{N} \Bigg[Z(n, j, k) \times M \times \frac{1}{(d^k)^*} \times h_j^k - \Big(1 - Z(n, j, k) \Big) \\
&\times M \times \frac{1}{1 - (d^k)^*} \times whj^k \Bigg] = 0
\end{aligned} \tag{15}
$$

Solving the above equations, we can obtain the results of M step presented in Equation (10).