ON QUANTIFYING THE QUALITY OF INFORMATION IN SOCIAL SENSING

BY

DONG WANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Professor Tarek Abdelzaher, Chair
Professor Jiawei Han
Professor Thomas Huang
Dr. Charu C. Aggarwal, IBM T. J. Watson Research Center

Abstract

This thesis develops the fundamental theory and methodology for quantifying the Quality of Information (QoI) in social sensing. We refer social sensing to the sensing applications where humans play a critical role in the sensing or data collection process. Social sensing has emerged as a new paradigm for sensory data collection, which is motivated by the proliferation of mobile platforms equipped with a variety of sensors (e.g., GPS, camera, microphone, motion and etc.) in the possession of common individuals, networking capabilities that enable fast and convenient data sharing (e.g., WiFi and 4G) and large-scale dissemination of opportunities (Twitter, Flicker and etc.). A significant challenge in social sensing applications lies in ascertaining the correctness of collected data and the reliability of information sources. We call this challenge *QoI quantification* in social sensing. Unlike the case with well-calibrated and well-tested infrastructure sensors, humans are less reliable. The term, participant (or source) *reliability* is used to denote the probability that the participant reports correct observations. Reliability may be impaired because of poor used sensor quality, lack of sensor calibration, lack of (human) attention to the task, or even intent to deceive. Moreover, data collection is often open to a large population, where it is impossible to screen all participants beforehand. The likelihood that a participant's measurements are correct is usually unknown *a priori*. Consequently, it is very challenging to ascertain the correctness of the collected data from unreliable sources with unknown reliability. Meanwhile, it is also challenging to ascertain the reliability of each information source without knowing whether their collected data are true or not. Therefore, the main questions posed in this thesis are: i) whether or not we can determine, in an optimal way, given only the measurements collected and without knowing the reliability of sources, which of the reported observations are true and which are not? ii) whether a source (participant) is reliable or not? iii) how to quantify the answers of the above questions?

The thesis answered the above questions by applying the key insights from estimation theory and data fusion to come up with new theories to accurately quantify both the participant reliability and correctness

of their observations for social sensing applications. Contrary to a large amount of literature in data mining and machine learning that use various kinds of heuristics whose inspiration can be traced back to Google's PageRank to solve a similar trust analysis problem in information networks, our approach provides the first *optimal* solution to the OoI quantification problem in social sensing by casting it as one of expectation maximization (EM) and quantifies the estimation *confidence* using the Cramer-Rao lower bound (CRLB) from estimation theory. More specifically, this thesis addressed the QoI quantification challenge of social sensing from the following perspectives.

First, we developed an analytically-founded Bayesian interpretation of the basic fact-finding scheme that is popularly used in data-mining literature to rank both sources and their asserted information based on credibility values. Our method offers the first probability based semantics to interpret the credibility results output by the fact-finders. It leads to a direct quantification on both participant reliability and correctness of the observations they asserted. The Bayesian interpretation is an approximation scheme based the linearity assumption made by the basic fact-finders, which motivates our further efforts to find the optimal solution to the QoI quantification problem in social sensing.

Second, we developed a maximum likelihood estimator by intelligently casting the QoI quantification problem in social sensing into an expectation maximization problem that can be solved *optimally* and efficiently. The EM scheme overcomes the approximate limitation of Bayesian interpretation and a large amount of heuristics in trust analysis of information networks. It provides the first *optimal* solution (in the sense of maximum likelihood estimation) to jointly estimate participant reliability and the correctness of their reported measured variables in the way that is most consistent with the data collected.

Third, a key quantification metric that is missing in all previous fact-finding literature is the *confidence* quantification of the estimation results. Without such a confidence metric, the estimation results lack important bounds to correctly characterize their accuracy. Thanks to the expectation maximization formulation of the problem, we are able to exactly quantify the *confidence* in the maximum likelihood estimation of EM scheme. Specifically, we obtained both real and asymptotic confidence bounds of the participant reliability estimation based on the Cramer-Rao lower bound in estimation theory and studied their scalability and robustness limitations for different application scenarios.

Fourth, considering some simplifying assumptions we made in our original model, we extended the model and the maximum likelihood approach to address them. In particular, we extended the maximum likelihood estimator to solve the QoI quantification problem when *conflicting observations* exists and the measured variables are *non-binary*.

Fifth, given the original iterative EM algorithm may not be efficient for the streaming data, we proposed a recursive EM algorithm that can compute the estimation results on the fly. We evaluate the performance of the recursive EM algorithm over different tradeoff dimensions such as trustworthiness of sources, freshness of input data and timeliness of algorithm execution.

Finally, the developed theory above has been implemented and built into the core of *Apollo*, a data distillation service for social sensing applications [1]. Apollo is designed to filter the noisy social sensing data by leveraging the developed theory to jointly estimate the credibility of information sources and the observations made by them, then remove less credible observations. We evaluated the performance of Apollo through the real world social sensing applications that report the progression of several real events (e.g., Egypt Unrest, Hurricane Irene and etc.). Results demonstrated that Apollo effectively cleans out the input data and correctly identifies the true and important information from a large crowd of unreliable human sources.

To Father, Mother and Family.

Acknowledgments

First I would thank sincerely my adviser, Prof. Tarek Abdelzaher, for his great guidance, continuous support, and persistent encouragement through my Ph.D study. I came to the CS department at UIUC with EE background. My advisor showed great patience and offered instructive advice for me to brush up the background for research. I also appreciate his free and open policy in research. He always let me to choose the topic that is most interesting to me and offer me chances to try and fail. Under this free research atmosphere, I eventually found my thesis topic and make efforts to have this thesis. Without his guidance and persistent help, this would not have been possible.

It is my great fortune to have as illustrious names as Prof. Jiawei Han, Prof. Tom Huang, Dr. Chara C. Aggarwal on my thesis committee. Their advice and comments have been tremendously constructive and helpful. I really appreciate all their help and support.

Special thanks goes to Dr. Lance Kaplan from Army Research Lab (ARL). Dr. Kaplan offered very insightful advice for the direction of the thesis as well as many detailed comments on the technique contents. I really appreciated his great guidance and help on the thesis.

Research in this thesis was sponsored by was sponsored by NSF grant CNS 1040380 and the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. This work was partially supported by the LCCC and eLLIIT centers at Lund University, Sweden. 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. On a separate note, some of my early Ph.D work on energy management on heterogeneous sensor platforms was sponsored in part by Microsoft Research and in part by NSF grants CNS 06-26825, CNS 06-15301, and CNS 09-16028.

Many faculty, colleges and researchers have helped me during my Ph.D study. I sincerely thank Prof.

Klara Nahrstedt for her great patience and advice on my first CS course taken at UIUC. I would like to thank Dr. Feng Zhao, Dr. Jie Liu, Dr. Bodhi Priyantha, Dr. Hossein Ahmadi, Dr. Nam Pham, Dr. Jin Heo, Dr. Yursuf Sarwar, Dr. Maifi Khan, Dr. Hieu Le who collaborated with me on various projects and papers.

And finally, sincere thanks to my parents, Guangsen Wang and Haixia Wang, my grandfather Xuezeng Wang, my girlfriend Na Wei and my U.S family of Jessie Johnson, Jeremiah Johnson, Ben Johnson and Cecilia Johnson who always offer support and love to me during this challenging process.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

QoI	Quality of Information
MLE	Maximum Likelihood Estimation
EM	Expectation Maximization
CRLB	Cramer-Rao Lower Bound
TPM	Trusted Platform Module

List of Symbols

s	second (of time, generally).

Chapter 1

Introduction

1.1 Motivation and Challenges

This thesis develops theory and methodology for quantifying Quality of Information (QoI) in social sensing. Social sensing has emerged as a new paradigm for collecting sensory measurements by means of "crowd-sourcing" sensory data collection tasks to a human population. Regarding human awareness and involvement in the sensing process, social sensing in this thesis broadly refers to both participatory sensing and opportunistic sensing. With participatory sensing people explicitly and actively get involved in the sensing process and choose to perform some critical operations (e.g., decide what data to share) to meet the application requirement. With opportunistic sensing people are more passively involved or may even not be aware of the ongoing sensing process. Instead, some devices people own (e.g., smart phone) perform the required sensing task on behalf of their owners to meet the application requirement [2]. The paradigm of social sensing is made possible by the proliferation of a variety of sensors in the possession of common individuals, together with networking capabilities that enable data sharing. Examples include cell-phone accelerometers, cameras, GPS devices, smart power meters, and interactive game consoles (e.g., Wii). Individuals who own such sensors can thus engage in data collection for some purpose of mutual interest. A classical example is geotagging campaigns, where participants report locations of conditions in their environment that need attention (e.g., litter in public parks).

A significant challenge in social sensing applications lies in ascertaining the correctness of collected data and the reliability of information sources. Unlike the case with well-calibrated and well-tested infrastructure sensors, humans are less reliable. The term, participant (or source) *reliability* is used to denote the probability that the participant reports correct observations. Reliability may be impaired because of poor used sensor quality, lack of sensor calibration, lack of (human) attention to the task, or even intent to deceive. Moreover, data collection is often open to a large population, where it is impossible to screen all participants (or information sources) beforehand. The likelihood that a participant's measurements are correct is usually unknown *a priori*. Consequently, it is very challenging to ascertain the correctness of the collected data from unreliable sources with unknown reliability. Meanwhile, it is also quite challenging to ascertain the reliability of each information source without knowing whether their collected data are true or not. Therefore, the main questions posed in this thesis are: i) whether or not we can determine, given only the measurements sent and without knowing the reliability of sources, which of the reported observations are true and which are not? ii) how reliable each source is without independent ways to verify the correctness of their measurements? iii) how to quantify the answers to the above questions?

Existing techniques attempted to solve the similar trust analysis problem in social sensing and information networks by using extra hardware, applying application specific design or adopting heuristics from related fields such as data mining and machine learning. For example, Trusted Platform Module (TPM), commonly used in commodity PCs, provides a platform attestation approach that guards whether the data produced by sensor has been maliciously modified by others at the expense of additional hardware [3, 4]. The LiveCompare [5], a participatory sensing application used for comparison shopping of grocery products, works by transmitting product photos taken by participants of competing products, but does not automatically extract the price information from the photos. This application specific design is because of the fact that the price extraction process is known to be error-prone. Other more general approaches use heuristics whose inspiration can be traced by to Google's PageRank [6]. PageRank iteratively ranks the credibility of sources on the Web, by iteratively considering the credibility of sources who link to them. Extensions of PageRank, known as fact-finders, iteratively compute the credibility of sources and claims. Specifically, they estimate the credibility of claims from the credibility of sources that make them, then estimate the credibility of sources based on the credibility of their claims. Several algorithms exist that feature modifications of the above basic heuristic scheme [7, 8, 9, 10, 11]. In contrast, the QoI quantification theory and method developed in this thesis provide the first *optimal* solution (in the sense of maximum likelihood estimation) to the QoI quantification problem in social sensing by casting it as one of expectation maximization (EM) and quantifies the estimation *confidence* using the Cramer-Rao lower bound (CRLB) from estimation theory. Moreover, the developed theory can be applied to a wide range of social sensing applications in that it does not need special hardware support or application specific design.

1.2 Contributions

The thesis has made the following contributions on addressing the above challenges of quantifying the Quality of Information (QoI) in social sensing:

1. We developed an analytically-founded Bayesian interpretation of the basic fact-finding scheme that is popularly used in data-mining literature to rank both sources and their asserted information based on credibility values. This interpretation enabled the calculation of correct *probabilities* of conclusions resulting from information network analysis. Such probabilities constitute a measure of QoI, which can be used to directly quantify the participant reliability and measurement correctness in social sensing context.

2. Considering the approximation nature of fact-finding schemes, we developed a maximum likelihood estimator by intelligently casting the QoI quantification problem into an Expectation Maximization (EM) problem that can be solved *optimally* and efficiently. The EM algorithm makes inferences regarding both source reliability and measurement correctness by observing which observations coincide and which don't. It was shown to be very accurate in assessing measurement correctness as long as sources, on average, make multiple observations, and as long as some sources make the same observation.

3. We not only developed the maximum likelihood estimation (MLE), but also derived real and asymptotic *confidence bounds* on the participant reliability estimation of EM scheme. Our quantification approach leveraged the asymptotic normality of maximum likelihood estimation and computed the Cramer-Rao lower bound (CRLB) of the estimation parameters used in the EM scheme. We also studied the scalability and robustness limitations of the confidence bounds we derived.

4. As the basic model of EM scheme assumed only *corroborating observations* from participants, we extended the maximum likelihood estimator to solve the above quantification problem where *conflicting observations* exist. This effort was motivated by the fact that observations from different participants with unknown reliability sometime appear to be contradicting to each other (e.g., on-line review system). We developed an extended maximum likelihood estimator in the context of conflicting observations to address this problem. Another assumption of the basic EM model is that the measured variables were assumed to be *binary*. We thus generalized the theory for conflicting observations to handle *non-binary* measured variables as well.

5. The iterative EM scheme is mainly designed to run on static data sets, where the computation overhead stays reasonable even when the dataset scales up. However, such computation may become less efficient for streaming data because we need to re-run the algorithm on the whole dataset from scratch every time the dataset gets updated. We designed a recursive EM algorithm for the streaming data that only runs upon the updated dataset and combines the results with previously computed ones in a recursive way. The recursive EM algorithm was shown to achieve nice performance over different tradeoffs dimensions including the trustworthiness of sources, the freshness of input data and the timeliness of the algorithm execution.

6. Finally, the theory developed in this thesis has been implemented as the core of Apollo [1], a new sensing information processing tool to uncover likely facts from the noisy social sensing data. We demonstrated through several datasets collected from real events (e.g., Hurricane Irene and Egypt Unrest) that our tool was able to find important and newsworthy information from a huge amount of noisy data generated by people.

The above contributions differ from the state of the art in several respects. First, in contrast to a large volume of past work, where different heuristics are used to iteratively compute the credibility of sources and their claims [6, 8, 9, 10], our work is the first attempt to *optimally* solve the QoI quantification problem in social sensing and find the maximum likelihood estimate of both participant reliability and correctness of measurements. Second, while rich literature exists on information network analysis for the purpose of fact-finding, prior work does not offer an assessment of *quality of analysis results*. Prior literature stops at producing a hypothesis regarding source and claim correctness [7, 11, 12]. Third, we are the first to perform sensitivity analysis of fact-finder accuracy. Given the analytic quantification of confidence in fact-finding results, we are able to rigorously analyze how such confidence changes as a function of information network topology. Such sensitivity analysis offers a fundamental understanding of the capabilities and limitations of fact-finders.

1.3 Outline

The rest of the thesis is organized as follows: In Chapter 2, we review the related work in the relevant research fields. We describe the Bayesian Interpretation scheme in Chapter 3. We describe the maximum likelihood estimation approach based on EM in Chapter 4. In Chapter 5, we describe the real and asymptotic confidence bounds derived based on CRLB. We present the extended model and MLE approach to handle conflicting observations and non-binary variables in Chapter 6. In Chapter 7, we propose the recursive EM algorithm and evaluate its performance through several tradeoffs studies. Finally we conclude the thesis with some directions for future research in Chapter 9.

Chapter 2

Related Work

2.1 Social Sensing

Social sensing which is also referred to as human-centric sensing [13, 14], is generally achieved by various kinds of sensors which are closely attached to humans, either in their wearable form or in their mobile devices (e.g., cell phones). This new paradigm of sensing has received significant attention due to the great increase in the number of mobile sensors owned by individuals (e.g., smart phones with GPS, camera, etc.) and the proliferation of Internet connectivity to upload and share sensed data (e.g., WiFi and 4G networks). A broad overview of social sensing applications is presented in [15]. Some early applications include CenWits [16], a participatory sensor network to search and rescue hikers in emergency situations, CarTel [17], a vehicular sensor network for traffic monitoring and mitigation, CabSense [18], an application that analyzes GPS data from NYC taxis and help you find the best corner to catch a cab [19] and BikeNet [20], a bikers sensor network for sharing cycling related data and mapping the cyclist experience. In recent years, social sensing applications in healthcare has also become very popular. Numerous medical devices have been built with embedded sensors that can be used to monitor the personal health of patients, or send alters to the clinic or through the patient's social network when something unexpected happens. This can be used for activity recognition for emergent response [21], long term prediction about diseases [22, 23, 24], or other life style change effect on health [25, 26].

More recent work has focused on addressing new challenges emerging in social sensing applications such as preserving privacy of participants [27, 28], improving energy efficiency of sensing devices [29, 30] and building general models in sparse and multi-dimensional social sensing space [31, 32]. Examples include privacy-aware regression modeling, a data fusion technique that produce the same model as that computed from raw data by properly computing non-invertible aggregates of samples [27]. Authors in [28] gave special attention to preserving privacy over time series data based on the observation that sensor data stream typically comprises a correlated series of sampled data from some continuous physical phenomena. Acquisitional Context Engine (ACE) is a middleware that infers the unknown human activity attribute from known ones by exploiting the observation that the values of various human context attribute are limited by physical constraints and hence highly correlated [29]. E-Gesture is an energy efficient gesture recognition architecture that significantly reduces the energy consumption of mobile sensing device while keeping the recognition accuracy acceptable [30]. Sparse regression cube is a modeling technique that combine estimation theory and data mining techniques to enable reliable modeling at multiple degrees of abstraction of sparse social sensing data [31]. A further improved model to consider the data collection cost was proposed in [32].

Social sensing is often organized as "sensing campaigns" where participants are recruited to contribute their personal measurements as part of a large-scale effort to collect data about a population or a geographical area. Examples include documenting the quality of roads [33], the level of pollution in a city [34], or reporting garbage cans on campus [35]. In addition, social sensing can also be triggered by human source spontaneously without prior coordination to report socially important events. Examples include reporting Egypt Unrest, London Riots, Japan Tsunami and etc on Twitter. The spread of social networks such as Twitter and You-tube offers a forum for global and real-time sharing of reported data, which makes the reporting especially powerful. This type of application represents a very broad, distributed and collaborative sensing paradigm that takes the most versatile mobile platform, *the human user*, as sensors. Recent research attempts to understand the fundamental factors that affect the behavior of these emerging social sensing applications, such as analysis of characteristics of social networks [36], information propagation [37] and tipping points [38].

A more critical question about trustworthiness arises when the data of social sensing applications are collected through the operations of humans. Due to the nature of the social sensing application, the ability to contribute information is open. Such openness is a coin of two sides: one on hand, it greatly increases the availability of the information and the diversity of its sources. On the other hand, it introduces the problem of understanding the reliability of the contributing sources and ensuring the quality of the information collected. Trusted Platform Module (TPM), commonly used in commodity PCs, can be used to provide a certain level of assurance at the expense of additional hardware [4]. YouProve is a recent technique that relies on the trust analysis of the derived data to allow un-trusted client applications to verify the meaning of the source data is preserved [39]. Our work in this thesis is to construct a consistent model based on the social sensing data to measure the source reliability as well as the correctness of their responses. Compared to the existing approaches of trust analysis in social sensing, our approach does not need special hardware support or extra plug-ins at client side to perform complex analysis. Instead, we perform the trust analysis at the server side by building the likelihood function of the sensing data and provide quantifiable and confident estimation on both source reliability and the correctness of their responses.

2.2 Trustworthy Analysis in Information Networks

To assess the credibility of facts reported in information networks, a relevant body of work in the machine learning and data mining communities performs trust analysis. Hubs and Authorities [40] used a basic fact-finder where the belief in a claim c is $B(c) = \sum_{s \in S_c} T(s)$ and the truthfulness of a source s is $T(s) = \sum_{c \in C_s} B(c)$, where S_c and C_s are the sources asserting a given claim and the claims asserted by a particular source, respectively. Pasternack et al. extended the fact-finder framework by incorporating prior knowledge into the analysis and proposed several extended algorithms: *Average.Log, Investment, and Pooled Investment* [7]. Yin et al. introduced *TruthFinder* as an unsupervised fact-finder for trust analysis on a providers-facts network [8]. Other fact-finders enhanced the basic framework by incorporating analysis on properties or dependencies within claims or sources. Galland et al. [9] took the notion of hardness of facts into consideration by proposing their algorithms: *Cosine, 2-Estimates, 3-Estimates*. The source dependency detection problem was discussed and several solutions proposed [10, 41, 42]. Bayesian analysis has been adapted to model the source trustworthiness in an explicit and probabilistic way and improved the accuracy of truth estimation. Wang et al. [43] proposed the Bayesian Interpretation scheme as an approximation approach to correctly quantify the conclusions obtained from the basic fact-finding scheme. More recent works came up with some new fact-finding algorithms designed to handle the background knowledge and multi-valued facts in their trust analysis models. Pasternack et al. [44] provided a generalized fact-finder framework to incorporate the background knowledge on source and claim similarity as well as the uncertainty in the information extraction. They assume different types of background knowledge and contextual details can be encoded as a unified link weight in a generalized k-partite graph. Zhao et al. [12] presented another approach to model different types of errors made by sources and merge multi-valued attribute types of entities in data integration systems. They assumed the possible true values of a fact are not unique and a source can claim multiple values of a fact at the same time. They took a Bayesian analysis scheme that needs the prior on both source reliability and fact truthfulness. Additionally, trust analysis was done both on a homogeneous network [45, 11] and a heterogeneous network [46]. Fact-finding in the case of social sensing is more challenging due to the highly dynamic nature of social sensing applications [47]. Moreover, the outputs of fact-finders are generally relative credibility scores of sources and facts, which cannot be used to directly *quantify* the participant reliability or measurement correctness for social sensing. Therefore, our work first established the relation between ranking outputs of fact-finders and posterior probabilities of participant reliability and measurement correctness by using Bayesian analysis. We then developed the maximum likelihood estimator based on Expectation Maximization (EM) scheme in estimation theory to provide the MLE on both participant reliability and measurement correctness.

There exists a good amount of literature in machine learning community to improve data quality and identify low quality labelers in a multi-labeler environment. Sheng et al. proposed a repeated labeling scheme to improve label quality by selectively acquiring multiple labels and empirically comparing several models that aggregate responses from multiple labelers [48]. Dekel et al. applied a classification technique to simulate aggregate labels and prune low-quality labelers in a crowd to improve the label quality of the training dataset [49]. However, all of the above approaches made explicit or implicit assumptions that are not appropriate in the social sensing context. For example, the work in [48] assumed labelers were known a priori and could be explicitly asked to label certain data points. The work in [49] assumed most of labelers were reliable and the simple aggregation of their labels would be enough to approximate the ground-truth. In contrast, participants in social sensing usually upload their measurements based on their own observations and the simple aggregation technique (e.g., majority voting) was shown to be inaccurate when the reliability of participant is not sufficient [7]. The maximum likelihood estimation approach studied in this thesis addressed these challenges by intelligently casting the QoI quantification problem in social sensing into an optimization problem that can be efficiently solved by the EM scheme.

2.3 Estimation Theory

In estimation theory, Expectation Maximization (EM)is a general optimization technique for finding the maximum likelihood estimation of parameters in a statistic model where the data are “incomplete” or involve latent variables in addition to estimation parameter and observed data [50]. That is, either there are some missing value among the data, or the model can be formulated more simply by assuming the existence of some unobserved data. The general EM algorithm iterates between two main steps: the Expectation step (E-step) and the Maximization step (M-step) until the estimation converges (i.e., the likelihood function reaches the maximum). In the E-step, the algorithm computes the expectation of the log-likelihood function (so-called Q-function) of complete data w.r.t the conditional distribution of the latent variables given the current settings of the parameters and the observed data. In the M-step, it re-estimates the parameters in the next iteration that maximizes the expectation of the log-likelihood function defined in the E-step. EM is frequently used for data clustering in data mining and machine learning. For language modeling, the EM is often used to estimate parameters of a mixed model where the exact model from which the data is generated is unobservable [51]. There are also many good tutorials on EM algorithms [52, 53]. In this thesis, we showed that social sensing applications lend themselves nicely to an EM formulation. The optimal solution, in the sense of maximum likelihood estimation, directly leads to an accurate quantification of measurement correctness as well as participant reliability.

The Cramer-Rao lower bound is a fundamental bound used in estimation theory to characterize the lower bound on the estimation variance of a deterministic parameter [54]. The bound states that the variance of any unbiased estimator is lower-bounded by the inverse of Fisher information. The partial derivative w.r.t to the estimation parameter of the log-likelihood function is called the score. The Fisher information is defined as the second moment of the score vector of random variable and estimation parameter [55]. The Fisher information is a way of measuring the amount of information that an observable random variable X carries about an unknown estimation parameter θ upon which the probability of X depends. Intuitively, if the Fisher information is large, the distribution with the θ_0 (i.e., true value) of the estimation parameter will be different and well distinguished from the distributions with parameter that is not so close to θ_0. This means we are able to estimate θ_0 well (hence a small variance) based on the data. If the Fisher information is small, our estimation will be worse due to the similar reason. One of the key properties of maximum likelihood estimation (MLE) is the asymptotic normality. This property basically states that the MLE estimator is asymptotically distributed with a normal distribution as the data sample size goes up [56]. The mean of the normal distribution is the MLE of the estimation parameter and the variance is given by the CRLB of the estimation. The EM scheme we developed in this thesis provides the maximum likelihood estimation of participant reliability for social sensing applications. We derived an quantification approach to compute the *confidence interval* for participant reliability estimation based on both the real and asymptotic CRLB by leveraging the asymptotic normality of our MLE estimator.

2.4 Outlier Analysis and Attack Detection

Several previous efforts on data cleaning and outlier analysis from data mining and noise removal from statistics addressed some notion of noisy data [57, 58, 59, 60, 61, 62]. They differ in the assumption made, the modeling approach applied and the objective targeted at. For example, Bayesian inference and decision tree induction techniques are applied to fill the missing values of data by predictions from their constructed model [57]. Binning and linear regression techniques are used to smooth the noisy data by either using bin means or fitting data into some linear functions [58, 59]. Clustering techniques are widely used to detect outliers by organizing similar data values into clusters and identifying the ones that fall outside the clusters as outliers [60]. Other approaches are used in statistics to filter noises from continuous data [61, 62]. Kalman filter is an efficient reclusive filter that estimates some latent variables of a linear dynamic system from a series of noisy measurements [61]. It produces estimates of the measurements by computing a weighted average of the predicted values based on their uncertainty. Particle filters are more sophisticated filters that are based on Sequential Monte Carlo methods. They are often used to determine the distribution of a latent variable whose state space is not restricted to Gaussian distribution [62]. Our work is complementary to the above efforts. On one hand, an appropriately cleaned and outlier-removed dataset will likely result in a better estimation of our scheme. On the other hand, outliers or noises may not be completely (or even possibly) removed by the data cleaning and outlier analysis techniques mentioned above due to their own limitations (e.g., linear model assumption, continuous data assumption, known data distribution assumption and etc.). The quantifiable and confident estimation provided by our approach on both information source and observed data could actually help the data cleaning and outlier analysis tools do a better job.

In intrusion detection, one critical task is to detect (or identify) the malicious nodes (or sources) accurately and confidently. Two main kinds of detection techniques exist: signature-based detection and anomaly-based detection [60, 63]. The signature-based detection takes the predefined attack patterns (by domain experts) as signatures and monitor the node's behavior (or network traffic) for matches to report the anomaly [60]. The anomaly-based detection builds profiles of normal node's (or network) behavior and use the profiles to detect new patterns that have remarkable deviation [63]. For the QoI quantification problem in our work, it is not obvious what behavior patterns the malicious (unreliable) sources will have without knowing the correctness of their measurements. Hence, there might not be an easy way to apply the intrusion techniques mentioned above to discover malicious sources for social sensing applications. Instead, given the maximum likelihood estimation on participant reliability and the corresponding confidence interval provided by our scheme, we are able to both identify unreliable sources and quantify their reliability with certain confidence without prior knowledge of their behavior patterns.

Since people are an indispensable element in social sensing, some popular attacks originated from human (or source) interactions are interesting to investigate. Collusion attack is carried out by a group of colluded attackers who collectively perform some malicious (sometimes illegal) actions based on their agreement to defraud honest sources or obtain objective forbidden by the system. This attack could be mitigated by monitoring the interactions or relationships among colluded attackers or identifying the abnormal behavior from the group [64]. Sybil attack is another related attack carried out by a single attacker who intentionally create a large number of pseudonymous entities and use them to gain a disproportionately large influence on the system. This attack could be mitigated by certifying trust of identity assignment, increasing the cost of creating identities, limiting the resource the attacker can use to create new identities and etc. [65]. By handling reports from colluded or duplicate sources in a way that takes care of the source dependency, we will be able to address the above attacks to some extent. For example, by identifying duplicate sources, we can remove them along with their reports from the observed dataset, which is expected to improve the estimation performance. Problems become more interesting when sources are not just duplicates but actually linked through some orthogonal information network (e.g., social network).

2.5 Recommender and Reputation Systems

Our work is related with a type of information filtering system called recommender systems, where the goal is usually to predict a user's rating or preference to an item using the model built from the characteristics of the item and the behavioral pattern of the user [66]. EM has been used in either collaborative recommender systems as a clustering module [67] to mine the usage pattern of users or in a content-based recommender systems as a weighting factor estimator [68] to infer the user context. However, in social sensing, the QoI quantification problem targets a different goal: we try to quantify how reliable a source is and identify whether a measured variable is true or not rather than predict how likely a user would choose one item compared to another. Moreover, users in recommender systems are commonly assumed to provide reasonably good data while the sources in social sensing are in general unreliable and the likelihood of the correctness of their measurements is unknown a priori. There appears no straightforward use of methods in the recommender systems regime for the target problem with unpredictably unreliable data. Additionally, the rating or preference we get from users in the recommender systems are sometimes *subjective* [69]. For example, some people may prefer Ford car to Toyota while others prefer exactly the opposite. It is hard to say who is right and who is wrong due to the fact that there is no universal ground truth on the items to be evaluated. We note that the work in this thesis may not be directly applicable to handle the above case due to a different assumption made in our model. In social sensing applications, we aim to leverage the data contributed by common individuals and reconstruct the *state of the physical world*, where we usually do have the universal ground truth associated with the assertions that describe those physical states (e.g, a building is either on fire or not). The QoI quantification theory developed in our thesis makes much more sense under this assumption of social sensing applications. It enables the application to not only obtain the optimal estimation (in MLE sense) on source and information reliability, but also assess the quality of the estimation compared to the ground truth.

Our work also bears resemblance to reputation systems. The basic idea of reputation systems is to let entities rate each other (e.g., after a transaction) or review some objects of common interests (e.g., products or dealers), and use the aggregated ratings or reviews to derive trust or reputation scores, which can help other entities in deciding whether or not to trust a given entity or purchase a certain object [70]. Different types of reputation systems are being used successfully in commercial online applications. For example, eBay is a type of reputation system based on homogeneous peer-to-peer systems, which allows peers to rate each other after each pair of them conduct a transaction [71, 72]. Our developed scheme may not be able to be directly applied to those systems. The reason is: the structure of a homogeneous peer-to-peer system is commonly represented by a *mesh* network graph while the structure of our scheme is represented by a *bipartite* network graph(i.e., sources and measures are in disjoint sets). Amazon on-line review system represents another type of reputation systems, where different sources offer reviews on products (or brands, companies) they have experienced. Customers are affected by those reviews (or reputation scores) in making purchase decisions. It turns out that our work fits better into this type of reputation systems and has the potential to provide more refined and confident results for the reputation computation. In this thesis, we also proposed the extended model to handle more general types of variables since reviews in real life may not necessarily be binary (i.e., like or dislike). By incorporating non-binary measurements in our model, we will be able to provide more accurate estimation results for those reputation systems. Additionally, reputation systems are in general vulnerable to several attacks: self-promoting, slandering, denial of service and etc [73]. Many of the attacks actually originate from collusion and Sybil attack we mentioned earlier. Hence, we can adopt similar techniques discussed before to address some of the attacks from reputation systems for the scheme we developed.

Chapter 3

Bayesian Interpretation of Basic Fact-finding

This chapter presents a foundation for quality of information analysis in information networks [43]. We are specifically interested in the network model used for deriving credibility of facts and sources. We call the iterative ranking algorithm used for analyzing source/assertion information networks, a *fact-finder*. The algorithm ranks a list of assertions and a list of sources by credibility. We first review the basic algorithm, then propose the Bayesian Interpretation that allows quantifying the actual probability that a source is truthful or that an assertion is true. The derived Bayesian interpretation is applied to a representative fact-finding problem, and is validated by extensive simulation where analysis shows significant improvement over past work and great correspondence with ground truth.

This chapter is organized as follows: In Section 3.1, we introduce the basic fact-finding scheme used in information network. We then derive the Bayesian Interpretation of the fact-finding scheme in Section 3.2. The evaluation results are presented in Section 3.3. We discuss the limitations and assumptions made in our model in Section 3.4

3.1 Basic Fact-finding in Information Networks

When information sources are unreliable, information networks have been used in data mining literature to uncover facts from large numbers of complex relations between noisy variables. The approach relies on topology analysis of graphs, where nodes represent pieces of (unreliable) information and links represent abstract relations. More specifically, let there be s sources, $S_1, ..., S_s$ who collectively assert c different pieces of information, $C_1, ..., C_c$. We call each such piece of information an assertion. We represent all sources and assertions by a network, where these sources and assertions are nodes, and where a claim, $C_{i,j}$ (denoting that a source S_i makes assertion C_j) is represented by a link between the corresponding source and assertion nodes. We assume that a claim can either be true or false. An example is "John Smith is CEO of Company X" or "Building Y is on Fire". We further define $Cred(S_i)$ as the credibility of source S_i, and $Cred(C_j)$ as the credibility of assertion C_j.

We define the $c \times 1$ vector, $\overline{C}_{cred}$, to be the assertion credibility vector $[Cred(C_1)...Cred(C_c)]^T$ and the $s \times 1$ vector, $\overline{S}_{cred}$, to be the source credibility vector $[Cred(S_1)...Cred(S_s)]^T$. We also define the $c \times s$ array CS such that element $CS(j,i) = 1$ if source S_i makes claim C_j, and is zero otherwise.

Now let us define $\overline{C}_{cred}^{est}$ as a vector of *estimated* assertion credibility, defined as $(1/\alpha)[CS]\overline{S}_{cred}$. One can pose the basic fact-finding problem as one of finding a least squares estimator (that minimizes the sum of squares of errors in source credibility estimates) for the following system:

$$\overline{C}_{cred}^{est} = \frac{1}{\alpha}[CS]\overline{S}_{cred} \qquad (3.1)$$

$$\overline{S}_{cred} = \frac{1}{\beta}[CS]^T\overline{C}_{cred}^{est} + \overline{e} \qquad (3.2)$$

where the notation X^T denotes the transpose of matrix X. It can further be shown that the condition for it to minimize the error is that α and β be chosen such that their product is an Eigenvalue of $[CS]^T[CS]$. The algorithm produces the credibility values $Cred(S_i)$ and $Cred(C_j)$ for every source S_i and for every assertion C_j. These values are used for ranking. The question is, does the solution have an interpretation that allows quantifying the actual probability that a given source is truthful or that a given assertion is true? The question is answered in the next section.

3.2 A Bayesian Interpretation

Let S_i^t denote the proposition that "Source S_i speaks the truth". Let C_j^t denote the proposition that "Assertion C_j is true". Also, let S_i^f and C_j^f denote the negation of the above propositions, respectively. Our objective, in this section, is to estimate the probabilities of these propositions. We further define S_iC_j to mean "Source S_i made assertion C_j".

It is useful to define $Claims_i$ as the set of all claims made by source S_i, and $Sources_j$ as the set of all sources who claimed assertion C_j. In the subsections below, we derive the posterior probability that an assertion is true, followed by the derivation of the posterior probability that a source is truthful.

3.2.1 Assertion Credibility

Consider some assertion C_j, claimed by a set of sources $Sources_j$. Let i_k be the kth source in $Sources_j$, and let $|Sources_j| = K_j$. (For notational simplicity, we shall occasionally omit the subscript j from K_j in the discussion below, where no ambiguity arises.) According to Bayes theorem:

$$P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) = \frac{P(S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j|C_j^t)}{P(S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j)} P(C_j^t) \tag{3.3}$$

The above equation makes the implicit assumption that the probability that a source makes any given assertion is sufficiently low that no appreciable change in posterior probability can be derived from the lack of a claim (i.e., lack of an edge between a source and an assertion). Hence, only existence of claims is taken into account. Assuming further that sources are conditionally independent (i.e., given an assertion, the odds that two sources claim it are independent), Equation (3.3) is rewritten as:

$$P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) = \frac{P(S_{i_1}C_j|C_j^t)...P(S_{i_K}C_j|C_j^t)}{P(S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j)} P(C_j^t) \tag{3.4}$$

Let us further assume that the change in posterior probability we get from any single source or claim is small. This is typical when using evidence collected from many individually unreliable sources. Hence:

$$\frac{P(S_{i_k}C_j|C_j^t)}{P(S_{i_k}C_j)} = 1 + \delta_{i_k j}^t \tag{3.5}$$

where $|\delta_{i_k j}^t| << 1$. Similarly:

$$\frac{P(S_{i_k}C_j|C_j^f)}{P(S_{i_k}C_j)} = 1 + \delta_{i_k j}^f \tag{3.6}$$

where $|\delta_{i_k j}^f| << 1$. Under the above assumptions, we prove in Appendix A that the denominator of the right hand side in Equation (3.4) can be rewritten as follows:

$$P(S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) \approx \prod_{k=1}^{K_j} P(S_{i_k}C_j) \tag{3.7}$$

Please see Appendix A for a proof of Equation (3.7). Note that, the proof does *not* rely on an independence assumption of the marginals, $P(S_{i_k}C_j)$. Those marginals are, in fact, not independent. The proof merely shows that, under the assumptions stated in Equation (3.5) and Equation (3.6), the above approximation holds true. Substituting in Equation (3.4):

$$\begin{aligned} P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) &= \\ \frac{P(S_{i_1}C_j|C_j^t)...P(S_{i_K}C_j|C_j^t)}{P(S_{i_1}C_j)...P(S_{i_K}C_j)} P(C_j^t) \end{aligned} \tag{3.8}$$

which can be rewritten as:

$$\begin{aligned} P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) &= \frac{P(S_{i_1}C_j|C_j^t)}{P(S_{i_1}C_j)} \\ &\times ... \\ &\times \frac{P(S_{i_K}C_j|C_j^t)}{P(S_{i_K}C_j)} \\ &\times P(C_j^t) \end{aligned} \tag{3.9}$$

Substituting from Equation (3.5):

$$\begin{aligned} P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) &= P(C_j^t) \prod_{k=1}^{K_j} (1 + \delta_{i_kj}^t) \\ &= P(C_j^t)(1 + \sum_{k=1}^{K_j} \delta_{i_kj}^t) \end{aligned} \tag{3.10}$$

The last line above is true because higher products of $\delta_{i_kj}^t$ can be neglected, since we assumed $|\delta_{i_kj}^t| << 1$. The above equation can be re-written as:

$$\frac{P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) - P(C_j^t)}{P(C_j^t)} = \sum_{k=1}^{K_j} \delta_{i_kj}^t \tag{3.11}$$

where, from Equation (3.5):

$$\delta_{i_kj}^t = \frac{P(S_{i_k}C_j|C_j^t) - P(S_{i_k}C_j)}{P(S_{i_k}C_j)} \tag{3.12}$$

3.2.2 Source Credibility

Next, consider some source S_i, who makes the set of claims $Claims_i$. Let j_k be the kth claim in $Claims_i$, and let $|Claims_i| = L_i$. (For notational simplicity, we shall occasionally omit the subscript i from L_i in the discussion below, where no ambiguity arises.) According to Bayes theorem:

$$\begin{aligned} P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L}) &= \\ \frac{P(S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L}|S_i^t)}{P(S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L})}P(S_i^t) \end{aligned} \tag{3.13}$$

As before, assuming conditional independence:

$$\begin{aligned} P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L}) &= \\ \frac{P(S_iC_{j_1}|S_i^t)...P(S_iC_{j_L}|S_i^t)}{P(S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L})}P(S_i^t) \end{aligned} \tag{3.14}$$

Once more we invoke the assumption that the change in posterior probability caused from any single claim is very small, we get:

$$\frac{P(S_iC_{j_k}|S_i^t)}{P(S_iC_{j_k})} = 1 + \eta_{ij_k}^t \tag{3.15}$$

where $|\eta_{ij_k}^t| << 1$. Similarly to the proof in Appendix A, this leads to:

$$\begin{aligned} P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L}) &= \frac{P(S_iC_{j_1}|S_i^t)}{P(S_iC_{j_1})} \\ &\times ... \\ &\times \frac{P(S_iC_{j_L}|S_i^t)}{P(S_iC_{j_L})} \\ &\times P(S_i^t) \end{aligned} \tag{3.16}$$

We can then re-write Equation (3.16) as follows:

$$\begin{aligned} P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L}) &= P(S_i^t)\prod_{k=1}^{L_i}(1 + \eta_{ij_k}^t) \\ &= P(S_i^t)(1 + \sum_{k=1}^{L_i}\eta_{ij_k}^t) \end{aligned} \tag{3.17}$$

The above equation can be further re-written as:

$$\frac{P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L}) - P(S_i^t)}{P(S_i^t)} = \sum_{k=1}^{L_i} \eta_{ij_k}^t \tag{3.18}$$

where, from Equation (3.15):

$$\eta_{ij_k}^t = \frac{P(S_iC_{j_k}|S_i^t) - P(S_iC_{j_k})}{P(S_iC_{j_k})} \tag{3.19}$$

3.2.3 The Iterative Algorithm

In the sections above, we derived the expressions of posterior probability that an assertion is true or that a source is truthful. These expressions were derived in terms of $\delta_{i_kj}^t$ and $\eta_{ij_k}^t$. It remains to show how these quantities are related. Let us first consider the terms in Equation (3.12) that defines $\delta_{i_kj}^t$. The first is $P(S_iC_j|C_j^t)$, the probability that S_i claims assertion C_j, given that C_j is true. (For notational simplicity, we shall use subscripts i and j to denote the source and the assertion.) We have:

$$P(S_iC_j|C_j^t) = \frac{P(S_iC_j, C_j^t)}{P(C_j^t)} \tag{3.20}$$

where:

$$\begin{aligned} P(S_iC_j, C_j^t) \quad = \quad & P(S_i\ speaks) \\ & P(S_i\ claims\ C_j|S_i\ speaks) \\ & P(C_j^t|S_i\ speaks, S_i\ claims\ C_j) \end{aligned} \tag{3.21}$$

In other words, the joint probability that link S_iC_j exists and C_j is true is the product of the probability that S_i speaks, denoted $P(S_i\ speaks)$, the probability that it claims C_j given that it speaks, denoted $P(S_i\ claims\ C_j|S_i\ speaks)$, and the probability that the assertion is true, given that it is claimed by S_i, denoted $P(C_j^t|S_i\ speaks, S_i\ claims\ C_j)$. Here, $P(S_i\ speaks)$ depends on the rate at which S_i makes assertions. Some sources may be more prolific than others. $P(S_i\ claims\ C_j|S_i\ speaks)$ is simply $1/c$, where c is the total number of assertions. Finally, $P(C_j^t|S_i\ speaks, S_i\ claims\ C_j)$ is the probability that S_i is truthful. Since we do not know ground truth, we estimate that probability by the best information we have, which is $P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L})$. Thus:

$$P(S_iC_j, C_j^t) = \frac{P(S_i\ speaks)P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L})}{c} \tag{3.22}$$

Substituting in Equation (3.20) from Equation (3.22) and noting that $P(C_j^t)$ is simply the ratio of true assertions, c_{true} to the total assertions, c, we get:

$$P(S_iC_j|C_j^t) = \frac{P(S_i\ speaks)P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L})}{c_{true}} \tag{3.23}$$

Similarly,

$$P(S_iC_j) = \frac{P(S_i\ speaks)}{c} \tag{3.24}$$

Substituting from Equation (3.23) and Equation (3.24) into Equation (3.12) and re-arranging, we get:

$$\begin{aligned} \delta_{i_kj}^t &= \frac{P(S_{i_k}C_j|C_j^t) - P(S_{i_k}C_j)}{P(S_{i_k}C_j)} \\ &= \frac{P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L})}{c_{true}/c} - 1 \end{aligned} \tag{3.25}$$

If we take the fraction of all true assertions to the total number of assertions as the prior probability that a source is truthful, $P(S_i^t)$ (which is a reasonable initial guess in the absence of further evidence), then the above equation can be re-written as:

$$\delta_{i_kj}^t = \frac{P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L})}{P(S_i^t)} - 1 \tag{3.26}$$

Substituting for $\delta_{i_kj}^t$ in Equation (3.11), we get:

$$\begin{aligned} \frac{P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) - P(C_j^t)}{P(C_j^t)} &= \\ \sum_{i=1}^{K_j} \frac{P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L}) - P(S_i^t)}{P(S_i^t)} \end{aligned} \tag{3.27}$$

We can similarly prove that:

$$\eta_{ij_k}^t = \frac{P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j)}{P(C_j^t)} - 1 \tag{3.28}$$

and:

$$\begin{array}{c} \frac{P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L}) - P(S_i^t)}{P(S_i^t)} \quad = \\ \sum_{j=1}^{L_i} \frac{P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) - P(C_j^t)}{P(C_j^t)} \end{array} \tag{3.29}$$

Comparing the above equations to the iterative formulation of the basic fact-finder, described in Section 3.1, we arrive at the sought interpretation of the credibility rank of sources $Rank(S_i)$ and credibility rank of assertions $Rank(C_j)$ in iterative fact-finding. Namely:

$$Rank(C_j) \quad = \quad \frac{P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) - P(C_j^t)}{P(C_j^t)} \tag{3.30}$$

$$Rank(S_i) \quad = \quad \frac{P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L}) - P(S_i^t)}{P(S_i^t)} \tag{3.31}$$

In other words, $Rank(C_j)$ is interpreted as the increase in the posterior probability that an assertion is true, normalized by the prior. Similarly, $Rank(S_i)$ is interpreted as the increase in the posterior probability that a source is truthful, normalized by the prior. Substituting from Equation (3.30) and Equation (3.31) into Equation (3.27) and Equation (3.29), we then get:

$$\begin{aligned} Rank(C_j) &= \sum_{k \in Sources_j} Rank(S_k) \\ Rank(S_i) &= \sum_{k \in Claims_i} Rank(C_k) \end{aligned} \tag{3.32}$$

Once the credibility ranks are computed such that they satisfy the above equations (and any other problem constraints), Equation (3.30) and Equation (3.31), together with the assumption that prior probability that an assertion is true is initialized to $p_a^t = c_{true}/c$, give us the main contribution of this effort, Namely[*]:

$$P(C_j^t|network) = p_a^t(Rank(C_j) + 1) \tag{3.33}$$

We can similarly show that if p_s^t is the prior probability that a randomly chosen source tells the truth, then:

$$P(S_i^t|network) = p_s^t(Rank(S_i) + 1) \tag{3.34}$$

Hence, the above Bayesian analysis presents, for the first time, a basis for estimating the probability that *each individual source*, S_i, is truthful and that *each individual assertion*, C_j, is true. These two vectors are computed based on two scalar constants: p_a^t and p_s^t, which represent estimated statistical averages over all assertions and all sources, respectively.

3.3 Evaluation

In this section, we carry out experiments to verify the correctness and accuracy of the probability that a source is truthful or an assertion is true predicted from the Bayesian interpretation of fact-finding in information networks. We then compare our techniques to previous algorithms in fact-finder literature.

We built a simulator in Matlab 7.8.0 to simulate the source and assertion information network. To test our results, we generate a random number of sources and assertions, and partition these assertions into true and false ones. A random probability, P_i, is assigned to each source S_i representing the ground truth probability that the source speaks the truth. For each source S_i, we then generate L_i claims. Each claim has a probability P_i of being true and a probability $1 - P_i$ of being false. A true claim links the source to a randomly chosen true assertion (representing that the source made that assertion). A false claim links the source to a randomly chosen false assertion. This generates an information network.

We let P_i be uniformly distributed between 0.5 and 1 in our experiments[†]. We then find an assignment of credibility values that satisfies Equation (3.32) for the topology of the generated information network. Finally, we compute the estimated probability that an assertion is true or a source is truthful from the result-

[*] The equations above are ambiguous with respect to a scale factor. To handle the ambiguity we impose the constraint that probabilities cannot exceed one.

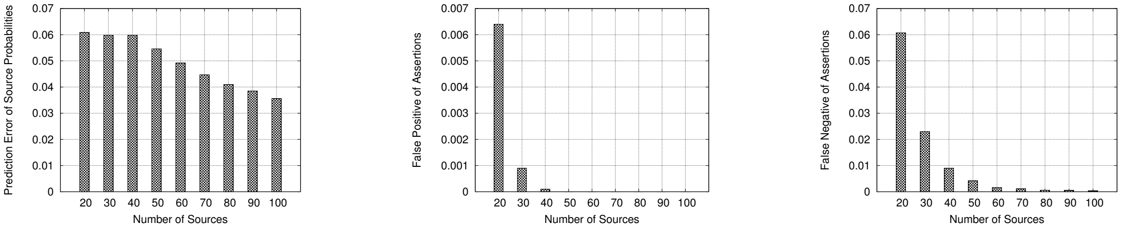
[†] In principle, there is no incentive for a source to lie more than 50% of the time, since negating their statements would then give a more accurate truth

ing credibility values of assertions and sources based on Equation (3.33) and (3.34). Since we assumed that claims are either true or false, we view each assertion as "true" or "false" based on whether the probability that it is true is above or below 50%. Then the computed results are compared against the ground truth to report the prediction accuracy.

For sources, we simply compare the computed probability to the ground truth probability that they tell the truth. For assertions, we define two metrics to evaluate prediction accuracy: false positives and false negatives. The false positives are defined as the ratio of the number of false assertions that are classified as true over the total number of assertions that are classified as true. The false negatives are defined as the ratio of the number of true assertions that are classified as false over the total number of assertions that are classified as false. For each given source correctness probability (i.e., ground truth) distribution, we average the results over 100 network topologies (e.g., datasets over a time series). Reported results are averaged over 100 random source correctness probability distributions.

In the first experiment, we show the effect of the number of sources on prediction accuracy. We fix the number of true and false assertions at 1000 respectively. We set the average number of claims per source to 100. The number of sources is varied from 20 to 100. The prediction accuracy for both sources and assertions is shown in Figure 3.1. We note that both false positives and false negatives decrease as the number of sources grows. For more than 40 sources less than 1% of assertions are misclassified. The source correctness probability prediction exhibits a relatively small error (between 3% and 6%). The error first increases and then decreases as the number of sources increases. The reason is that there are two conflicting factors that affect the credibility prediction accuracy of sources: i) average number of assertions per source and ii) average number of sources per assertion. As the the number of sources increases, the first factor decreases (reduce source credibility prediction accuracy) and the second factor increases (improve assertion and eventually source credibility prediction accuracy). When the number of sources is small, the change of the first factor is more significant than the second, thus its effect dominates. As the number of sources increases, the effect of the second factor overweights the first one and makes source correctness probability prediction error reduce.

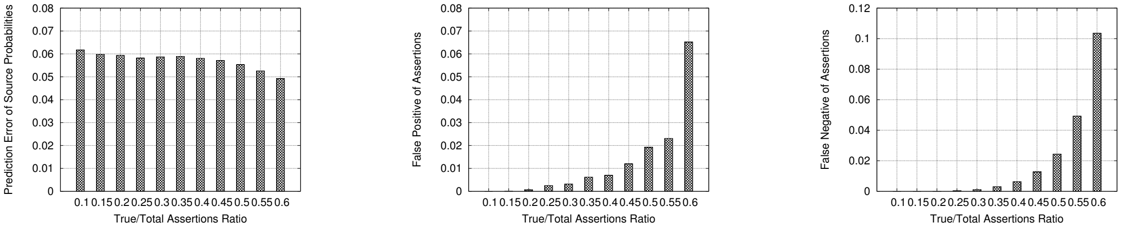
Note that, the source correctness probability prediction is especially accurate (e.g., error is around 0.03) when the number of sources is relatively large. At the same time, both the false positives and false negatives in assertion classification are near zero under those conditions, illustrating that the approach has good scalability properties. Its usefulness increases for large networks.



(a) Source Prediction Accuracy (b) Assertion Prediction False Positives (c) Assertion Prediction False Negatives

Figure 3.1: Prediction Accuracy vs Varying Number of Sources

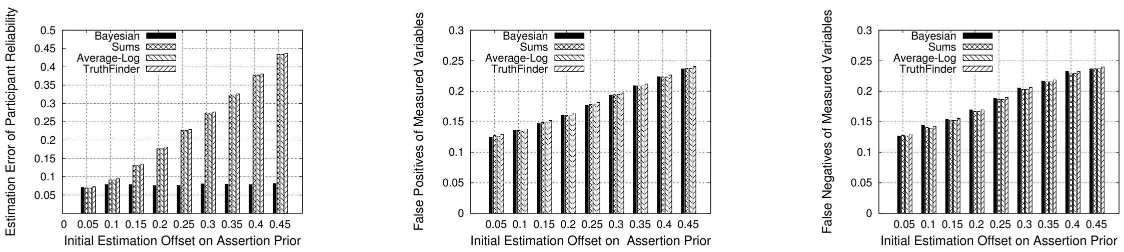
The next experiment shows the effect of changing the assertion mix on prediction accuracy. We vary the ratio of the number of true assertions to the total number of assertions in the network. Assuming that there is usually only one variant of the truth, whereas rumors have more versions, one might expect the set of true assertions to be smaller than the set of false ones. Hence, we fix the total number of assertions to be 2000 and change the ratio of true to total assertions from 0.1 to 0.6. The number of sources in the network is set to 30. The prediction accuracy for both sources and assertions is shown in Figure 3.2. Observe that the source correctness probability prediction error decreases as the ratio of true assertions increases. This is intuitive: more independent true assertions can be used to improve credibility estimates of sources. Additionally, the false positives and false negatives increase because the true assertion set becomes less densely claimed and more true and false assertions are misclassified as each other as the number of true assertions grows.



(a) Source Prediction Accuracy (b) Assertion Prediction False Positives (c) Assertion Prediction False Negatives

Figure 3.2: Prediction Accuracy vs Varying True/Total Assertions

Finally, we compared our proposed Bayesian interpretation scheme to four other fact-finder schemes: Average-Log [7], Sums(Hubs and Authorities) [40], an adapted PageRank [74] where claims are bidirectional "links" between source and asserted "documents", and TruthFinder [8]. We selected these because, unlike other state-of-art fact-finders (e.g., 3-Estimates [9]), these do not require knowing what mutual ex-



(a) Source Prediction Accuracy (b) Assertion Prediction False Positives (c) Assertion Prediction False Negatives

Figure 3.3: Prediction Accuracy Comparison with Baseline Fact-finders

clusion, if any, exists among the assertions. In this experiment, the number of true and false assertions is 1000 respectively, the number of claims per source is 100, and the number of sources is set to 50. We vary the initial estimation offset on prior assertion correctness from 0.05 to 0.45. Here we don't average results for multiple topologies due to the fact that most of the selected baselines are mainly designed for a single topology scenario. Using the initial assertion beliefs suggested by [7], we ran each baseline fact-finder for 20 iterations, and then selected the 1000 highest-belief assertions as those predicted to be correct. The estimated probability of each source making a true claim was thus calculated as the proportion of predicted-correct claims asserted relative to the total number of claims asserted by source.

The compared results are shown in Figure 3.3. Observe that the prediction error of source correctness probability by the Bayesian interpretation scheme is significantly lower than all baseline fact-finder schemes. The reason is that Bayesian analysis estimates the source correctness probability more accurately based on Equation (3.34) derived in this chapter rather than using heuristic methods adopted by the baseline schemes that depends on the correct estimation on prior assertion correctness. We also note that the prediction performance for assertions in the Bayesian scheme is generally as good as the baselines. This is good since the other techniques excel at ranking, which (together with the hint on the number of correct assertions) is sufficient to identify which ones these are. The results confirm the advantages of the Bayesian approach over previous ranking-based work at what the Bayesian analysis does best: estimation of probabilities of conclusions from observed evidence.

3.4 Discussion

This chapter presented a Bayesian interpretation of the most basic fact-finding algorithm. The question was to understand why the algorithm is successful at ranking, and to use that understanding to translate the ranking into actual probabilities. Several simplifying assumptions were made that offer opportunities for future extensions.

No dependencies were assumed among different sources or different claims. In reality, sources could be influenced by other sources. Claims could fall into mutual exclusion sets, such as when one is the negation of the other. Taking such relations into account can further improve quality of fact-finding. The change in posterior probabilities due to any single edge in the source-assertion network was assumed to be very small. In other words, we assumed that $|\delta^t_{i_k j}| << 1$ and $|\eta^t_{i j_k}| << 1$. It is interesting to extend the scheme to situations where a mix of reliable and unreliable sources is used. In this case, assertions from reliable sources can help improve the determination of credibility of other sources.

The probability that any one source makes any one assertion was assumed to be low. Hence, the lack of an edge between a source and an assertion did not offer useful information. There may be cases, however, when the absence of a link between a source and an assertion is important. For example, when a source is expected to bear on an issue, a source that "withholds the truth" exhibits absence of a link that needs to be accounted for.

In this chapter, we presented a novel analysis technique for information networks that uses a Bayesian interpretation of the network to assess the credibility of facts and sources. Prior literature that uses information network analysis for fact-finding aims at computing the credibility *rank* of different facts and sources. This work, in contrast, proposes an analytically founded technique to convert rank to a probability that a fact is true or that a source is truthful. This chapter therefore lays out a foundation for quality of information assurances in iterative fact-finding, a common branch of techniques used in data mining literature for analysis of information networks. The fact-finding techniques addressed in this chapter are particularly useful in environments where a large number of sources are used whose reliability is not *a priori* known (as opposed collecting information from a small number of well-characterized sources). Such situations are common when, for instance, crowd-sourcing is used to obtain information, or when information is to be gleaned from informal sources such as Twitter messages. Our work shows that accurate information may indeed be obtained regarding facts and sources even when we do not know the credibility of each source in advance, and where individual sources may generally be unreliable.

Chapter 4

A Maximum Likelihood Estimation Approach

Considering the heuristic nature of fact-finding techniques, the quantified source truthfulness and assertion correctness from Bayesian Interpretation remain to be linear approximations. Moreover, results of Bayesian Interpretation are shown to be sensitive to the priors given to the algorithm [43]. To overcome these limitations, this chapter presents the first *optimal solution* to the QoI quantification problem in social sensing. Optimality, in the sense of maximum likelihood estimation, is attained by solving an expectation maximization problem that returns the best guess regarding the source reliability as well as the correctness of each measurement [75]. The algorithm makes inferences regarding both source reliability and measurement correctness by observing which observations coincide and which don't. The approach is shown to outperform the state of the art fact-finding heuristics, as well as simple baselines such as majority voting.

This chapter is organized as follows: In Section 4.1, we describe problem of quantifying source reliability and their measurement correctness in social sensing applications. We then cast the formalized quantification problem into an optimization problem that can be solved by Expectation Maximization (EM) scheme in Section 4.2. The evaluation results are presented in Section 4.3. We discuss the limitations of our model and possible extensions in Section 4.4

4.1 The Problem Formulation of Social Sensing

In the context of social sensing, we adopt some sensor community friendly terminologies to denote several concepts we have defined in the previous chapter with the background of information network analysis. In particular, we denote sources as participants, assertions as measured variables, claims as observations, source truthfulness as participant reliability.

To formulate the QoI quantification problem in social sensing in a manner amenable to rigorous optimization, we consider a social sensing application model where a group of M participants, $S_1, ..., S_M$, make individual observations about a set of N measured variables $C_1, ..., C_N$ in their environment. For example, a group of individuals interested in the appearance of their neighborhood might join a sensing campaign to report all locations of offensive graffiti. Alternatively, a group of drivers might join a campaign to report freeway locations in need of repair. Hence, each measured variable denotes the existence or lack thereof of an offending condition at a given location[*]. In this effort, we consider only binary variables and assume, without loss of generality, that their "normal" state is negative (e.g., no offending graffiti on walls, or no potholes on streets). Hence, participants report only when a positive value is encountered.

Each participant generally observes only a subset of all variables (e.g., the conditions at locations they have been to). Our goal is to determine which observations are correct and which are not. As mentioned in the introduction, we differ from a large volume of previous sensing literature in that we assume no prior knowledge of source reliability, as well as no prior knowledge of the correctness of individual observations.

Let S_i represent the i^{th} participant and C_j represent the j^{th} measured variable. S_iC_j denotes an observation reported by participant S_i claiming that C_j is true (e.g., that graffiti is found at a given location, or that a given street is in disrepair). Let $P(C_j^t)$ and $P(C_j^f)$ denote the probability that the actual variable C_j is indeed true and false, respectively. Different participants may make different numbers of observations. Let the probability that participant S_i makes an observation be s_i. Further, let the probability that participant S_i is right be t_i and the probability that it is wrong be $1 - t_i$. Note that, this probability depends on the participant's reliability, which is not known *a priori*. Formally, t_i is defined as the odds of a measured variable to be true given that participant S_i reports it:

$$t_i = P(C_j^t | S_iC_j) \tag{4.1}$$

Let us also define a_i as the (unknown) probability that participant S_i reports a measured variable to be true when it is indeed true, and b_i as the (unknown) probability that participant S_i reports a measured variable to be true when it is in reality false. Formally, a_i and b_i are defined as follows:

$$\begin{aligned} a_i &= P(S_iC_j | C_j^t) \\ b_i &= P(S_iC_j | C_j^f) \end{aligned} \tag{4.2}$$

[*] We assume that locations are discretized, and therefore finite. For example, they are given by street addresses or mile markers.

From the definition of t_i, a_i and b_i, we can determine their relationship using the Bayesian theorem:

$$\begin{aligned} a_i &= P(S_iC_j|C_j^t) = \frac{P(S_iC_j, C_j^t)}{P(C_j^t)} = \frac{P(C_j^t|S_iC_j)P(S_iC_j)}{P(C_j^t)} \\ b_i &= P(S_iC_j|C_j^f) = \frac{P(S_iC_j, C_j^f)}{P(C_j^f)} = \frac{P(C_j^f|S_iC_j)P(S_iC_j)}{P(C_j^f)} \end{aligned} \tag{4.3}$$

The only input to our algorithm is the social sensing topology represented by a matrix SC, where $S_iC_j = 1$ when participant S_i reports that C_j is true, and $S_iC_j = 0$ otherwise. Let us call it the *observation matrix*.

The goal of the algorithm is to compute (i) the best estimate h_j on the correctness of each measured variable C_j and (ii) the best estimate e_i of the reliability of each participant S_i. Let us denote the sets of the estimates by vectors H and E, respectively. Our goal is to find the optimal H^* and E^* vectors in the sense of being most consistent with the observation matrix SC. Formally, this is given by:

$$< H^*, E^* >= \underset{<H,E>}{\operatorname{argmax}}\, p(SC|H, E) \tag{4.4}$$

We also compute the background bias d, which is the overall probability that a randomly chosen measured variable is true. For example, it may represent the probability that any street, in general, is in disrepair. It does not indicate, however, whether any particular claim about disrepair at a particular location is true or not. Hence, one can define the prior of a claim being true as $P(C_j^t) = d$. Note also that, the probability that a participant makes an observation (i.e., s_i) is proportional to the number of measured variables observed by the participant over the total number of measured variables observed by all participants, which can be easily computed from the observation matrix. Hence, one can define the prior $P(S_iC_j) = s_i$. Plugging these, together with t_i into the definition of a_i and b_i, we get the relationship between the terms we defined above:

$$\begin{aligned} a_i &= \frac{t_i \times s_i}{d} \\ b_i &= \frac{(1 - t_i) \times s_i}{1 - d} \end{aligned} \tag{4.5}$$

4.2 Expectation Maximization

In this section, we solve the problem formulated in the previous section using the Expectation-Maximization (EM) algorithm. EM is a general algorithm for finding the maximum likelihood estimates of parameters in a statistic model, where the data are "incomplete" or the likelihood function involves latent variables [50]. Intuitively, what EM does is iteratively "completes" the data by "guessing" the values of hidden variables then re-estimates the parameters by using the guessed values as true values.

4.2.1 Background

Much like finding a Lyapunov function to prove stability, the main challenge in using the EM algorithm lies in the mathematical formulation of the problem in a way that is amenable to an EM solution. Given an observed data set X, one should judiciously choose the set of latent or missing values Z, and a vector of unknown parameters θ, then formulate a likelihood function $L(\theta; X, Z) = p(X, Z|\theta)$, such that the maximum likelihood estimate (MLE) of the unknown parameters θ is decided by:

$$L(\theta; X) = p(X|\theta) = \sum_{Z} p(X, Z|\theta) \tag{4.6}$$

Once the formulation is complete, the EM algorithm finds the maximum likelihood estimate by iteratively performing the following steps:

- E-step: Compute the expected log likelihood function where the expectation is taken with respect to the computed conditional distribution of the latent variables given the current settings and observed data.

$$Q\left(\theta|\theta^{(t)}\right) = E_{Z|X,\theta^{(t)}}[\log L(\theta; X, Z)] \tag{4.7}$$

- M-step: Find the parameters that maximize the Q function in the E-step to be used as the estimate of θ for the next iteration.

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, Q\left(\theta|\theta^{(t)}\right) \tag{4.8}$$

4.2.2 Mathematical Formulation

Our social sensing problem fits nicely into the Expectation Maximization (EM) model. First, we introduce a latent variable Z for each measured variable to indicate whether it is true or not. Specifically, we have a corresponding variable z_j for the j^{th} measured variable C_j such that: $z_j = 1$ when C_j is true and $z_j = 0$ otherwise. We further denote the observation matrix SC as the observed data X, and take $\theta = (a_1, a_2, ...a_M; b_1, b_2, ...b_M; d)$ as the parameter of the model that we want to estimate. The goal is to get the maximum likelihood estimate of θ for the model containing observed data X and latent variables Z.

The likelihood function $L(\theta; X, Z)$ is given by:

$$\begin{aligned} L(\theta; X, Z) &= p(X, Z|\theta) \\ &= \prod_{j=1}^{N} \left\{ \prod_{i=1}^{M} a_i^{S_iC_j} (1-a_i)^{(1-S_iC_j)} \times d \times z_j \right. \\ &+ \left. \prod_{i=1}^{M} b_i^{S_iC_j} (1-b_i)^{(1-S_iC_j)} \times (1-d) \times (1-z_j) \right\} \end{aligned} \tag{4.9}$$

where, as we mentioned before, a_i and b_i are the conditional probabilities that participant S_i reports the measured variable C_j to be true given that C_j is true or false (i.e., defined in Equation (4.2)). $S_iC_j = 1$ when participant S_i reports that C_j is true, and $S_iC_j = 0$ otherwise. d is the background bias that a randomly chosen measured variable is true. Additionally, we assume participants and measured variables are independent respectively. The likelihood function above describes the likelihood to have current observation matrix X and hidden variable Z given the estimation parameter θ we defined.

4.2.3 Deriving the E-step and M-step

Given the above formulation, substitute the likelihood function defined in Equation (4.9) into the definition of Q function given by Equation (4.7) of Expectation Maximization. The Expectation step (E-step) becomes:

$$
\begin{aligned}
&Q\left(\theta|\theta^{(t)}\right) = E_{Z|X,\theta^{(t)}}[\log L(\theta; X, Z)] \\
&= \sum_{j=1}^{N} \Bigg\{ p(z_j = 1|X_j, \theta^{(t)}) \\
&\times \left[\sum_{i=1}^{M} (S_iC_j \log a_i + (1 - S_iC_j) \log(1 - a_i) + \log d) \right] \\
&+ p(z_j = 0|X_j, \theta^{(t)}) \\
&\times \left[\sum_{i=1}^{M} (S_iC_j \log b_i + (1 - S_iC_j) \log(1 - b_i) + \log(1 - d)) \right] \Bigg\}
\end{aligned}
\tag{4.10}
$$

where X_j represents the j^{th} column of the observed SC matrix (i.e., observations of the j^{th} measured variable from all participants) and $p(z_j = 1|X_j, \theta^{(t)})$ is the conditional probability of the latent variable z_j to be true given the observation matrix related to the j^{th} measured variable and current estimate of θ, which is given by:

$$
\begin{aligned}
&p(z_j = 1|X_j, \theta^{(t)}) \\
&= \frac{p(z_j = 1; X_j, \theta^{(t)})}{p(X_j, \theta^{(t)})} \\
&= \frac{p(X_j, \theta^{(t)}|z_j = 1)p(z_j = 1)}{p(X_j, \theta^{(t)}|z_j = 1)p(z_j = 1) + p(X_j, \theta^{(t)}|z_j = 0)p(z_j = 0)} \\
&= \frac{A(t,j) \times d^{(t)}}{A(t,j) \times d^{(t)} + B(t,j) \times (1 - d^{(t)})}
\end{aligned}
\tag{4.11}
$$

where $A(t,j)$ and $B(t,j)$ are defined as:

$$
\begin{aligned}
A(t,j) &= p(X_j, \theta^{(t)}|z_j = 1) \\
&= \prod_{i=1}^{M} a_i^{(t)S_iC_j}(1 - a_i^{(t)})^{(1-S_iC_j)} \\
B(t,j) &= p(X_j, \theta^{(t)}|z_j = 0) \\
&= \prod_{i=1}^{M} b_i^{(t)S_iC_j}(1 - b_i^{(t)})^{(1-S_iC_j)}
\end{aligned}
\tag{4.12}
$$

$A(t,j)$ and $B(t,j)$ represent the conditional probability regarding observations about the j^{th} measured variable and current estimation of the parameter θ given the j^{th} measured variable is true or false respectively.

Next we simplify Equation (4.10) by noting that the conditional probability of $p(z_j = 1|X_j, \theta^{(t)})$ given by Equation (4.11) is only a function of t and j. Thus, we represent it by $Z(t,j)$. Similarly, $p(z_j = 0|X_j, \theta^{(t)})$ is simply:

$$\begin{aligned} & p(z_j = 0|X_j, \theta^{(t)}) = 1 - p(z_j = 1|X_j, \theta^{(t)}) \\ & = \frac{B(t,j) \times (1 - d^{(t)})}{A(t,j) \times d^{(t)} + B(t,j) \times (1 - d^{(t)})} \\ & = 1 - Z(t,j) \end{aligned} \tag{4.13}$$

Substituting from Equation (4.11) and (4.13) into Equation (4.10), we get:

$$\begin{aligned} & Q\left(\theta|\theta^{(t)}\right) \\ & = \sum_{j=1}^{N} \left\{ Z(t,j) \right. \\ & \times \left[\sum_{i=1}^{M} (S_i C_j \log a_i + (1 - S_i C_j) \log(1 - a_i) + \log d) \right] \\ & + (1 - Z(t,j)) \\ & \times \left. \left[\sum_{i=1}^{M} (S_i C_j \log b_i + (1 - S_i C_j) \log(1 - b_i) + \log(1 - d)) \right] \right\} \end{aligned} \tag{4.14}$$

The Maximization step (M-step) is given by Equation (4.8). We choose θ^* (i.e., $(a_1^*, a_2^*, ...a_M^*; b_1^*, b_2^*, ...b_M^*; d^*)$) that maximizes the $Q\left(\theta|\theta^{(t)}\right)$ function in each iteration to be the $\theta^{(t+1)}$ of the next iteration.

To get θ^* that maximizes $Q\left(\theta|\theta^{(t)}\right)$, we set the derivatives $\frac{\partial Q}{\partial a_i} = 0$, $\frac{\partial Q}{\partial b_i} = 0$, $\frac{\partial Q}{\partial d} = 0$ which yields:

$$\begin{aligned} & \sum_{j=1}^{N} \left[Z(t,j)(S_i C_j \frac{1}{a_i^*} - (1 - S_i C_j) \frac{1}{1 - a_i^*}) \right] = 0 \\ & \sum_{j=1}^{N} \left[(1 - Z(t,j))(S_i C_j \frac{1}{b_i^*} - (1 - S_i C_j) \frac{1}{1 - b_i^*}) \right] = 0 \\ & \sum_{j=1}^{N} \left[Z(t,j) M \frac{1}{d^*} - (1 - Z(t,j)) M \frac{1}{1 - d^*}) \right] = 0 \end{aligned} \tag{4.15}$$

Let us define SJ_i as the set of measured variables the participant S_i actually observes in the observation matrix SC, and $\bar{SJ}_i$ as the set of measured variables participant S_i does not observe. Thus, Equation (4.15) can be rewritten as:

$$
\begin{aligned}
&\sum_{j \in SJ_i} Z(t,j)\frac{1}{a_i^*} - \sum_{j \in \bar{SJ}_i} Z(t,j)\frac{1}{1-a_i^*} = 0 \\
&\sum_{j \in SJ_i} (1 - Z(t,j))\frac{1}{b_i^*} - \sum_{j \in \bar{SJ}_i} (1 - Z(t,j))\frac{1}{1-b_i^*} = 0 \\
&\sum_{j=1}^{N} \left[Z(t,j)\frac{1}{d^*} - (1 - Z(t,j))\frac{1}{1-d^*}) \right] = 0
\end{aligned}
\tag{4.16}
$$

Solving the above equations, we can get expressions of the optimal a_i^*, b_i^* and d^*:

$$
\begin{aligned}
a_i^{(t+1)} &= a_i^* = \frac{\sum_{j \in SJ_i} Z(t,j)}{\sum_{j=1}^{N} Z(t,j)} \\
b_i^{(t+1)} &= b_i^* = \frac{K_i - \sum_{j \in SJ_i} Z(t,j)}{N - \sum_{j=1}^{N} Z(t,j)} \\
d_i^{(t+1)} &= d_i^* = \frac{\sum_{j=1}^{N} Z(t,j)}{N}
\end{aligned}
\tag{4.17}
$$

where K_i is the number of measured variables observed by participant S_i and N is the total number of measured variables in the observation matrix. $Z(t,j)$ is defined in Equation (4.11).

Given the above, The E-step and M-step of EM optimization reduce to simply calculating Equation (4.11) and Equation (4.17) iteratively until they converge. The convergence analysis has been done for EM scheme and it is beyond the scope of this chapter [76]. In practice, we can run the algorithm until the difference of estimation parameter between consecutive iterations becomes insignificant. Since the measured variable is binary, we can compute the optimal decision vector H^* from the converged value of $Z(t,j)$. Specially, h_j is true if $Z(t,j) \geq 0.5$ and false otherwise. At the same time, we can also compute the optimal estimation vector E^* of participant reliability from the converged values of $a_i^{(t)}$, $b_i^{(t)}$ and $d^{(t)}$ based on their relationship given by Equation (4.5). This completes the mathematical development. We summarize the resulting algorithm in the subsection below.

4.2.4 The Final Algorithm

Algorithm 1 Expectation Maximization Algorithm

```
Initialize $\theta$ with random values between 0 and 1
while $\theta^{(t)}$ does not converge do
    for $j = 1 : N$ do
        compute $Z(t, j)$ based on Equation (4.11)
    end for
    $\theta^{(t+1)} = \theta^{(t)}$
    for $i = 1 : M$ do
        compute $a_i^{(t+1)}, b_i^{(t+1)}, d^{(t+1)}$ based on Equation (4.17)
        update $a_i^{(t)}, b_i^{(t)}, d^{(t)}$ with $a_i^{(t+1)}, b_i^{(t+1)}, d^{(t+1)}$ in $\theta^{(t+1)}$
    end for
    $t = t + 1$
end while
Let $Z_j^c$ = converged value of $Z(t, j)$
Let $a_i^c$ = converged value of $a_i^{(t)}$; $b_i^c$ = converged value of $b_i^{(t)}$; $d^c$ = converged value of $d^{(t)}$
for $j = 1 : N$ do
    if $Z_j^c \geq 0.5$ then
        $h_j^*$ is true
    else
        $h_j^*$ is false
    end if
end for
for $i = 1 : M$ do
    calculate $e_i^*$ from $a_i^c$, $b_i^c$ and $d^c$ based on Equation (4.5)
end for
Return the computed optimal estimates of measured variables $C_j = h_j^*$ and source reliability $e_i^*$.
```

In summary of the EM scheme derived above, the input is the observation matrix SC from social sensing data, and the output is the maximum likelihood estimation of participant reliability and measured variable correctness (i.e., E^* and H^* vector defined in Equation (4.4)). In particular, given the observation matrix SC, our algorithm begins by initializing the parameter θ with random values between 0 and 1[†]. The algorithm then performs the E-steps and M-steps iteratively until θ converges. Specifically, we compute the conditional probability of a measured variable to be true (i.e., $Z(t, j)$) from Equation (4.11) and the estimation parameter (i.e., $\theta^{(t+1)}$) from Equation (4.17). After the estimated value of θ converges, we compute the optimal decision vector H^* (i.e., decide whether each measured variable C_j is true or not) based on the converged value of $Z(t, j)$ (i.e., Z_j^c). We can also compute the optimal estimation vector E^* (i.e., the esti-

[†] In practice, if the a rough estimate of the average reliability of participants is known *a priori*, EM will converge faster

mated t_i of each participant) from the converged values of $\theta^{(t)}$ (i.e., a_i^c, b_i^c and d^c) based on Equation (4.5) as shown in the pseudocode of Algorithm 1.

One should note that a theoretical quantification of accuracy of maximum likelihood estimation (MLE) using the EM scheme is well-known in literature, and can be done using the Cramer-Rao lower bound (CRLB) on estimator variance[54]. In estimation theory, if the estimation variance of an unbiased estimator reaches the Cramer-Rao lower bound, the estimator provides the maximum likelihood estimation and the CRLB quantifies the minimum estimation variance. The estimator proposed in this chapter is shown to operate at this bound and hence reach the maximum likelihood estimation [77]. This observation makes it possible to quantify estimation accuracy, or confidence in results generated from our scheme, using the Cramer-Rao lower bound.

4.3 Evaluation

In this section, we carry out experiments to evaluate the performance of the proposed EM scheme in terms of estimation accuracy of the probability that a participant is right or a measured variable is true compared to other state-of-art solutions. We begin by considering algorithm performance for different abstract observation matrices (SC), then apply it to both an emulated participatory sensing scenario and a real world social sensing application. We show that the new algorithm outperforms the state of the art.

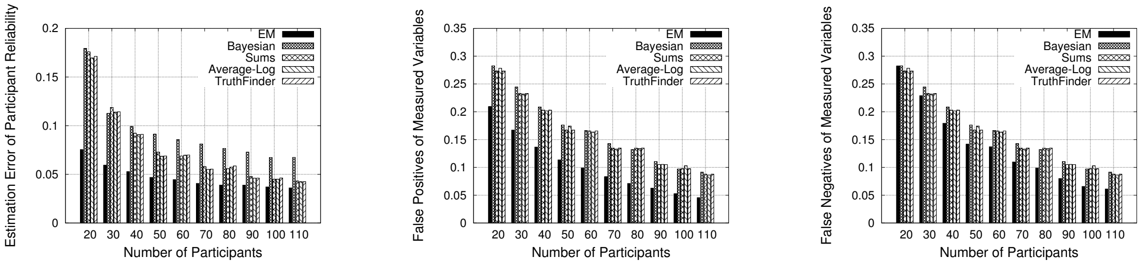
4.3.1 A Simulation Study

We built a simulator in Matlab 7.10.0 that generates a random number of participants and measured variables. A random probability t_i is assigned to each participant S_i representing his/her reliability (i.e., the ground truth probability that they report correct observations). For each participant S_i, L_i observations are generated. Each observation has a probability t_i of being true (i.e., reporting a variable as true correctly) and a probability $1 - t_i$ of being false (reporting a variable as true when it is not). Remember that, as stated in our application model, participants do not report "lack of problems". Hence, they never report a variable to be false. We let t_i be uniformly distributed between 0.5 and 1 in our experiments[‡]. For initialization, the initial values of participant reliability (i.e., t_i) in the evaluated schemes are set to the mean value of its

[‡] In principle, there is no incentive for a participant to lie more than 50% of the time, since negating their statements would then give a more accurate truth

definition range.

In recent work, a heuristic called *Bayesian Interpretation* was demonstrated to outperform all contenders from prior literature [43]. Bayesian Interpretation takes a linear approximation approach to convert the credibility ranks of fact-finders into a Bayesian probability that a participant reports correctly or the measured variable is true. In Bayesian Interpretation, the performance evaluation results were averaged over multiple observation matrices for a given participant reliability distribution. This is intended to approximate performance where highly connected sensing topologies are available (e.g., observations from successive time intervals involving the same set of sources and measured variables). In this chapter, we consider more challenging conditions not investigated in [43], where only a *single observation matrix* is taken as the input into the algorithm. This is intended to understand the algorithm's performance in more realistic scenarios where the sensing topologies are sparsely connected. We compare EM to Bayesian Interpretation and three state-of-art fact-finder schemes from prior literature that can function using only the inputs offered in our problem formulation [40, 7, 8]. Results show a significant performance improvement of EM over all heuristics compared.



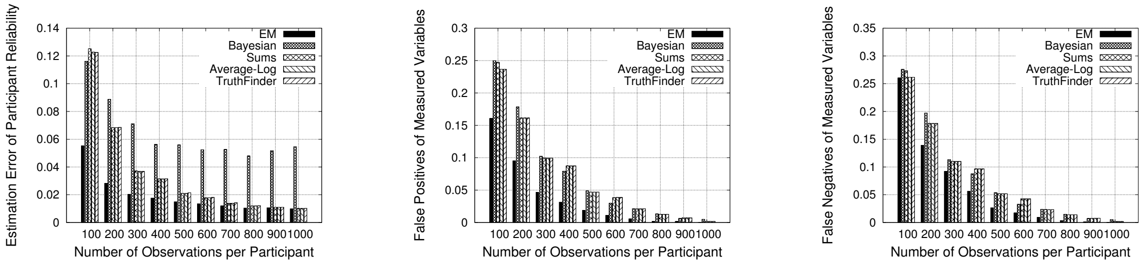
(a) Participant Reliability Estimation Accuracy

(b) Measured Variable Estimation: False Positives

(c) Measured Variable Estimation: False Negatives

Figure 4.1: Estimation Accuracy versus Number of Participants

In the first experiment, we compare the estimation accuracy of EM and the baseline schemes by varying the number of participants in the system. The number of reported measured variables was fixed at 2000, of which 1000 variables were reported correctly and 1000 were misreported. To favor our competition, we "cheat" by giving the other algorithms the correct value of bias d (in this case, $d = 0.5$). The average number of observations per participant was set to 100. The number of participants was varied from 20 to 110. Reported results are averaged over 100 random participant reliability distributions. Results are shown in Figure 4.1. Observe that EM has the smallest estimation error on participant reliability and the least false positives among all schemes under comparison. For false negatives, EM performs similarly to other schemes when the number of participants is small and starts to gain improvements when the number of participants becomes large. Note also that the performance gain of EM becomes large when the number of participants is small, illustrating that EM is more useful when the observation matrix is sparse.



(a) Participant Reliability Estimation Accuracy

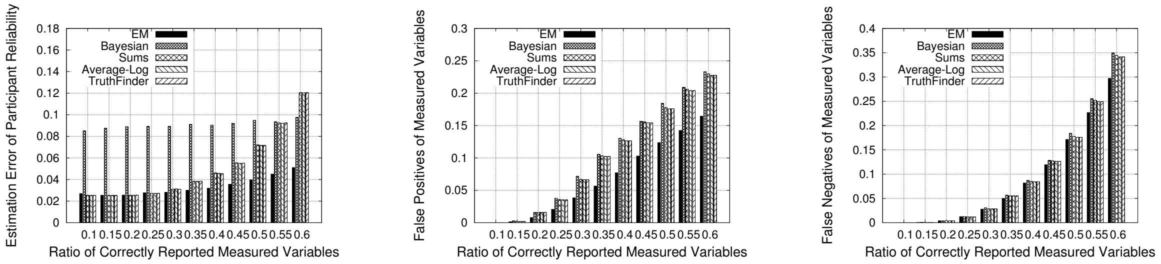
(b) Measured Variable Estimation: False Positives

(c) Measured Variable Estimation: False Negatives

Figure 4.2: Estimation Accuracy versus Average Number of Observations per Participant

The second experiment compares EM with baseline schemes when the average number of observations per participant changes. As before, we fix the number of correctly and incorrectly reported variables to 1000 respectively. Again, we favor our competition by giving their algorithms the correct value of background bias d (here, $d = 0.5$). We also set the number of participants to 30. The average number of observations per participant is varied from 100 to 1000. Results are averaged over 100 experiments. The results are shown in Figure 4.2. Observe that EM outperforms all baselines in terms of both participant reliability estimation accuracy and false positives as the average number of observations per participant changes. For false negatives, EM has similar performance as other baselines when the average number of observations per participant is small and starts to gain advantage as the average number of observations per participant becomes large. As before, the performance gain of EM is higher when the average number of observations per participant is low, verifying once more the high accuracy of EM for sparser observation matrices.

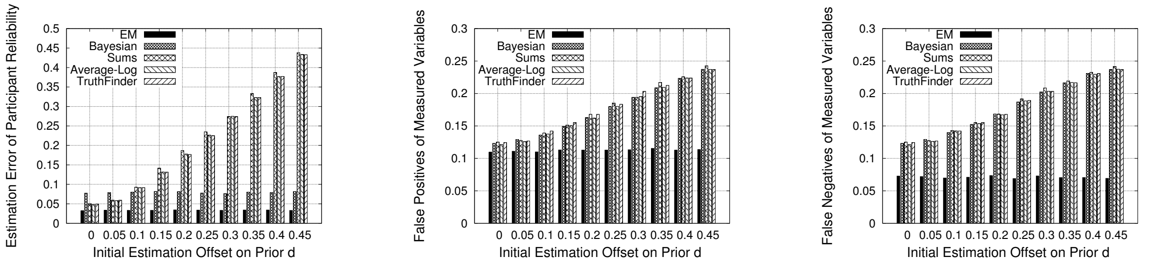
The third experiment examines the effect of changing the measured variable mix on the estimation accuracy of all schemes. We vary the ratio of the number of correctly reported variables to the total number of reported variables from 0.1 to 0.6, while fixing the total number of such variables to 2000. To favor the competition, the background bias d is given correctly to the other algorithms (i.e., $d = varying \quad ratio$). The number of participants is fixed at 30 and the average number of observations per participant is set to 150. Results are averaged over 100 experiments. These results are shown in Figure 4.3. We observe that



(a) Participant Reliability Estimation Accuracy (b) Measured Variable Estimation: False Positives (c) Measured Variable Estimation: False Negatives

Figure 4.3: Estimation Accuracy versus Ratio of Correctly Reported Measured Variables

EM has almost the same performance as other fact-finder baselines when the fraction of correctly reported variables is relatively small. The reason is that the small amount of true measured variables are densely observed and most of them can be easily differentiated from the false ones by both EM and baseline fact-finders. However, as the number of variables (correctly) reported as true grows, EM is shown to have a better performance in both participant reliability and measured variable estimation. Additionally, we also observe that the Bayesian interpretation scheme predicts less accurately than other heuristics under some conditions. This is because the estimated posterior probability of a participant to be reliable or a measured variable to be true in Bayesian interpretation is a linear transform of participant and measured variable credibility values. Those values obtained from a single or sparse observation matrix may not be very accurate and refined [43].

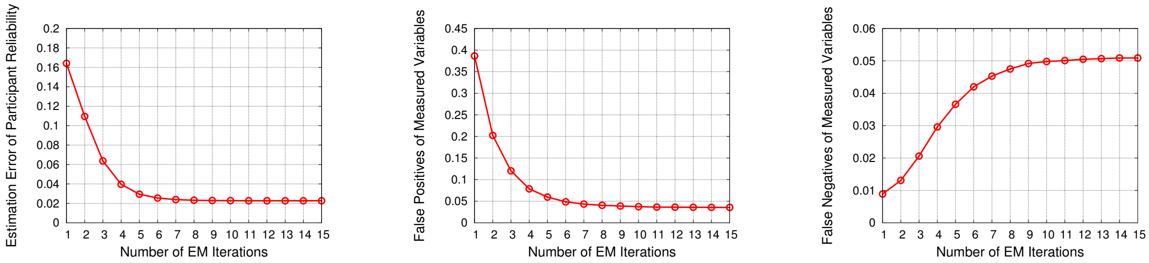


(a) Participant Reliability Estimation Accuracy (b) Measured Variable Estimation: False Positives (c) Measured Variable Estimation: False Negatives

Figure 4.4: Estimation Accuracy versus Initial Estimation Offset on Prior d

The fourth experiment evaluates the performance of EM and other schemes when the offset of the initial estimation on the background bias d varies. The offset is defined as the difference between initial estimation on d and its ground-truth. We fix the number of correctly and incorrectly reported variables to 1000 respectively (i.e., $d = 0.5$). We vary the absolute value of the initial estimate offset on d from 0 to

0.45. The reported results are averaged for both positive and negative offsets of the same absolute value. The number of participants is fixed at 50 and the average number of observations per participant is set to 150. Reported results are averaged over 100 experiments. Figure 4.4 shows the results. We observe that the performance of EM scheme is stable as the offset of initial estimate on d increases. On the contrary, the performance of other baselines degrades significantly when the initial estimate offset on d becomes large. This is because the EM scheme incorporates the d as part of its estimation parameter and provides the MLE on it. However, other baselines depend largely on the correct initial estimation on d (e.g., from the past history) to find out the right number of correctly reported measured variables. These results verify the robustness of the EM scheme when the accurate estimate on the prior d is not available to obtain.



(a) Participant Reliability Estimation Accuracy

(b) Measured Variable Estimation: False Positives

(c) Measured Variable Estimation: False Negatives

Figure 4.5: Convergence Property of the EM Algorithm

The fifth experiment shows the convergence property of the EM iterative algorithm in terms of the estimation error on participant reliability, as well as the false positives and false negatives on measured variables. We fix the number of correctly and incorrectly reported variables to 1000 respectively and set the initial estimate offset on d to 0.3. The number of participants is fixed at 50 and the average number of observations per participant is set to 250. Reported results are averaged over 100 experiments. Figure 4.5 shows the results. We observe that both the estimation error on participant reliability and false positives/negatives on measured variable converge reasonably fast (e.g., less than 10 iterations) to stable values as the number of iterations of EM algorithm increases. It verifies the efficiency of applying EM scheme to solve the maximum likelihood estimation problem formulated.

4.3.2 A Geotagging Case Study

In this subsection, we applied the proposed EM scheme to a typical social sensing application: Geotagging locations of litter in a park or hiking area. In this application, litter may be found along the trails (usually proportionally to their popularity). Participants visiting the park geotag and report locations of litter. Their reports are not reliable however, erring both by missing some locations, as well as misrepresenting other objects as litter. The goal of the application is to find where litter is actually located in the park, while disregarding all false reports.

To evaluate the performance of different schemes, we define two metrics of interest: (i) *false negatives* defined as the ratio of litter locations missed by a scheme to the total number of litter locations in the park, and (ii) *false positives* defined as the ratio of the number of incorrectly labeled locations by a scheme, to the total number of locations in the park. We compared the proposed EM scheme to the Bayesian Interpretation scheme and to voting, where locations are simply ranked by the number of times people report them.

We created a simplified trail map of a park, represented by a binary tree as shown in Figure 4.6. The entrance of the park (e.g., where parking areas are usually located) is the root of the tree. Internal nodes of the tree represent forking of different trails. We assume trails are quantized into discretely labeled locations (e.g., numbered distance markers). In our emulation, at each forking location along the trails, participants have a certain probability P_c to continue walking and $1 - P_c$ to stop and return. Participants who decide to continue have equal probability to select the left or right path. The majority of participants are assumed to be reliable (i.e., when they geotag and report litter at a location, it is more likely than not that the litter exists at that location).

In the first experiment, we study the effect of the number of people visiting the park on the estimation accuracy of different schemes. We choose a binary tree with a depth of 4 as the trail map of the park. Each segment of the trail (between two forking points) is quantized into 100 potential locations (leading to 1500 discrete locations in total on all trails). We define the pollution ratio of the park to be the ratio of the number of littered locations to the total number of locations in the park. The pollution ratio is fixed at 0.1 for the first experiment. The probability that people continue to walk past a fork in the path is set to be 95% and the percent of reliable participants is set to be 80%. We vary the number of participants visiting the park from 5 to 50. The corresponding estimation results of different schemes are shown in Figure 4.7. Observe that both false negatives and false positives decrease as the number of participants increases for all

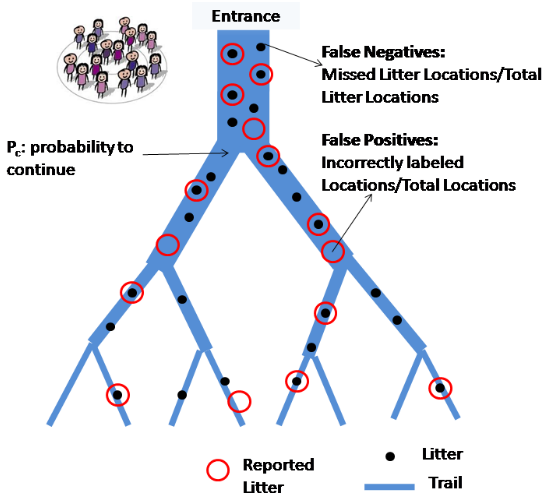
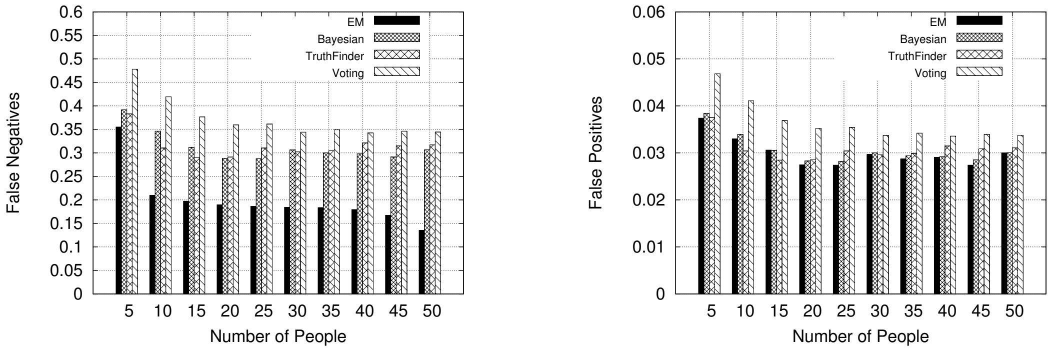

Entrance
False Negatives:
Missed Litter Locations/Total Litter Locations
False Positives:
Incorrectly labeled Locations/Total Locations
P_c: probability to continue
Reported Litter
Litter
Trail

Figure 4.6: A Simplified Trail Map of Geotagging Application



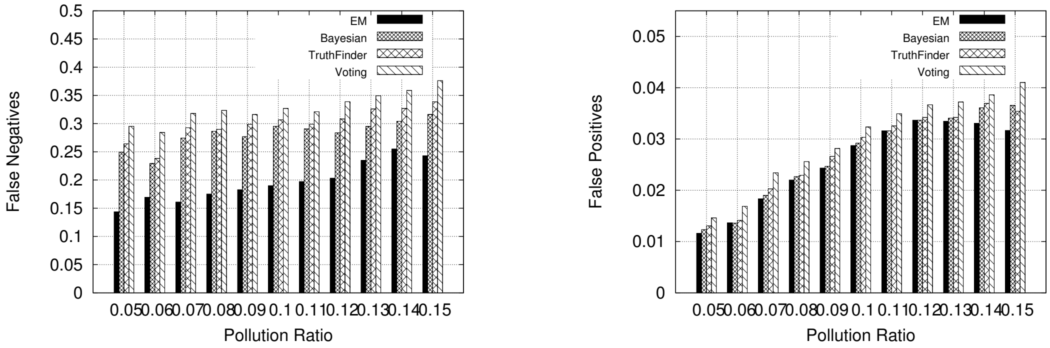
(a) False Negatives (missed/total litter)

(b) False Positives (false/total locations)

Figure 4.7: Litter Geotagging Accuracy versus Number of People Visiting the Park

schemes. This is intuitive: the chances of finding litter on different trails increase as the number of people visiting the park increases. Note that, the EM scheme outperforms others in terms of false negatives, which means EM can find more pieces of litter than other schemes under the same conditions. The improvement becomes significant (i.e., around 20%) when there is a sufficient number of people visiting the park. For the false positives, EM performs similarly to Bayesian Interpretation and Truth Finder scheme and better than voting. Generally, voting performs the worst in accuracy because it simply counts the number of reports complaining about each location but ignores the reliability of individuals who make them.

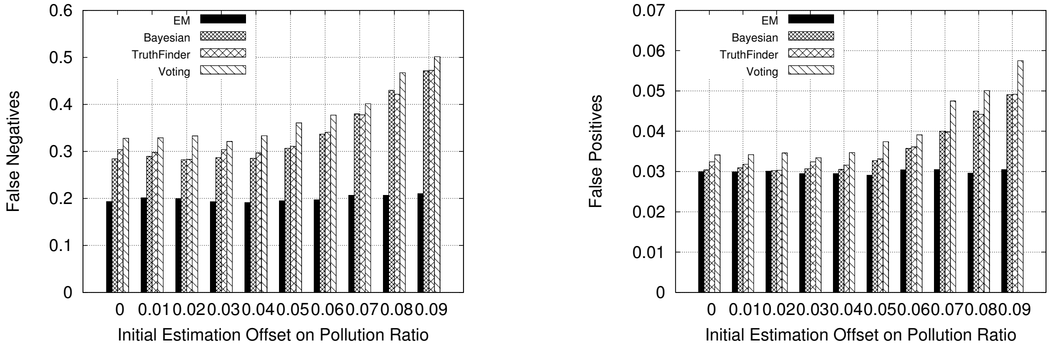
In the second experiment, we show the effect of park pollution ratio (i.e, how littered the park is) on



(a) False Negatives (missed/total litter)

(b) False Positives (false/total locations)

Figure 4.8: Litter Geotagging Accuracy versus Pollution Ratio of the Park



(a) False Negatives (missed/total litter)

(b) False Positives (false/total locations)

Figure 4.9: Litter Geotagging Accuracy versus Initial Estimation Offset on Pollution Ratio of Park

the estimation accuracy of different schemes. The number of individuals visiting the park is set to be 40. We vary the pollution ratio of the park from 0.05 to 0.15. The estimation results of different schemes are shown in Figure 4.8. Observe that both the false negatives and false positives of all schemes increase as the pollution ratio increases. The reason is that: litter is more frequently found and reported at trails that are near the entrance point. The amount of unreported litter at trails that are far from entrance increases more rapidly compared to the total amount of litter as the pollution ratio increases. Note that, the EM scheme continues to find more actual litter compared to other baselines. The performance of false positives is similar to other schemes.

In the third experiment, we evaluate the effect of the initial estimation offset of the pollution ratio on the performance of different schemes. The pollution ratio is fixed at 0.1 and the number of individuals visiting the park is set to be 40. We vary the absolute value of initial estimation offset of the pollution ratio from 0 to 0.09. Results are averaged over both positive and negative offsets of the same absolute value. The estimation results of different schemes are shown in Figure 4.9. Observe that EM finds more actual litter locations and reports less falsely labeled locations than other baselines as the initial estimation offset of pollution ratio increases. Additionally, the performance of EM scheme is stable while the performance of other baselines drops substantially when the initial estimation offset of the pollution ratio becomes large.

The above evaluation demonstrates that the new EM scheme generally outperforms the current state of the art in inferring facts from social sensing data. This is because the state of the art heuristics infer the reliability of participants and correctness of facts based on the hypothesis that their relationship can be approximated *linearly* [7, 8, 43]. However, EM scheme makes its inference based on a maximum likelihood hypothesis that is most consistent with the observed sensing data, thus it provides an optimal solution.

4.3.3 A Real World Application

In this subsection, we evaluate the performance of the proposed EM scheme through a real-world social sensing application, based on Twitter. The objective was to see whether our scheme would distill from Twitter feeds important events that may be newsworthy and reported by media. Specifically, we followed the news coverage of Hurricane Irene and manually selected, as ground truth, 10 important events reported by media during that time. Independently from that collection, we also obtained more than 600,000 tweets originating from New York City during Hurricane Irene using the Twitter API (by specifying keywords as “hurricane”, “Irene” and “flood”, and the location to be New York). These tweets were collected from August 26 until September 2nd, roughly when Irene struck the east coast. Retweets were removed from the collected data to keep sources as independent as possible.

We then generated an observation matrix from these tweets by clustering them based on the Jaccard distance metric (a simple but commonly used distance metric for micro-blog data [78]). Each cluster was taken as a statement of claim about current conditions, hence representing a measured variable in our model. Sources contributing to the cluster were connected to that variable forming the observation matrix. In the formed observation matrix, participants are the twitter users who provided tweets during the observation period, measured variables are represented by the clusters of tweets and the element S_iC_j is set to 1 if

#	Media	Tweet found by EM
1	East Coast Braces For Hurricane Irene; Hurricane Irene is expected to follow a path up the East Coast	@JoshOchs A #hurricane here on the east coast
2	Hurricane Irene's effects begin being felt in NC, The storm, now a Category 2, still has the East Coast on edge.	Winds, rain pound North Carolina as Hurricane Irene closes in http://t.co/0gVOSZk
3	Hurricane Irene charged up the U.S. East Coast on Saturday toward New York, shutting down the city, and millions of Americans sought shelter from the huge storm.	Hurricane Irene rages up U.S. east coast http://t.co/u0XiXow
4	The Wall Street Journal has created a way for New Yorkers to interact with the location-based social media app Foursquare to find the nearest NYC hurricane evacuation center.	Mashable - Hurricane Irene: Find an NYC Evacuation Center on Foursquare ... http://t.co/XMtpH99
5	Following slamming into the East Coast and knocking out electricity to more than a million people, Hurricane Irene is now taking purpose on largest metropolitan areas in the Northeast.	2M lose power as Hurricane Irene moves north - Two million homes and businesses were without power ... http://t.co/fZWkEU3
6	Irene remains a Category 1, the lowest level of hurricane classification, as it churns toward New York over the next several hours, the U.S. National Hurricane Center said on Sunday.	Now its a level 1 hurricane. Let's hope it hits NY at Level 1
7	Blackouts reported, storm warnings issued as Irene nears Quebec, Atlantic Canada.	DTN Canada: Irene forecast to hit Atlantic Canada http://t.co/MjhmeJn
8	President Barack Obama declared New York a disaster area Wednesday, The New York Times reports, allowing the release of federal aid to the state's government and individuals.	Hurricane Irene: New York State Declared A Disaster Area By President Obama
9	Hurricane Irene's rampage up the East Coast has become the tenth billion-dollar weather event this year, breaking a record stretching back to 1980, climate experts said Wednesday.	Irene is 10th billion-dollar weather event of 2011.
10	WASHINGTON- On Sunday, September 4, the President will travel to Paterson, New Jersey, to view damage from Hurricane Irene.	White House: Obama to visit Paterson, NJ Sunday to view damage from Hurricane Irene

Table 4.1: Ground truth events and related tweets found by EM in Hurricane Irene

the tweets of participant S_i belong to cluster C_j, or to 0 otherwise. The matrix was then fed to our EM scheme. We ran the scheme on the collected data and picked the top (i.e., most credible) tweet in each hour. We then checked if our 10 "ground truth" events were reported among the top tweets. Table 4.1 compares the ground truth events to the corresponding top hourly tweets discovered by EM. The results show that indeed all events were reported correctly, demonstrating the value of our scheme in distilling key important information from large volumes of noisy data.

4.4 Discussion

Participants (sources) are assumed to be independent from each other in the current EM scheme. However, sources can sometimes be dependent. That is, they copy observations from each other in real life (e.g., retweets of Twitter). Regarding possible solutions to this problem, one possibility is to remove duplicate observations from dependent sources and only keep the original ones. This can be achieved by applying copy detection schemes between sources [41, 42]. Another possible solution is to cluster dependent sources based on some *source-dependency* metric [10]. In other words, sources in the same cluster are closely related with each other but independent from sources in other clusters. Then we can apply the developed algorithm on top of the clustered sources.

Observations from different participants on a given measured variable are assumed to be *corroborating* in this chapter. This happens in social sensing applications where people do not report "lack of problems". For example, a group of participants involved in a geotagging application to find litter of a park will only report locations where they observe litter and ignore the locations they don't find litter. However, sources can also make conflicting observations in other types of applications. For example, comments from different reviewers in an on-line review system on the same product often contradict with each other. Fortunately, our current model can be flexibly extended to handle conflicting observations. The idea is to extend the estimation vector to incorporate the conflicting states of a measured variable and rebuild the likelihood function based on the extended estimation vector. The general outline of the EM derivation still holds.

The current EM scheme is mainly designed to run on static data sets, where the computation overhead stays reasonable even when the dataset scales up (e.g., the Irene dataset). However, such computation may become less efficient for streaming data because we need to re-run the algorithm on the whole dataset from scratch every time the dataset gets updated. Instead, it will be more technically sound that the algorithm only runs on the updated dataset and combines the results with previously computed ones in an optimal (or suboptimal) way. One possibility is to develop a scheme that can compute the estimated parameters of interest recursively over time using incoming measurements and a mathematical process model. The challenge here is that the relationship between the estimation from the updated dataset and the complete dataset may not be linear. Hence, linear regression might not be generally plausible. Rather, recursive estimation schemes, such as the Recursive EM estimation, would be a better fit.

The developed EM scheme is currently an unsupervised scheme, where we don't assume any data samples to be used to train our model. What happens if we do have some training samples available? For example, we might have some prior knowledge on either source reliability or the correctness of measured variables from other independent ways of data verification. One possible way to incorporate such prior knowledge into our model is to reset the known variables in each iteration of EM to their correct values, which may help the algorithm to converge much faster and also reduce the estimation error. Some techniques exist in machine learning community that try to incorporate the prior knowledge (e.g., source and claim similarity, common-sense reasoning, etc.) into the fact-finding framework by using linear programming [7] or k-partite graph generalization [44]. It would be interesting to investigate if it would be possible to borrow some of these techniques and leverage the the training data to further improve the accuracy of our estimation.

This chapter described a maximum likelihood estimation approach to accurately discover the truth in social sensing applications. The approach can determine the correctness of reported observations given only the measurements sent without knowing the trustworthiness of participants. The optimal solution is obtained by solving an expectation maximization problem and can directly lead to an analytically founded quantification of the correctness of measurements as well as the reliability of participants. Evaluation results show that non-trivial estimation accuracy improvements can be achieved by the proposed maximum likelihood estimation approach compared to other state of the art solutions.

Chapter 5

Real and Asymptotic Confidence Bounds on the Maximum Likelihood Estimation

In the previous chapter, we developed a maximum likelihood estimator based on EM to estimate the reliability of participants and determine the correctness of facts concluded from the data. However, an important problem that remains unanswered from the EM scheme is: what is the *confidence* of the resulting participant reliability estimation? Only by answering this question, can we completely characterize estimation performance, and hence participant reliability in social sensing applications. This chapter presents analytically-founded bounds that quantify the accuracy of such maximum likelihood estimation in social sensing [79]. It is shown that the estimation confidence can be quantified accurately based on both real and asymptotic Cramer-Rao lower bound (CRLB). Additionally, this chapter also proposes an estimator on the accuracy of measured variable classification without knowing the ground truth values of the variables. The results of this chapter are important because they allow social sensing applications to assess the reliability of un-vetted sources (like human participants) to a desired confidence level and estimate the accuracy of measured variable classification under the maximum likelihood hypothesis, in the absence of independent means to verify the data and in the absence of prior knowledge of reliability of sources. This is attained via a well-founded analytic problem formulation and a solution that leverages well-known results in estimation theory.

This chapter is organized as follows: In Section 5.1, we briefly go over the maximum likelihood estimation (MLE) approach and the problem of quantifying source reliability and estimating accuracy of measured variable classification in social sensing applications. We then derive the real and asymptotic CRLBs to compute the confidence interval on source reliability and propose the accuracy estimator on measured variable classification in Section 5.2. The evaluation results are presented in Section 5.3. We discuss the limitations of our model and possible extensions for future work in Section 5.4.

5.1 Problem Statement

In this chapter, we propose a *confidence interval* quantification of the maximum likelihood estimation of participant reliability from EM scheme. In particular, the goal is to demonstrate, in an analytically-founded manner, how to compute the confidence interval of each participant's reliability. Formally, this is given by:

$$(\hat{t}_i^{MLE} - c_p^{lower}, \hat{t}_i^{MLE} + c_p^{upper}) \qquad c\% \qquad i = 1, 2, ..., M \tag{5.1}$$

where $\hat{t}_i^{MLE}$ is the maximum likelihood estimation (MLE) on the reliability of participant S_i ,$c\%$ is the confidence level of the estimation interval, c_p^{lower} and c_p^{upper} represent the lower and upper bound on the estimation deviation from the MLE $\hat{t}_i^{MLE}$ respetively. We target to find c_p^{lower} and c_p^{upper} for a given $c\%$ and an observation matrix SC. It turns out that we need to compute the CRLB of the MLE on the participant reliability in order to obtain the c_p^{lower} and c_p^{upper}. Therefore, our goal in this chapter is to:(i) derive the actual and asymptotic error bounds that characterize the accuracy of the maximum likelihood estimator and compute its confidence interval; (ii) estimate the accuracy of measured variable classification without knowing the ground truth values of the variables; and (iii) derive the dependency of the accuracy of maximum likelihood estimation on parameters of the problem space.

5.2 Confidence Interval Derivation from CRLB

In this section, we show that the confidence interval on source reliability is derived by computing the Cramer-Rao lower bound (CRLB) for the estimation parameters (i.e., θ) and leveraging the asymptotic normality of maximum likelihood estimation. We start with the real CRLB derivation and identify its scalability limitation. We then derive the asymptotic CRLB that works for the sensing topology with a large number of sources. We compute the confidence interval on source reliability based on the derived CRLBs. Additionally, we also derive the expected number of misclassified measured variables (i.e., false measured variable classified as true and true measured variable classified as false).

5.2.1 Real Cramer Rao Lower Bound

We first derive the real CRLB that characterizes the estimation performance of the maximum likelihood estimation of source reliability in social sensing. In estimation theory, the CRLB expresses a lower bound on the estimation variance of a minimum-variance unbiased estimator. In its simplest form, the bound states the variance of any unbiased estimator is at least as high as the inverse of the Fisher information [55]. The estimator that reaches this lower bound is said to be *efficient*. For notational convenience, we denote the observation matrix SC as the observed data X and use $X_{ij} = S_i C_j$ for the following derivation.

The likelihood function (containing hidden variable Z) of the maximum likelihood estimation we get from EM can be expressed as [75]:

$$\begin{aligned} L(\theta; X, Z) &= p(X, Z|\theta) \\ &= \prod_{j=1}^{N} \left\{ \prod_{i=1}^{M} a_i^{X_{ij}} (1 - a_i)^{(1-X_{ij})} \times d \times z_j \right. \\ &\left. + \prod_{i=1}^{M} b_i^{X_{ij}} (1 - b_i)^{(1-X_{ij})} \times (1 - d) \times (1 - z_j) \right\} \end{aligned} \tag{5.2}$$

where z_j is the hidden variable. The EM scheme is used to handle the hidden variable and aims to find:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(X|\theta) \tag{5.3}$$

where

$$\begin{aligned} p(X|\theta) = &\prod_{j=1}^{N} \left\{ \prod_{i=1}^{M} a_i^{X_{ij}} (1 - a_i)^{(1-X_{ij})} \times d \right. \\ &\left. + \prod_{i=1}^{M} b_i^{X_{ij}} (1 - b_i)^{(1-X_{ij})} \times (1 - d) \right\} \end{aligned} \tag{5.4}$$

By definition of CRLB, it is given by

$$CRLB = J^{-1} \tag{5.5}$$

where

$$J = E[\bigtriangledown_\theta \ln p(X|\theta) \bigtriangledown_\theta^H \ln p(X|\theta)] \tag{5.6}$$

where J is the Fisher information of the estimation parameter, $\bigtriangledown_\theta = (\frac{\partial}{\partial a_1}, \ldots \frac{\partial}{\partial a_M}, \frac{\partial}{\partial b_1}, \ldots., \frac{\partial}{\partial b_M})^H$ and H denotes the conjugate transpose operation. In information theory, the Fisher information is a way of measuring the amount of information that an observable random variable X carries about an estimated parameter θ upon which the probability of X depends. The expectation in Equation (5.6) is taken over all values for X with respect to the probability function $p(X|\theta)$ for any given value of θ. Let $\mathcal{X}$ represent the set of all possible values of $X_{ij} \in \{0, 1\}$ for $i = 1, 2...M; j = 1, 2, ...N$. Note $|\mathcal{X}| = 2^{MN}$. Likewise, let $\mathcal{X}_j$ represent the set of all possible values of $X_{ij} \in \{0, 1\}$ for $i = 1, 2...M$ and a given value of j. Note $|\mathcal{X}_j| = 2^M$. Taking the expectation, Equation (5.6) can be rewritten as follows:

$$J = \sum_{X \in \mathcal{X}} \bigtriangledown_\theta \ln p(X|\theta) \bigtriangledown_\theta^H \ln p(X|\theta) p(X|\theta) \tag{5.7}$$

Then, the fisher information matrix can be represented as:

$$J = \begin{bmatrix} A & C \\ C^T & B \end{bmatrix}$$

where submatrices A, B and C contain the elements related with the estimation parameter a_i, b_i and their cross terms respectively. The representative elements A_{kl}, B_{kl} and C_{kl} of A, B and C can be derived as follows:

$$\begin{aligned} A_{kl} &= E\Big[\frac{\partial}{\partial a_k} \ln p(X|\theta) \frac{\partial}{\partial a_l} \ln p(X|\theta)\Big] \\ &= E\Big[\Big(\sum_j \frac{(2X_{kj}-1)Z_j}{a_k^{X_{kj}}(1-a_k)^{(1-X_{kj})}} \sum_q \frac{(2X_{lq}-1)Z_q}{a_l^{X_{lq}}(1-a_l)^{(1-X_{lq})}}\Big)\Big] \\ &= \sum_j \sum_q E\Big[\frac{(2X_{kj}-1)Z_j(2X_{lq}-1)Z_q}{a_k^{X_{kj}}(1-a_k)^{(1-X_{kj})} a_l^{X_{lq}}(1-a_l)^{(1-X_{lq})}}\Big] \end{aligned} \tag{5.8}$$

where

$$Z_j = p(z_j = 1|X) = \frac{A_j \times d}{A_j \times d + B_j \times (1-d)}$$

where

$$A_j = \prod_{i=1}^{M} a_i^{X_{ij}} (1-a_i)^{(1-X_{ij})} \quad B_j = \prod_{i=1}^{M} b_i^{X_{ij}} (1-b_i)^{(1-X_{ij})} \tag{5.9}$$

Z_j is the conditional probability of the measured variable C_j to be true given the observation matrix. After further simplification as shown in Appendix B, A_{kl} can be expressed as the summation of only the expectation terms where $j = q$:

$$
\begin{aligned}
A_{kl} &= \sum_j E\Big[\frac{(2X_{kj}-1)(2X_{lj}-1)Z_j^2}{a_k^{X_{kj}}(1-a_k)^{(1-X_{kj})}a_l^{X_{lj}}(1-a_l)^{(1-X_{lj})}}\Big] \\
&= \sum_{j=1}^{N} \sum_{X \in \mathcal{X}j} \frac{(2X_{kj}-1)(2X_{lj}-1)\prod_{\substack{i=1\\ i\neq k}}^{M} A_{ij} \prod_{\substack{i=1\\ i\neq l}}^{M} A_{ij} d^2}{\prod_{i=1}^{M} A_{ij} d + \prod_{i=1}^{M} B_{ij}(1-d)}
\end{aligned}
\tag{5.10}
$$

where

$$
A_{ij} = a_i^{X_{ij}}(1-a_i)^{(1-X_{ij})} \quad B_{ij} = b_i^{X_{ij}}(1-b_i)^{(1-X_{ij})} \tag{5.11}
$$

Since the inner sum in (5.10) is invariant to the claim index j, we can rewrite $A_{k,l} = N\bar{A}_{k,l}$ where $\bar{A}_{kl}$ is:

$$
\bar{A}_{kl} = \sum_{x \in \mathcal{X}j} \frac{(2X_{kj}-1)(2X_{lj}-1)\prod_{\substack{i=1\\ i\neq k}}^{M} A_{ij} \prod_{\substack{i=1\\ i\neq l}}^{M} A_{ij} d^2}{\prod_{i=1}^{M} A_{ij} d + \prod_{i=1}^{M} B_{ij}(1-d)} \tag{5.12}
$$

It should also be noted that the summation in Equation (5.12) is the same for all j.

By similar calculations, we can obtain the inverse of the Fisher information matrix as follows:

$$
J^{-1} = \frac{1}{N}\begin{bmatrix} \bar{A} & \bar{C} \\ \bar{C}^T & \bar{B} \end{bmatrix}^{-1}
$$

where we define the kl^{th} element of $\bar{B}$, $\bar{C}$ as:

$$
\begin{aligned}
&\bar{B}_{kl} = \\
&\sum_{x \in \mathcal{X}j} \frac{(2X_{kj}-1)(2X_{lj}-1)\prod_{\substack{i=1\\ i\neq k}}^{M} B_{ij} \prod_{\substack{i=1\\ i\neq l}}^{M} B_{ij}(1-d)^2}{\prod_{i=1}^{M} A_{ij} d + \prod_{i=1}^{M} B_{ij}(1-d)}
\end{aligned}
\tag{5.13}
$$

$$
\begin{aligned}
&\bar{C}_{kl} = \\
&\sum_{x \in \mathcal{X}j} \frac{(2X_{kj}-1)(2X_{lj}-1)\prod_{\substack{i=1\\ i\neq k}}^{M} A_{ij} \prod_{\substack{i=1\\ i\neq l}}^{M} B_{ij} d(1-d)}{\prod_{i=1}^{M} A_{ij} d + \prod_{i=1}^{M} B_{ij}(1-d)}
\end{aligned}
\tag{5.14}
$$

Note that the sum of $\bar{A}_{kl}$, $\bar{B}_{kl}$ and $\bar{C}_{kl}$ are over the 2^M different permutations of X_{ij} for $i = 1, 2, ...M$ at a given j. This is much smaller than the 2^{MN} permutations of $\mathcal{X}$.

This gives us the real CRLB. Note that more measured variables simply lead to better estimates for θ as the variance decreases as $\frac{1}{N}$. The decrease in variance for the estimates as a function of M is more complicated. We can only compute it numerically.

5.2.2 Asymptotic Cramer Rao Lower Bound

Observe that the complexity of the real CRLB computation in the above subsection is exponential with respect to the number of sources (i.e., M) in the system. Therefore, it is inefficient (or infeasible) to compute the real CRLB when the number of sources becomes large. In this subsection, we outline the asymptotic CRLB for efficient computation in the sensing topology with a large number of sources. The asymptotic CRLB is derived based on the assumption that the correctness of the hidden variable (i.e., z_j) can be correctly estimated from EM. This is a reasonable assumption when the number of sources is sufficient [75]. Under this assumption, the log-likelihood function of the maximum likelihood estimation we get from EM can be expressed as follows:

$$
\begin{aligned}
& l_{em}(x;\theta) = \sum_{j=1}^{N} \Bigg\{ \\
& z_j \times \left[\sum_{i=1}^{M} (X_{ij} \log a_i + (1 - X_{ij}) \log(1 - a_i) + \log d) \right] \\
& + (1 - z_j) \\
& \times \left[\sum_{i=1}^{M} (X_{ij} \log b_i + (1 - X_{ij}) \log(1 - b_i) + \log(1 - d)) \right] \Bigg\}
\end{aligned} \tag{5.15}
$$

We first compute the Fisher Information Matrix at the MLE from the log-likelihood function given by Equation (5.15). According to prior work [75], the maximum likelihood estimator $\hat{\theta}_{MLE}$ is given by:

$$
\hat{a}_i^{MLE} = \frac{\sum_{j=1}^{N} X_{ij} Z_j^c}{\sum_{j=1}^{N} Z_j^c} \quad \hat{b}_i^{MLE} = \frac{\sum_{j=1}^{N} X_{ij}(1 - Z_j^c)}{N - \sum_{j=1}^{N} Z_j^c} \tag{5.16}
$$

where Z_j^c is the converged probability of the j^{th} measured variable to be true from EM algorithm. Observe that each $\hat{a}_i^{MLE}$ or $\hat{b}_i^{MLE}$ is computed from N independent samples (i.e., measured variables).

Plugging $l_{em}(x;\theta)$ given by Equation (5.15) into the Fisher information defined in Equation (5.6), we have the representative element of Fisher Information Matrix from N measured variables as:

$$(J(\hat{\theta}_{MLE}))_{i,j} \tag{5.17}$$
$$= \begin{cases} 0 & i \neq j \\ -E_X\left[\frac{\partial^2 l_{em}(x;a_i)}{\partial a_i^2}|_{a_i=\hat{a}_i^{MLE}}\right] & i = j \in [1, M] \\ -E_X\left[\frac{\partial^2 l_{em}(x;b_i)}{\partial b_i^2}|_{b_i=\hat{b}_i^{MLE}}\right] & i = j \in (M, 2M] \end{cases}$$

Substituting the log-likelihood function in Equation (5.15) and MLE in Equation (5.16) into Equation (5.17), the asymptotic CRLB (i.e., the inverse of the Fisher Information Matrix) can be written as:

$$(J^{-1}(\hat{\theta}_{MLE}))_{i,j} = \begin{cases} 0 & i \neq j \\ \frac{\hat{a}_i^{MLE}\times(1-\hat{a}_i^{MLE})}{N\times d} & i = j \in [1, M] \\ \frac{\hat{b}_i^{MLE}\times(1-\hat{b}_i^{MLE})}{N\times(1-d)} & i = j \in (M, 2M] \end{cases} \tag{5.18}$$

Note that the asymptotic CRLB is independent of M under the assumption that M is sufficient, and it can be quickly computed from the MLE of the EM scheme.

5.2.3 Confidence Interval

In this subsection, we show that the confidence interval of source reliability can be obtained by using the CRLB we derived in previous sections and leveraging the asymptotic normality of the maximum likelihood estimation.

The maximum likelihood estimator posses a number of attractive asymptotic properties. One of them is called *asymptotic normality*, which basically states the MLE estimator is asymptotically distributed with Gaussian behavior as the data sample size goes up, in particular[56]:

$$(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N(0, J^{-1}(\hat{\theta}_{MLE})) \tag{5.19}$$

where J is the Fisher Information Matrix computed from all samples, θ_0 and $\hat{\theta}_{MLE}$ are the true value and the maximum likelihood estimation of the parameter θ respectively. The Fisher information at the MLE is used to to estimate its true (but unknown) value [55]. Hence, the asymptotic normality property means that in a regular case of estimation and in the distribution limiting sense, the maximum likelihood estimator $\hat{\theta}_{MLE}$ is unbiased and its covariance reaches the Cramer-Rao lower bound (i.e., an efficient estimator).

From the asymptotic normality of the maximum likelihood estimator [75], the error of the corresponding estimation on θ follows a normal distribution with zero mean and the covariance matrix given by the CRLB we derived in previous subsections. Let us denote the variance of estimation error on parameter a_i as $var(\hat{a}_i^{MLE})$. Recall the relation between source reliability (i.e., t_i) and estimation parameter a_i and b_i is $t_i = \frac{a_i \times d}{a_i \times d + b_i \times (1-d)}$. For a sensing topology with small values of M and N, the estimation of t_i has a complex distribution and its estimation variance can be approximated [54]. For a sensing topology with sufficient M and N (i.e., under asymptotic condition), the denominator of t_i can be approximated as s_i based on Equation (4.5).* Therefore, $(\hat{t}_i^{MLE} - t_i^0)$ also follows a normal distribution with 0 mean and variance given by:

$$var(\hat{t}_i^{MLE}) = \left(\frac{d}{s_i}\right)^2 var(\hat{a}_i^{MLE}) \tag{5.20}$$

Hence, we are now able to obtain the confidence interval that can be used to quantify the estimation accuracy of the maximum likelihood estimation on source reliability. The confidence interval of the reliability estimation of source S_i (i.e., $\hat{t}_i^{MLE}$) at confidence level p is given by the following:

$$(\hat{t}_i^{MLE} - c_p\sqrt{var(\hat{t}_i^{MLE})}, \hat{t}_i^{MLE} + c_p\sqrt{var(\hat{t}_i^{MLE})}) \tag{5.21}$$

where c_p is the standard score (z-score) of the confidence level p. For example, for the 95% confidence level, $c_p = 1.96$. Therefore, the derived confidence interval of the source reliability MLE, as we demonstrated, can be computed by using the CRLB derived in this section.

5.2.4 Estimation of Measured Variable Classification Accuracy

In previous subsections, we discussed how to compute the CRLB and the confidence interval in source reliability from the maximum likelihood estimation (MLE) of the EM algorithm. However, one problem remains to be answered is how to estimate the accuracy of the measured variable classification (i.e, false positives and false negatives) without having the ground truth values of the measured variables at hand. In

*The value of s_i can be estimated as $\frac{L_i}{N}$, where L_i is the number of observations reported by source S_i

this subsection, we propose a quick and effective method to answer the above question under the maximum likelihood hypothesis.

The results of the EM algorithm not only offered the MLE on the estimation parameters (i.e., θ) but also the probability of each measured variable to be true, which is given by [75]:

$$Z_j^* = p(z_j = 1 | X_j, \theta^*) \tag{5.22}$$

where X_j is the observed data of the measured variable C_j and θ^* is the maximum likelihood estimation of the parameter. Since the measured variable is binary, it is judged as true if $Z_j^* \geq 0.5$ and false otherwise. Based on the above definition, the false positives and false negatives of the measured variable classification can be estimated as follows:

$$\begin{aligned} FP &= \sum_{j:Z_j^* \geq 0.5}^{N} \{Z_j^* \times 0 + (1 - Z_j^*) \times 1\} \\ &= \sum_{j:Z_j^* \geq 0.5}^{N} (1 - Z_j^*) \end{aligned} \tag{5.23}$$

$$\begin{aligned} FN &= \sum_{j:Z_j^* < 0.5}^{N} \{Z_j^* \times 1 + (1 - Z_j^*) \times 0\} \\ &= \sum_{j:Z_j^* < 0.5}^{N} Z_j^* \end{aligned} \tag{5.24}$$

where FP and FN stand for false positives and false negatives respectively. From above equations, we can compute the estimated false positives and false negatives of the measured variable classification under the maximum likelihood hypothesis. This enables us to estimate the accuracy of the measured variable classification without knowing the ground truth values a priori.

In this section, we derived a confidence interval in source reliability and proposed an accuracy estimator on the measured variable classification. This allows social sensing applications to assess the quality of their estimation on source reliability as well as the accuracy of measured variable classification. In the following section, we evaluate the performance of the computed confidence bounds on source reliability and the estimated false positives and false negatives on the measured variable classification.

5.3 Evaluation

In this section, we present the evaluation of the performance of the computed confidence interval of source reliability, the derived CRLBs, and the accuracy estimation of the measured variable classification in social sensing. The reported CRLB results are computed upon the estimated a's and b's instead of the ground truth. In practice, it provides a sense of the sensitivity (or significance) of the estimated values. We built a simulator in Matlab 7.10.0 that generates a random number of sources and measured variables. A random probability P_i is assigned to each source S_i representing his/her reliability (i.e., the ground truth probability that they report correct observations). For each source S_i, L_i observations are generated. Each observation has a probability P_i of being true (i.e., reporting a variable as true correctly) and a probability $1-P_i$ of being false (reporting a variable as true when it is not). One can think of these variables as observed "problems". Sources do not report "lack of problems". Hence, they never report a variable to be false. We let P_i be uniformly distributed between 0.5 and 1 in our experiments. The background prior d is set to be 0.5 unless otherwise specified.

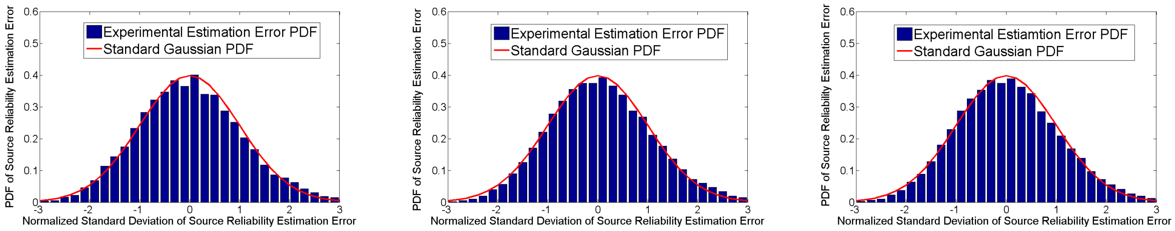
5.3.1 Evaluation of Confidence Interval

In this subsection, we evaluate the performance of the confidence interval in source reliability derived in the previous section. We carried out experiments over three different observation matrix scales: small, medium and large. The simulation parameters of three observation matrix scales are listed in Table 5.1. The average observations reported by each source is set to 100. For each observation matrix scale, we run the EM algorithm and compute the confidence interval in source reliability based on Equation (5.21). We repeat the experiments 100 times for each observation matrix scale. Three representative confidence levels (i.e., 68%, 90%, 95%) are used in our evaluation.

Figure 5.1 shows the normalized probability density function (PDF) of source reliability estimation error over three observation matrix scales. We computed the experimental PDF by leveraging the actual

Observation Matrix Scale	Number of Sources	Number of True Measured Variables	Number of False Measured Variables
Small	100	500	500
Medium	200	1000	1000
Large	300	2000	2000

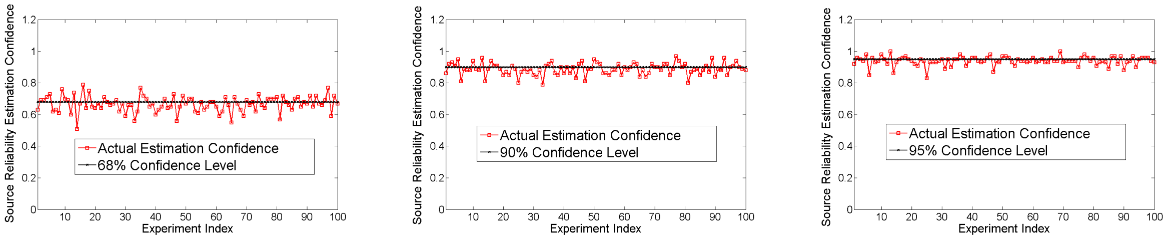
Table 5.1: Parameters of Three Typical Observation Matrix Scale



(a) Small Observation Matrix (b) Medium Observation Matrix (c) Large Observation Matrix

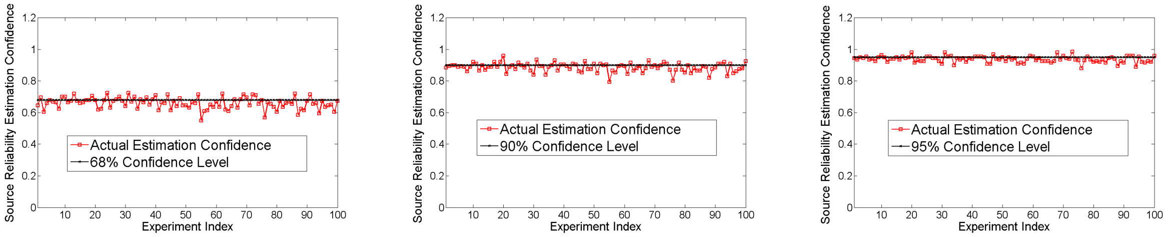
Figure 5.1: Normalized Source Reliability Estimation Error PDF

estimation error (i.e., compare to the ground truth) and the confidence interval derived in Section 5.2. We compared the experimental PDF with the standard Gaussian distribution to verify the asymptotic normality property of estimation results. We observe the experimental PDF match well with the theoretical Gaussian distribution over three observation matrix scales.



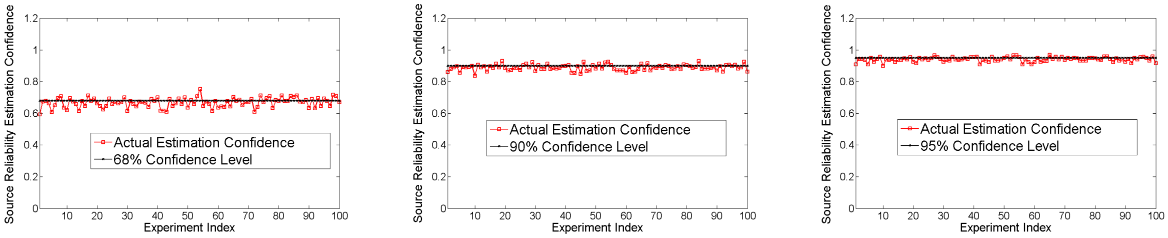
(a) 68% Confidence Level (b) 90% Confidence Level (c) 95% Confidence Level

Figure 5.2: Source Reliability Estimation Confidence for Small Observation Matrix



(a) 68% Confidence Level (b) 90% Confidence Level (c) 95% Confidence Level

Figure 5.3: Source Reliability Estimation Confidence for Medium Observation Matrix



(a) 68% Confidence Level (b) 90% Confidence Level (c) 95% Confidence Level

Figure 5.4: Source Reliability Estimation Confidence for Large Observation Matrix

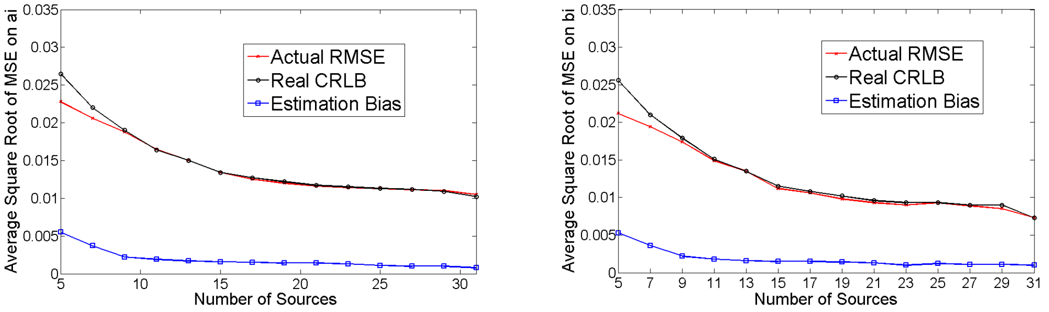
Figure 5.2 shows the comparison between the actual estimation confidence and three different confidence levels we set for the small observation matrix scenario. The actual estimation confidence is computed as the percentage of sources whose estimation error stay within the corresponding confidence bound for every experiment. This percentage represents the probability that a randomly chosen source keeps its reliability estimation error within the confidence bound. We observe that the actual estimation confidence of using 3 different confidence bounds stays close to the corresponding confidence levels we used for the experiment. Moreover, at higher confidence levels, a lower fluctuation of the actual estimation confidence is observed. Similar results are observed for the medium and large observation matrices as well, which are shown in Figure 5.3 and Figure 5.4. Additionally, we also note that the fluctuation of the actual estimation confidence decreases as the observation matrix scale increases. This is because the estimation variance characterized by CRLB is inversely proportional to the number of measured variables in the system, which will be further evaluated in the next subsection.

5.3.2 Evaluation of CRLB

In this subsection, we evaluate the performance of derived CRLBs (both real and asymptotic) in Section 5.2.1 and 5.2.2 by comparing them to the actual estimation variance of the estimation parameter (i.e., a_i, b_i). The actual estimation variance is characterized by the average RMSE (square root of the mean squared error) of all sources.

Scalability Study

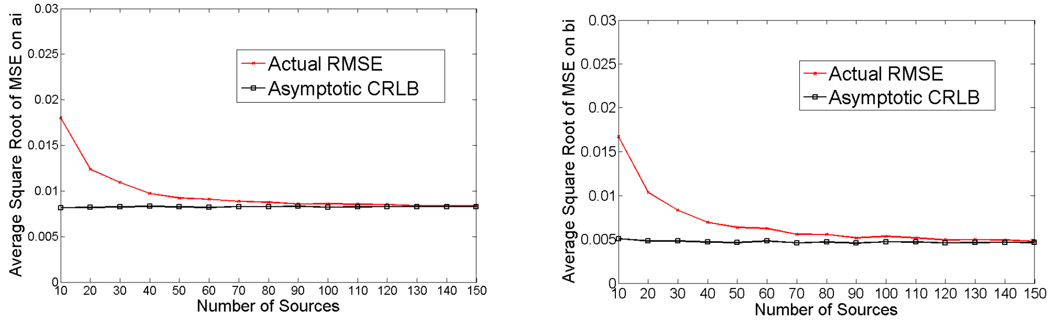
We first evaluate the scalability of CRLB performance with respect to the sensing topology (i.e, M and N). The first experiment evaluates the effect of the number of sources (i.e., M) in the system on the CRLB performance. We start with the real CRLB evaluation. We fix the true and false measured variables to be



(a) Real CRLB of a_i

(b) Real CRLB of b_i

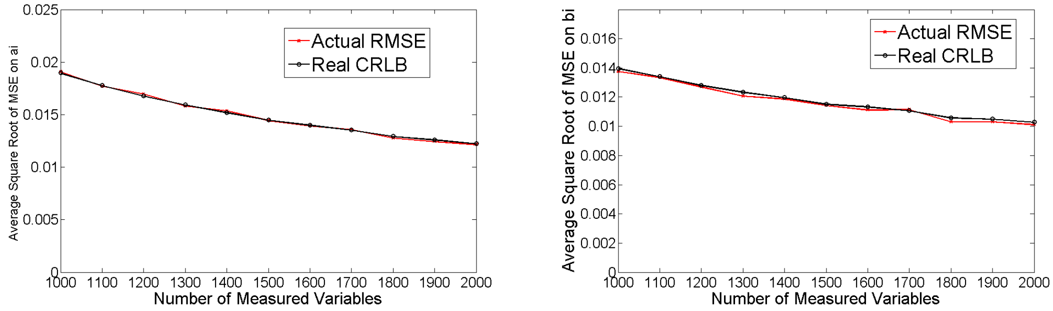
Figure 5.5: Real CRLB of a_i and b_i versus Varying M



(a) Asymptotic CRLB of a_i

(b) Asymptotic CRLB of b_i

Figure 5.6: Asymptotic CRLB of a_i and b_i versus Varying M

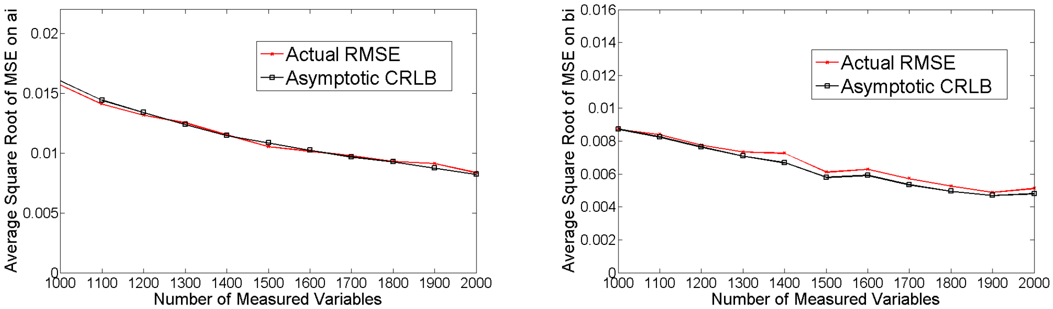


(a) Real CRLB of a_i

(b) Real CRLB of b_i

Figure 5.7: Real CRLB of a_i and b_i versus Varying N

1000 respectively, the average observations per source is set to 100. We vary the number of sources from 5 to 31. Reported results are averaged over 100 experiments and are shown in Figure 5.5. Observe that the real CRLB tracks the actual estimation variance of estimation parameters accurately even when the number of sources is small (e.g., $M \leq 20$) in the system. We also observe that the RMSE is smaller than the Real CRLB when there are too few sources. This is because the MLE is biased on those points due to the small dataset. As illustrated in Section 5.2.1, the computation of real CRLB does not scale with the number of



(a) Asymptotic CRLB of a_i

(b) Asymptotic CRLB of b_i

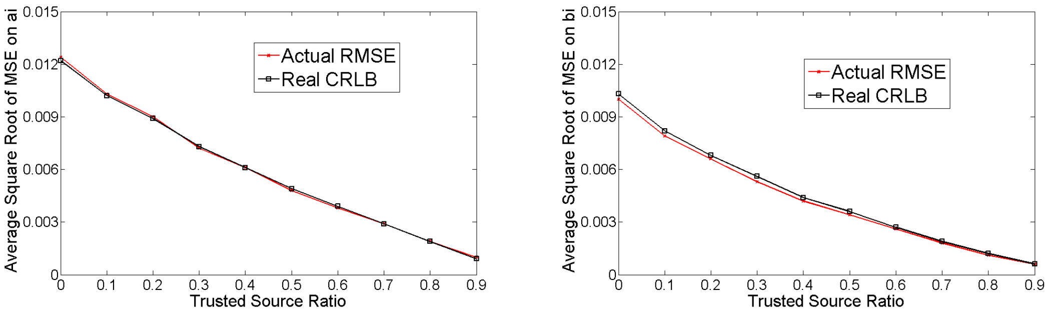
Figure 5.8: Asymptotic CRLB of a_i and b_i versus Varying N

sources in the system. Hence, we also evaluate the the performance of asymptotic CRLB when the number of sources becomes large. We keep the experimental configuration the same as above, but change the number of sources from 10 to 150. Results are shown in Figure 5.6. We observe that the asymptotic CRLB deviates from the actual estimation variance when the number of sources is small (e.g., $M \leq 20$). However, as the number of sources becomes sufficient in the network, the actual RMSE converges to the asymptotic CRLB quickly and the difference between the two becomes insignificant.

The second experiment compares the derived CRLBs (both real and asymptotic) to the actual RMSE of estimation parameters when the number of measured variables (i.e., N) changes. As shown in Section 5.2, both the real and asymptotic CRLB decrease as $\frac{1}{N}$. As before, we first evaluate the performance of real CRLB. We fix the number of sources as 20, the average number of observations per source is set to 100. We also keep the number of true and false measured variables the same. The number of measured variables varies from 1000 to 2000. Reported results are averaged over 100 experiments and are shown in Figure 5.7. We observe that the real CRLB is able to track the actual RMSE on estimation parameter correctly and they both decrease approximately as $\frac{1}{N}$ when the number of measured variable increases. Similarly, we carry out the experiment to evaluate the performance of asymptotic CRLB. We keep the experimental configuration the same as above, but set the number of sources to be 100. Results are shown in Figure 5.8. We observe that the asymptotic CRLB also follows closely on the actual RMSE of the estimation parameter and they reduce approximately as $\frac{1}{N}$ when the number of measured variable increases.

Trustworthiness and Freshness Study

In the trustworthiness study, we evaluate the estimation performance of CRLB when the ratio of trusted sources in the system changes. The trusted sources are the sources who always make correct observations (i.e, their reliability is 1) and the ratio of trusted sources is the ratio of the number of trusted sources over the total number of sources in the system. We start with the real CRLB evaluation. We fix the true and false measured variables to be 1000 respectively, the number of sources is set to 20 and each source reports 100 observations on average. We vary the trusted source ratio from 0 to 0.9. Reported results are averaged over 100 experiments and shown in Figure 5.9. Observe that that the real CRLB tracks the actual estimation variance tightly when the trusted source ratio changes. We also note that both the real CRLB and actual estimation variance of estimation parameters improve as the trusted source ratio increases. The reason is: the estimation error decreases as the ratio of sources with $t_i = 1$ increases. This is also reflected by the fact that $b_i = 0$ for trusted sources and the asymptotic variance goes to zero as one can see in (5.18). Similarly, we carry out experiments to evaluate the performance of the asymptotic CRLB. We keep the experiment configuration the same as above, but set the number of sources to be 100. Results are shown in Figure 5.10. We observe that the asymptotic CRLB also follows the actual estimation variance of the estimation parameters correctly and they improve as the trusted source ratio increases.

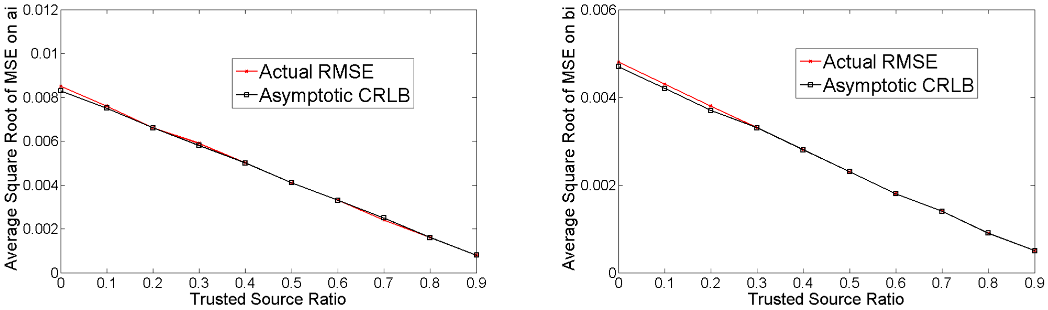


(a) Real CRLB of a_i

(b) Real CRLB of b_i

Figure 5.9: Real CRLB of a_i and b_i versus Trusted Sources Ratio

In the freshness study, we evaluate the estimation performance of CRLB when the freshness ratio of the data that the algorithm takes as input changes. The freshness ratio is defined as the ratio of the input data size (in terms of the number of observations) normalized by a pre-defined data size. This ratio reflects the sparsity of the sensing topology when the algorithm starts to run. We start with the real CRLB evaluation. We fix the true and false measured variables to be 1000 respectively. The number of sources is set to 20.



(a) Asymptotic CRLB of a_i

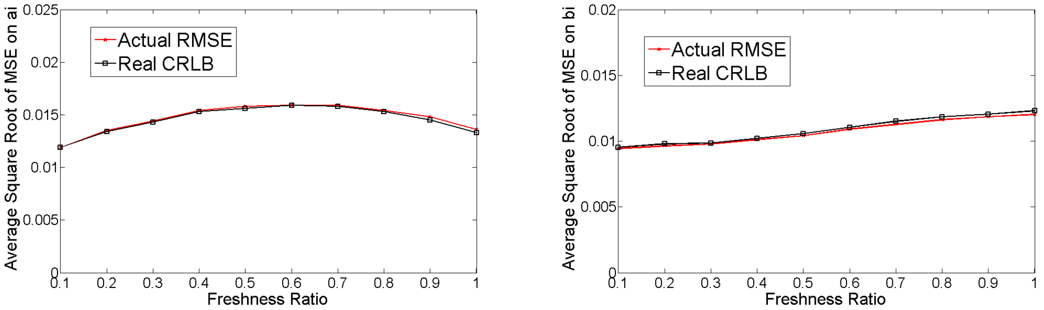
(b) Asymptotic CRLB of b_i

Figure 5.10: Asymptotic CRLB of a_i and b_i versus Trusted Sources

The input data size that is used for the freshness ratio normalization (i.e, having freshness ratio of 1) is set to 1000 observations per source. We vary the freshness ratio from 0.1 to 1. Reported results are averaged over 100 experiments and shown in Figure 5.11. We observe that the real CRLB tracks the actual RMSE of the estimation parameters correctly as the freshness ratio changes. We also note that the estimation variance of parameter a_i first increases and then decreases while the estimation variance of parameter b_i increases as the freshness ratio increases. The reason is: two factors affect the variance of the estimation parameters in different directions when the freshness ratio changes. One factor is the probability a source reports a measured variable (i.e, s_i). This factor increases as the freshness ratio increases, which will enlarge the estimation variance of a_i and b_i based on (4.5). The other factor is the estimation variance of the source reliability (i.e, t_i), which decreases as the freshness ratio increases. Hence, the estimation variance of a_i is first dominated by the first factor and then by the second one while the estimation variance of b_i is dominated by the first factor in the evaluation range as the freshness ratio increases. We then carry out similar experiments to evaluate the performance of the asymptotic CRLB. We keep the experiment configuration the same as above, but set the number of sources to be 100. Results are shown in Figure 5.12. We observe that the asymptotic CRLB also follows the actual estimation variance of the estimation parameters tightly and their trends are similar as those of real CRLB.

Robustness Study

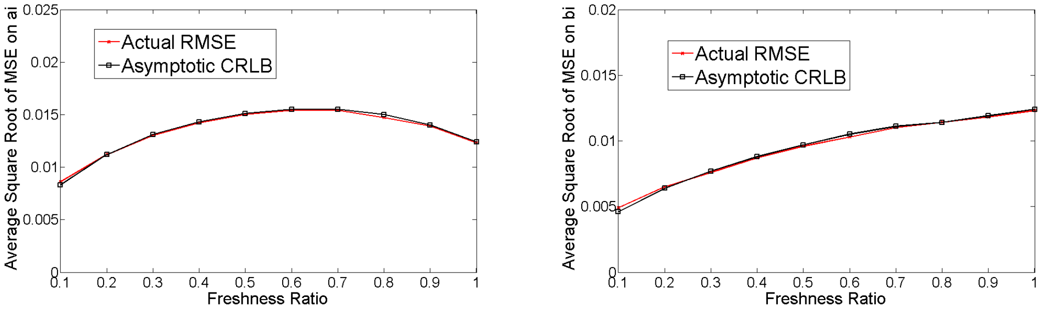
In the robustness study, we evaluate the robustness (or sensitivity) of the estimation performance and the derived CRLBs when the number of sources changes under different source reliability distributions. The key characteristic that determines the resilience of a network is the network topology. The social sensing



(a) Real CRLB of a_i

(b) Real CRLB of b_i

Figure 5.11: Real CRLB of a_i and b_i versus Freshness



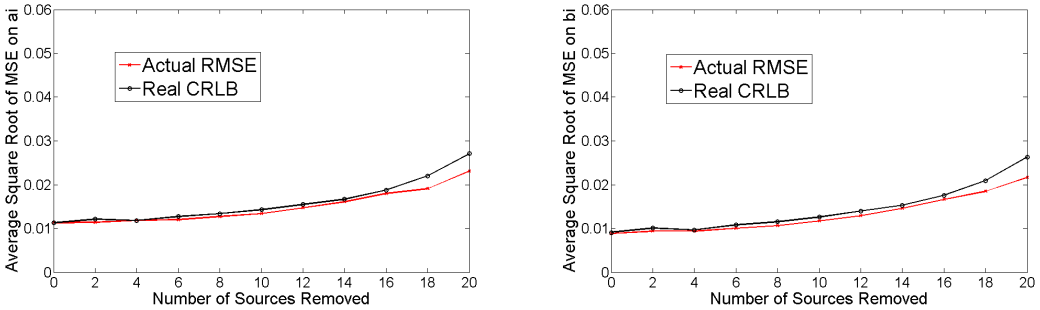
(a) Asymptotic CRLB of a_i

(b) Asymptotic CRLB of b_i

Figure 5.12: Asymptotic CRLB of a_i and b_i versus Freshness

topology is characterized by the link connections between sources and two sets of measured variables (i.e., true and false). The link connection skew is mainly determined by the source reliability distribution. We consider two representative network topologies: scale-free and exponential topologies in our evaluation. For scale-free topology, sources have diverse reliability and the probability for sources to have different reliability is similar. For exponential topology, sources have similar reliability and nodes with higher reliability are exponentially less probable. Our experiments were done by source removal (i.e., sources are randomly selected and removed from the system). This represents the scenario where random sources decide to quit the sensing application or their sensing devices fail. However, it is equivalent to reversing the steps and investigating the addition of sources.

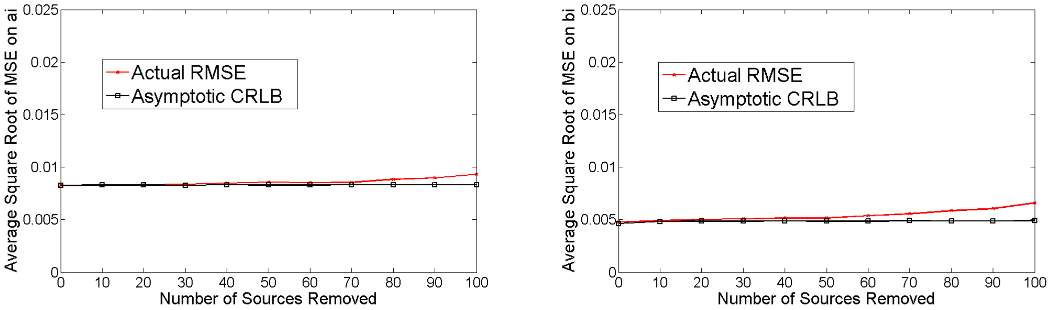
In the first experiment, we evaluate the estimation performance and the derived CRLBs of the scale-free network topology. To generate the scale-free network topology, we let the source reliability follow a uniform distribution on its definition range. We first evaluate the performance of the real CRLB compared to the actual RMSE on the estimation parameter. We fix both the number of true and false measured variables to 1000. The average number of observations per source is set to 100. We start with 25 sources and gradually remove sources from the system. Figure 5.13 shows the real CRLB and actual RMSE of the estimation parameter. Observe that the estimation performance (i.e., actual RMSE) degrades gracefully and the real CRLB tracks the actual RMSE reasonably well as the number of removed sources increases. Also note that the real CRLB deviates slightly from the RMSE when majority of sources are removed from the system. We then repeat similar experiments for the asymptotic CRLB as well. We start with 150 sources and gradually remove the sources from the system. Results are shown in Figure 5.14. The results for asymptotic CRLB are similar to real CRLB.



(a) Real CRLB of a_i

(b) Real CRLB of b_i

Figure 5.13: Real CRLB of a_i and b_i versus Source Removal of Scale-free Topology

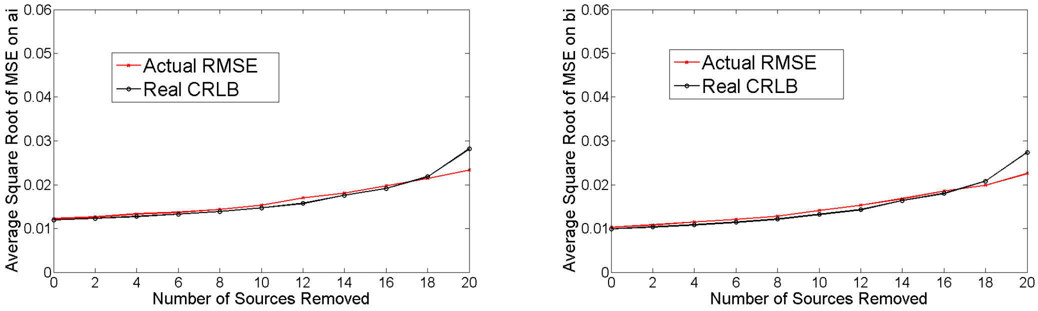


(a) Asymptotic CRLB of a_i

(b) Asymptotic CRLB of b_i

Figure 5.14: Asymptotic CRLB of a_i and b_i versus Source Removal of Scale-free Topology

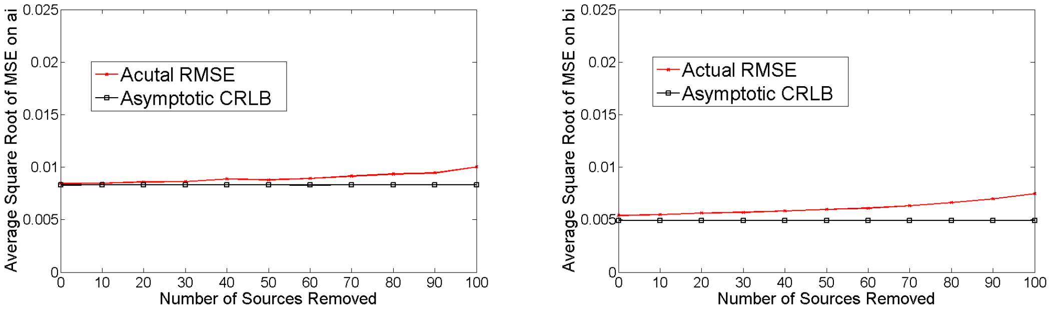
In the second experiment, we evaluate the estimation performance and the derived CRLBs of the exponential network topology. To generate the exponential network topology, we let the source reliability follow a normal distribution (with the mean value as the mean of its definition range and a reasonably small variance). As we did before, we first evaluate the performance of the real CRLB compared to the actual RMSE on the estimation parameter. The standard deviation of the normal distribution of source reliability is set to 0.02, other settings are kept the same as the first experiment. Figure 5.15 shows the real CRLB and actual RMSE of the estimation parameter. Observe that actual RMSE increases gradually as the number of removed sources grows and the real CRLB tracks the actual RMSE well. We then repeat similar experiments for the asymptotic CRLB as well. The experimental settings are kept the same as the first experiment. Results are shown in Figure 5.16. Similar results as we have for the real CRLB are observed for the asymptotic CRLB.



(a) Real CRLB of a_i

(b) Real CRLB of b_i

Figure 5.15: Real CRLB of a_i and b_i versus Source Removal of Exponential Topology



(a) Asymptotic CRLB of a_i

(b) Asymptotic CRLB of b_i

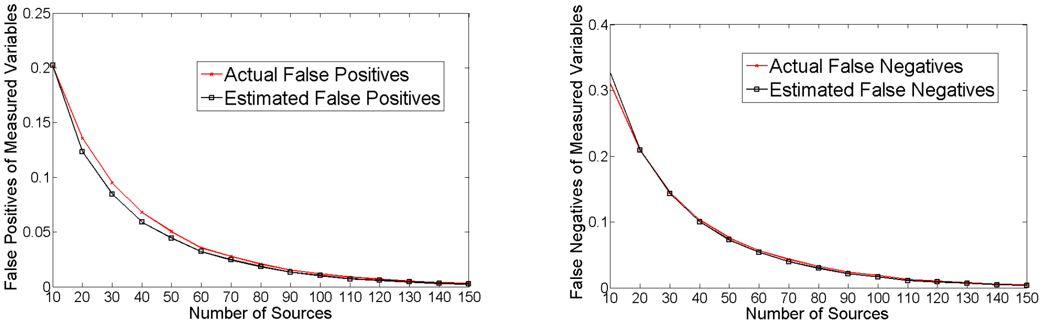
Figure 5.16: Asymptotic CRLB of a_i and b_i versus Source Removal of Exponential Topology

For both the scale-free and exponential topology of social sensing, the above results show that the estimation performance is relatively robust (or insensitive) to changes in the number of sources in the network. Both real and asymptotic CRLBs are able to track the estimation performance as long as a limited number of sources stay in the system.

5.3.3 Evaluation of Estimated False Positives/Negatives on Measured Variables

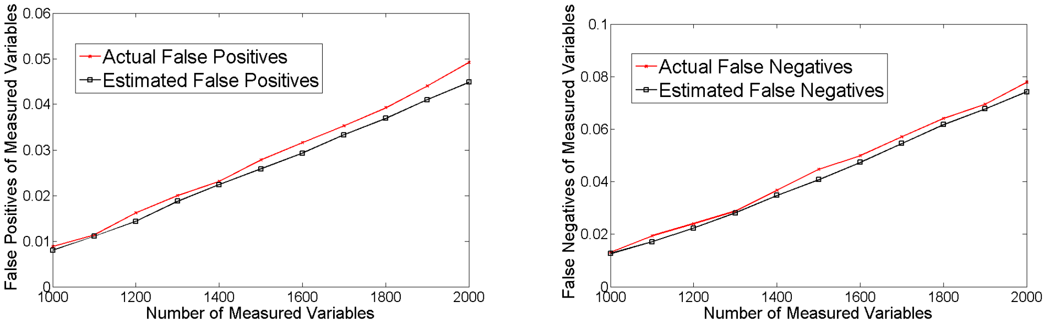
In this subsection, we evaluate the estimated false positives/negatives performance on measured variables derived in Section 5.2.4 by comparing them to the actual false positives/negatives (i.e, the ones that are computed from the ground truth). We carried out similar experiments in the previous subsection and evaluated its performance through scalability, trustworthiness, freshness and robustness studies.

Scalability Study



(a) False Positives (b) False Negatives

Figure 5.17: Estimation of Measured Variable Classification Accuracy versus Varying M



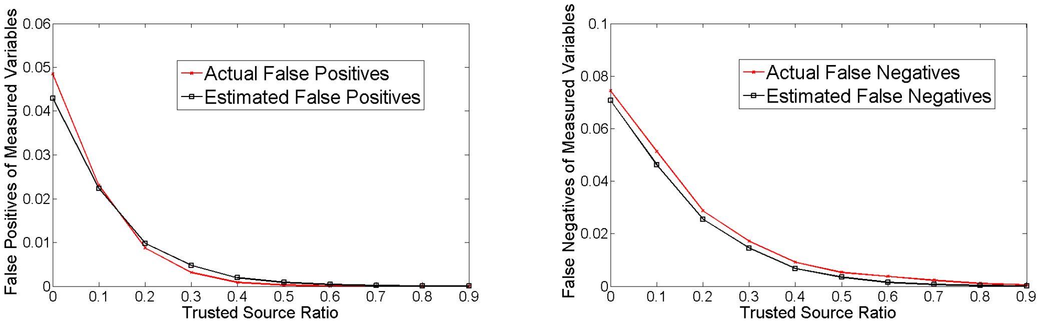
(a) False Positives (b) False Negatives

Figure 5.18: Estimation of Measured Variable Classification Accuracy versus Varying N

We first evaluate the scalability of the estimated false positives/negatives with respect to the sensing topology. The first experiment evaluated the performance when the number of sources (i.e, M) in the system changes. We fix the number of true and false measured variables to be 1000 respectively, the average number of observations per source is set to 200. We vary the number of sources from 10 to 150. Reported results are averaged over 100 experiments and are shown in Figure 5.17. Observe that both estimated false positives and false negatives track the actual values accurately as the number of sources changes. We also note that the false positives/negatives decrease as the number of sources increases. The second experiment compared the estimated false positives/negatives to the actual values when the number of measured variables (i.e, N) changes. We fix the number of sources as 50, the average number of observations per source is set to 200. We also keep the number of true and false measured variables the same. We vary the number of measured variables from 1000 to 2000. Reported results are averaged over 100 experiments and shown in Figure 5.18. Observe that the estimated false positives/negatives are able to track the actual values correctly when the number of measured variables changes. We also note that the estimation performance degrades as the number of measured variables increases. The reason is: the sensing topology becomes sparse as the number of measured variables increases while the number of sources and observations per source stay the same.

Trustworthiness and Freshness Study

In the trustworthiness study, we evaluate the estimated false positives/negatives when the ratio of trusted sources changes in the system. In the experiment, we fix the number of sources to be 50. The number of true and false measured variables are set to be 1000 respectively and the observations per source is set to be 200. We vary the trusted source ratio from 0 to 0.9. The reported results are averaged over 100 experiments and shown in Figure 5.19. Observe that the estimated false positives/negatives track the actual values correctly and both of them decrease as the trusted source ratio increases. The reason is: trusted sources always provide correct observations (i.e, their reliability is 1), which helps the algorithm to estimate the truthfulness of measured variables more accurately.

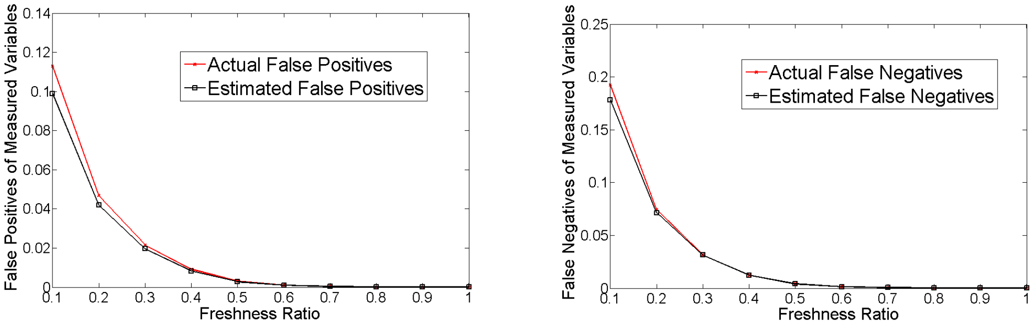


(a) False Positives

(b) False Negatives

Figure 5.19: Estimation of Measured Variable Classification Accuracy versus Trusted Sources Ratio

In the freshness study, we evaluate the estimated false positives/negatives when the freshness ratio changes in the system. In the experiment, we fix the number of sources to be 50. The number of true and false measured variables are set to be 1000 respectively. The data size that is used for the freshness ratio normalization is set to 1000 observations per source. We vary the freshness ratio from 0.1 to 1. The reported results are averaged over 100 experiments and shown in Figure 5.20. Observe that the estimated false positives/negatives track the actual values correctly and both of them decrease as the freshness source ratio increases. The reason is: the sensing topology becomes more densely connected and offers a better chance for the algorithm to correctly judge the truthfulness of the measured variables as the freshness ratio increases.

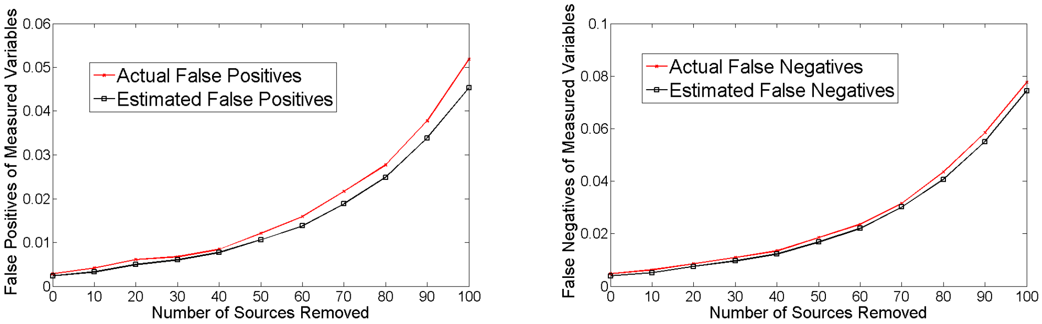


(a) False Positives

(b) False Negatives

Figure 5.20: Estimation of Measured Variable Classification Accuracy versus Freshness

Robustness Study



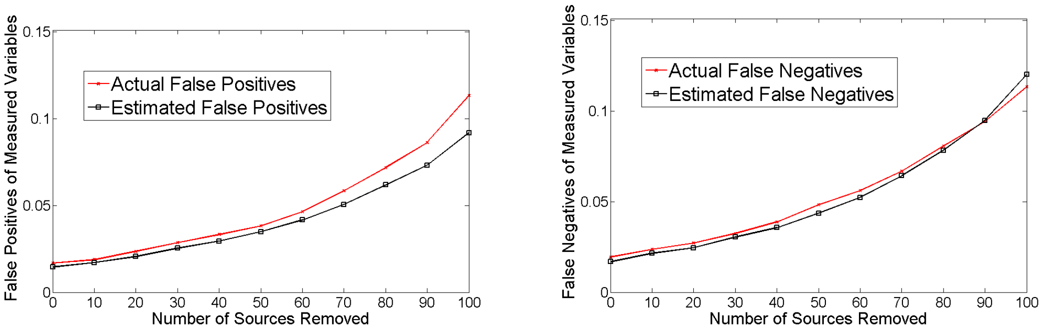
(a) False Positives

(b) False Negatives

Figure 5.21: Estimation of Measured Variable Classification Accuracy versus Source Removal of Scale-free Topology

In the robustness study, similarly as before, we evaluate the estimated false positives/negatives when the number of sources changes under different source reliability distribution. [†] In the first experiment, we

[†] The source reliability distribution parameters for scale-free and exponential topology generation are set the same as the previous subsection



(a) False Positives

(b) False Negatives

Figure 5.22: Estimation of Measured Variable Classification Accuracy versus Source Removal of Exponential Topology

evaluate the estimation performance of the scale-free network topology. We fix both the number of true and false measured variables to be 1000. The average number of observations per source is set to 200. We start with 150 sources and gradually remove sources from the system. Reported results are averaged over 100 experiments and shown in Figure 5.21. Observe that the estimation performance degrades and the estimated false positives/negatives track the actual values reasonably well as the number of removed sources increases. Also note that the estimated values deviate slightly from the actual values when majority of the sources are removed from the system. In the second experiment, we evaluate the estimation performance of the exponential network topology. We change the source reliability distribution to be normal distribution and keep other settings the same as the first experiment. Reported results are averaged over 100 experiments and shown in Figure 5.22. We observe the estimated false positives/negatives track the actual values well when a reasonable number of sources stay in the system. However, we also note that the estimation performance degrades compared to the results of scale-free topology. The reason is: sources are more likely to have similar reliability in the exponential topology. Such similarity makes it a more challenging scenario for our algorithm to accurately pinpoint the source reliability and identity the correctness of measured variables. For both scale-free and exponential topology, the above results show that the estimated false positives/negatives are able to track the actual values as long as limited number of sources stay in the system. The estimation performance on measured variables is relatively robust to changes in the number of sources in the network.

5.4 Discussion

This chapter studies the the confidence intervals in source reliability and estimated classification accuracy of measured variables in social sensing. Several simplifying assumptions were made that offer opportunities for future work.

Sources were assumed to be independent. In reality, sources could be influenced by each other (i.e., copy observations, forward rumor, and etc.) or even collude to misrepresent the truth. Recent work has proposed techniques to detect the dependency and copying relationship between sources [80]. Other methods are proposed to mitigate the source collusion attack by analyzing the network or interaction pattern of colluding sources [64]. The above techniques can be used together with our quantification scheme to handle source dependency. Moreover, authors are also working on extending the current model to handle non-independent sources. For example, one could cluster dependent sources into approximately independent ones according to some source similarity metric and run our scheme on top of the clustered sources. Additionally, sources are sometimes experts in specific domains. It would be interesting to assess the estimation performance on source reliability by taking source expertise into consideration. One possibility is to weight observations differently depending on the source's expertise in the confidence calculation.

No dependencies were assumed among different measured variables. There may be cases, however, observations on one measured variable could imply observations on another (e.g., "flooding" at city B may imply "raining" at city A). The background knowledge of the observation dependency can thus be integrated with our scheme to pre-process the observation matrix (e.g., add or remove links) based on the reported observations and their relationship. Moreover, all observations are treated equally in our model. It is interesting to extend the model to handle the hardness of different observations. In other words, the source reliability and confidence estimation will be computed not only based on whether those observations from the source are true or not but also based on whether such observations are trivial to make. This extension prevents sources from obtaining high reliability and confidence in estimation by simply making many trivially true observations. There are techniques that analyze the hardness of observations, which is possible to be integrated with our scheme [9]. In this chapter, sources are assumed to report positive states of measured variables (e.g., litter found) only and ignore the negative states. This is a reasonable assumption for some typical social sensing applications (e.g., geotagging). However, sources can also make contradicting observations in other types of applications (e.g., on-line review system). Our model can be extended to handle contradicting observations by expanding the estimation parameter vector that covers only positive states to both positive and negative states and rebuilding the likelihood function. The general outline of the proof still holds true in this scenario.

This chapter presents new confidence bounds on source reliability estimation error as well as estimated classification accuracy of measured variables in social sensing applications. It allows the applications to not only assess the reliability of sources and measured variables, given neither in advance, but also estimate the accuracy of such assessment. The confidence bounds are computed based on the Cramer-Rao lower bound (CRLB) of the maximum likelihood estimation of source reliability. The accuracy of measured variable classification is estimated by computing the probability that each measured variable is correct. The derived accuracy results are shown to predict actual errors very well.

Chapter 6

Extensions of the Model and MLE Approach

As we discussed in previous chapters, there are some limitations of the model we established. Two of them are: 1) the observations from different sources on the same measured variable are assumed to be *corroborating*, they don't contradict with each other. 2) the measured variables are assumed to be *binary* only. However, the above assumptions may not always hold in various kinds of social sensing applications. In this chapter, we proposed the extended EM model and derived the maximum likelihood estimation (MLE) approach to remove the above two limitations from our model. It turns out the extended model and MLE scheme remained to be optimal and outperformed the start-of-art heuristics in the presence of *conflicting* observations and *non-binary* measured variables. In the remaining of this chapter, we first demonstrated how to extend the model to handle conflicting observations and then we generalized the model to incorporate the measured variables with non-binary values.

6.1 Extended Model and MLE approach for Conflicting Binary Observations

In this section, we extended our scheme to handle conflicting binary observations (e.g., positive or negative assertion) from different sources on the same measured variable. An important assumption made in the original EM model is that observations from different participants on a given measured variable are *corroborating* (i.e., no conflicting observations exist). However, this is not always true in reality. For example, comments from different reviewers in an on-line review system often contradict with each other, making it difficult for readers to make a decision. This section addresses the challenge of having *conflicting observations* in QoI quantification of social sensing. An extended EM scheme is developed to provide maximum likelihood estimation on participant reliability and measured variable correctness while taking care of conflicting observations from different participants on the same measured variable.

6.1.1 Extended Model

In the extended model to handle conflicting binary observations, we assume that observations are either *positive* or *negative* assertion of the corresponding measured variable. As we mentioned before, the measured variable is assumed to be binary. Let the probability that participant S_i makes a positive observation be s_i^T, probability that participant S_i makes a negative observation be s_i^F. Furthermore, t_i still denotes the odds that participant S_i is right, but it is redefined as the probability that the participant's observation *matches* the ground truth of the measured variable and $1 - t_i$ denotes the probability that it is wrong. Note that, this probability depends on the participant's reliability, which is not known *a priori*. Formally, t_i is redefined in the context of conflicting observations as:

$$t_i = P(C_j^t | S_i C_j^t) = P(C_j^f | S_i C_j^f) \tag{6.1}$$

where $S_i C_j^t$ denotes participant S_i reports the measured variable C_j to be true (i.e., S_i makes positive observation on C_j) and $S_i C_j^f$ denotes participant S_i reports the measured variable C_j to be false (i.e., S_i makes negative observation on C_j). C_j^t and C_j^f denote the measured variable C_j is indeed true or false as we mentioned before.

Let us also define a_i^T and a_i^F as the (unknown) probability that participant S_i reports a variable to be true or false when it is indeed true respectively. Formally, a_i^T and a_i^F are defined as follows:

$$\begin{aligned} a_i^T &= P(S_i C_j^t | C_j^t) \\ a_i^F &= P(S_i C_j^f | C_j^t) \end{aligned} \tag{6.2}$$

b_i^T and b_i^F are defined as the (unknown) probability that participant S_i reports a variable to be true or false when it is in reality false. Formally, b_i^T and b_i^F are defined as follows:

$$\begin{aligned} b_i^T &= P(S_i C_j^t | C_j^f) \\ b_i^F &= P(S_i C_j^f | C_j^f) \end{aligned} \tag{6.3}$$

Let us redefine the observation matrix SC to handle conflicting observations as well: $S_i C_j = 1$ when participant S_i reports that C_j is true, $S_iC_j = -1$ when participant S_i reports that C_j is false and $S_iC_j = 0$ when participant S_i does not observe C_j . Let us call this observation matrix the *conflicting observation matrix*. As we mentioned before, d represents the overall prior probability that a randomly chosen measured variable is true. Additionally, we denote $P(C_j^t) = d$ and $P(S_iC_j^t) = s_i^T$, $P(S_iC_j^f) = s_i^F$. Plugging these, together with t_i into the definition of a_i^T, a_i^F, b_i^T and b_i^F, we get the relations between the terms defined above by using the Bayesian theorem:

$$\begin{aligned} a_i^T &= \frac{t_i \times s_i^T}{d} & a_i^F &= \frac{(1-t_i) \times s_i^F}{d} \\ b_i^T &= \frac{(1-t_i) \times s_i^T}{1-d} & b_i^F &= \frac{t_i \times s_i^F}{1-d} \end{aligned} \tag{6.4}$$

The goal of QoI quantification for the extended model for conflicting observations is the same as the regular model in Section 4.1. That is we target to find the optimal estimation (in the maximum likelihood sense) of the participant reliability and the correctness of the measured variables. Formally, it is given by Equation (4.4).

6.1.2 Re-derive the E-step and M-step

Fortunately, it turns out that we are able to extend the MLE approach in Chapter 4 to solve the optimization problem with conflicting observations. The estimation parameter now becomes: θ=($a_1^T, a_2^T, ...a_M^T$; $a_1^F, a_2^F, ...a_M^F$; $b_1^T, b_2^T, ...b_M^T$; $b_1^F, b_2^F, ...b_M^F$; d). We make corresponding changes to the likelihood function and re-derive the E-step and M-step of the EM scheme accordingly to incorporate this new estimation parameter. The converged results of the extended approach offers the maximum likelihood estimate of θ for the model that is most consistent with the conflicting observation matrix. From there, we adopt similar procedures as discussed in Section 4.2 to compute the H^* and E^* in Equation (4.4).

The likelihood function $L(\theta; X, Z)$ is given by:

$$\begin{aligned}
&L(\theta; X, Z) = p(X, Z|\theta) \\
&= \prod_{j=1}^{N} \left\{ \prod_{i=1}^{M} \Big[{a_i^T}^{S_iC_j^T} \times {a_i^F}^{S_iC_j^F} \right. \\
&\times (1 - a_i^T - a_i^F)^{(1 - S_iC_j{}^T - S_iC_j{}^F)} \Big] \times d \times z_j \\
&+ \prod_{i=1}^{M} \Big[{b_i^T}^{S_iC_j^T} \times {b_i^F}^{S_iC_j^F} \\
&\left. \times (1 - b_i^T - b_i^F)^{(1 - S_iC_j{}^T - S_iC_j{}^F)} \Big] \times (1 - d) \times (1 - z_j) \right\}
\end{aligned} \tag{6.5}$$

where $S_iC_j^T = 1$ when participant S_i claims the measured variable C_j to be true and $S_iC_j^T = 0$ otherwise. Similarly, $S_iC_j^F = 1$ when participant S_i claims the measured variable C_j to be false and $S_iC_j^F = 0$ otherwise.

Given the above formulation, we can derive the E-Step as follows:

$$\begin{aligned}
&Q\left(\theta|\theta^{(t)}\right) \\
&= \sum_{j=1}^{N} \left\{ Z(t, j) \right. \\
&\times \Big[\sum_{i=1}^{M} \Big(S_iC_j^T \log a_i^T + S_iC_j^F \log a_i^F \\
&+ (1 - S_iC_j^T - S_iC_j^F) \log(1 - a_i^T - a_i^F) + \log d \Big) \Big] \\
&+ (1 - Z(t, j)) \\
&\times \Big[\sum_{i=1}^{M} \Big(S_iC_j^T \log b_i^T + S_iC_j^F \log b_i^F \\
&\left. + (1 - S_iC_j^T - S_iC_j^F) \log(1 - b_i^T - b_i^F) + \log(1 - d) \Big) \Big] \right\}
\end{aligned} \tag{6.6}$$

where $Z(t,j)$ is given by:

$$\begin{aligned} Z(t,j) =& p(z_j = 1 | X_j, \theta^{(t)}) \\ =& \frac{A(t,j) \times d^{(t)}}{A(t,j) \times d + B(t,j) \times (1 - d^{(t)})} \end{aligned} \tag{6.7}$$

where $A(t,j)$ and $B(t,j)$ are defined as:

$$\begin{aligned} A(t,j) &= p(X_j, \theta^{(t)} | z_j = 1) \\ &= \prod_{i=1}^{M} \Big\{ a_i^{T(t) S_i C_j^T} \times a_i^{F(t) S_i C_j^F} \\ &\times (1 - a_i^{T(t)} - a_i^{F(t)})^{(1 - S_i C_j^T - S_i C_j^F)} \Big\} \\ B(t,j) &= p(X_j, \theta^{(t)} | z_j = 0) \\ &= \prod_{i=1}^{M} \Big\{ b_i^{T(t) S_i C_j^T} \times b_i^{F(t) S_i C_j^F} \\ &\times (1 - b_i^{T(t)} - b_i^{F(t)})^{(1 - S_i C_j^T - S_i C_j^F)} \Big\} \end{aligned} \tag{6.8}$$

The Maximization step (M-Step) is given by Equation (4.8). We choose θ^* (i.e., ($a_1^{T*}, ...a_M^{T}{}^*$; $a_1^{F*}, ...a_M^{F}{}^*$; $b_1^{T*}, ...b_M^{T}{}^*$; $b_1^{F*}, ...b_M^{F}{}^*$; d^*)) that maximizes the $Q\left(\theta | \theta^{(t)}\right)$ function in each iteration to be the $\theta^{(t+1)}$ of the next iteration.

To get θ^* that maximizes $Q\left(\theta | \theta^{(t)}\right)$, we set the derivatives $\frac{\partial Q}{\partial a_i^T} = 0$, $\frac{\partial Q}{\partial a_i^F} = 0$, $\frac{\partial Q}{\partial b_i^T} = 0$, $\frac{\partial Q}{\partial b_i^F} = 0$ and $\frac{\partial Q}{\partial d} = 0$. Solving the above equations, we can get expressions of the optimal a_i^{T*}, a_i^{F*}, b_i^{T*}, b_i^{F*} and d^*:

$$\begin{aligned} a_i^{T(t+1)} &= a_i^{T*} = \frac{\sum_{j \in SJ_i^T} Z(t,j)}{\sum_{j=1}^{N} Z(t,j)} \\ a_i^{F(t+1)} &= a_i^{F*} = \frac{\sum_{j \in SJ_i^F} Z(t,j)}{\sum_{j=1}^{N} Z(t,j)} \end{aligned} \tag{6.9}$$

$$b_i^{T(t+1)} = b_i^{T*} = \frac{K_i^T - \sum_{j \in SJ_i^T} Z(t,j)}{N - \sum_{j=1}^{N} Z(t,j)}$$
$$b_i^{F(t+1)} = b_i^{F*} = \frac{K_i^F - \sum_{j \in SJ_i^F} Z(t,j)}{N - \sum_{j=1}^{N} Z(t,j)} \tag{6.10}$$

$$d^{(t+1)} = d^* = \frac{\sum_{j=1}^{N} Z(t,j)}{N} \tag{6.11}$$

where K_i^T and K_i^F are the number of true and false observations made by participant S_i respectively and N is the total number of claims in the conflicting observation matrix. SJ_i^T and SJ_i^F are the sets of claims the participant S_i actually observes as true and false respectively in the conflicting observation matrix (i.e, SC). $Z(t,j)$ is defined in (6.7). For details of deriving the above solution, please refer to Appendix C. This completes the mathematical development. We summarize the EM algorithm to handle conflicting observations in the next subsection.

We call the EM scheme derived above to handle conflicting observations the conflict EM algorithm. The input to the conflict EM algorithm is the conflicting observation matrix (i.e., SC) and the output is the maximum likelihood estimation of participant reliability and corresponding judgment on the correctness of claims in the context of conflicting observations. The E-step and M-step of the conflict EM algorithm reduce to simply calculating (6.7) and (6.9)- (6.11) iteratively until they converge. The convergence analysis has been done for EM scheme and it is beyond the scope of this paper [76]. In practice, we can run the algorithm until the difference of estimation parameter between consecutive iterations becomes insignificant. Since the claim is binary, C_j is true if $Z(t,j) \geq 0.5$ and false otherwise. At the same time, we can also compute the maximum likelihood estimation on participant reliability from the converged values of $\theta^{(t)}$ based on (6.4). We summarize the resulting algorithm as shown in Algorithm 2.

Algorithm 2 Expectation Maximization Algorithm for Conflicting Observations

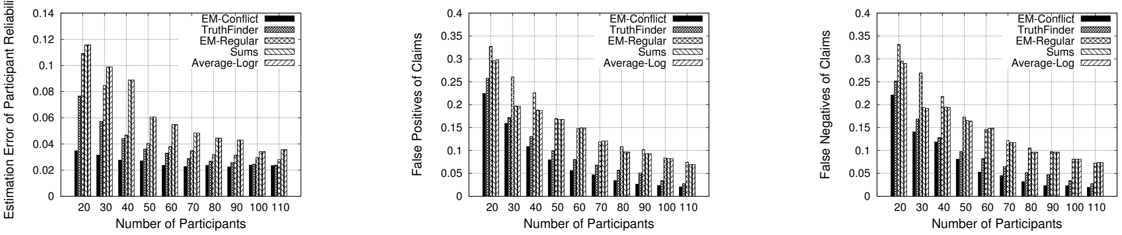
```
Initialize $\theta$ with random values between 0 and 1
while $\theta^{(t)}$ does not converge do
    for $j = 1 : N$ do
        compute $Z(t, j)$ based on Equation (4.11)
    end for
    $\theta^{(t+1)} = \theta^{(t)}$
    for $i = 1 : M$ do
        compute $a_i^{T(t+1)}, a_i^{F(t+1)}, b_i^{T(t+1)}, b_i^{F(t+1)}, d^{(t+1)}$ based on Equation (6.9), (6.10) and (6.11)
        update $a_i^{T(t)}, a_i^{F(t)}, b_i^{T(t)}, b_i^{F(t)}, d^{(t+1)}$ with $a_i^{T(t+1)}, a_i^{F(t+1)}, b_i^{T(t+1)}, b_i^{F(t+1)}, d^{(t)}$in $\theta^{(t+1)}$
    end for
    $t = t + 1$
end while
Let $Z_j^c$ = converged value of $Z(t, j)$
Let $a_i^{Tc}$ = converged value of $a_i^{T(t)}$;
    $b_i^{Fc}$ = converged value of $b_i^{F(t)}$;
    $d^c$ = converged value of $d^{(t)}$
for $j = 1 : N$ do
    if $Z_j^c \geq 0.5$ then
        $h_j^*$ is true
    else
        $h_j^*$ is false
    end if
end for
for $i = 1 : M$ do
    calculate $e_i^*$ from $a_i^{Tc}$, $a_i^{Fc}$, $b_i^{Tc}$, $b_i^{Fc}$, $d^c$ based on Equation (6.4).
end for
Return the computed optimal estimates of measured variables $C_j = h_j^*$ and source reliability $e_i^*$.
```

6.2 Evaluation of Extended Model for Conflicting Binary Observations

6.2.1 Simulation

In this subsection, we repeated the five experiments of Chapter 4 for the scenarios where conflicting observations exist. We applied the extended model and EM approach derived in section 6.1 to the sensing topology with conflicting observations and showed that the extended EM continue to outperform all state-of-art baselines. In the context of conflicting observations, each observation has a probability t_i of being matched to the correctness of the measured variable (i.e., reporting a variable to be the same as its ground truth) and a probability $1 - t_i$ of being mismatched. Remember that participants can report contradicting observations for the same measured variable in this scenario. Note that Bayesian Interpretation was designed to handle corroborating observations only, so we did not use it as a baseline in this subsection. Instead, we added the regular EM scheme studied in Chapter 4 as an additional baseline in these experiments. The regular EM scheme can be adapted in a similar way as fact-finders to handle the conflicting observations of the same measured variable [7]. Specifically, it takes conflicting observations of the same measured variable as two independent measured variables and pick the one with highest probability to believe after the algorithm terminates. The remaining simulation configurations are kept the same as section 4.3. Reported results are averaged over 100 experiments.

We start by repeating the first experiment to show the performance comparison between the extended EM scheme (i.e, EM-Conflict) and other baselines by varying the number of participants in the network. The number of reported measured variables was fixed at 2000, of which 1000 variables were reported correctly and 1000 were misreported. The average number of observations per participant was set to 200. The number of participants was varied from 20 to 110. Results are shown in Figure 6.1. Observe that the extended EM has both smaller estimation error on participant reliability and less false positives/negatives on measured variables among all schemes under comparison. Note also that the performance gain of the extended EM is large when the number of participants is small.



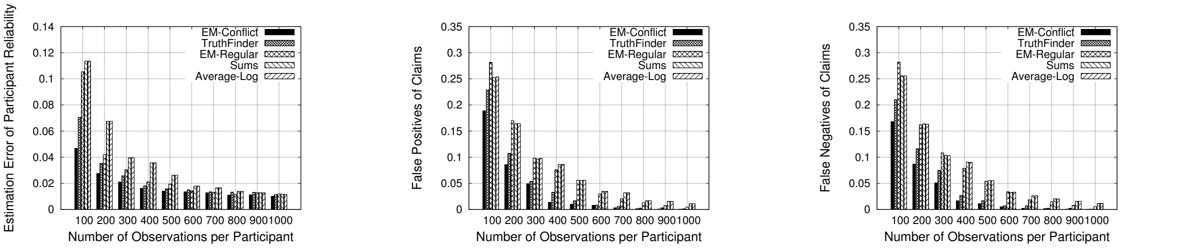
(a) Participant Reliability Estimation Accuracy

(b) Measured Variable Estimation: False Positives

(c) Measured Variable Estimation: False Negatives

Figure 6.1: Estimation Accuracy versus Number of Participants for Conflicting Observations

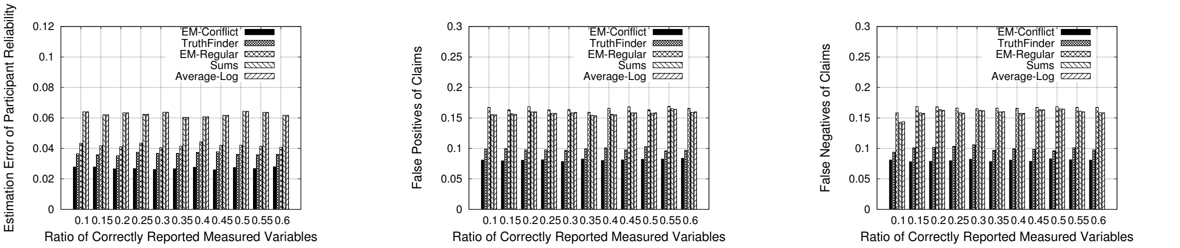
We then repeated the second experiment to compare the extended EM scheme with other baselines while varying the average number of observations per participant. The number of participants was fixed at 50. We vary the average number of observations per participant from 100 to 1000. The results are shown in Figure 6.2. Observe that the extended EM outperforms all baselines in terms of both participant reliability estimation accuracy and false positives/negatives of measured variables as the average number of observations per participant changes. As before, the performance gain of the extended EM is higher when the average number of observations per participant is low.



(a) Participant Reliability Estimation Accuracy (b) Measured Variable Estimation: False Positives (c) Measured Variable Estimation: False Negatives

Figure 6.2: Estimation Accuracy versus Average Number of Observations per Participant for Conflicting Observations

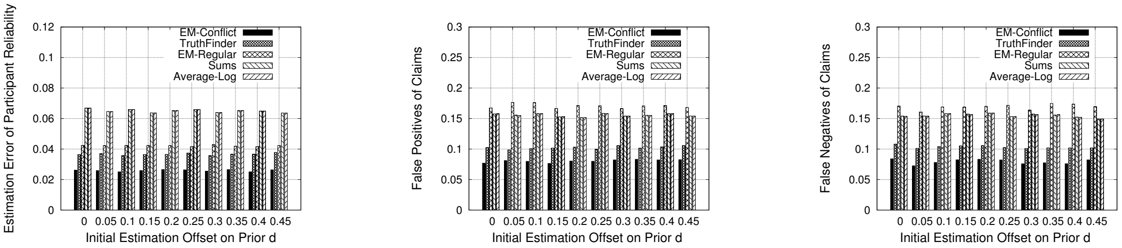
We repeated the third experiment to evaluate the performance of the extended EM scheme compared to baselines when the ratio of correctly reported measured variable changes. The number of participants was fixed to 50 and the average number of observations reported by a participant was set to 200. The total number of measured variable was kept as 2000 and the ratio of correctly reported ones varied from 0.1 to 0.6. Reported results are shown in Figure 6.3. We observe that the extended EM has less error in both participant reliability estimation and false positives/negatives on measured variables under different mix of correct and false measured variables. Moreover, the estimation performance of the extended EM scheme is also more stable compared to other baselines when the correct reported measured variable ratio changes.



(a) Participant Reliability Estimation Accuracy (b) Measured Variable Estimation: False Positives (c) Measured Variable Estimation: False Negatives

Figure 6.3: Estimation Accuracy versus Ratio of Correctly Reported Measured Variables for Conflicting Observations

We repeated the fourth experiment to evaluate the performance of the extended EM and other schemes when the offset of the initial estimation on the background bias d varies. We vary the absolute value of the initial estimate offset on d from 0 to 0.45. The number of participants is fixed at 50 and the average number of observations per participant is set to 200. Figure 6.4 shows the results. We observe that the performance of EM scheme is better than other baselines in terms of both participant reliability estimation and false positives/negatives on measured variables when the initial estimate offset on d changes. We also observe the performance of baselines are relatively stable when offset on d increases. The reason is the baselines mainly depend on the mutual exclusive property of the reports (rather than correct estimation on prior d) to figure out the correctness of measured variables in the context of conflicting observations.



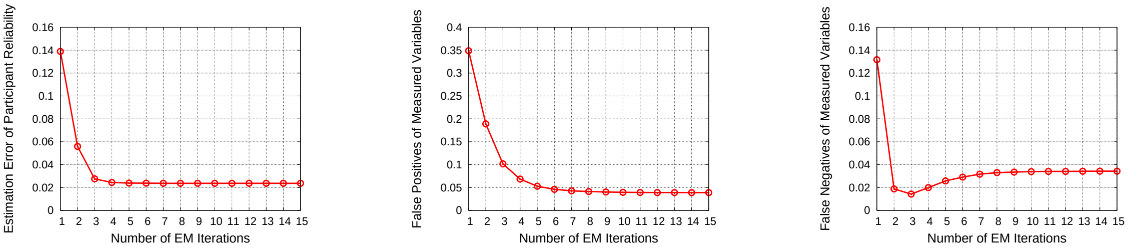
(a) Participant Reliability Estimation Accuracy

(b) Measured Variable Estimation: False Positives

(c) Measured Variable Estimation: False Negatives

Figure 6.4: Estimation Accuracy versus Initial Offset on Prior d for Conflicting Observations

Finally, we repeated the fifth experiment to show the convergence property of the extended EM algorithm. We fix the number of correctly and incorrectly reported variables to 1000 respectively and set the initial estimate offset on d to 0.3. The number of participants is fixed at 50 and the average number of observations per participant is set to 200. Reported results are averaged over 100 experiments. Figure 6.5 shows the results. We observe similar fast convergence of the extended EM as the regular EM scheme in the previous chapter.



(a) Participant Reliability Estimation Accuracy

(b) Measured Variable Estimation: False Positives

(c) Measured Variable Estimation: False Negatives

Figure 6.5: Convergence Property of the EM Algorithm for Conflicting Observations

6.2.2 A Real World Application

In this subsection, we evaluate the performance of the proposed extended EM scheme for conflicting observations through a real world application, finding free parking lots on UIUC campus. “Free parking lots”’ refer to the parking lots that is free of charge after 5pm on weekday as well as weekends. The goal was to see if our scheme can find the free parking lots most accurately compared to other state-of-art baselines. Specifically, we selected 106 parking lots of our interests around the campus and asked volunteers to mark them as either "Free"' or "Not Free"'. Participants mark those parking lots they have been to or are familiar with. We note that there are actually various types of parking lots on campus: enforced parking lots with time limits, parking meters, permit parking, street parking, and etc. Different parking lots have different regulations for free parking. Moreover, instructions and permit signs sometimes read similar and are easy to miss. Hence, people are prone to generate both false positives and false negatives in their reports. For evaluation purpose, we went to those selected parking lots and manually collected the ground truth.

In the experiment, 30 participants were invited to offer their marks on the 106 parking lots (46 of which are indeed free). There were 901 marks collected from participants in total. We then generated the observation matrix by taking the participants as sources and different parking lots as measured variables. The free parking lots map to the true measured variables while the non-free ones map to the false measured variables. The corresponding element S_iC_j is set according to the marks each participant placed on those parking lots. We applied the extended EM scheme discussed in Section 6.1 to handle conflicting observations and other state-of-art baselines (including the regular EM scheme adapted for conflicting claims) as well as the simple voting scheme to the data we collected. We then compared the false positives and false negatives of different schemes in identifying the free parking lots among all places selected. The result is shown in Table 6.1. We observe that the extended EM scheme to handle conflicting observations (i.e., EM-Conflict) achieved the least false positives and false negatives among all schemes under comparison. The reason is the extended EM scheme modeled the conflicting observations explicitly and used the MLE approach to find the value of each measured variable that is most consistent with the observations we had.

Schemes	False Positives	False Negatives
EM-Conflict	**6.67%**	**10.87%**
EM-Regular	11.67%	17.39%
Average-Log	16.67%	19.57%
Truth-Finder	18.33%	15.22%
Voting	21.67%	23.91%

Table 6.1: Accuracy of Finding Free Parking Lots on Campus

6.3 Generalized Model and MLE approach for Non-Binary Measured Variables

Recall the value of the measured variable is assumed to be *binary* in our original model of EM formulation. Though this covers a wide range of social sensing applications where the status of the measured variable is either “true” or “false”. For example, a building is either on fire or not. Jeffery is either the CEO of company X or not. However, there are also some other application scenarios that the values of measured variables may not necessarily be binary. For example, the types of parking lots in a district can be “parking meters”, “permit only” or “private”. In this effort, we consider how to generalize our model to incorporate the *non-binary* values of measured variables and extend the MLE approach we derived earlier .

The generalized model for non-binary measured variables is the same as the original model except there can be more than two values reported by different participants about the same measured variable. Hence, the measured variable in this section is assumed to have K ($K \geq 2$) mutually exclusive possible values and only one of them represents the true value of the measured variable. In the model to handle conflicting observations, we assume that observations from participants assert one of the K values of the corresponding measured variable, thus can be potentially conflicting. Let S_i represent the i^{th} participant and C_j represent the j^{th} measured variable. Each participant generally observes only a subset of all measured variables (e.g., the conditions at locations they visited). Let $S_iC_j = k$ denote participant S_i reports the measured variable C_j to be of value k for $k = 0, ..., K$. Note that $S_iC_j = 0$ means that participant S_i does not report an observation for measured variable C_j. Let probability that participant S_i reports the measured variable to be of value k be s_i^k (i.e, $s_i^k = P(S_iC_j = k)$ for $k = 0, ..., K$). Let $s_i^{\bar{k}}$ represent the probability that S_i reports a measured variable to be of value other than k (i.e., $s_i^{\bar{k}} = \sum_{k' \neq 0,k} s_i^{k'}$).

Further, t_i denotes the probability that participant S_i is right (i.e., probability that the participant’s observation *matches* the ground truth of the measured variable) and $1 - t_i$ denotes the probability that it is wrong. Note that, this probability depends on the participant’s reliability, which is not known *a priori*. Our goal is to determine which observations are correct and which are not as well as the reliability of each participant. As mentioned in the introduction, we differ from a large volume of previous sensing literature in that we assume no prior knowledge of source reliability, as well as no prior knowledge of the correctness of individual observations.

Let us also define $a^T_{k,i}$ and $a^F_{k,i}$ as the (unknown) probability that participant S_i reports a measured variable to be of value k and value other than k when the measured variable is indeed of value k respectively. Formally, $a^T_{k,i}$ and $a^F_{k,i}$ are defined as follows:

$$\begin{aligned} a^T_{k,i} &= P(S_iC_j = k|C_j = k) \\ a^F_{k,i} &= \sum_{k' \neq 0,k}^{K} P(S_iC_j = k'|C_j = k) \end{aligned} \tag{6.12}$$

where $C_j = k$ denotes the measured variable C_j is indeed of value k for $k = 1, ..., K$. We assume that participant S_i can report one (and only one) of the K mutually exclusive values for measured variable C_j (i.e., a source is not self-contradictory on its assertion for a measured variable). Since a source may not assert a measured variable ($k = 0$), $a^T_{k,i} + a^F_{k,i} \leq 1$.

Let us define the observation matrix SC to handle conflicting observations: $S_iC_j = k$ when participant S_i reports that C_j is of value k, $S_iC_j = 0$ when no participants reports C_j. Let us call this observation matrix the *conflicting observation matrix*. Let d_k represent the overall prior probability that an arbitrary measured variable is of value k.

Plugging these, together with t_i into the definition of $a^T_{k,i}$ and $a^F_{k,i}$, we get the relations between the terms defined above by using the Bayesian theorem:

$$\begin{aligned} a^T_{k,i} &= \frac{t_i \times s^k_i}{d_k} \\ a^F_{k,i} &= \frac{(1 - t_i) \times s^{\bar{k}}_i}{d_k} \end{aligned} \tag{6.13}$$

6.3.1 Generalized E and M steps for Non-Binary Measured Variables

In this subsection, we solve the problem formulated in the previous subsection for non-binary measured variables using a generalized version of the Expectation-Maximization algorithm. Similarly as before, we have a latent variable Z for each measured variable to indicate the value of the measured variable. However the definition of Z is generalized for non-binary measured variables: we have a corresponding variable z_j for the j^{th} measured variable C_j such that: $z_j = k$ when C_j is of value k. We further denote the observation matrix SC as the observed data X, and take $\theta = (\theta_1, \theta_2, ..., \theta_K)$ where $\theta_k = (a^T_{k,1}, a^F_{k,1}, a^T_{k,2}, a^F_{k,2}...a^T_{k,M}, a^F_{k,M}, d_k)$ as the parameters of the model that we want to estimate. The goal is to get the maximum likelihood estimate of θ for the model containing observed data X and latent variables Z.

Given the estimation parameter and hidden variables defined above, the likelihood function $L(\theta; X, Z)$ for conflicting observations is:

$$\begin{aligned}
L(\theta; X, Z) &= p(X, Z|\theta) \\
&= \prod_{j=1}^{N} \Bigg\{ \sum_{k=1}^{K} \Big[\prod_{i=1}^{M} {a_{k,i}^{T}}^{S_iC_j^k} \times {a_{k,i}^{F}}^{S_iC_j^{\bar{k}}} \\
&\times (1 - a_{k,i}^{T} - a_{k,i}^{F})^{(1-{S_iC_j}^{k}-{S_iC_j}^{\bar{k}})} \times d_k \times z_j^k \Big] \Bigg\}
\end{aligned} \tag{6.14}$$

where $S_iC_j^k = 1$ when participant S_i asserts the measured variable C_j to be of value k (i.e., $S_iC_j = k$) and 0 otherwise, $S_iC_j^{\bar{k}} = 1$ when participant S_i asserts the measured variable C_j to be of value other than k (i.e., $S_iC_j \neq k$ or 0) and 0 otherwise, and $z_j^1, z_j^2, ..., z_j^K$ is a set of indicator variables for measured variable C_j where $z_j^k = 1$ when C_j is of value k and $z_j^k = 0$ otherwise. Additionally, the values of S_iC_j are statistically independent over the M participants and N measured variables. The likelihood function above describes the likelihood to have current observation matrix X and hidden variable Z given the estimation parameter θ we defined.

Given the above formulation, we can derive the E-Step as

$$\begin{aligned}
&Q\left(\theta|\theta^{(t)}\right) = \\
&\sum_{j=1}^{N} \Bigg\{ \sum_{k=1}^{K} Z_k(t, j) \times \Big[\sum_{i=1}^{M} \Big(S_iC_j^k \log a_{k,i}^{T} + S_iC_j^{\bar{k}} \log a_{k,i}^{F} \\
&+ (1 - S_iC_j^k - S_iC_j^{\bar{k}}) \log(1 - a_{k,i}^{T} - a_{k,i}^{F}) + \log d_k \Big) \Big] \Bigg\}
\end{aligned} \tag{6.15}$$

where $Z_k(t, j)$ is given by:

$$\begin{aligned}
Z_k(t, j) =& p(z_j = k|X_j, \theta^{(t)}) \\
&= \frac{A_k(t, j) \times d_k^{(t)}}{\sum_{k=1}^{K} A_k(t, j) \times d_k^{(t)}}
\end{aligned} \tag{6.16}$$

where $A_k(t, j)$ is defined as:

$$\begin{aligned}
A_k(t, j) &= p(X_j, \theta^{(t)} | z_j = k) \\
&= \prod_{i=1}^{M} \Big\{ {a_{k,i}^{T}}^{(t) S_i C_j^k} \times {a_{k,i}^{F}}^{(t) S_i C_j^{\bar{k}}} \\
&\times (1 - {a_{k,i}^{T}}^{(t)} - {a_{k,i}^{F}}^{(t)})^{(1 - S_i C_j^k - S_i C_j^{\bar{k}})} \Big\}
\end{aligned} \tag{6.17}$$

where $Z_k(t, j)$ is the conditional probability of the measured variable C_j to have value k given the observation matrix related to the j^{th} measured variable and current estimate of θ. X_j represents the j^{th} column of the observed SC matrix (i.e., observations of the j^{th} measured variable from all participants). $A_k(t, j)$ represents the conditional probability regarding observations about the j^{th} measured variable and current estimation of the parameter θ given the j^{th} measured variable is of value k.

The Maximization step (M-Step) is given by (4.8). We choose θ^* (i.e., $({a_{k,1}^{T}}^*, ...{a_{k,M}^{T}}^*; {a_{k,1}^{F}}^*, ...{a_{k,M}^{F}}^*;$ $d^*) \quad k = 1, 2, ...K$) that maximizes the $Q\left(\theta | \theta^{(t)}\right)$ function in each iteration to be the $\theta^{(t+1)}$ of the next iteration.

To get θ^* that maximizes $Q\left(\theta | \theta^{(t)}\right)$, we set the derivatives $\frac{\partial Q}{\partial a_{k,i}^T} = 0$, $\frac{\partial Q}{\partial a_{k,i}^F} = 0$ and $\frac{\partial Q}{\partial d_k} = 0$.

Solving the above equations, we can get expressions of the optimal ${a_{k,i}^{T}}^*$, ${a_{k,i}^{F}}^*$ and d_k^*:

$$\begin{aligned}
{a_{k,i}^{T}}^{(t+1)} &= {a_{k,i}^{T}}^* = \frac{\sum_{j \in SJ_i^k} Z_k(t, j)}{\sum_{j=1}^{N} Z_k(t, j)} \\
{a_{k,i}^{F}}^{(t+1)} &= {a_{k,i}^{F}}^* = \frac{\sum_{j \in SJ_i^{\bar{k}}} Z_k(t, j)}{\sum_{j=1}^{N} Z_k(t, j)} \\
{d_k}^{(t+1)} &= {d_k}^* = \frac{\sum_{j=1}^{N} Z_k(t, j)}{N}
\end{aligned} \tag{6.18}$$

where N is the total number of measured variables in the conflicting observation matrix. SJ_i^k are the sets of measured variables the participant S_i actually observes to have value k and $SJ_i^{\bar{k}}$ are the ones S_i observes to have value other than k in the conflicting observation matrix (i.e, SC). $Z_k(t, j)$ is defined in (6.16). For details of deriving the above solution, please refer to Appendix C. Note that the case where the value of the measured variable is binary (i.e., $K = 2$) can be considered as a special case of the algorithm derived in this section. The E-step and M-step of the algorithm for binary measured variables can be written as in (6.19) and (6.20) respectively:

$$
\begin{aligned}
& Q\left(\theta|\theta^{(t)}\right) = \\
& \sum_{j=1}^{N} \Bigg\{ Z_1(t,j) \times \Big[\sum_{i=1}^{M} \Big(S_iC_j^1 \log a_{1,i}^T + S_iC_j^2 \log a_{1,i}^F \\
& + (1 - S_iC_j^1 - S_iC_j^2) \log(1 - a_{1,i}^T - a_{1,i}^F) + \log d_1 \Big) \Big] \\
& + (1 - Z_1(t,j)) \times \Big[\sum_{i=1}^{M} \Big(S_iC_j^2 \log a_{2,i}^T + S_iC_j^1 \log a_{2,i}^F \\
& + (1 - S_iC_j^1 - S_iC_j^2) \log(1 - a_{2,i}^T - a_{2,i}^F) + \log(1 - d_1) \Big) \Big] \Bigg\}
\end{aligned} \tag{6.19}
$$

where $S_iC_j^k = 1, k = 1, 2$ when S_i reports C_j to have value k and 0 otherwise. Note that $Z_2(t,j) = 1 - Z_1(t,j)$ and $d_2 = 1 - d_1$ for the binary case.

$$
\begin{aligned}
{a_{1,i}^T}^{(t+1)} &= {a_{1,i}^T}^* = \frac{\sum_{j \in SJ_i^1} Z_1(t,j)}{\sum_{j=1}^N Z_1(t,j)} \\
{a_{1,i}^F}^{(t+1)} &= {a_{1,i}^F}^* = \frac{\sum_{j \in SJ_i^2} Z_1(t,j)}{\sum_{j=1}^N Z_1(t,j)} \\
{a_{2,i}^T}^{(t+1)} &= {a_{2,i}^T}^* = \frac{K_i^1 - \sum_{j \in SJ_i^1} Z_1(t,j)}{N - \sum_{j=1}^N Z_1(t,j)} \\
{a_{2,i}^F}^{(t+1)} &= {a_{2,i}^F}^* = \frac{K_i^2 - \sum_{j \in SJ_i^2} Z_1(t,j)}{N - \sum_{j=1}^N Z_1(t,j)} \\
{d_1}^{(t+1)} &= {d_1}^* = \frac{\sum_{j=1}^N Z_1(t,j)}{N}
\end{aligned} \tag{6.20}
$$

where SJ_i^1 and SJ_i^2 are the sets of measured variables S_i reports to have one of the binary values respectively and K_i^1 and K_i^2 are the number of measured variables in the above two sets. Note that the results for the binary measured variable are essentially the same as we derived in Section 6.1.

This completes the mathematical development. We summarize the EM algorithm to handle conflicting observations in the next subsection.

6.3.2 The Generalized EM Algorithm for Non-Binary Measured Variables

Algorithm 3 Generalized Expectation Maximization Algorithm for Non-Binary Measured Variables

```
Initialize $\theta$ with random values between 0 and 1
while $\theta^{(t)}$ does not converge do
    for $j = 1 : N$ do
        for $k = 1 : K$ do
            compute $Z_k(t, j)$ based on (6.16)
        end for
    end for
    $\theta^{(t+1)} = \theta^{(t)}$
    for $k = 1 : K$ do
        for $i = 1 : M$ do
            compute ${a_{k,i}^T}^{(t+1)}$, ${a_{k,i}^F}^{(t+1)}$ and $d_k^{(t+1)}$ based on (6.18)
            update ${a_{k,i}^T}^{(t)}$, ${a_{k,i}^F}^{(t)}$, $d_k^{(t)}$ with ${a_{k,i}^T}^{(t+1)}$, ${a_{k,i}^F}^{(t+1)}$ and $d_k^{(t+1)}$ in $\theta^{(t+1)}$
        end for
    end for
    $t = t + 1$
end while
Let $Z_{k,j}^c$ = converged value of $Z_k(t, j)$
Let $\theta^c$ = converged value of $\theta^{(t)}$
for $j = 1 : N$ do
    $max = 0; k^* = 0$
    for $k = 1 : K$ do
        if $Z_{k,j}^c \geq max$ then
            $max = Z_{k,j}^c$ and $k^* = k$
        end if
    end for
    measured variable $C_j$ is of value $k^*$
end for
for $i = 1 : M$ do
    calculate $t_i^*$ from $\theta^c$ based on (6.13).
end for
Return the computed maximum likelihood estimation on source reliability $t_i^*$ and corresponding judgment on the correctness of measured variable $C_j$.
```

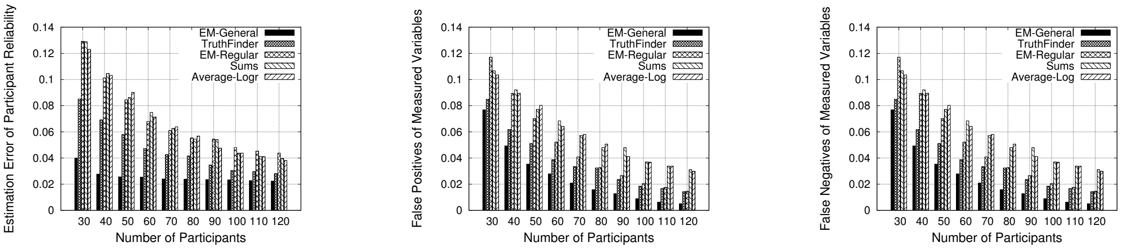
We call the EM scheme derived above to handle conflicting observations the general EM algorithm. The input to the general EM algorithm is the conflicting observation matrix (i.e., SC) and the output is the maximum likelihood estimation of participant reliability and corresponding judgment on the correctness of measured variables in the context of conflicting observations. The E-step and M-step of the conflict EM algorithm reduce to simply calculating (6.16) and (6.18) iteratively until they converge. The convergence analysis has been done for EM scheme and it is beyond the scope of this paper [76]. In practice, we can run the algorithm until the difference of estimation parameter between consecutive iterations becomes insignificant. We can then decide the value of measured variable C_j as the one that has the highest $Z_k(t, j)$ value for $k = 1, 2, ...K$. In the special case where the measured variable is binary, C_j is true if $Z_k(t, j) \geq 0.5$ and false otherwise. At the same time, we can also compute the maximum likelihood estimation on participant reliability from the converged values of $\theta^{(t)}$ based on (6.13). We summarize the resulting algorithm as shown in Algorithm 3.

6.4 Evaluation of Generalized Model for Non-Binary Measured Variables

In this section, we carried out similar experiments as the previous sections for the scenarios where non-binary measured variables exist. We applied the generalized model and EM approach derived in section 6.3 to the sensing topology with non-binary measured variables and showed that the generalized EM scheme outperform all state-of-art baselines. In the context of non-binary measured variables, each observation of participant S_i has a probability t_i of being matched to the real value of the measured variable (i.e., reporting a variable to be the same as its ground truth) and a probability $1 - t_i$ of being mismatched. Note that participants can report every possible value of the measured variable in this scenario. For simplicity, we study the case where the measured variable can have three different values a, b, c. The remaining simulation configurations are kept the same as before. Reported results are averaged over 100 experiments.

We started with the first experiment to show the performance comparison between the generalized EM scheme for non-binary measured variables and other baselines by varying the number of participants in the network. The number of reported measured variables was fixed at 1500. The number of the a, b and c valued measured variables is 500 each. The average number of observations per participant was set to 200. The number of participants was varied from 30 to 120. Results are shown in Figure 6.6. Observe that the generalized EM has both smaller estimation error on participant reliability and less false positives/negatives on measured variables among all schemes under comparison. Note also that the performance gain of the generalized EM is large when the number of participants is small.

We then carried out the second experiment to compare the generalized EM scheme for non-binary measured variables with other baselines while varying the average number of observations per participant. The number of participants was fixed at 50. We vary the average number of observations per participant from 100 to 1000. The results are shown in Figure 6.7. Observe that the generalized EM outperforms all base-



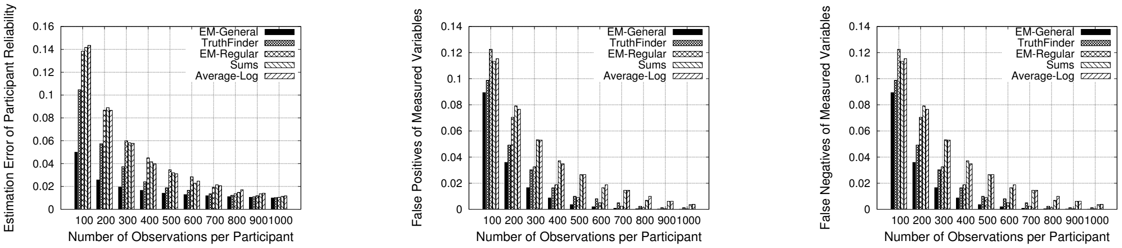
(a) Participant Reliability Estimation Accuracy

(b) Measured Variable Estimation: False Positives

(c) Measured Variable Estimation: False Negatives

Figure 6.6: Estimation Accuracy versus Number of Participants for Non-Binary Variables

lines in terms of both participant reliability estimation accuracy and false positives/negatives of measured variables as the average number of observations per participant changes. As before, the performance gain of the generalized EM is higher when the average number of observations per participant is low.



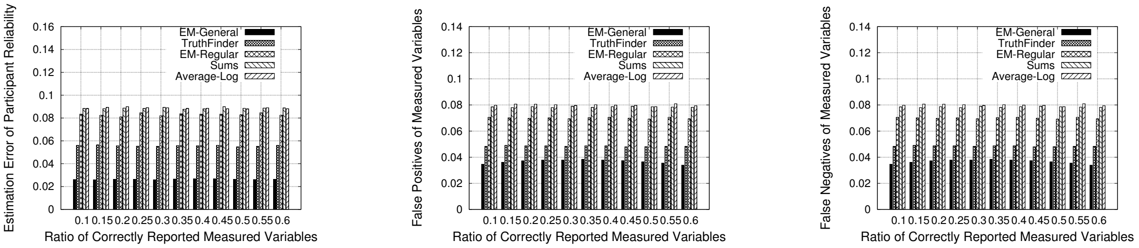
(a) Participant Reliability Estimation Accuracy

(b) Measured Variable Estimation: False Positives

(c) Measured Variable Estimation: False Negatives

Figure 6.7: Estimation Accuracy versus Average Number of Observations per Participant for Non-Binary Variables

Finally, we did the third experiment to evaluate the performance of the generalized EM scheme for non-binary measured variables compared to baselines when the ratio of the measured variable with one value (e.g., a in this study) changes. The number of participants was fixed to 50 and the average number of observations reported by a participant was set to 200. The total number of measured variable was kept as 1500 and the ratio of a valued measured variables varied from 0.1 to 0.6. Reported results are shown in Figure 6.8. We observe that the generalized EM has less error in both participant reliability estimation and false positives/negatives on measured variables under different ratio of value a measured variables. Moreover, the estimation performance of the generalized EM scheme is also more stable compared to other baselines when the ratio of a valued measured variable changes.



(a) Participant Reliability Estimation Accuracy

(b) Measured Variable Estimation: False Positives

(c) Measured Variable Estimation: False Negatives

Figure 6.8: Estimation Accuracy versus Ratio of Value a Measured Variables for Non-binary Variables

6.5 Discussion

In this chapter, we extended our basic MLE model to handle more general cases in social sensing applications where the claims from different participants can be conflicting and the value of measured variables can be non-binary. Several future research directions exist by further generalizing some assumptions we made on the current model.

First, the values of the measured variable are assumed to be discrete and the true value is unique. What happens if there are degrees of truth on the measured variable? For example, the true value of a measured variable may not simply be true or false but could stay within a spectrum between the two. In such case, we may not be able to use the current discrete indication vectors to represent all possible values of the measured variable. Instead, some continuous variables could be used to reconstruct the likelihood function and reflect the actual degrees of truth on the measured variable. In statistics, some filtering algorithms (e.g., Kalman filter, particle filter, etc.) are designed to remove the noise from continuous variables [61, 62]. Their models usually make different assumptions on the underlying distribution of variable state space and the sample size. For example, Kalman filter assumes the measurements have a Gaussian distribution and the underlying system is linear. These assumptions may not necessarily hold for the data collected from social sensing applications. However, it would be interesting to see if it is possible to adapt some of the filtering techniques and further extend our model to better handle continuous measured variables.

Second, the observations participants make on measured variables are assumed to be binary (i.e., either positive or negative). What happens if there are degrees of support on the observations reported by participants on the measured variables? For example, a participant may report a measurement with a certain degree of confidence. In our current model, we assume a source either report an observation (positive or negative) or not, we do not consider the degrees of support on the reported observations. One possible way to incorporate such degrees of support on observations into our model is to associate links with different weights in the bipartite graph of the sensing topology. Some prior work applied similar ideas to incorporate the prior knowledge into their fact-finding framework [44]. However, the challenge for our model is that we need to keep the probability semantics of the estimation parameters defined when we add weights of links into the graph. Another possibility is to model the degrees of support on the observations separately and relate them to the estimation parameters we defined for the MLE approach.

Chapter 7

Recursive EM Algorithm and its Tradeoffs Study

7.1 Recursive EM Algorithm

As we discussed earlier in Chapter 4, EM is an iterative algorithm that provides the maximum likelihood estimation when the iteration converges. However, running the iterative algorithm is not necessarily efficient for the streaming data, especially when the estimation parameter θ remains stable over time. This observation motivates us to develop a recursive version of the EM algorithm for streaming data to achieve a better tradeoff between the estimation accuracy and running time.

In estimation theory, a recursive formula of the EM scheme estimates parameters of the model in consecutive time intervals as follows [81]:

$$\hat{\theta}_{k+1} = \hat{\theta_k} + \{(k+1)I_c(\hat{\theta_k})\}^{-1}\psi(X_{k+1}, \hat{\theta_k}) \tag{7.1}$$

where $\hat{\theta_k}$ is the estimation parameter by observing the data up to the time interval k, $I_c^{-1}(\hat{\theta_k})$ represents the inverse of the Fisher information (i.e., Cramer Rao lower bound (CRLB)) of the estimation parameter at time k and $\psi(X_{k+1}, \hat{\theta_k})$ is the score vector of the observed data at time interval $k+1$ w.r.t the estimation parameter $\hat{\theta_k}$. This formula basically provides us a recursive way to compute the estimation parameter in the new time interval (i.e., $\hat{\theta}_{k+1}$) based on its estimation value in the previous time interval (i.e., $\hat{\theta}_k$), the CRLB of the estimation (i.e., $I_c^{-1}(\hat{\theta_k})$) and the score vector of the updated data observed in the new interval (i.e., $\psi(X_{k+1}, \hat{\theta_k})$). Based on our previous results of the EM scheme, $\hat{\theta_k}$ is the estimation vector defined as $\hat{\theta_k} = (\hat{a}_1^k, \hat{a}_2^k, ...\hat{a}_M^k; \hat{b}_1^k, \hat{b}_2^k, ...\hat{b}_M^k)$. $I_c^{-1}(\hat{\theta_k})$ and $\psi(X_{k+1}, \hat{\theta_k})$ are given by [79]:

$$I_c^{-1}(\hat{\theta_k})_{i,j} \\ = \begin{cases} 0 & i \neq j \\ \frac{\hat{a}_i^k \times (1-\hat{a}_i^k)}{N \times d} & i = j \in [1, M] \\ \frac{\hat{b}_i^k \times (1-\hat{b}_i^k)}{N \times (1-d)} & i = j \in (M, 2M] \end{cases} \tag{7.2}$$

$$\psi(X_{k+1}, \hat{\theta_k})_{i,j} \\ = \begin{cases} 0 & i \neq j \\ \sum_{j=1}^{N} \hat{Z_j}^{k+1} (\frac{S_i C_j}{\hat{a}_i^k} - \frac{1 - S_i C_j}{1 - \hat{a}_i^k}) & i = j \in [1, M] \\ \sum_{j=1}^{N} (1 - \hat{Z_j}^{k+1}) (\frac{S_i C_j}{\hat{b}_i^k} - \frac{1 - S_i C_j}{1 - \hat{b}_i^k}) & i = j \in (M, 2M] \end{cases} \tag{7.3}$$

where $\hat{Z_j}^{k+1}$ is the probability of the j^{th} measured variable to be true in the $k+1$ time interval. Plugging Equation (7.2) and (7.3) into (7.1), the recursive formula to update the estimation parameters is given by:

$$\begin{aligned} \hat{a_i}^{k+1} &= \hat{a_i}^k + \frac{1}{Nd(k+1)} \times \\ &\Big[\sum_{j \in SJ_i^{k+1}} \hat{Z_j}^{k+1} (1 - \hat{a_i}^k) - \sum_{j \in S\bar{J_i}^{k+1}} \hat{Z_j}^{k+1} \hat{a_i}^k \Big] \\ \hat{b_i}^{k+1} &= \hat{b_i}^k + \frac{1}{Nd(k+1)} \times \\ &\Big[\sum_{j \in SJ_i^{k+1}} (1 - \hat{Z_j}^{k+1})(1 - \hat{b_i}^k) - \sum_{j \in S\bar{J_i}^{k+1}} (1 - \hat{Z_j}^{k+1}) \hat{b_i}^k \Big] \end{aligned} \tag{7.4}$$

From above equations, we observe that the estimation of the parameters related with reliability of each source in current time interval can be computed from their estimations in the past and the observed data in the new interval. Moreover, $\hat{Z_j}^{k+1}$ is unknown and can be estimated by its approximation $\tilde{Z_j}^{k+1}$, which can be computed as follows:

$$
\begin{aligned}
\tilde{Z_j}^{k+1} &= f(\tilde{a_i}^{k+1}, \tilde{b_i}^{k+1}, X_{k+1}) \\
&= \frac{A_j^{k+1} \times d}{A_j^{k+1} \times d + B_j^{k+1} \times (1-d)}
\end{aligned}
$$

where

$$
\begin{aligned}
A_j^{k+1} &= \prod_{i=1}^{M} (\tilde{a_i}^{(k+1)})^{S_i {C_j}^{k+1}} (1-\tilde{a_i}^{(k+1)})^{(1-S_i C_j^{k+1})} \\
B_j^{k+1} &= \prod_{i=1}^{M} (\tilde{b_i}^{(k+1)})^{S_i {C_j}^{k+1}} (1-\tilde{b_i}^{(k+1)})^{(1-S_i C_j^{k+1})} \\
\tilde{a_i}^{k+1} &= \hat{a_i}^{k} \times \frac{s_i^{k+1}}{s_i^{k}} \quad \tilde{b_i}^{k+1} = \hat{b_i}^{k} \times \frac{s_i^{k+1}}{s_i^{k}}
\end{aligned}
\tag{7.5}
$$

where s_i^{k+1} and s_i^k are the probabilities of source S_i to report a measured variable at time interval $k+1$ and k. For the above equation to hold, we assume source reliability changes slowly over time and can be treated unchanged over two consecutive time intervals.

Based on the above approximate estimation definition, we can represent $\tilde{Z_j}^{k+1}$ as a function of $\hat{a_i}^k$,$\hat{b_i}^k$,X_k, and X_{k+1}, the values of which we know at time interval $k+1$:

$$
\begin{aligned}
\tilde{Z_j}^{k+1} &= g(\hat{a_i}^{k}, \hat{b_i}^{k}, X_k, X_{k+1}) \\
&= \frac{C_j^{k+1} \times d}{C_j^{k+1} \times d + D_j^{k+1} \times (1-d)}
\end{aligned}
$$

where

$$
\begin{aligned}
C_j^{k+1} &= \prod_{i=1}^{M} (\hat{a_i}^{k} \times \frac{s_i^{k+1}}{s_i^{k}})^{S_i {C_j}^{k+1}} (1-\hat{a_i}^{k} \times \frac{s_i^{k+1}}{s_i^{k}})^{(1-S_i C_j^{k+1})} \\
D_j^{k+1} &= \prod_{i=1}^{M} (\hat{b_i}^{k} \times \frac{s_i^{k+1}}{s_i^{k}})^{S_i {C_j}^{k+1}} (1-\hat{b_i}^{k} \times \frac{s_i^{k+1}}{s_i^{k}})^{(1-S_i C_j^{k+1})}
\end{aligned}
\tag{7.6}
$$

Plugging Equation (7.6) into Equation (7.4), we can get the following recursive computation of the estimation parameters:

$$
\begin{aligned}
&\hat{a_i}^{k+1} = \hat{a_i}^{k} + \frac{1}{Nd(k+1)} \times \\
&\Big[\sum_{j \in SJ_i^{k+1}} g(\hat{a_i}^{k}, \hat{b_i}^{k}, X_k, X_{k+1})(1 - \hat{a_i}^{k}) \\
&- \sum_{j \in S\bar{J_i}^{k+1}} g(\hat{a_i}^{k}, \hat{b_i}^{k}, X_k, X_{k+1})\hat{a_i}^{k} \Big] \\
&\hat{b_i}^{k+1} = \hat{b_i}^{k} + \frac{1}{Nd(k+1)} \times \\
&\Big[\sum_{j \in SJ_i^{k+1}} (1 - g(\hat{a_i}^{k}, \hat{b_i}^{k}, X_k, X_{k+1}))(1 - \hat{b_i}^{k}) \\
&- \sum_{j \in S\bar{J_i}^{k+1}} (1 - g(\hat{a_i}^{k}, \hat{b_i}^{k}, X_k, X_{k+1}))\hat{b_i}^{k} \Big]
\end{aligned}
\tag{7.7}
$$

Additionally, we can also compute the updated correctness of measured variables (i.e, $\hat{Z_j}^{k+1}$) as follows:

$$
\hat{Z_j}^{k+1} = f(\hat{a_i}^{k+1}, \hat{b_i}^{k+1}, X_{k+1}) \tag{7.8}
$$

where function f is the same as the one in Equation (7.5).

This gives us the recursive equations to compute the estimation parameters of our model in the current time interval based on the estimations from the previous time interval and the observed data up to now. Therefore, we can utilize (7.7) to keep track of the estimation parameter of the sources that report new observations consecutively over time. We also note that the estimation parameter change of the updated sources will affect the credibility of measured variables they report, which in turn will affect the credibility of other sources asserting the same measured variable. We call this credibility update prorogation "Ripple Effect". To capture such an effect, we do a simple trick: only run one EM iteration after applying the recursive formula (as compared to running the full version of EM from scratch). This turns out to be an efficient heuristic based on the following observations: i) the recursive estimation already offers us a reasonably good initialization on the estimation parameter; ii) the credibility change of sources by a few updates in a short time interval is usually slight. This allows the recursive EM to converge much faster than the batch algorithm that starts from a random point. We will further evaluate the performance of the recursive EM algorithm through several tradeoffs studies in the next section.

7.2 Tradeoffs Studies of the Recursive EM Algorithm

In this section we presented the tradeoffs study of the recursive EM algorithm described in the previous section for streaming data. We evaluated the tradeoffs of the algorithm through three dimensions: trustworthiness of sources, freshness of input data and timeliness of the algorithm execution. For trustworthiness study, we vary the ratio of trustworthy sources in the system. The trustworthy sources are defined as the sources who always assert the correct measured variables (i.e, their reliability is 1). For freshness study, we vary the freshness ratio, which is defined as the normalized input size used by the recursive EM algorithm to calculate the initial estimation of parameters over a pre-defined data size (i.e, the one of freshness ratio 1) . For the timeliness, we vary the number of iterations in the recursive EM algorithm to capture the “ripple effect” of an updated observation. The performance metrics we chose to evaluate the algorithm are the false positives and false negatives of the measured variables asserted by sources. We evaluate the above performance tradeoffs of the recursive EM algorithm when different factors affecting the sensing topology generation change. These factors include: the source chat rate, the source node degree skew and its distributions, the source reliability skew and its distribution and the roles of trusted sources.

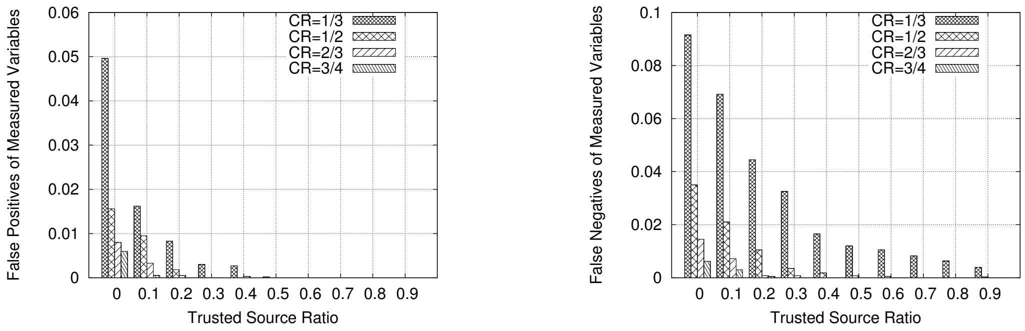
We built a simulator in Matlab 7.10.0 that generates a random number of participants and measured variables. A random probability P_i is assigned to each participant S_i representing his/her reliability (i.e., the ground truth probability that they report correct observations). Each participant S_i report the observations of the measured variables based on its own reliability and chat rate. The chat rate (CR) is defined as the probability a source reports an observation at a time slot, representing his willingness to report observations over time. Observations from different sources are generated as a data stream as time passes by. Each reported observation from S_i has a probability t_i of being true (i.e., reporting a variable as true correctly) and a probability $1 - t_i$ of being false (reporting a variable as true when it is not). We let t_i be uniformly distributed between 0.5 and 1 in our experiments. By definition, we set the reliability of trustworthy sources to be 1. For initialization, the initial values of participant reliability (i.e., t_i) in the experiments are set to the mean value of its definition range. The number of simulated time slots is set to 120. The reported results are averaged over 100 experiments.

7.2.1 Performance Tradeoffs under Different Source Chat Rate

In this subsection, we studied the algorithm tradeoffs over the three dimensions (i.e, trustworthiness, freshness and timeliness) under different chat rates (CR) of sources. We set the number of sources to be 50. The number of true and false measured variables are set to 200 respectively. In the first experiment, we vary the trusted source ratio and evaluate the performance of the recursive algorithm under different source chat rates. The trusted source ratio changes from 0 to 0.9 and the freshness ratio is set to 0.6. The number of EM iterations used to capture the "ripple effect" is set to 1. The results are shown in Figure 7.1. We observe that both false positives and false negatives of the algorithm reduce as the ratio of trusted sources increases in the system. The reason is intuitive: trusted sources only assert the true measured variables, which will make them more distinguishable from the false ones. We also observe that the algorithm performs better when sources become more chatty. This is because the sensing topology becomes more densely connected when more observations are reported by the sources.

In the second experiment, we vary the freshness ratio and evaluate the performance of the recursive algorithm under different chat rates of sources. The ratio of trusted source is set to 0 and the freshness ratio changes from 0.1 to 0.9. The number of EM iterations used to capture the "ripple effect" is set to 1. The results are shown in Figure 7.2. We observe that false positives/negatives decrease as the freshness ratio increases. This is because the more observations are available to be used by the algorithm, the better the initial estimation of the parameters can be computed for the recursive algorithm. Similarly, we also observe that the algorithm has better performance under higher source chat rate.

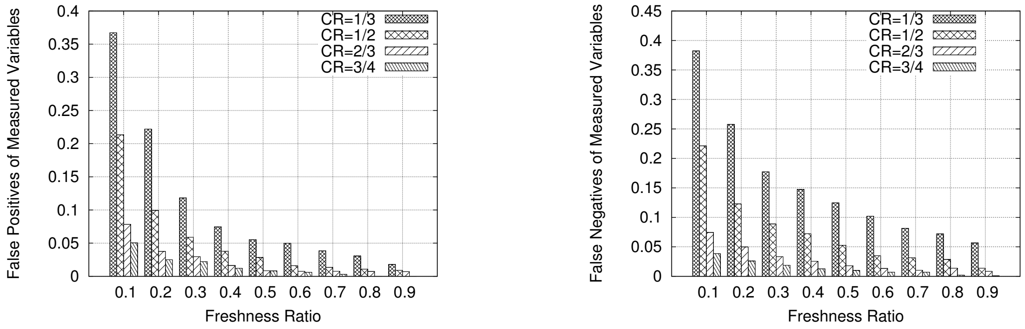
In the third experiment, we vary the number of iterations run by the recursive EM algorithm to catch the "ripple effect" (which affects the average execution time of the algorithm to process an update) and evaluate the performance of the recursive EM algorithm under different chat rates of sources. The ratio of trusted source is set to 0.2 and the freshness ratio is set to 0.6. We vary the number of iterations from 0 to 9. The results are shown in Figure 7.3. We observe that running one EM iteration in the recursive algorithm significantly reduces false positives/negatives of the algorithm compared to directly applying the recursive algorithm without running any EM iterations. However, not much improvement can be gained by further increasing the number of EM iterations over one. As expected, the algorithm also performs better when source chat rate is higher.



(a) False Positives of Measured Variables

(b) False Negatives of Measured Variables

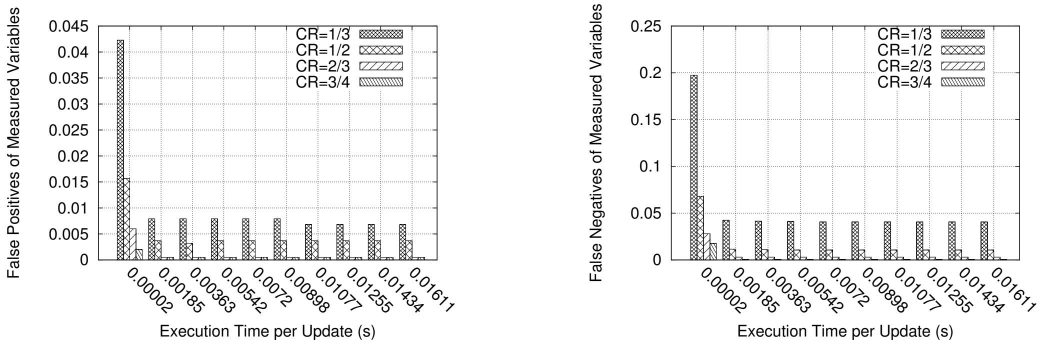
Figure 7.1: Algorithm Performance versus Trusted Source Ratio under Different Chat Rate



(a) False Positives of Measured Variables

(b) False Negatives of Measured Variables

Figure 7.2: Algorithm Performance versus Freshness Ratio under Different Chat Rate



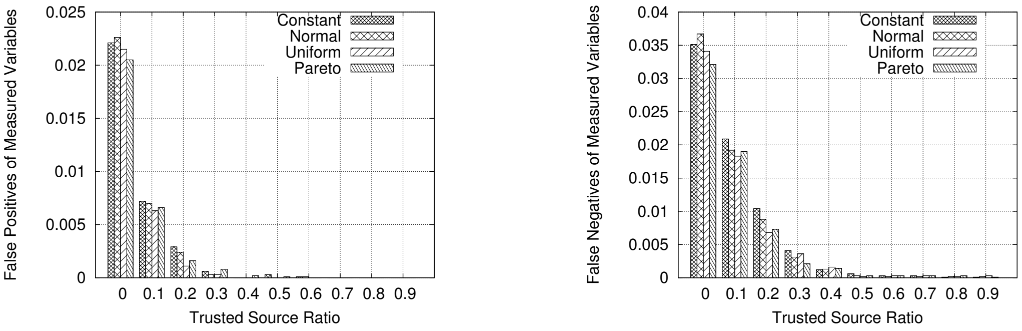
(a) False Positives of Measured Variables

(b) False Negatives of Measured Variables

Figure 7.3: Algorithm Performance versus Timeliness under Different Chat Rate

7.2.2 Performance Tradeoffs under Different Source Degree Distributions

In this subsection, we studied the algorithm performance over the three tradeoff dimensions under different distributions of source node degree. We apply different distributions on the source node degree and repeat the three experiments of the tradeoffs study for the recursive EM algorithm. We set the number of sources to be 50. The number of true and false measured variables are set to 200 respectively. The average number of observations per source is set to 60. The source node degree distributions we choose to examine include constant (i.e., all sources have the same node degree), normal, uniform and pareto (a power-law distribution). The mean value of the number of observations per source of different distributions are kept the same (i.e., 60). The results are shown in Figure 7.4 to Figure 7.6. We observe that the basic trends of the algorithm performance over the three tradeoff dimensions (i.e., trustworthiness, freshness and timeliness) are the same as we discussed in the previous subsection under different source node degree distributions. Furthermore, we don't observe any significant performance differences of the algorithm between different source node degree distributions. This observation verifies the robustness of the algorithm against the source node degree distribution.



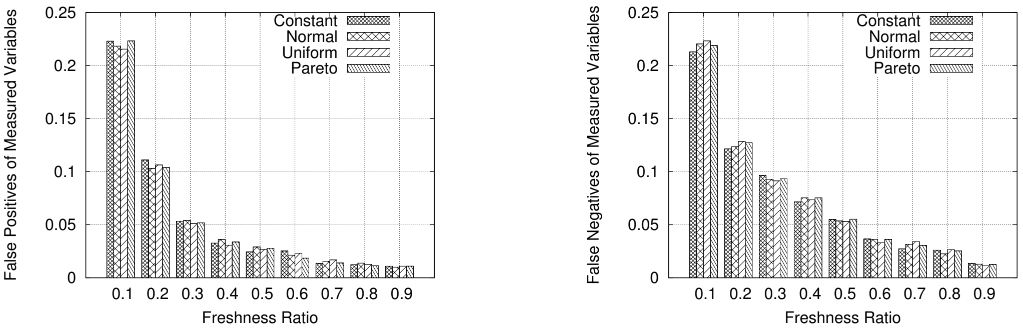
(a) False Positives of Measured Variables

(b) False Negatives of Measured Variables

Figure 7.4: Algorithm Performance versus Trusted Source Ratio under Different Source Degree Distributions

7.2.3 Performance Tradeoffs under Different Source Degree Skewness

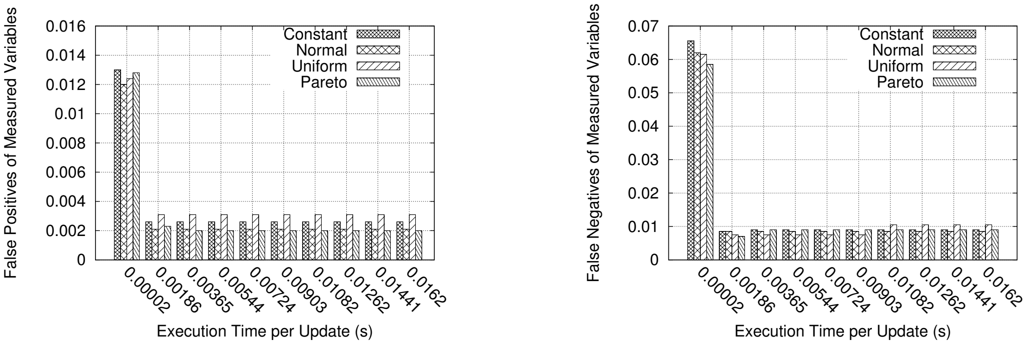
In this subsection, we studied the algorithm performance over three tradeoff dimensions of different source node degree skewness under the normal distribution. We set the number of sources to be 50. The number of true and false measured variables are set to 200 respectively. The average number of observations per



(a) False Positives of Measured Variables

(b) False Negatives of Measured Variables

Figure 7.5: Algorithm Performance versus Freshness Ratio under Different Source Degree Distributions .

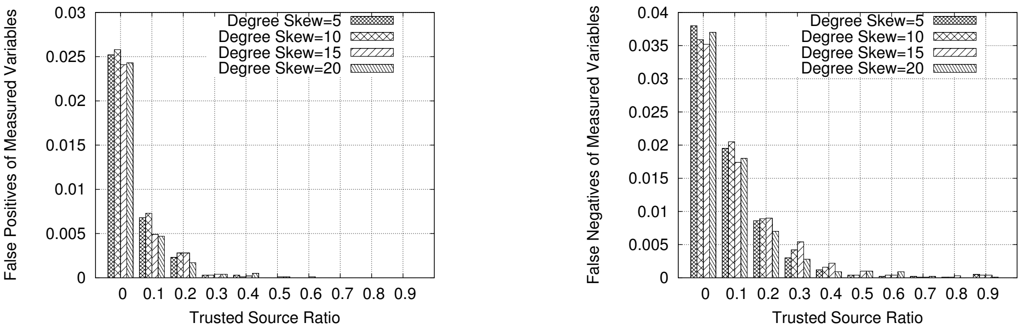


(a) False Positives of Measured Variables

(b) False Negatives of Measured Variables

Figure 7.6: Algorithm Performance versus Timeliness under Different Source Degree Distributions

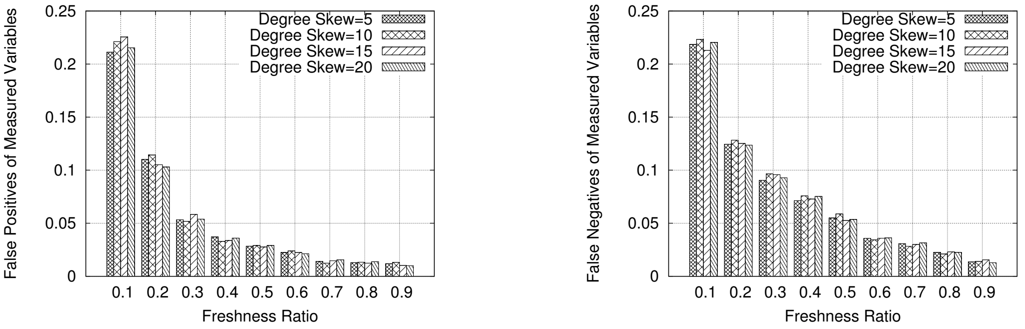
source is set to 60. We vary the standard deviation of the source node degree from 5 to 20. The results are shown in Figure 7.7 to Figure 7.9. We observe that the basic trends of the algorithm performance over the three tradeoff dimensions (i.e., trustworthiness, freshness and timeliness) are the same as we discussed before under different skewness of the source node degree. Moreover, we don't observe any significant performance differences of the algorithm between different skewness of the source node degree. This illustrates the robustness of the algorithm against the skewness of the source node degree under the normal distribution.



(a) False Positives of Measured Variables

(b) False Negatives of Measured Variables

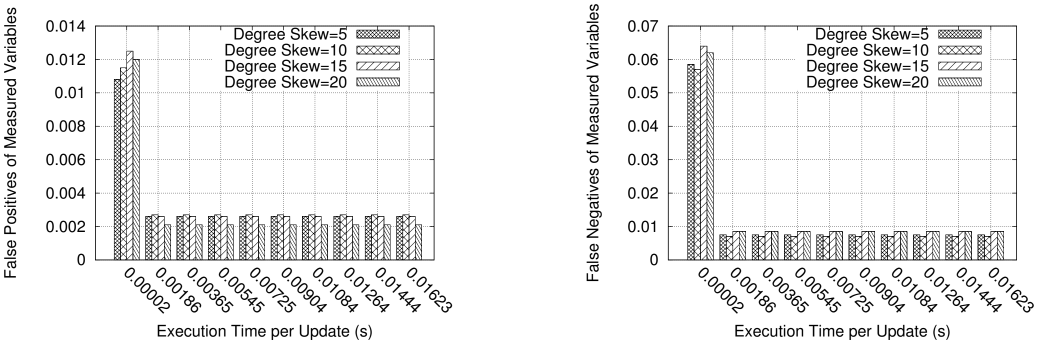
Figure 7.7: Algorithm Performance versus Trusted Source Ratio under Source Degree Skewness



(a) False Positives of Measured Variables

(b) False Negatives of Measured Variables

Figure 7.8: Algorithm Performance versus Freshness Ratio under Source Degree Skewness



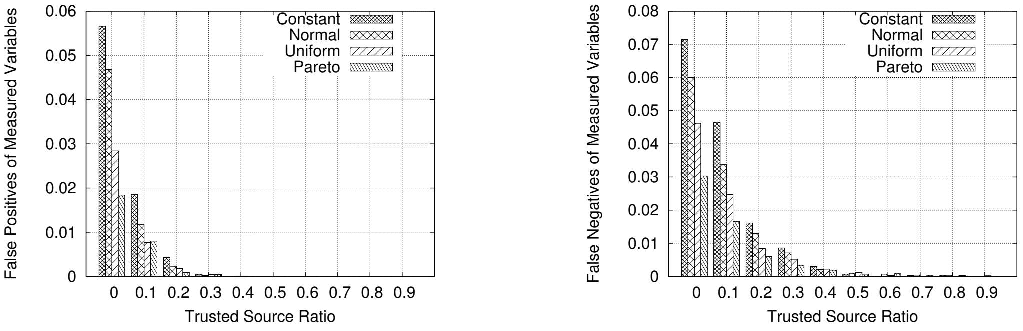
(a) False Positives of Measured Variables

(b) False Negatives of Measured Variables

Figure 7.9: Algorithm Performance versus Timeliness under Source Degree Skewness

7.2.4 Performance Tradeoffs under Different Source Reliability Distributions

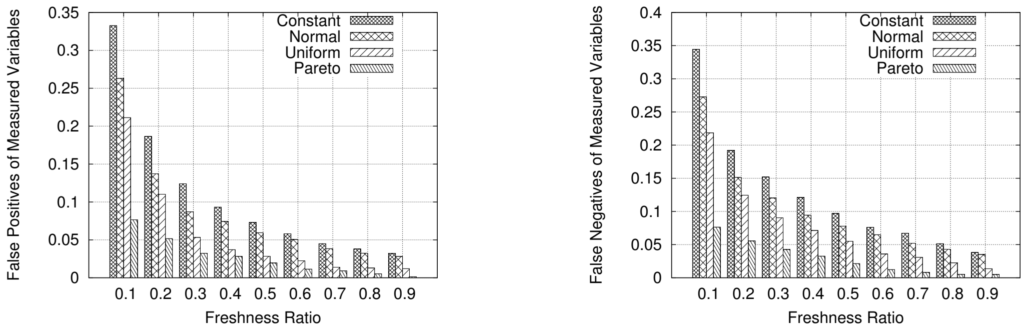
In this subsection, we studied the algorithm performance over the three tradeoff dimensions under different source reliability distributions. We apply different distributions on the source reliability and repeat the three experiments as we did before. We set the number of sources to be 50. The number of true and false measured variables are set to 200 respectively. The average number of observations per source is set to 60. The source reliability degree distributions we choose to examine include constant (i.e., all sources have the same reliability), normal, uniform and pareto (a power-law distribution). The mean value of the average source reliability of different distributions is kept the same (i.e, the mean of the reliability definition range). The standard deviation of the source reliability for the normal distribution is set to 0.08. The results are shown in Figure 7.10 to Figure 7.12. We observe that the basic trends of the algorithm performance over the three tradeoff dimensions (i.e., trustworthiness, freshness and timeliness) are the same as we discussed before under different source reliability distributions. However, we also observe that algorithm performance improves as the source reliability distribution becomes less concentrated over its definition range. The reason is: our algorithm estimation accuracy depends on the correct estimation of the source reliability. Hence, the flatter the source reliability distribution is, the more easily the algorithm can distinguish the reliability of one source from another.



(a) False Positives of Measured Variables

(b) False Negatives of Measured Variables

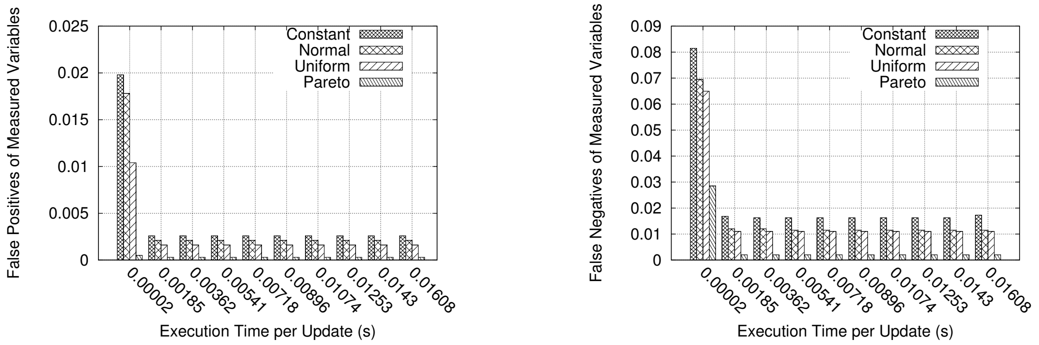
Figure 7.10: Algorithm Performance versus Trusted Source Ratio under Different Source Reliability Distributions



(a) False Positives of Measured Variables

(b) False Negatives of Measured Variables

Figure 7.11: Algorithm Performance versus Freshness Ratio under Different Source Reliability Distributions



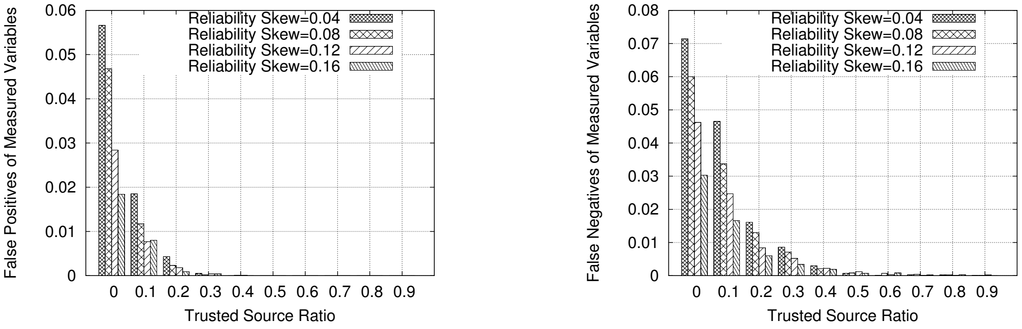
(a) False Positives of Measured Variables

(b) False Negatives of Measured Variables

Figure 7.12: Algorithm Performance versus Timeliness under Different Source Reliability Distributions

7.2.5 Performance Tradeoffs under Different Source Reliability Skewness

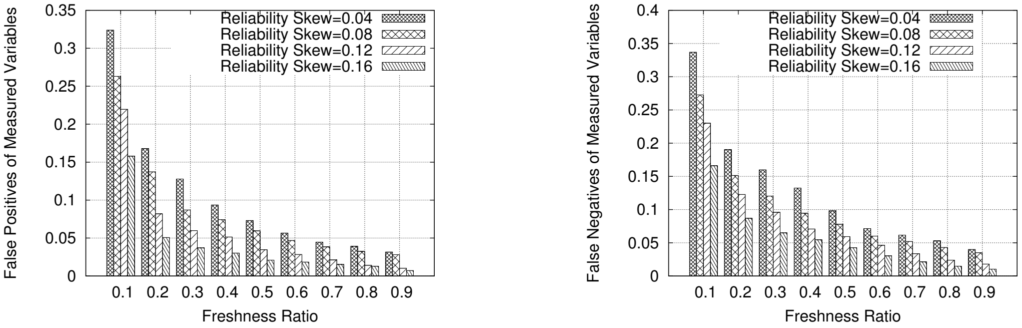
In this subsection, we studied the algorithm performance over the three tradeoff dimensions of different source reliability skewness under the normal distribution. We set the number of sources to be 50. The number of true and false measured variables are set to 200 respectively. The average number of observations per source is set to 60. We set the mean value of the normal distribution as the mean of the source reliability definition range. We vary the standard deviation of the source reliability from 0.04 to 0.16. The results are shown in Figure 7.13 to Figure 7.15. We observe that the basic trends of the algorithm performance over the three tradeoff dimensions (i.e., trustworthiness, freshness and timeliness) are the same as we observed before under different skewness of the source reliability. Moreover, we also observe that the performance of the algorithm improves as the source reliability deviation increases. The reason is similar as we discussed for the source reliability distribution in the previous subsection.



(a) False Positives of Measured Variables

(b) False Negatives of Measured Variables

Figure 7.13: Algorithm Performance versus Trusted Source Ratio under Source Reliability Skewness



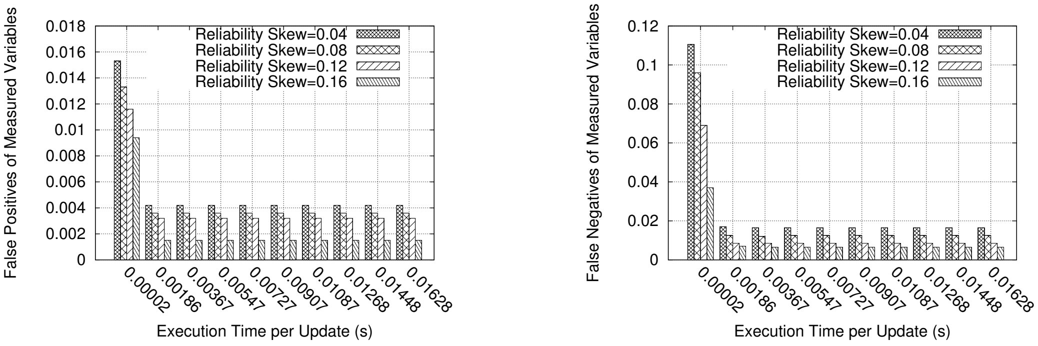
(a) False Positives of Measured Variables

(b) False Negatives of Measured Variables

Figure 7.14: Algorithm Performance versus Freshness Ratio under Source Reliability Skewness

7.2.6 Performance Tradeoffs under Different Roles of Trusted Source

In this subsection, we studied the algorithm performance over three tradeoff dimensions under different roles the trusted sources may play in the system. Here, we classify the roles of the trusted sources into 3 categories depending on the node degree: High Degree, Medium Degree and Low Degree. We set the number of sources to be 50. The number of true and false measured variables are set to 200 respectively. The source node degree for High, Medium and Low are set to be 80, 60 and 40 respectively. The trusted source ratio for freshness and timeliness study is set to 0.2. We varied the role of trusted sources in the system. The results are shown in Figure 7.16 to Figure 7.18. We observe that the basic trends of the

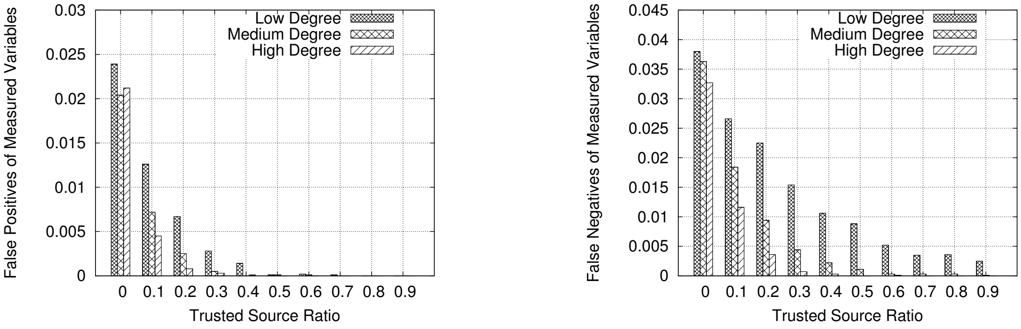


(a) False Positives of Measured Variables

(b) False Negatives of Measured Variables

Figure 7.15: Algorithm Performance versus Timeliness under Source Reliability Skewness

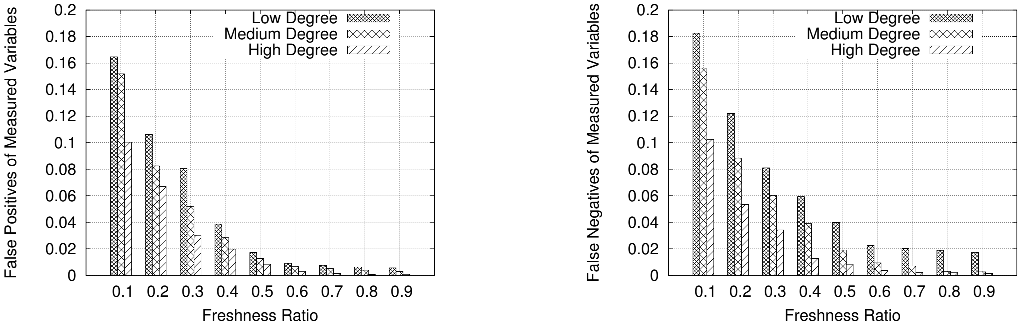
algorithm performance over the three tradeoff dimensions (i.e., trustworthiness, freshness and timeliness) are the same as we discussed before. We also note that the performance of the algorithm improves as the trusted source node degree increases from low to high. The reason is: more observations from trusted sources who always report correctly will help the algorithm to make better decisions on which measured variable is true and which is not.



(a) False Positives of Measured Variables

(b) False Negatives of Measured Variables

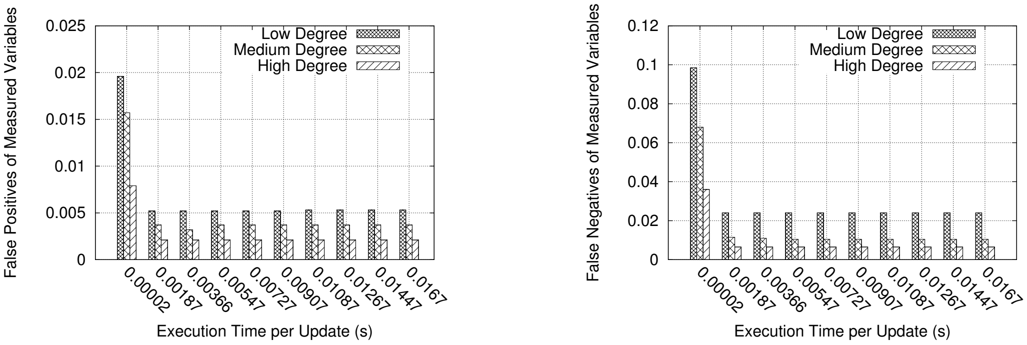
Figure 7.16: Algorithm Performance versus Trusted Source Ratio under Different Trusted Source Roles



(a) False Positives of Measured Variables

(b) False Negatives of Measured Variables

Figure 7.17: Algorithm Performance versus Freshness Ratio under Different Trusted Source Roles



(a) False Positives of Measured Variables

(b) False Negatives of Measured Variables

Figure 7.18: Algorithm Performance versus Timeliness under Different Trusted Source Roles

Chapter 8

Apollo: A Data Distillation Tool for Social Sensing Applications

In this chapter, we describe Apollo, a data distillation tool for social sensing applications that is built on top of the theories developed in this thesis [82]. We first introduce the complete architecture of Apollo and then demonstrate that Apollo can be used as a general tool to clean the noisy data from a large crowd of potentially unreliable sources through several real world social sensing applications.

8.1 Architecture of Apollo

The Apollo architecture is shown in Figure 8.1. One nice feature of the model described in this thesis is that it provides a well-defined and clean abstraction as a bipartite graph containing the sources and the observed measured variables. This allows Apollo to cleanly separate application dependent part that pre-processes the application specific data from the application independent core that does the credibility analysis task. The applicant dependent part is able to handle different types of data by using application specific plug-ins (e.g., distance metrics) to appropriately cluster the measured variables. For example, Apollo can take the inputs as numerical data from sensor readings (e.g., GPS traces), images from photo centric applications (e.g., Flickr) and textual data from text based applications (e.g., Twitter). Essentially, Apollo can also take heterogeneous data representing states of the same measured variables (e.g., a piece of tweet and an image describing the same event). After the measured variables is clustered, Apollo applies the credibility analysis technique on top of the source-to-cluster graph (i.e, the observation matrix) to jointly assess the credibility of both sources and their claims. The results can be shown as the top sources and claims of the measured variables ranked by their credibility values. As shown in the figure, the algorithms developed in this thesis (i.e, the Bayesian interpretation and the EM-CRLB estimation) have been implemented (using Python and C++) and integrated as the core of credibility assessment module in the Apollo architecture.

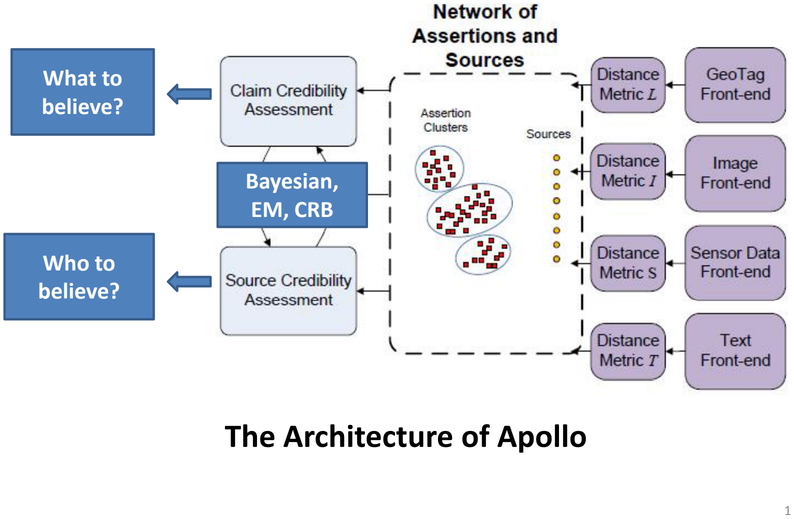
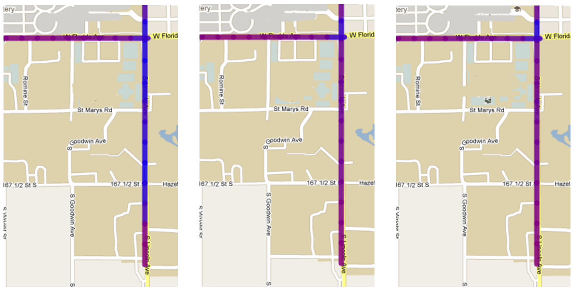


Figure 8.1: The Architecture of Apollo

8.2 Real World Social Sensing Applications of Apollo

In this section, we present two real world sensing applications where Apollo is used to distill the useful information from a large amount of noisy data.

1 Speed Mapping: An important role human could play in social sensing applications is to perform as sensor carriers where they are not directly considered as a source of sensing data. We carried out a traffic monitoring application, Speed Mapping, to validate the performance of our scheme in this type of social sensing application. The goal of Speed Mapping is to create a map of different streets and show the average speed of cars on different segments of the road. People carry phones and share the GPS traces through an application server. The Apollo backend cleans the noisy data and compute the average driving speed on roads. We investigated the data cleaning performance of our scheme in three common causes of sharing noisy data: i) inconsistent context; ii) faulty sensors; iii) incorrect calibration. The inconsistent context refers to the case where the sensing data are collected when people perform the task in an incorrect manner (e.g., speed data of pedestrian are shared instead of drivers.). Faulty sensors are a common cause of bad data in all kinds of sensing applications, where sensors either stop working or provide incorrect readings. Finally, by incorrect calibration we refer to cases where sensors report in different units (e.g., speed in km/h versus mph). We evaluate the performance of our scheme, in each of the above causes of



(a) Averaging (b) Apollo (c) GroundTruth

Figure 8.2: Output of the speed mapping application.

noisy data, by comparing the average speed computed from the data cleaned by Apollo against the ground truth and report the corresponding error.

Implementation

We implemented an actual speed mapping application, where the client side runs on Android phones and reports its data to Apollo, whereas the server side takes its input from Apollo and computes the average traffic speed. Streets are segmented into 500-feet segments identified by a `segment_id`. Each claim consists of a `segment_id`, an average speed value and the number of samples used for averaging.

The *distance metric* is defined such that the distance is infinity if the `segment_id`s differ. Otherwise, it is the absolute difference in the speed value. After distillation with Apollo, the application takes surviving speed measurements and creates the speed map by simply averaging them for each road segment. A color-coded map of the area is produced using Google Map static API. Figure 8.2 is a map of an area that highlights the speed value at a particular location using color-coded marks. The deep blue color means zero speed and a complete red represents the maximum speed ($50mph$ in our experiments).

On the client side, GPS-equipped Android smart-phones (in particular Nexus One and Nexus S phones) do the data collection. An android application is developed to sample GPS location, time, speed, and bearing every 5 seconds. Two identifiers are added to each sample: a NodeID (the cellphone unique IMEI identifier), and a SampleID (cellphone local timestamp value). Each sample is then formatted as a string of key-values and shared via the sever. A total of 15 hours of driving data was collected that covers 10 streets and around 180 miles.

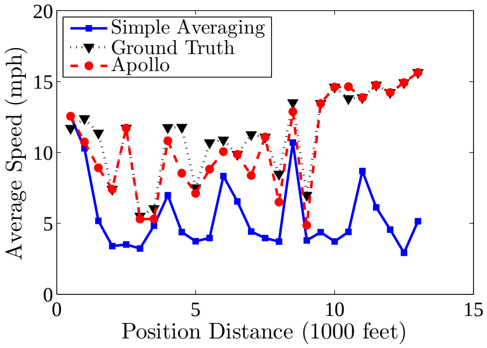


Figure 8.3: The average speed on Main street in the inconsistent context scenario.

Average Error (%)	Simple Averaging	Apollo
Main Street	41.89	9.23
Oak Street	7.71	6.17
First Street	6.89	6.38
Lake Street	0.0	0.53
All Streets	15.0	6.2

Table 8.1: Average error in the inconsistent context scenario.

Inconsistent Context

In the first experiment with the speed mapping application, we investigate the ability of Apollo to remove data that are shared from an inconsistent context. Here, we insert sources who share GPS speed traces while walking instead of driving. When looking at the average speed value at each road segment for a particular street, we observe in Figure 8.3 that the lower pedestrian speed significantly reduces the average traffic speed values for the segments. However, our Apollo implementation of speed mapping removes the pedestrians from averaging and results in a curve much closer to ground truth. The ground truth represents the best possible performance of a well-designed application-specific scheme (e.g., [19])
.

Table 8.1 shows the average percentage error in speed estimation on four different streets. As the results suggest, using Apollo reduces the error compared to a simple averaging scheme.

Faulty Sensors

In the second speed mapping experiment, we investigate the performance of Apollo in the presence of faulty sensors. This example naturally occurred during testing, as some of the phones (particularly Nexus

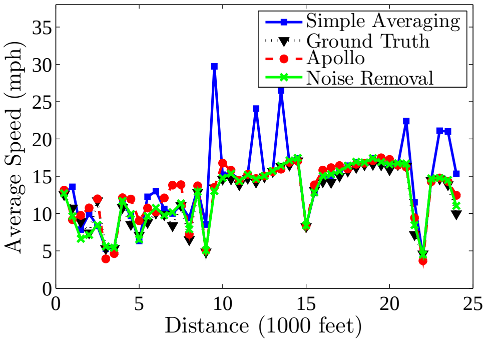


Figure 8.4: The average speed on Oak street in the faulty sensor scenario.

Average Error (%)	Simple Averaging	Noise Removal	Apollo
Main Street	21.76	8.95	14.04
Oak Street	13.36	5.56	5.29
First Street	25.25	5.32	16.18
Lake Street	29.55	1.54	2.4
All Streets	25.31	6.87	9.48

Table 8.2: Average error in the faulty sensor scenario.

Ones) produced very noisy speed values (in ranges from 0 to 150 miles per hour). Application-specific knowledge can be used here to remove the outliers from the data set. By doing this experiment, we aim to show how Apollo can achieve similar performance without using the application-specific knowledge. Again, we plot the speed curve for First street in Figure 8.4 based on both the ground truth and results from simple averaging and Apollo.

The application-specific noise removal scheme uses knowledge of speed limit values, removing samples that are larger than 30% above the street speed limit. We compare the average error over all street segments and report the results in Table 8.2. Note that, the result from Apollo is quite comparable with the application-specific noise removal scheme.

Incorrect Calibration

The final experiment for the speed mapping application focuses on the case where the sensors are incorrectly calibrated. We emulate incorrectly calibrated devices by incorrectly reporting the speed in km/h instead of mph on some phones (without telling the right units to the receiver). Again, we compare both simple averaging and Apollo in computing the average speed values in each street segment. Figure 8.5 compares speed curves of different schemes on the First street. The higher speed values of the simple av-

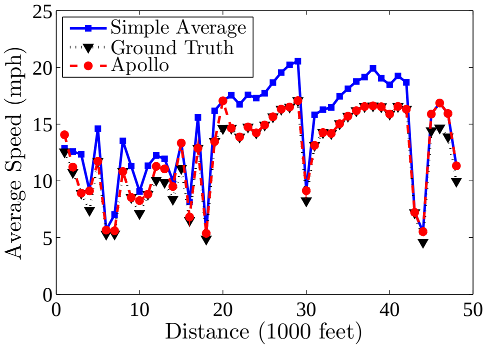


Figure 8.5: The average speed on First street in the calibration error scenario.

Average Error (%)	Simple Averaging	Apollo
Main Street	21.22	5.4
Oak Street	14.85	3.97
First Street	21.02	12.08
Lake Street	17.47	5.74
All Streets	19.02	6.63

Table 8.3: Average error in the incorrect calibration scenario.

erage scheme is caused by taking incorrect reports in km/h. We show the average error over all segments in Table 8.3. The results show a significant improvement when using Apollo.

2 Human as Sensors: This application focuses on human as sensors. With the fast development of social media (e.g., twitter, facebook, Youtube, etc.), human themselves start to perform the role as sensors in social sensing applications. The social media and in particular Twitter can be considered as a huge source of human-generated sensing data. Considering the large population of human as the information sources and their unknown reliability, some challenges exist for this application: i) Since a person can actually tweet anything at anytime, anybody can produce false or irrelevant information. ii) Verifying the trustworthiness of a source is non-trivial as anybody can create a Twitter account.

For validation, we have collected several datasets of tweets using the Twitter API during the time of real events (e.g., Egypt Unrest, Japan Tsunami, Hurricane Irene). The application-dependent module of Apollo (e.g., Parser, Clustering Module, etc.) is used to process the tweets and generate the observation matrix as the input into the credibility assessment module we designed. More specifically, for the *distance metric*, we use the Jaccard distance [78], a simple yet widely used distance function for the content of the tweets. Jaccard distance is the ratio of the size of the set of shared keywords and the size of the set of all keywords of two tweets. All stop-words are discarded since they do not contribute any noticeable meaning to the tweets' content. Since a tweet is a relatively short string of text, we observed that Jaccard distance is sufficient for the purpose.

The application simply reports the most credible tweet of every hour in chronological order. From this presentation, users are able to capture or reconstruct the history of events hour-by-hour. For evaluation purpose, we pick up the top-hourly tweet recommended by Apollo and compare that with the ground truth events reported by the traditional media. The following shows the results we obtained by applying Apollo on tweets collected during Egypt Unrest last year. The data were collected from February 1st until February 18th by by specifying keywords as "Egypt", "Cairo" and "Tahrir", and the location to be Cairo. More than a million tweets have been collected. The results of the experiment is shown in Table 8.4. We observe that all 10 important events selected as ground truth has been covered by the top hourly tweets found by the EM scheme, and their contents match well.

We repeated the experiment with the voting scheme, returning only the tweet with the highest number of votes (i.e., number of sources) every hour. These tweets did not cover all the ground truth events. We then increased the reported tweets to top 2, 3, and so on. Using the voting algorithm, we eventually needed to look down to the 6th ranked tweet of every hour to find all of the ground truth events. This means that voting is not as good at ranking ground truth highly. While the difference between 1 and 6 might seems small as a ranking, the implication is that the user will need to wade through *six times more data* to find relevant information. Figure 8.6 shows the coverage of ground truth events (truth coverage) versus the lowest tweet tank one needs to consider to attain that coverage.

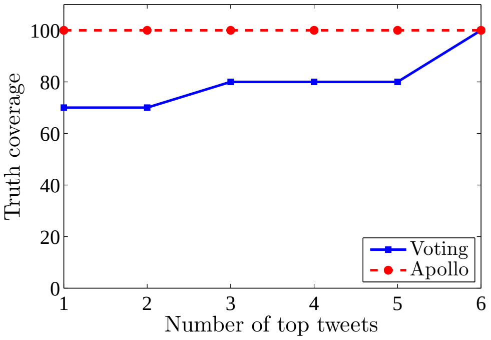


Figure 8.6: Truth coverage vs top results needed for Egypt Unrest dataset

The above results help us understand the performance of our developed schemes in real and massive datasets.

#	Media	Tweet found by EM
1	Google says one of its Middle East managers has gone missing in Cairo, where violent protests against the ruling regime have embroiled Egypt's capital for the past week.	Google says cant find manager last seen in Cairo (AP)
2	Number of protesters in Cairo's Tahir Square are revised to more than a million people.	Aljazeera: Protesters flood Egypt streets: Up to two million gather in Cairo's Tahrir Square as massive protests...
3	Egypt faces another day of protests on Wednesday as protestors dismissed President Hosni Mubarak's pledge to not seek office again after his current term and continued their demand for him to step down immediately.	CBC.ca Protests continue in Egypt after Mubarak vows not to seek office again By the CNN Wire Staff Cairo- http://bit.ly/dWkt6s
4	Internet services partially restored in Cairo.	Egypt Internet Back Up As Protests Turn Violent In Cairo: Internet access in Egypt was restored Wednesday as pro... http://bit.ly/eqY10i
5	Bursts of heavy gunfire early aimed at anti-government demonstrators in Tahrir leave at least five people dead and several wounded.	Heavy gunfire rings out in Cairo protest square: http://apne.ws/fYsV5i.
6	Hundred of thousands of anti-government protesters gather in Tahrir Square for what they have termed the "Day of Departure".	Egypt set for 'Day of Departure': Thousands of Egyptian protesters are again expected in Cairo's Tahrir Squar... http://bit.ly/i9t9rM
7	The leadership of Egypt's ruling National Democratic Party resign, including Gamal Mubarack, the son of Hosni Mubarak. Hossam Badrawi, a member of the liberal wing of the party, became the new secretary-general.	Leadership of Egypt's ruling party resigns http://bit.ly/hebSGm
8	Wael Ghonim, a Google executive and political activist arrested by the state authorities since Jan 28 is released.	CNN: Google executive Wael Ghomin, who went missing in protests in Cairo, Egypt, has been released, Goo... http://bit.ly/eC4LOj
9	Egypt's military rulers yesterday dissolved both houses of parliament, suspended the constitution and pledged presidential and parliamentary elections in six months.	Egypt's army pledges new elections within 6 months: CAIRO -
10	A sea of Egyptians from all walks of life packed every meter in and around Cairo's Tahrir Square on Friday for a "Day of Victory".	Egyptians gather for 'Day of Victory': Waving flags and beating drums, thousands gathered at Cairo's Tahrir Square on Friday...

Table 8.4: Ground truth events and related tweets found by EM in Egypt Unrest

Chapter 9

Conclusions and Future Work

In this thesis, we developed a set of theories and methodologies to quantify the Quality of Information in social sensing. Social sensing has emerged as a new paradigm of sensing and data collection due to the proliferation of mobile devices owned by common individuals, fast data sharing and large scale information dissemination opportunities. A key challenge in social sensing applications lies in quantifying the correctness of collected data and reliability of participants. Solutions to address this key challenge is non-trivial given the reliability of participants is usually unknown *a priori* and there is no independent way to verify the correctness of their measurements.

We first proposed a Bayesian Interpretation that offered a probability semantic to interpret the ranking outputs of the basic fact-finder used in trust analysis in information networks. This interpretation leads to a direct quantification of the accuracy of the conclusions obtained from information network analysis. Hence we provide a general foundation for using information network analysis not only to heuristically extract likely facts, but also to quantify, in analytical founded manner, the probability each source is correct. Such probability constitutes a measure of QoI. We also note the Bayesian Interpretation remain to be an approximation scheme due to the heuristic nature of fact-finders and is sensitive to the priors of initialization.

Considering the limitation of Bayesian Interpretation as mentioned above, we proposed a maximum likelihood estimator to obtain the optimal estimation on participant reliability and the correctness of their measurements. In this effort, we showed that the social sensing applications lend themselves nicely to an Expectation Maximization (EM) formulation. The maximum likelihood estimator we developed makes inference regarding both source reliability and measurement correctness by observing which observations coincide and which don't. The optimal solution, in the sense of maximum likelihood estimation, directly leads to an accurate quantification of the QoI measure for social sensing applications. The EM based approach was shown to outperform the state of the art fact-finding heuristics as well as simple baseline such as majority voting.

An important problem remains unanswered from the maximum likelihood estimation of the EM scheme is: what is the confidence bound of the resulting participant reliability estimation? To answer this question, we derived both real and asymptotic confidence bounds for participant reliability estimation in EM scheme. The confidence bounds are obtained by leveraging the asymptotic normality of the maximum likelihood estimation and computing the Cramer-Rao Lower Bound (CRLB) for the estimation parameters. We studied the limitation of the real and asymptotic CRLBs and demonstrated the trade-offs they offer between computation complexity and estimation scalability. We also examined the robustness of these bounds to changes in the number of sources. The results offered us an understanding of attainable estimation accuracy of source reliability in social sensing applications that rely on un-vetted sources whose reliability is not known in advance.

A few assumptions were made in the maximum likelihood estimation (MLE) model for the EM scheme. Two important ones are : 1) observations from different participants on the same measured variable are assumed to be corroborating; 2) measured variables are assumed to be binary (i.e., true or false). We generalized the MLE model to remove the above two assumptions and make the EM scheme applicable to a much wider range of social sensing applications. First, we extended the estimation parameter to incorporate the conflicting observations from different participants on the same measured variable. The corresponding likelihood function and E-step and M-Step were re-derived to obtain the extended maximum likelihood estimator to handle conflicting observations. Second, we generalized the model for conflicting observations to incorporate non-binary values of measured variables and similarly derived the maximum likelihood estimator for non-binary measured variables.

The iterative EM algorithm is mainly designed for static dataset and not necessarily efficient for the streaming data. We proposed a recursive EM algorithm that computes the updated estimations from the previous results and the new observed data in a recursive way. We studied the performance of the recursive EM algorithm over different tradeoff dimensions such as the trustworthiness of sources, the freshness of input data and timeliness of the algorithm execution. The recursive EM algorithm was shown to achieve nice performance trade-offs.

Finally, the QoI quantification theory and method for social sensing developed in this thesis have been implemented as the credibility assessment core of Apollo. We demonstrated through several real social sensing applications that Apollo can be used as a general and effective tool to extract important information from large amount of noisy data generated from potentially unreliable sources.

For the future work, we note that the theory and methodology developed in this thesis can also be applied to other research area such as the trust analysis in information networks. As we mentioned earlier in Chapter 3, the trust analysis shares a similar problem prototype as the QoI quantification problem we studied for social sensing. Our work could be leveraged to offer not only the relative ranking on the information sources and their claims, but also the quantifiable and confident estimation on the credibility of sources and claims. It is also possible to combine our scheme with several state-of-art trust analysis tools to further improve their ranking results. In this thesis, we mainly focus on the bipartite network topology that consists only sources and their measured variables. However, it is possible to generalize our maximum likelihood estimation approach to a more general graph where each node has a possibility to connect to other nodes in the network, where the connections constitute the constraint between nodes. The remaining challenge is to find the appropriate likelihood function that allows a rigorous optimization problem formulation. The network topology (i.e., all nodes and links between them) will be represented by the likelihood function. Finding the MLE solution of such likelihood function will find the assignment of node attributes (e.g., probability of correctness of sources and claims in the bipartite graph) that is maximally consistent with all constraints.

Finally, the theory and analysis techniques developed in this thesis can be used beyond social sensing to build the next generation of information search engines. Data feeds for future search engines will not be limited to static and slowly updated web pages. Instead, a huge volume of real-time information streams coming from various sources (e.g., sensors, Twitter users, GPS traces, etc) will need to be processed and analyzed in a unified framework to meet some quality of information requirements. Our thesis is a step towards better distillation of useful information from a large crowd. The new search capability will offer a more predictable, reliable and timely summaries of dynamic events that leverage connectivity and observations of the common individual. It will contribute to applications that save time (e.g, traffic monitoring and prediction), money (e.g, stock analysis and prediction) and even lives (e.g., event reports of disasters). Together, they will help materialize a vision of a smarter planet where crowd-sourcing makes lives easier, more efficient and safer than before.

References

[1] H. Le, D. Wang, H. Ahmadi, M. Y. S. Uddin, B. Szymanski, R. Ganti, T. Abdelzaher, O. Fatemieh, H. Wang, J. Pasternack, J. Han, D. Roth, S. Adali, and H. Lei, "Demo: Distilling likely truth from noisy streaming data with apollo," in *The 11th ACM/IEEE Sensys 11 (Demo)*, April 2011.

[2] N. D. Lane, S. B. Eisenman, M. Musolesi, E. Miluzzo, and A. T. Campbell, "Urban sensing systems: Opportunistic or participatory," in *In Proc. ACM 9th Workshop on Mobile Computing Systems and Applications (HOTMOBILE 08)*, 2008.

[3] A. Dua, N. Bulusu, W.-C. Feng, and W. Hu, "Towards trustworthy participatory sensing," in *Proceedings of the 4th USENIX conference on Hot topics in security*, ser. HotSec'09. Berkeley, CA, USA: USENIX Association, 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=1855628.1855636 pp. 8–8.

[4] P. Gilbert, L. P. Cox, J. Jung, and D. Wetherall, "Toward trustworthy mobile sensing," in *Proceedings of the Eleventh Workshop on Mobile Computing Systems &#38; Applications*, ser. HotMobile '10. New York, NY, USA: ACM, 2010, pp. 31–36.

[5] L. Deng and L. P. Cox, "Livecompare: Grocery bargain hunting through participatory sensing," in *HotMobile*, 2009.

[6] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *7th international conference on World Wide Web (WWW'07)*, 1998. [Online]. Available: http://portal.acm.org/citation.cfm?id=297805.297827 pp. 107–117.

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

[8] 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

[9] A. Galland, S. Abiteboul, A. Marian, and P. Senellart, "Corroborating information from disagreeing views," in *WSDM*, 2010, pp. 131–140.

[10] 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'09*, 2009.

[11] X. Yin and W. Tan, "Semi-supervised truth discovery," in *WWW*. New York, NY, USA: ACM, 2011.

[12] 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.*, vol. 5, no. 6, pp. 550–561, Feb. 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2168651.2168656

[13] A. T. Campbell, S. B. Eisenman, N. D. Lane, E. Miluzzo, R. A. Peterson, H. Lu, X. Zheng, M. Musolesi, K. Fodor, and G.-S. Ahn, "The rise of people-centric sensing," *IEEE Internet Computing*, vol. 12, no. 4, pp. 12–21, July 2008. [Online]. Available: http://dx.doi.org/10.1109/MIC.2008.90

[14] E. Miluzzo, N. D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. B. Eisenman, X. Zheng, and A. T. Campbell, "Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application," in *Proceedings of the 6th ACM conference on Embedded network sensor systems*, ser. SenSys '08. New York, NY, USA: ACM, 2008. [Online]. Available: http://doi.acm.org/10.1145/1460412.1460445 pp. 337–350.

[15] T. Abdelzaher, Y. Anokwa, P. Boda, J. Burke, D. Estrin, L. Guibas, A. Kansal, S. Madden, and J. Reich, "Mobiscopes for human spaces," *IEEE Pervasive Computing*, vol. 6, no. 2, pp. 20–29, 2007.

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

[17] 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 *Proceedings of the 4th international conference on Embedded networked sensor systems*, ser. SenSys '06. New York, NY, USA: ACM, 2006. [Online]. Available: http://doi.acm.org/10.1145/1182807.1182821 pp. 125–138.

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

[19] A. Thiagarajan, J. Biagioni, T. Gerlich, and J. Eriksson, "Cooperative transit tracking using smartphones," in *SenSys'10*, 2010, pp. 85–98.

[20] 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 *Proceedings of the 5th international conference on Embedded networked sensor systems*, ser. SenSys '07. New York, NY, USA: ACM, 2007. [Online]. Available: http://doi.acm.org/10.1145/1322263.1322273 pp. 87–101.

[21] B. Longstaff, S. Reddy, and D. Estrin, "Improving activity classification for health applications on mobile devices using active and semi-supervised learning," 6 2010.

[22] A. Helal, D. J. Cook, and M. Schmalz, "Smart home-based health platform for behavioral monitoring and alteration of diabetes patients." *Journal of diabetes science and technology*, vol. 3, no. 1, pp. 141–148, Jan. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2769843/

[23] A. Madan, M. Cebrian, D. Lazer, and A. Pentland, "Social sensing for epidemiological behavior change," in *Proceedings of the 12th ACM international conference on Ubiquitous computing*, ser. Ubicomp '10. New York, NY, USA: ACM, 2010. [Online]. Available: http://doi.acm.org/10.1145/1864349.1864394 pp. 291–300.

[24] A. Madan, S. T. Moturu, D. Lazer, and A. Pentland, "Social sensing: obesity, unhealthy eating and exercise in face-to-face networks," in *Wireless Health*, 2010, pp. 104–110.

[25] D. J. Cook and L. B. Holder, "Sensor selection to support practical use of health-monitoring smart environments," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 4, pp. 339–351, 2011. [Online]. Available: http://dx.doi.org/10.1002/widm.20

[26] R. K. Ganti, S. Srinivasan, and A. Gacic, "Multisensor fusion in smartphones for lifestyle monitoring," in *Proceedings of the 2010 International Conference on Body Sensor Networks*, ser. BSN '10. Washington, DC, USA: IEEE Computer Society, 2010. [Online]. Available: http://dx.doi.org/10.1109/BSN.2010.10 pp. 36–43.

[27] H. Ahmadi, N. Pham, R. Ganti, T. Abdelzaher, S. Nath, and J. Han, "Privacy-aware regression modeling of participatory sensing data," in *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '10. New York, NY, USA: ACM, 2010. [Online]. Available: http://doi.acm.org/10.1145/1869983.1869994 pp. 99–112.

[28] N. Pham, R. K. Ganti, Y. S. Uddin, S. Nath, and T. Abdelzaher, "Privacy-preserving reconstruction of multidimensional data maps in vehicular participatory sensing," in *Proceedings of the 7th European conference on Wireless Sensor Networks*, ser. EWSN'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 114–130.

[29] 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.

[30] T. Park, J. Lee, I. Hwang, C. Yoo, L. Nachman, and J. Song, "E-gesture: a collaborative architecture for energy-efficient gesture recognition with hand-worn sensor and mobile devices," in *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '11. New York, NY, USA: ACM, 2011. [Online]. Available: http://doi.acm.org/10.1145/2070942.2070969 pp. 260–273.

[31] 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.

[32] D. Wang, H. Ahmadi, T. Abdelzaher, H. Chenji, R. Stoleru, and C. Aggarwal, "Optimizing quality-of-information in cost-sensitive sensor data fusion," in *IEEE 7th International Conference on Distributed Computing in Sensor Systems (DCoSS 11)*, June 2011.

[33] S. Reddy, K. Shilton, G. Denisov, C. Cenizal, D. Estrin, and M. Srivastava, "Biketastic: sensing and mapping for better biking," in *Proceedings of the 28th international conference on Human factors in computing systems*, ser. CHI '10. New York, NY, USA: ACM, 2010. [Online]. Available: http://doi.acm.org/10.1145/1753326.1753598 pp. 1817–1820.

[34] M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard, R. West, and P. Boda, "Peir, the personal environmental impact report, as a platform for participatory sensing systems research," in *Proceedings of the 7th international conference on Mobile systems, applications, and services*, ser. MobiSys '09. New York, NY, USA: ACM, 2009. [Online]. Available: http://doi.acm.org/10.1145/1555816.1555823 pp. 55–68.

[35] S. Reddy, D. Estrin, and M. Srivastava, "Recruitment framework for participatory sensing data collections," in *Proceedings of the 8th International Conference on Pervasive Computing*. Springer Berlin Heidelberg, May 2010, pp. 138–155.

[36] S. A. Delre, W. Jager, and M. A. Janssen, "Diffusion dynamics in small-world networks with heterogeneous consumers," *Comput. Math. Organ. Theory*, vol. 13, pp. 185–202, June 2007. [Online]. Available: http://portal.acm.org/citation.cfm?id=1210317.1210335

[37] C. Hui, M. K. Goldberg, M. Magdon-Ismail, and W. A. Wallace, "Simulating the diffusion of information: An agent-based modeling approach." *IJATS*, pp. 31–46, 2010.

[38] J. Xie, S. Sreenivasan, G. Korniss, W. Zhang, C. Lim, and B. K. Szymanski, "Social consensus through the influence of committed minorities," *CoRR*, vol. abs/1102.3931, 2011.

[39] P. Gilbert, J. Jung, K. Lee, H. Qin, D. Sharkey, A. Sheth, and L. P. Cox, "Youprove: authenticity and fidelity in mobile sensing," in *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '11. New York, NY, USA: ACM, 2011. [Online]. Available: http://doi.acm.org/10.1145/2070942.2070961 pp. 176–189.

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

[41] X. Dong, L. Berti-Equille, and D. Srivastava, "Truth discovery and copying detection in a dynamic world," *VLDB*, vol. 2, no. 1, pp. 562–573, 2009. [Online]. Available: http://portal.acm.org/citation.cfm?id=1687627.1687691

[42] 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.

[43] 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.

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

[45] R. Balakrishnan, "Source rank: Relevance and trust assessment for deep web sources based on inter-source agreement," in *20th World Wide Web Conference (WWW'11)*, 2011.

[46] Y. Sun, Y. Yu, and J. Han, "Ranking-based clustering of heterogeneous information networks with star network schema," in *15th SIGKDD international conference on Knowledge discovery and data mining (KDD'09)*, 2009. [Online]. Available: http://doi.acm.org/10.1145/1557019.1557107 pp. 797–806.

[47] C. Aggarwal and T. Abdelzaher, "Integrating sensors and social networks," *Social Network Data Analytics, Springer*, 2011.

[48] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '08. New York, NY, USA: ACM, 2008. [Online]. Available: http://doi.acm.org/10.1145/1401890.1401965 pp. 614–622.

[49] O. Dekel and O. Shamir, "Vox populi: Collecting high-quality labels from a crowd," in *In Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

[50] 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*, vol. 39, no. 1, pp. 1–38, 1977.

[51] C. Zhai, "A note on the expectation maximization (em) algorithm," *Department of Computer Scinece, University of Illinois at Urbana Champaign*, 2007.

[52] J. Bilmes, "A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," *Technical Report, University of Berkeley, ICSI-TR-97-021*, 1997.

[53] G. J. McLachlan and T. Krishnan, "The em algorithm and extensions." *John Wiley and Sons, Inc.,*, 1997.

[54] H. Cramer, *Mathematical Methods of Statistics*. Princeton Univ. Press., 1946.

[55] R. V. Hogg and A. T. Craig, *Introduction to mathematical statistics*. Prentice Hall, 1995.

[56] G. Casella and R. Berger, *Statistical Inference*. Duxbury Press, 2002.

[57] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*, 2nd ed. Wiley-Interscience, Nov. 2001. [Online]. Available: http://www.worldcat.org/isbn/0471056693

[58] U. T. Inc and U. T. I. Staff, *Solving Data Mining Problems Using Pattern Recognition Software with Cdrom*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1997.

[59] R. A. Johnson and D. W. Wichern, *Applied multivariate statistical analysis*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2002.

[60] J.Han, M.Kamber, and J. Pei, *Data Mining: Concepts and Techniques, Third Edition*. Morgan Kaufman, 2011.

[61] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Transactions of the ASME Journal of Basic Engineering*, no. 82 (Series D), pp. 35–45, 1960. [Online]. Available: http://www.cs.unc.edu/˜welch/kalman/media/pdf/Kalman1960.pdf

[62] A. Doucet, N. De Freitas, and N. Gordon, Eds., *Sequential Monte Carlo methods in practice*, 2001. [Online]. Available: http://www.worldcatlibraries.org/wcpa/top3mset/839aaf32b6957a10a19afeb4da09e526.html

[63] M. E. Whitman and H. J. Mattord, *Principles of Information Security*. Boston, MA, United States: Course Technology Press, 2004.

[64] Q. Lian, Z. Zhang, M. Yang, B. Y. Zhao, Y. Dai, and X. Li, "An empirical study of collusion behavior in the maze p2p file-sharing system," in *Proceedings of the 27th International Conference on Distributed Computing Systems*, ser. ICDCS '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 56–.

[65] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman, "Sybilguard: defending against sybil attacks via social networks," *SIGCOMM Comput. Commun. Rev.*, vol. 36, pp. 267–278, August 2006.

[66] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, vol. 17, no. 6, pp. 734–749, 2005.

[67] N. Mustapha, M. Jalali, and M. Jalali, "Expectation maximization clustering algorithm for user modeling in web usage mining systems," *European Journal of Scientific Research*, vol. 32, no. 4, pp. 467–476, 2009.

[68] D. Pomerantz and G. Dudek, “Context dependent movie recommendations using a hierarchical bayesian model,” in *Proceedings of the 22nd Canadian Conference on Artificial Intelligence: Advances in Artificial Intelligence*, ser. Canadian AI ’09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 98–109.

[69] G. Adomavicius and Y. Kwon, “New recommendation techniques for multicriteria rating systems,” *IEEE Intelligent Systems*, vol. 22, no. 3, pp. 48–55, May 2007. [Online]. Available: http://dx.doi.org/10.1109/MIS.2007.58

[70] A. Jøsang, R. Ismail, and C. Boyd, “A survey of trust and reputation systems for online service provision,” *Decision Support Systems*, vol. 43, no. 2, pp. 618–644, Mar. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.dss.2005.05.019

[71] K. Aberer and Z. Despotovic, “Managing trust in a peer-2-peer information system,” in *CIKM ’01: Proceedings of the tenth international conference on Information and knowledge management*. New York, NY, USA: ACM, 2001. [Online]. Available: http://dx.doi.org/10.1145/502585.502638 pp. 310–317.

[72] D. Houser and J. Wooders, “Reputation in auctions: Theory, and evidence from ebay,” *Journal of Economics & Management Strategy*, vol. 15, no. 2, pp. 353–369, 2006. [Online]. Available: http://dx.doi.org/10.1111/j.1530-9134.2006.00103.x

[73] K. Hoffman, D. Zage, and C. N. Rotaru, “A survey of attack and defense techniques for reputation systems,” *ACM Computing Surveys*, vol. 42, no. 1, pp. 1–31, Dec. 2009. [Online]. Available: http://dx.doi.org/10.1145/1592451.1592452

[74] 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.

[75] 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.

[76] C. F. J. Wu, “On the convergence properties of the EM algorithm,” *The Annals of Statistics*, vol. 11, no. 1, pp. 95–103, 1983. [Online]. Available: http://dx.doi.org/10.2307/2240463

[77] D. Wang, T. Abdelzaher, L. Kaplan, and C. C. Aggarwal, “On quantifying the accuracy of maximum likelihood estimation of participant reliability in social sensing,” in *DMSN11: 8th International Workshop on Data Management for Sensor Networks*, August 2011.

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

[79] D. Wang, L. Kaplan, T. Abdelzaher, and C. C. Aggarwal, “On scalability and robustness limitations of real and asymptotic confidence bounds in social sensing,” in *The 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON 12)*, June 2012.

[80] X. L. Dong, L. Berti-Equille, and D. Srivastava, “Integrating conflicting data: the role of source dependence,” *Proc. VLDB Endow.*, vol. 2, pp. 550–561, August 2009. [Online]. Available: http://portal.acm.org/citation.cfm?id=1687627.1687690

[81] D. M. Titterington, “Recursive parameter estimation using incomplete data,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 46, no. 2, pp. pp. 257–267, 1984. [Online]. Available: http://www.jstor.org/stable/2345509

[82] 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.

Appendix A

Consider an assertion C_j made be several sources $S_{i_1}, ..., S_{i_K}$. Let $S_{i_k}C_j$ denote the fact that source S_{i_k} made assertion C_j. We further assume that Equation (3.5) and Equation (3.6) hold. In other words:

$$\frac{P(S_{i_k}C_j|C_j^t)}{P(S_{i_k}C_j)} = 1 + \delta^t_{i_kj}$$
$$\frac{P(S_{i_k}C_j|C_j^f)}{P(S_{i_k}C_j)} = 1 + \delta^f_{i_kj}$$

where $|\delta^t_{i_kj}| << 1$ and $|\delta^f_{i_kj}| << 1$.

Under these assumptions, we prove that the joint probability $P(S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j)$, denoted for simplicity by $P(Sources_j)$, is equal to the product of marginal probabilities $P(S_{i_1}C_j), ..., P(S_{i_K}C_j)$.

First, note that, by definition:

$$\begin{aligned} P(Sources_j) &= P(S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) \\ &= P(S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j|C_j^t)P(C_j^t) \\ &+ P(S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j|C_j^f)P(C_j^f) \end{aligned} \tag{9.1}$$

Using the conditional independence assumption, we get:

$$\begin{aligned} P(Sources_j) &= P(C_j^t)\prod_{k=1}^{K} P(S_{i_k}C_j|C_j^t) \\ &+ P(C_j^f)\prod_{k=1}^{K} P(S_{i_k}C_j|C_j^f) \end{aligned} \tag{9.2}$$

Using Equation (3.5) and Equation (3.6), the above can be rewritten as:

$$\begin{aligned} P(Sources_j) &= P(C_j^t)\prod_{k=1}^{K_j}(1+\delta_{i_kj}^t)\prod_{k=1}^{K_j}P(S_{i_k}C_j) \\ &+ P(C_j^f)\prod_{k=1}^{K_j}(1+\delta_{i_kj}^f)\prod_{k=1}^{K_j}P(S_{i_k}C_j) \end{aligned} \tag{9.3}$$

and since $|\delta_{i_kj}^t| << 1$ and $|\delta_{i_kj}^f| << 1$, any higher-order terms involving them can be ignored. Hence, $\prod_{k=1}^{K_j}(1+\delta_{i_kj}^t) = 1 + \sum_{k=1}^{K_j}\delta_{i_kj}^t$, which results in:

$$\begin{aligned} P(Sources_j) &= P(C_j^t)(1+\sum_{k=1}^{K_j}\delta_{i_kj}^t)\prod_{k=1}^{K}P(S_{i_k}C_j) \\ &+ P(C_j^f)(1+\sum_{k=1}^{K_j}\delta_{i_kj}^f)\prod_{k=1}^{K}P(S_{i_k}C_j) \end{aligned} \tag{9.4}$$

Distributing multiplication over addition in Equation (9.4), then using the fact that $P(C_j^t) + P(C_j^f) = 1$ and rearranging, we get:

$$P(Sources_j) = \prod_{k=1}^{K_j}P(S_{i_k}C_j)(1+Terms_j) \tag{9.5}$$

where:

$$Terms_j = P(C_j^t)\sum_{k=1}^{K_j}\delta_{i_kj}^t + P(C_j^f)\sum_{k=1}^{K_j}\delta_{i_kj}^f \tag{9.6}$$

Next, it remains to compute $Terms_j$.

Consider $\delta_{i_kj}^t$ as defined in Equation (3.5). We can rewrite the equation as follows:

$$\delta_{i_kj}^t = \frac{P(S_{i_k}C_j|C_j^t) - P(S_{i_k}C_j)}{P(S_{i_k}C_j)} \tag{9.7}$$

where by definition, $P(S_{i_k}C_j) = P(S_{i_k}C_j|C_j^t)P(C_j^t) + P(S_{i_k}C_j|C_j^f)P(C_j^f)$. Substituting in Equation (9.7), we get:

$$\delta^t_{i_kj} \quad = \quad \frac{P(S_{i_k}C_j|C^t_j)(1-P(C^t_j)) - P(S_{i_k}C_j|C^f_j)P(C^f_j)}{P(S_{i_k}C_j|C^t_j)P(C^t_j) + P(S_{i_k}C_j|C^f_j)P(C^f_j)} \tag{9.8}$$

Using the fact that $1 - P(C^t_j) = P(C^f_j)$ in the numerator, and rearranging, we get:

$$\delta^t_{i_kj} = \frac{(P(S_{i_k}C_j|C^t_j) - P(S_{i_k}C_j|C^f_j))P(C^f_j)}{P(S_{i_k}C_j|C^t_j)P(C^t_j) + P(S_{i_k}C_j|C^f_j)P(C^f_j)} \tag{9.9}$$

We can similarly show that:

$$\begin{aligned}
\delta^f_{i_kj} &= \frac{P(S_{i_k}C_j|C^f_j) - P(S_{i_k}C_j)}{P(S_{i_k}C_j)} \\
&= \frac{P(S_{i_k}C_j|C^f_j)(1-P(C^f_j)) - P(S_{i_k}C_j|C^t_j)P(C^t_j)}{P(S_{i_k}C_j|C^t_j)P(C^t_j) + P(S_{i_k}C_j|C^f_j)P(C^f_j)} \\
&= \frac{(P(S_{i_k}C_j|C^f_j) - P(S_{i_k}C_j|C^t_j))P(C^t_j)}{P(S_{i_k}C_j|C^t_j)P(C^t_j) + P(S_{i_k}C_j|C^f_j)P(C^f_j)}
\end{aligned} \tag{9.10}$$

Dividing Equation (9.9) by Equation (9.10), we get:

$$\frac{\delta^t_{i_kj}}{\delta^f_{i_kj}} = -\frac{P(C^f_j)}{P(C^t_j)} \tag{9.11}$$

Substituting for $\delta^t_{i_kj}$ from Equation (9.11) into Equation (9.6), we get $Terms_j = 0$. Substituting with this result in Equation (9.5), we get:

$$P(Sources_j) = \prod_{k=1}^{K_j} P(S_{i_k}C_j) \tag{9.12}$$

The above result completes the proof. We have shown that the joint probability $P(S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j)$, denoted for simplicity by $P(Sources_j)$, is well approximated by the product of marginal probabilities $P(S_{i_1}C_j), ..., P(S_{i_K}C_j)$. Note that, the proof did not assume independence of the marginals. Instead, it proved the result under the small δ_{i_kj} assumption.

Appendix B

The expectation term in Equation (5.8) can be further simplified:

$$
\begin{aligned}
&E\Big[\frac{(2X_{kj}-1)Z_j(2X_{lq}-1)Z_q}{a_k^{X_{kj}}(1-a_k)^{(1-X_{kj})}a_l^{X_{lq}}(1-a_l)^{(1-X_{lq})}}\Big] = \\
&\sum_{x\in\mathcal{X}} \frac{(2X_{kj}-1)Z_j(2X_{lq}-1)Z_q}{a_k^{X_{kj}}(1-a_k)^{(1-X_{kj})}a_l^{X_{lq}}(1-a_l)^{(1-X_{lq})}} p(X|\theta) \\
&= \sum_{x\in\mathcal{X}} \frac{(2X_{kj}-1)(2X_{lq}-1)Z_jZ_q}{a_k^{X_{kj}}(1-a_k)^{(1-X_{kj})}a_l^{X_{lq}}(1-a_l)^{(1-X_{lq})}} \times \\
&\left(\prod_{j'=1}^{N}\left\{\prod_{i=1}^{M} a_i^{X_{ij'}}(1-a_i)^{(1-X_{ij'})} \times d\right.\right. \\
&\left.\left. + \prod_{i=1}^{M} b_i^{X_{ij'}}(1-b_i)^{(1-X_{ij'})} \times (1-d)\right\}\right) \qquad (9.13)
\end{aligned}
$$

When $j \neq q$, plugging the expressions of Z_j and Z_q, we can prove the expectation term in Equation (5.8) is zero:

$$
\begin{aligned}
&E\Big[\frac{(2X_{kj}-1)Z_j(2X_{lq}-1)Z_q}{a_k^{X_{kj}}(1-a_k)^{(1-X_{kj})}a_l^{X_{lq}}(1-a_l)^{(1-X_{lq})}}\Big] = \\
&\sum_{x\in\mathcal{X}}(2X_{kj}-1)(2X_{lq}-1)\times \\
&\Big(\prod_{\substack{i=1\\ i\neq k}}^{M} a_i^{X_{ij}}(1-a_i)^{(1-X_{ij})}\times d\prod_{\substack{i=1\\ i\neq l}}^{M} a_i^{X_{iq}}(1-a_i)^{(1-X_{iq})}\times d\Big) \\
&\times\Bigg(\prod_{\substack{j^{'}=1\\ j^{'}\neq j \text{ or } q}}^{N}\Bigg\{\prod_{i=1}^{M} a_i^{X_{ij}{'}}(1-a_i)^{(1-X_{ij}{'})}\times d \\
&+\prod_{i=1}^{M} b_i^{X_{ij}{'}}(1-b_i)^{(1-X_{ij}{'})}\times(1-d)\Bigg\}\Bigg) \\
&=\sum_{x\in\mathcal{X}j\times\mathcal{X}_q}(2X_{kj}-1)(2X_{lq}-1)\times \\
&\Big(\prod_{\substack{i=1\\ i\neq k}}^{M} a_i^{X_{ij}}(1-a_i)^{(1-X_{ij})}\times d\prod_{\substack{i=1\\ i\neq l}}^{M} a_i^{X_{iq}}(1-a_i)^{(1-X_{iq})}\times d\Big) \\
&=\sum_{X_{kj}=0}^{1}\sum_{X_{lq}=0}^{1}(2X_{kj}-1)(2X_{lq}-1)=0 \qquad j\neq q
\end{aligned}
\tag{9.14}
$$

Appendix C

The following derivation demonstrates the details to obtain the results in (6.9) to (6.11). The derivation that maximizes the $Q\left(\theta|\theta^{(t)}\right)$ in the M-step in Section 6.1 yields:

$$\begin{aligned}
&\sum_{j=1}^{N} Z(t,j)\left[\frac{S_iC_j^T}{a_i^{T*}} - \frac{(1-S_iC_j^T-S_iC_j^F)}{1-a_i^{T*}-a_i^{F*}}\right]=0\\
&\sum_{j=1}^{N} Z(t,j)\left[\frac{S_iC_j^F}{a_i^{F*}} - \frac{(1-S_iC_j^T-S_iC_j^F)}{1-a_i^{T*}-a_i^{F*}}\right]=0
\end{aligned}\tag{9.15}$$

$$\begin{aligned}
&\sum_{j=1}^{N} (1-Z(t,j))\left[\frac{S_iC_j^T}{b_i^{T*}} - \frac{(1-S_iC_j^T-S_iC_j^F)}{1-b_i^{T*}-b_i^{F*}}\right]=0\\
&\sum_{j=1}^{N} (1-Z(t,j))\left[\frac{S_iC_j^F}{b_i^{F*}} - \frac{(1-S_iC_j^T-S_iC_j^F)}{1-b_i^{T*}-b_i^{F*}}\right]=0
\end{aligned}\tag{9.16}$$

$$\sum_{j=1}^{N}\left[Z(t,j)\sum_{i=1}^{M}\frac{1}{d^*} - (1-Z(t,j))\sum_{i=1}^{M}\frac{1}{1-d^*}\right]=0\tag{9.17}$$

As we defined earlier, SJ_i^T and SJ_i^F represent the sets of claims the participant S_i actually reports as true and false respectively in the conflicting observation matrix (i.e, SC). Let us also define $\bar{SJ}_i$ as the set of claims participant S_i does not report in the conflicting observation matrix. Thus, (9.15) and (9.16) can be rewritten as:

$$\begin{aligned}
&\sum_{j\in SJ_i^T} Z(t,j)\frac{1}{a_i^{T*}} - \sum_{j\in \bar{SJ}_i} Z(t,j)\frac{1}{1-a_i^{T*}-a_i^{F*}} = 0\\
&\sum_{j\in SJ_i^F} Z(t,j)\frac{1}{a_i^{F*}} - \sum_{j\in \bar{SJ}_i} Z(t,j)\frac{1}{1-a_i^{T*}-a_i^{F*}} = 0
\end{aligned}\tag{9.18}$$

$$\begin{aligned}
&\sum_{j\in SJ_i^T} (1-Z(t,j))\frac{1}{b_i^{T*}} - \sum_{j\in \bar{SJ}_i} (1-Z(t,j))\frac{1}{1-b_i^{T*}-b_i^{F*}} = 0\\
&\sum_{j\in SJ_i^F} (1-Z(t,j))\frac{1}{a_i^{F*}} - \sum_{j\in \bar{SJ}_i} (1-Z(t,j))\frac{1}{1-b_i^{T*}-b_i^{F*}} = 0
\end{aligned}\tag{9.19}$$

Solving the above equations, we can obtain the expressions of the optimal a_i^{T*}, a_i^{F*}, b_i^{T*}, b_i^{F*} and d^* as shown in (6.9) to (6.11).

The following derivation demonstrates the details to obtain the results in (6.18). The derivation that maximizes the $Q\left(\theta|\theta^{(t)}\right)$ in the M-step in Section 6.3 yields:

$$\begin{aligned}
&\sum_{j=1}^{N} Z_k(t,j)\left[\frac{S_iC_j^k}{a_{k,i}^{T\ *}} - \frac{(1-S_iC_j^k-S_iC_j^{\bar{k}})}{1-a_{k,i}^{T\ *}-a_{k,i}^{F\ *}}\right] = 0\\
&\sum_{j=1}^{N} Z_k(t,j)\left[\frac{S_iC_j^{\bar{k}}}{a_{k,i}^{F\ *}} - \frac{(1-S_iC_j^k-S_iC_j^{\bar{k}})}{1-a_{k,i}^{T\ *}-a_{k,i}^{F\ *}}\right] = 0 \quad k=1,2,..K
\end{aligned}\tag{9.20}$$

$$\sum_{j=1}^{N}\left[Z_k(t,j)\frac{1}{d_k^*} - Z_K(t,j))\frac{1}{1-\sum_{i=1}^{K-1} d_i^*}\right] = 0 \quad k=1,2,..K-1 \tag{9.21}$$

As we defined earlier, SJ_i^k and $SJ_i^{\bar{k}}$ represent the sets of claims the participant S_i actually reports as value k and value other than k respectively in the conflicting observation matrix (i.e, SC). Let us also define $\bar{SJ}_i$ as the set of claims participant S_i does not report in the conflicting observation matrix. Thus, (9.20) can be rewritten as:

$$\sum_{j \in SJ_i^k} Z_k(t,j) \frac{1}{{a_{k,i}^{T}}^*} - \sum_{j \in S\bar{J}_i} Z_k(t,j) \frac{1}{1 - a_i^{T*} - {a_{k,i}^{F}}^*} = 0$$
$$\sum_{j \in SJ_i^{\bar{k}}} Z_k(t,j) \frac{1}{{a_{k,i}^{F}}^*} - \sum_{j \in S\bar{J}_i} Z_k(t,j) \frac{1}{1 - {a_{k,i}^{T}}^* - {a_{k,i}^{F}}^*} = 0 \tag{9.22}$$

Solving the above equations, we can obtain the expressions of the optimal ${a_{k,i}^{T}}^*$, ${a_{k,i}^{F}}^*$ and d_k^* as shown in (6.18).