WIRELESS SOCIAL SENSOR NETWORKS

by

Kardelen Çepni

A Thesis Submitted to the
Graduate School of Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Electrical and Electronics Engineering

Koç University

August, 2015

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Kardelen Çepni

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Prof. Dr. Özgür B. Akan (Advisor)

_____

Prof. Dr. Fatih Alagöz

_____

Assoc. Prof. Öznur Özkasap

Date: _____

To my parents and my brother,
Handan Çepni, Nedim Çepni and Karayel Çepni.

ABSTRACT

Online Social Networks (OSNs) have recently become the essential means of communication, networking and entertainment. One of the prominent applications of OSNs has originated from their frequent use during major events, known as *social sensing*, which is the utilization of the information shared in OSNs to estimate an observed yet unknown phenomenon. In this thesis, we analytically investigate social sensing capabilities of OSNs. To this end, we introduce Wireless Sensor Network (WSN) paradigm, Wireless Social Sensor Network (WSSN) and explore the WSSNs within the most widely used OSNs, i.e., Twitter and Facebook. First, we develop communication theoretical models for the mechanisms of information propagation in Twitter and then analytically model the social sensing with Twitter. The accuracy of the estimated signal is investigated with mean square error analysis. Later, using a simple observation model by considering the features of Twitter, i.e., tweet and retweet, we extend the performance analysis (in terms of mean square error) of social sensing by comparing with fundamental estimators in classical and Bayesian estimation theory for various factors such as user behavior, number of people participating to social sensing and geotag use percentage. Lastly, we model and investigate the main social sensing mechanism in Facebook, i.e., Facebook Comment Thread Network (FCTN). By developing an analytical model for user observations in CTN, we analyze the reliability of social sensing with Facebook CTN for varying user behaviors and relationships, event characteristics, Facebook features and network size. The results indicate that, in addition to network conditions, the reliability of social sensing, i.e., the accuracy of the estimated signal, is affected by the features of the OSN, user behavior patterns and source event characteristics.

ÖZETÇE

Günümüzde sosyal ağlar temel haberleşme ve eğlence araçlarından biridir. Bunların yanı sıra, sosyal ağların diğer fonksiyonu ise majör olaylardaki sık kullanımıyla oluşmuştur. Bu fonksiyon *sosyal algılama*, sosyal ağlardaki kullanıcılar tarafından paylaşılan bilgilerin gözlemlenen bir olayın algılanması için kullanılması, yakın zamanda bir araştırma alanı olarak literatürde yer almıştır. Bu tezde, sosyal ağlardaki sosyal algılama mekanizmaları modellenerek sosyal ağların algılama kapasiteleri analitik olarak incelenmiştir. Bu amaçla, Kablosuz Sensör Ağ (KSA) yaklaşımı olan Kablosuz Sosyal Sensör Ağ (KSSA) önerilmiş ve yaygın olarak kullanılan sosyal ağlardaki (Twitter ve Facebook) KSSA'lar bu yaklaşım ile incelenmiştir. Öncelikle, Twitter için temel bir sosyal algılama modeli geliştirilip, bu mekanizmanın algılama doğruluğu ortalama karesel uzaklık analizi ile incelenmiştir. Devamında, Twitter'daki temel özellikleri içeren gözlem modeli referans olarak alınıp performans analizi tahmin teorisindeki klasik ve bayesci kestiriciler ile değişken faktörler (kullanıcı davranışı, katılımcı kullanıcı sayısı, lokasyon bildirimi kullanımı) için karşılaştırılmıştır. Son olarak, Facebook içindeki temel sosyal algılama mekanizması olan Facebook Gönderi Zinciri Ağı incelenmiştir. Kullanıcı gözlemleri analitik olarak modellenerek Facebook Gönderi Zinciri Ağı'nın algılama performansı değişen kullanıcı sayısı, Facebook özellikleri kullanımı, kullanıcı davranışı ve kullanıcı ilişkileri için incelenmiştir. Bu çalışma sosyal algılama alanındaki öncü çalışmalardan biri olup, çalışmanın genel sonuçları sosyal algılama performansının ağ kullanıcılarına, kullanıcı davranışlarına, sosyal ağ özelliklerine ve olay özelliklerine bağlı olarak değiştiğini göstermektedir.

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my advisor Dr. Özgür B. Akan for the continuous support of my MSc study and research, for his patience, enthusiasm, and immense knowledge. His guidance helped me during all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor, who has inspired me in countless ways.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Öznur Özkasap and Dr. Fatih Alagöz for all of their guidance through this process; their discussion, ideas, and feedback have been absolutely invaluable.

I thank my colleagues in Next-generation and Wireless Communications Laboratory (NWCL) for stimulating discussions and for their support in the last two years. In particular, my special thanks are due to Murat Kuşcu who has been more of a good friend than a colleague.

Besides NWCL, I gratefully acknowledge Koç University for supporting me during my undergraduate/graduate studies and for being my first home in Istanbul.

I should also acknowledge the profound love and support of my beloved friends, Gizem, İpek, Suay and Çiğdem, who have stood by me at my highest and at my lowest points during these two years.

Last but foremost, I am eternally grateful to my family. Without their unconditional support and love, I could not have made it this far.

TABLE OF CONTENTS

List of Tables **x**

List of Figures **xi**

Nomenclature **xii**

Chapter 1: Introduction **1**

1.1 Social Sensing with Wireless Social Sensor Networks 2

1.2 Scope of the Thesis 5

Chapter 2: Social Sensing Model and Analysis for Event Detection and Estimation with Twitter **6**

2.1 Introduction 6

2.2 Twitter as a Sensor Network 8

2.2.1 Tweet Propagation Model 8

2.2.2 Tweet Delivery Failures 13

2.3 Event Estimation in Twitter 13

2.4 Error Analysis 14

2.4.1 Spatially-correlated Users 16

2.4.2 α Spatially-correlated Users 17

Chapter 3: Estimation Theoretical Analysis for Social Sensing with Twitter **18**

3.1 Introduction 18

3.2 Observation Model 20

3.3 Estimation via Twitter 22

3.3.1 Classical Approach 23

3.3.2 Bayesian Approach . . . 25
3.4 Numerical Analysis . . . 27
3.4.1 Correlated Users . . . 28
3.4.2 α-spatially Correlated Users . . . 29

Chapter 4: Event Estimation Accuracy of Social Sensing with Facebook Comment Thread Network 33

4.1 Introduction . . . 33
4.2 Preliminaries on Facebook CTN . . . 36
4.2.1 User Actions in Facebook CTN. . . . 37
4.2.2 Model Assumptions . . . 38
4.3 Comment Thread Network Model . . . 38
4.3.1 User Observations . . . 39
4.3.2 Comment and Like Process . . . 41
4.3.3 Facebook CTN Estimator . . . 43
4.4 Numerical Analysis . . . 44
4.4.1 Estimation Accuracy Analysis for Varying Network Size and Post Characteristics . . . 45
4.4.2 Estimation Accuracy Analysis for Varying User Behaviors and Relationships . . . 46

Chapter 5: Conclusion and Future Research Directions 50

5.1 Contributions . . . 51
5.1.1 Social Sensing Model and Analysis for Event Detection and Estimation with Twitter . . . 51
5.1.2 Estimation Theoretical Analysis for Social Sensing with Twitter . . . 51
5.1.3 Event Estimation Accuracy of Social Sensing with Facebook Comment Thread Network . . . 52
5.2 Future Research Directions . . . 52
5.2.1 Source Signal . . . 53
5.2.2 Information Source . . . 54

5.2.3 Information Transmission 56
5.2.4 The OSN in Use 57
5.2.5 Event Estimation 58

Bibliography **60**

Vita **65**

LIST OF TABLES

4.1 Numerical Analysis Parameters 45

LIST OF FIGURES

1.1 Social sensing with OSNs. 3

2.1 Information flow model in Twitter. 9
2.2 MSE for varying (a) p_h, p_s and p_{tw} (b) n_d and p_{tw}, (c) N, (d) α. 15

3.1 MSE for varying p_{tw}. 29
3.2 MSE for varying N. 30
3.3 MSE for varying α. 31

4.1 Comment Thread Network Model. 37
4.2 Social Consensus Formation in CTN. 37
4.3 $D(\hat{\theta})$ for varying N and σ_θ. 46
4.4 $D(\hat{\theta})$ for varying percentage of text-based posts in comment thread. 47
4.5 $D(\hat{\theta})$ for varying like behaviors. 47
4.6 $D(\hat{\theta})$ for varying opinion weighting behaviors. 48
4.7 $D(\hat{\theta})$ for varying θ_F. 49

5.1 Information generation and propagation in the OSN. 53

NOMENCLATURE

CTN	Comment Thread Network
FCTN	Facebook Comment Thread Network
OSN	Online Social Network
WSN	Wireless Sensor Network
WSSN	Wireless Social Sensor Network

Chapter 1

INTRODUCTION

Online Social Networks (OSNs) have become the essential means of networking and entertainment. OSNs provide web-based services, in which users create their profiles and communicate with their friends in the network [5]. Many OSNs worldwide provide distinct functionalities and applications [19]. One of the prominent applications of OSNs has originated from their frequent use during major events, known as *social sensing*, which is the utilization of the information shared in OSNs to estimate an observed yet unknown phenomenon. This application has fostered the research on OSNs since social sensing has a big societal impact in many aspects of human lives. For instance, during social disasters the authorities may feel the need of investigating users' perceptions, i.e., learning out of collected observations, and they can design their emergency plan accordingly to maintain control. More examples can be given in other fields including online marketing, politics and financial markets. Along with the utilization of OSNs in sensing applications, understanding information delivery and reliability characteristics of OSNs in social sensing becomes an important problem. To the best of our knowledge, there exists no analytical study on the accuracy of event detection and estimation of event signal with OSNs using observations collected.

In this thesis, we investigate social sensing capabilities of OSNs by introducing tools from signal processing and communications theory. To this end, we introduce the WSN paradigm, Wireless Social Sensor Network (WSSN) and model information formation and propagation processes in the OSNs from a sensor networking perspective. We develop communication theoretical models for social sensing with Twitter and Facebook, and investigate the accuracy of the estimated signal with mean square error analysis. The results indicate that, in addition to network conditions, the accuracy of the estimation is affected by the features of the OSN, user behavior patterns and event characteristics.

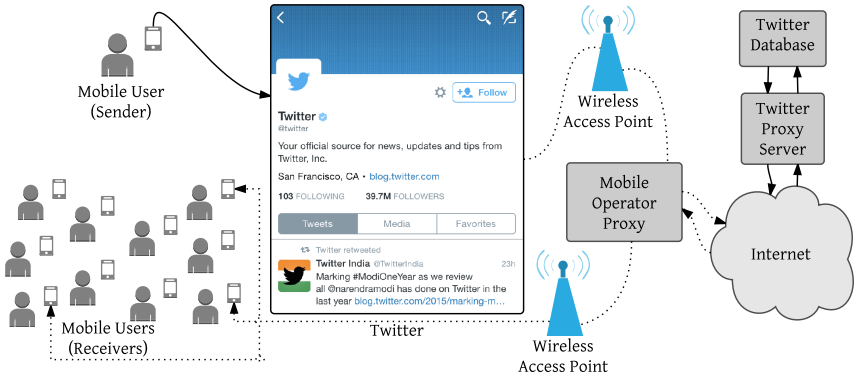
1.1 Social Sensing with Wireless Social Sensor Networks

Social sensing involves an estimation problem, i.e., given a collection of users' observations, an unknown parameter of interest is aimed to be estimated. In this section, we give an overview of the state-of-the-art in social sensing and introduce the WSN paradigm, Wireless Social Sensor Networks (WSSNs).

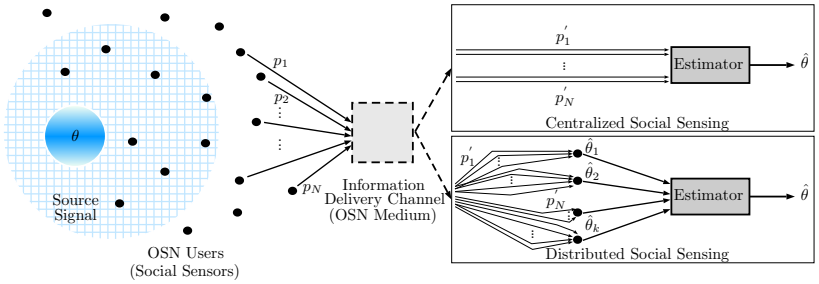
Mainly, there are two research directions in social sensing. First line considers social sensing as an application-specific estimation problem and aims to develop event estimation/detection algorithms using the data in OSNs. For example, Twitter messages are used for a real-time heavy traffic detection application [40], and an earthquake detection system is proposed based on the sentiment analysis of Twitter messages [41]. The approach is likely to be beneficial in developing social sensing applications in real life, however, it does not reflect the real nature of social sensing since the algorithms used in social sensing may modify or eliminate the existing data after analysis, e.g., sentiment analysis. If there are repetitive, inaccurate or irrelevant information in the data set, then the algorithms eliminate them to improve the estimation.

Second line introduces social sensing as an abstract concept by defining the users as *social sensors* with unreliable sensing capabilities. Several studies build an introductory framework for social sensing and introduce its future applications [17, 38, 7]. Moreover, social sensing problem is formulated from an estimation theoretical perspective, and binary event estimation with unreliable human observations is studied in [48], and the interactions between the users are investigated in [28] from a game theoretical perspective. We believe that this approach reflects the real nature of social sensing. However, there is still a lack of theoretical understanding of social sensing. To the best of our knowledge, none of these studies have explored the impact of users and also the information sharing processes in OSNs on the reliability of social sensing.

In social sensing, OSNs resemble to wireless sensor networks (WSNs), in which users exchange/share information with their contacts and shared information is used in estimation of a phenomenon. Features of OSNs define how the information is exchanged, and the mechanisms that facilitate the collection of the information. Moreover, the characteristics of the observation signals may change during information transmission in OSNs, and hence influencing the reliability of social sensing. Therefore, further studying underlying processes



(a) Information share/retrieval in Twitter.



(b) Centralized and distributed WSSNs.

Figure 1.1: Social sensing with OSNs.

in OSNs is imperative in characterization of the social sensing capabilities of OSNs. At this point, it may be advantageous to investigate social sensing with OSNs from a WSN perspective to explore communication theoretical aspects of OSNs and to fully reveal the factors in social sensing. In this thesis, we introduce the WSN paradigm, Wireless Social Sensor Network (WSSN), which can be in general defined as a *partition of an OSN, which consists of social sensor nodes that are capable of sensing an event with their sensory mechanisms and reporting their observation signals through the information dissemination mechanisms in the OSN wirelessly.*

WSSNs are formed with the communication schemes in OSNs. In the most basic form, the users may communicate directly with each other, i.e., one-to-one communications. Although this scheme does not exactly form a WSSN, hence a social sensing mechanism, it is intuitively useful for understanding the transmission of the information from one point to another in the OSN. Let us consider Twitter as an example.

Twitter is a directed OSN based on following/follower relationships, in which users simply send and read 140-character limited short messages called *tweets* [45]. A mobile user creates a message using Twitter application on its mobile device and sends it to its followers as in Fig. 1.1(a). The message is delivered to Twitter database passing through the proxy of the user's mobile operator and Twitter's server via Internet. Afterwards, some users, i.e., the followers of the sender, in the network receive the message as they are connected to the sender. Essentially, the message is delivered to the end users through the OSN application that operates in Internet.

When the users try to communicate their information to a single destination using the features of the OSN, i.e., many-to-one communications, this communication scheme models a *centralized WSSN*. In this scheme, the users forward their observations to a central receiver in OSNs as shown in Fig. 1.1(b). Here, there are OSN users and there is a source event θ. The users sense the source signal, each user's experience may be different in the sense that the source signal may have an influence range, represented with the hatched area in the model. Furthermore, the users may observe different forms of the source signal. After making their observations, the users encode their observations $(\hat{\theta}_1, ..., \hat{\theta}_N)$ as messages $(p_1, ...p_N)$ and send them to a common receiver using the OSN, which resembles to a sink, i.e., fusion center in WSNs, in which users' observations are collected.

Moreover, we may consider the communication scheme in which the information is transmitted from one set of users to another set of users, i.e., many-to-many communications. This scheme is useful in modeling of the *distributed WSSN*, in which local estimate signals are used in the estimation of the source signal. Distributed WSSN is shown in the right bottom part of Fig. 1.1(b). Similar to the centralized WSSN, we assume that there are N users in the OSN, they sense a physical phenomenon θ, and they share their messages, e.g., $p_1, ...p_N$, with their friends in the OSN. In distributed setting, k users in the OSN may act as local estimators by making their own estimation on the source signal, i.e., $\hat{\theta}_1, ..., \hat{\theta}_N$, based on the messages they receive. Local estimate signals $\hat{\theta}_1, ..., \hat{\theta}_N$ are utilized in estimation of the source signal. This application of social sensing may be used in the scenarios, in which the complete collection of user observations is not available to the estimator.

1.2 Scope of the Thesis

As we describe in the previous section, social sensing can be studied in different communication schemes, yet in this thesis, we focus on the centralized WSSNs in two most widely used OSNs, e.g., Twitter and Facebook. The thesis is organized as follows.

In Chapter 2, we study social sensing with Twitter from a communication theoretical perspective [8]. By modeling Twitter as a sensor network and information transmission as an on-off channel with constant delay, we investigate the effect of the features of Twitter and channel characteristics on the reliability of social sensing. Our results indicate that different user behaviors (tweet or retweet) and network characteristics, e.g., delay levels and on-off characteristics, have an effect on the performance of social sensing.

In Chapter 3, we extend our analysis for Twitter from an estimation theoretical perspective. Using a simple observation model by considering the features of Twitter, i.e., tweet and retweet, the performance (in terms of mean square error) of social sensing, defined as social learning, i.e., social consensus, is analyzed and compared with the fundamental estimators in classical and Bayesian estimation theory for various factors such as user behavior, number of people participating to social sensing and geotag use percentage. The results indicate that when phenomenon is a random process, the estimator, defined as social consensus, is not sufficient to make an accurate estimation and attain the fundamental error bounds. The error level of social sensing and the fundamental bounds depend on the user behavior. Hence, it is crucial to investigate the behavior patterns of users in Twitter to further understand the social sensing capabilities of Twitter.

In Chapter 4, we develop social sensing model for Facebook Comment Thread Network (FCTN) and investigate the reliability characteristics of FCTN with a similar approach to Chapter 2. Furthermore, we include the effect of various post types available in Facebook into FCTN model. The results indicate that while the characteristics of the event signal affect the accuracy of the social sensing, also the polarized weighting of the observations and use of less reliable post types in FCTN during social sensing are important in the sense that they deteriorate the accuracy of the estimate signal, i.e., social consensus. Furthermore, user relationships inside the community may alter the reliability of social sensing implying that the selection of OSN users in social sensing is essential in obtaining an accurate estimation. Chapter 5 concludes the thesis by outlining the contributions and future directions.

Chapter 2

SOCIAL SENSING MODEL AND ANALYSIS FOR EVENT DETECTION AND ESTIMATION WITH TWITTER

2.1 Introduction

Beside their networking aspects, Online Social Networking services recently came into use as a platform for collecting data from people. People share information with their contacts through mediums such as Facebook, Twitter, Flickr, LinkedIn, Instagram and other similar services. Active use of OSN and resulting accumulated data leads to a recently new field called *participatory sensing* or *social sensing*, which is based on the utilization of the information shared in OSNs for sensing a physical phenomenon [3]. Main attributes of participatory sensing are the type of observed phenomenon, type of information shared in OSN and the features of OSN. Then, aggregated data in OSN can be used to estimate the phenomenon considering those aspects.

Twitter is one of most favorable OSN for participatory sensing, in which users simply share short text messages with their followers. Two main uses of Twitter are broadly defined [23]. First, people use it to share instances about their lives. Second, both individual users and news companies use it for disseminating news or observations about an event. The latter use of Twitter makes it an efficient platform for information sharing and spreading news online during major events [23]. Twitter use in mass protests in Iran, Tunisia and Egypt during 2011 Arab Spring has been investigated in [24]. It is found that people used Twitter to instantly learn the latest updates about the mass protests, and act accordingly. Furthermore, they used it to share their observations with others during the protests.

With its popularity and significant role in participatory sensing, research on Twitter exists in some directions. User behavior statistics and network characteristics are investigated in [23] and [29], respectively. Several event detection systems are proposed in [31], [40], [41] and [44], where event detection is mainly realized via sentiment analysis of tweets along with their location and time information. Although they are proposed for event detection, to the

best of our knowledge, their detection and estimation accuracy have not been investigated based on a theoretical communication network model of Twitter.

In fact, Twitter may be considered as a Wireless Sensor Network (WSN), where users are sensors and they share observations about the event, i.e., samples of the event signal, to be observed. Observations including same hashtag are collected in hashtag timeline in Twitter. Here, timelines refer to webpages where various tweets are collected, i.e., tweets with the same hashtag are collected in hashtag timeline or a user reads the tweets created by its followees in its home timeline. Hashtag timeline is similar to the sink node in sensor networks since both are many-to-one receivers, where observations are collected.

Beside similarities, there are also some differences between Twitter and WSNs. In general, sensors nodes in WSNs interact with each other only when relaying information from source nodes to the sink nodes. However, the sensors interact by communicating with each other in Twitter. Namely, a user communicates with other users by following and receiving their tweets. Furthermore, while sensors send their own observations to the sink in WSNs, users do not have to send their own observations in Twitter. They may simply relay other users' observations by retweeting their tweets, which is an important feature of Twitter.

The main objective of this chapter is to investigate the accuracy of event detection and estimation based on the utilizing the tweets collected in Twitter. The estimate is prone to errors since there may be fake users in the network or users' biased observations may hamper the accuracy of the information shared. In addition, network structure of Twitter, user behavior patterns and network conditions may change the accuracy of information. For the first time in the literature, we develop an analytical model for tweet propagation and information fusion in Twitter by incorporating all these factors, and we investigate the accuracy of the estimated signal.

The remainder of the chapter is organized as follows. In Section 2.2, we develop an analytical model for tweet propagation in Twitter. In Section 2.3, we define the estimate signal and derive distortion function. In Section 2.4, error analysis is performed and the results are presented.

2.2 Twitter as a Sensor Network

Twitter is a social network in which users simply send and read short messages called tweets. Interactions are directed in Twitter based on following/follower relationships. Hashtag use during major events turns Twitter into a natural sensor network by enabling Twitter to collect observations in a single pool. The function of Twitter as a sensor network depends on user preferences to include hashtag in their tweets or not. There may be users sharing their observations without a hashtag. However, in [32], it is shown that majority of Twitter users add hashtag during major events in order to increase the visibility of their tweets.

During major events, the observations collected in hashtag timeline in Twitter are widely used by users. Unlike the event detection systems, observations are available to all users. A user, who does not directly observe the event, may make an estimate about the event reading through the tweets in the hashtag timeline, and take decisions based on its estimate. Hence, consensus formed in Twitter may govern people's actions during critical situations. Therefore, it is vital to understand the accuracy of estimated signal in Twitter.

While modeling, we use a discrete time framework and make the following assumptions. Although there are various features of Twitter such as reply, mention and favorite, in the model, we assume that users may only tweet, retweet and use geotag/hashtag in their tweets. Users may take a single action during each time interval. For this study, we approximate the average time a user spends in Twitter including reading tweets in its home timeline and/or creating a tweet as a minute assuming that a dynamic event is occurring, which needs to be observed in minute intervals. For simplicity, we assume that tweets include text messages and users do not have a memory. Namely, their current observations are not affected by the previous observations. In the model, all accounts are taken to be public. Namely, user may follow people without requiring an approval. Lastly, we assume that if a user inserts a hashtag, then it tweets about the event related to that hashtag.

2.2.1 Tweet Propagation Model

We investigate the information propagation in Twitter by modeling the path of a tweet after it is created until it reaches follower timelines and hashtag timeline. Suppose there exists a social event or a disaster in the region R. The source of the event signal to be observed is located at (x_0, y_0), and i^{th} user, i.e., U_i resides at the location (x_i, y_i). Every $U_i \in E$

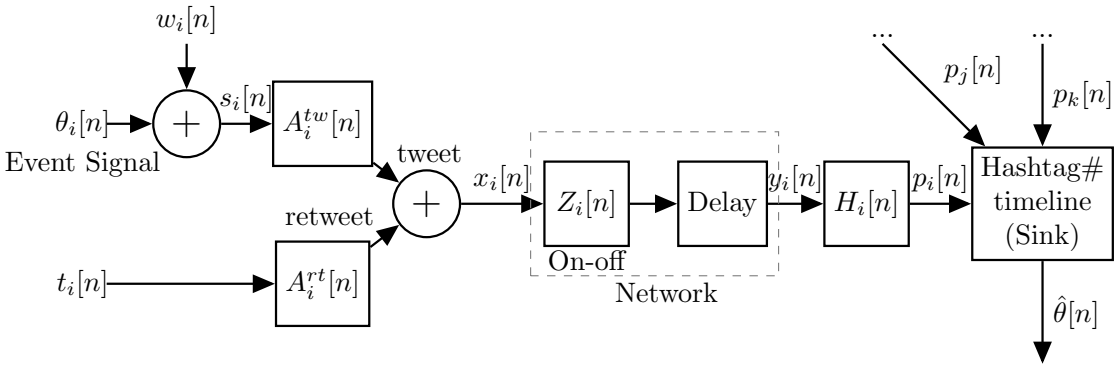


Figure 2.1: Information flow model in Twitter.

for $i = 1...N$, where E is the set of Twitter users observing the event or interested in the observations. Note that E does not include all Twitter users but only the ones participating in social sensing. Furthermore, users do not need to be around the event region R. They may be following the event in a far location from R. The set of people followed by U_i is named as *followees of* U_i and defined as

$$F_i = \{U_x \in E, U_x \neq U_i \mid U_i \rightarrow U_x\}, \tag{2.1}$$

where $\rightarrow$ means *follows* and $|F_i|$ is the number of followees of U_i at time n. F_i may change with time as U_i follows and unfollows users. Thus, the observation of U_i at time n is denoted as

$$s_i[n] = \theta_i[n] + w_i[n], \tag{2.2}$$

where $\theta_i[n]$ is event signal at U_i's location (x_i, y_i), and $w_i[n]$ is the observation noise. The noise can be considered as user's bias on the source event or it may be due to sensual impairments, i.e., visual, hearing impairments. The noise prevents U_i from observing the pure event signal $\theta_i[n]$.

If U_i wants to share the observation $s_i[n]$ with its followers, it has several options. $x_i[n]$ is signal sent to the Twitter network after U_i's action, i.e., tweet or retweet, as shown in Fig. 2.1. User's possible actions are discussed in the following part.

Users and their actions

When a user logs in to Twitter, it has several options. First, it may not take any action, i.e., log off after checking its account. In this case, sensor is considered as inactive. Second, it can take only the following actions based on our assumptions:

Tweet: U_i only tweets its observation, and the message $s_i[n]$ may appear in the timelines of its followers depending on the network availability.

Retweet: U_i prefers to share one of the messages in its home timeline, where the messages from U_i's followees are stored. Note that the observations do not have to be created by its followees. The reason is that the messages in U_i's home timeline may also include retweets. A message may be retweeted several times before it reaches to U_i's home timeline, but it appears with its original author name and original creation time in U_i's home timeline.

Regarding retweets, we make the following assumptions. At time n, a user reads all messages in its timeline that are created at time $n-1$, and retweets one of these messages. Here, all messages are equally likely to be selected by the user. We also assume that user retweets messages created up to time $n-a$. For example, if a=3, then user may retweet observations created within 3 minutes from the current time n. This assumption enables us to incorporate the possibility of a user retweeting a previously retweeted message into our model. Retweet signal depends on the previous observations created by users. We assume that history of user actions are not explicitly known for simplicity. If U_i decides to retweet, then it selects one of the possible messages that appear in its home timeline at time n. We define the set S_i of all possible messages that may appear in U_i's home timeline at n as

$$\begin{aligned} S_i =& P_1 \cup P_2 ... \cup P_a \\ =& \{s_k[n-1] \mid k \in F_i\} \cup \{s_l[n-2] \mid l \in F_k,\, k \in F_i\} \\ & ... \cup \{s_t[n-a] \mid t \in F_l, ..., k \in F_i\} \text{ and } |S_i| = X_i, \end{aligned} \tag{2.3}$$

where P_a is the set of all tweets created by the users in F_i at time $n-a$ appearing on the home timeline of user U_i at time n. S_i consists of subsets P_1,P_2,...,P_a, which include all possible messages created at various time intervals, which may appear in U_i's home timeline at time n. In order observation $s_k[n-2]$ to appear in U_i's home timeline at time n, $s_k[n-2]$ must be retweeted by one of the followees of U_i at time $n-1$. If U_i selects a message from P_1, then it means that U_i retweets one of its followees' observations. Furthermore, if

U_i selects a message from either P_2 or P_3, it means that U_i retweets a message, which is previously retweeted by its followee.

For the sake of clarity, we order the elements of S_i in ascending order with respect to user indices and denote the ordered set

$$M_i = \{m_1, m_2, ..., m_{X_i}\} \tag{2.4}$$

where each m_k corresponds to the ordered element in S_i. If user indices are the same, then the elements are ordered with respect to tweet time indices, i.e., $n-1$, $n-2$. Note that the same signal may appear twice since some users might be following the same users. They are counted as distinct elements of the set.

Since a user equally likely retweets one of the messages in M_i, U_i's retweet signal $t_i[n]$ is obtained as

$$t_i[n] = m_k,\ m_k \in M_i. \tag{2.5}$$

Briefly, at time n user sends either its observation or others' observation to Twitter. We incorporate a probabilistic model for user behavior. We define three binary random variables $A_i^{tw}[n]$ (tweet), $A_i^{rt}[n]$(retweet) and $A_i^{na}[n]$ (no action), which are jointly distributed. $(A_i^{tw}[n], A_i^{rt}[n], A_i^{na}[n]) = (1, 0, 0)$ means that U_i just tweets, and then logs off. Here, binary 1 represents that user takes the corresponding action. The probability density function of joint random variables is

$$p_{A_i^{tw}[n]A_i^{rt}[n]A_i^{na}[n]}(x, y, z) = \begin{cases} 1,0,0 & \text{, with } p = p_i^{tw}[n] \\ 0,1,0 & \text{, with } p = p_i^{rt}[n] \\ 0,0,1 & \text{, with } p = p_i^{na}[n] \\ \text{otherwise} & , p = 0 \end{cases}$$

From the probability density function, it may be seen that U_i can only take a single action at each interval, and $p_i^{tw}[n]+p_i^{rt}[n]+p_i^{na}[n] = 1$ since there are no other possible actions for U_i. At time n, user's signal to Twitter Network is written as

$$x_i[n] = A_i^{tw}[n]s_i[n]+A_i^{rt}[n]t_i[n]+A_i^{na}[n]0. \tag{2.6}$$

For simplicity, we assume that user behavior characteristics stay constant during the period of observation transmission and event detection. Thus, (2.6) reduces to

$$x_i[n] = A_i^{tw}s_i[n]+A_i^{rt}t_i[n]. \tag{2.7}$$

Additional Features of Tweets

Beside the content of the tweet, a user may utilize functionalities of Twitter to increase the impact of its tweet in the network, i.e., by using hashtag and geotag. As discussed in Section 2.1, hashtag use enables collecting tweets under the same category. Geotag is another feature of Twitter, which lets user to share its location information while tweeting. This is an optional feature. User may or may not prefer to share its location with its followers due to privacy concerns. If user locations are known, a spatio-temporal model can be used to explore relationship between the observations and event signal. If user locations are not known, then we assume that user observations are uncorrelated to event signal.

Event signal is modeled as joint Gaussian random variable at each user location

$$E\{\theta_i[n]\} = 0,\ var\{\theta_i[n]\} = \sigma_\theta^2,\ i = 0, 1, 2, 3, ...N$$

where σ_θ^2 is the variance and N is the number of the users in the network. Here, $\theta_0[n] = \theta[n]$ is the event signal.

Assuming that the set of users U_i, $i = 1..K$ used geotag

$$cov\{\theta_i[n_i], \theta_j[n_j]\} = \sigma_\theta^2 corr\{\theta_i[n_i], \theta_j[n_j]\}$$

$$E[\theta_i[n_i]\theta_j[n_j]] = \sigma_\theta^2 K_v(d_{i,j}, n_i - n_j)$$

for $i = 0..K$ where n, n_i, n_j are time indices, respectively. K_v is the spatio-temporal correlation function, which is a function of the distance between the sensors

$$d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2},$$

and the time difference between observations $n_i - n_j$. The form of K_v depends on the type of the physical phenomenon. We model the event signal as a field source, in which source signal is distributed over a region. Using the corresponding correlation function for field sources given in [46], $K_V(d_{i,j}, \delta_{i,j}) = e^{-d_{i,j}/c_s} e^{-|\delta_{i,j}|/c_t}$, where $\delta_{i,j} = (n_i - n_j)/(f)$, f is observation frequency, c_s and c_t are spatial and temporal correlation coefficients. Different correlation functions may be used in the model depending on the type of observed events.

We assume that observations without geotag are uncorrelated with other observations and the event signal. Assuming that the first K users $(U_1, U_2, ..., U_K)$ use geotag, then for

$i = 0..N$ and $k = K{+}1, K{+}2, ..., N$

$$E\Big[\theta_i[n_i]\theta_k[n_k]\Big] = \begin{cases} 0 & , i \neq k \\ \sigma_\theta^2 & , i = k \wedge n_i = n_k \end{cases}$$

2.2.2 Tweet Delivery Failures

Tweets may not be delivered to users (followers) and the hashtag timeline because of network failures. Low Internet speed may incur delay in the delivery of tweets. Furthermore, tweet of a user may be blocked by the following factors: (i) most Twitter users are mobile during major events [24], and if the battery of a mobile device runs out, the user may not send its observation; (ii) due to imperfections in the Internet network, the connection may be lost suddenly. These factors can be modeled as an on/off channel. U_i's message $x_i[n]$ may be blocked or delayed due to imperfections in the network as shown in Fig. 2.1. Thus, the resulting signal of U_i at time n is

$$y_i[n] = Z_i[n]x_i\Big[n{-}n_d[n]\Big], \tag{2.8}$$

where $n_d[n]$ is the amount of delay such that $n_d[n] \in Z$. Here, $Z_i[n]$ is an i.i.d. Bernoulli process, which either allows message to pass through the network or blocks it such that $Z_i[n] = 1$ with probability $p_i^s[n]$ and $Z_i[n] = 0$ with probability $1{-}p_i^s[n]$.

2.3 Event Estimation in Twitter

An estimate about the event signal at time n can be made by reading the tweets accumulated in hashtag timeline. For a tweet to appear in hashtag timeline, a hashtag should be added. As shown in Fig. 2.1, the output signal to sink becomes

$$p_i[n] = H_i[n]y_i[n] \tag{2.9}$$

where $H_i[n]$ is a Bernoulli random process, which corresponds to U_i's action of adding the event related hashtag to its tweet at n, and $H_i[n] = 1$ with probability $p_i^{ht}[n]$ and $H_i[n] = 0$ with probability $1{-}p_i^{ht}[n]$.

The estimate $\hat{\theta}[n]$ is formed by individuals who read the tweets in hashtag timeline and learn from these tweets about the event signal. Various learning models for social networks exist in literature [1]. We use De Groot's learning model, in which users utilize a weighted

$$mse(\hat{\theta})=\sigma_\theta^2+\sum_{i=1}^{N}\frac{p_i^{ht}p_i^s}{N^2}\left[(p_i^{tw}+p_i^{rt})(\sigma_\theta^2+\sigma_w^2)-2N\left(p_i^{tw}\sigma_\theta^2(e^{\frac{-d_{i,0}}{c_s}}e^{\frac{-|n_d|}{c_t}})+p_i^{rt}\sigma_\theta^2(e^{\frac{-d_{m,0}}{c_s}}e^{\frac{-|n_d+n_m|}{c_t}})\right)\right]$$
$$+\sum_{\substack{i,j=1\\ i\neq j}}^{N}\frac{p_i^{ht}p_j^{ht}p_i^sp_j^s}{N^2}\left[p_i^{tw}p_j^{tw}\left(\sigma_\theta^2(e^{\frac{-d_{i,j}}{c_s}})+\sigma_w^2\delta(i-j)\right)+p_i^{rt}p_j^{rt}\left(\sigma_\theta^2(e^{\frac{-d_{m,u}}{c_s}}e^{\frac{-|n_m-n_u|}{c_t}})\right.\right.$$
$$\left.\left.+\sigma_w^2\delta(m-u)\delta(n_m-n_u)\right)+p_i^{tw}p_j^{rt}\sigma_\theta^2(e^{\frac{-d_{i,u}}{c_s}}e^{\frac{-|n_u|}{c_t}})+p_i^{rt}p_j^{tw}\sigma_\theta^2(e^{\frac{-d_{j,m}}{c_s}}e^{\frac{-|n_m|}{c_t}})\right] \quad (2.11)$$

average. The weights may be described as the trust factor of corresponding users. Namely, if a user cares more about its contacts' observations than strangers' observations, then the user does not pay attention to them. Broadly, in this work, we assume that users assign equal weights on the observations of the other users, and thus

$$\hat{\theta}[n]=\frac{1}{N}\sum_{n=1}^{N}p_i[n]. \quad (2.10)$$

In order to measure the accuracy of the event signal estimation based on tweets, we define mean square error (MSE) of the estimate signal as $D[n]=E[(\hat{\theta}-\theta_0)^2]$ and derive it. In the derivation, user observation noise is taken as i.i.d. zero-mean Gaussian noise with variance σ_n^2, and it is assumed all $(A_i^{tw}[n], A_i^{rt}[n], A_i^{na}[n])$, $Z_i[n]$, $H_i[n]$ are i.i.d. uncorrelated processes for every U_i. User behavior pattern, network characteristics and delay stay constant during period of observation transmission and event detection, therefore time index n is dropped in expressions, and (2.11) is obtained for $D[n]$, that is, MSE of the estimation of event signal at time n. Here, $\delta(.)$ is Kronecker delta function and two sets of variables $\{m, n_m\}$, $\{u, n_u\}$ appear in (2.11). These sets of variables originate from $t_i[n]$ and $t_j[n]$, which are defined in Section 2.2.1, respectively. Considering $t_i[n]$, i.e., retweet signal of U_i at n, m is the user index of the selected message and n_m determines the creation time of the message such that $t_i[n]=s_m[n-n_m]$. Similar explanation holds for the set of variables $\{u, n_u\}$ for $t_j[n]$.

2.4 Error Analysis

In analysis, we calculate MSE in the estimation of the event signal (2.11) in MATLAB for various cases. During evaluations, an iterative approach is used since MSE depends on both directed network structure, i.e., who follows whom, and the locations of the users. Most of the social networks including Twitter [29] have the characteristics of scale free random

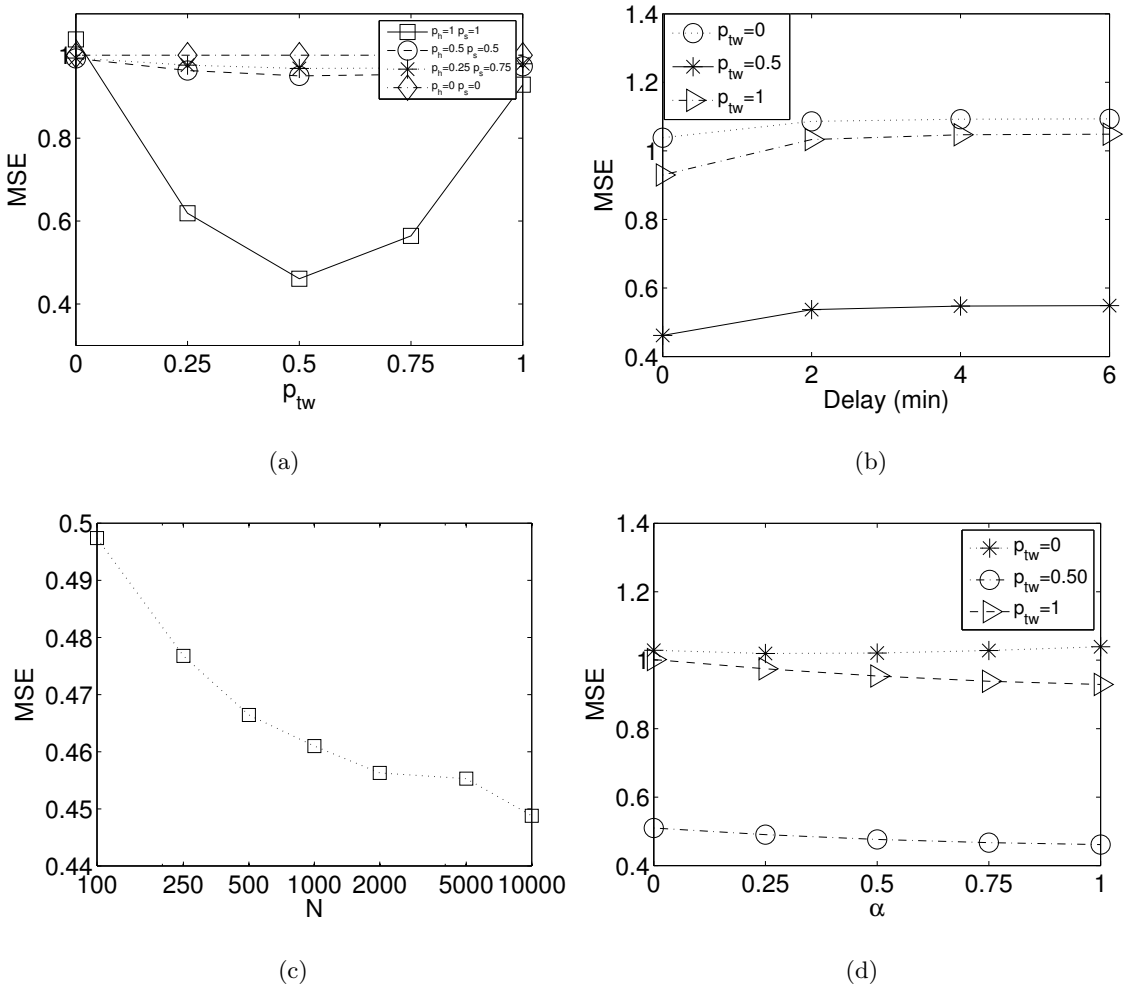


Figure 2.2: MSE for varying (a) p_h, p_s and p_{tw} (b) n_d and p_{tw}, (c) N, (d) α.

graphs. From the perspective of Twitter, these graphs imply that while a large number of users have less number of followers, a small number of users in the network have high number of followers. We generate random scale free graphs using [34] and randomly locate users in 100 x 100 m^2 square grid area. Source signal is assumed to be located at the center of grid area. For each analysis, we iterate MSE over for 1000 different network and deployment configurations, and then take average in order to obtain the final MSE. The variables related to retweet signal u, n_u, m and n_m are chosen randomly out of possible message sets X_i, X_j defined by network graph and a=2 min. In the correlation model, we set $f = 1$ observation/minute, $c_s = 10$ and $c_t = 10$. Note that MSE is unitless in the results as the event signal to be observed is not physically defined in the model. In order to observe

the effect of behavior patterns explicitly, we set p_i^{na}=0 for all users since the estimated signal is formed by the observations of active users. During analysis, unless opposite is stated we assume that (i) default network size is 1000 users ($N = 1000$), (ii) all users behave the same and (iii) network conditions are the same for all users, these conditions imply that $p_i^{ht}[n]$=p_h, $p_i^{s}[n]$=p_s, $p_i^{tw}[n]$=p_{tw}, and $p_i^{rt}[n]$=p_{rt} for all $U_i \in E$. Default values for p_h, p_s and n_d are taken as p_h=1, p_s=1 and n_d=0 for all users.

2.4.1 Spatially-correlated Users

In the first analysis, we assume that all users insert geotag in their tweets and investigate MSE for varying tweeting probabilities (p_{tw}), hashtag use (p_h) and network blockage probabilities (p_s).

In Fig. 2.2(a), we observe that unless both p_h and p_s are high, MSE stays close to 1 since path of tweet to hashtag timeline may also be blocked if hashtag is not used. Furthermore, MSE does not fall below 0.5 all the time, which is likely due to noisy user observations. When p_h=1 and p_s=1, MSE decreases as p_{tw} increases, reaches its minimum by decreasing 50% when p_{tw}=0.5, and then increases with p_{tw}. Note that when p_{tw}=1 MSE is better (by 9%) than the case when p_{tw}=0. Thus, we deduce that only retweeting does not improve MSE since randomly retweeted observations may include many repeated tweets, but distinct observations are required for a better estimate of the event signal. However, the number of distinct observations are likely to be bounded since the case p_{tw}=0.5 outperforms the case p_{tw}=1, which shows that all distinct user observations are not required for improving accuracy, repeated observations may also improve MSE. In Fig. 2.2(b), MSE is calculated for varying tweeting probabilities (p_{tw}) and delays (n_d). It is observed that when p_{tw}=0.5, MSE increases by approximately 20% with increasing delay, but still remains below the values when p_{tw}=1 and p_{tw}=0 for n_d=0. In the following analysis, we investigate the effect of number of the active users on MSE for $p_{tw} = 0.5$. MSE improves approximately by 10% as N goes from 100 to 10000 as shown in Fig. 2.2(c). Increasing the number of active sensors ultimately improves MSE, however, the number of users participating to social sensing may change depending on the type of physical phenomenon to be observed.

2.4.2 α Spatially-correlated Users

In the last analysis, we investigate the effect of geotag use on MSE. Being α spatially-correlated means that the first αN of the users, i.e., $U_i \in E$ for i=1..αN, use geotag in their tweets. In Fig. 2.2(d), MSE is calculated for varying tweeting probabilities and α values. It is observed that as α increases as MSE decreases. As users use geotag, the accuracy of information aggregated at hashtag timeline improves approximately 7-10% for $p_{tw} = 0.5$ and $p_{tw} = 1$ since correlated observations reduce the effect of noise in MSE expression (2.11). When all users retweet, the result indicates that geotag use does not improve MSE.

Chapter 3

ESTIMATION THEORETICAL ANALYSIS FOR SOCIAL SENSING WITH TWITTER

3.1 Introduction

While Twitter is a recently popular Online Social Networking (OSN) Service for social interactions, it is also a platform for collecting data from people. *Participatory* or *Social Sensing*, simply sensing any phenomenon based on data collected from people in Online Social Networks, emerged as a new research field in the past several years, as a consequence of increased use of OSN Services in major social events [3].

Twitter is one of the most favorable Online Social Networks for participatory sensing, in which users instantly share short text messages with their followers. Beside the individual interactions in Twitter, its use during major events shows that Twitter acts like sensing system, e.g., similar to a wireless sensor network (WSN). In [24] and [27], Twitter use in mass protests in Iran, Tunisia and Egypt during 2011 Arab Spring and 2011 Thai Flood is analyzed respectively and the studies have found that people indeed use Twitter to learn recent updates and share information with others. In addition, several features of tweets in Twitter such as hashtag and geotag enable users to canalize the information into the collection, in which tweets on same topic are collected, and add location information, respectively. These features bring additional sensor characteristics to Twitter users.

Sensing a phenomenon through Twitter may be considered in mainly two lines. First, all tweets published in a time period (sensor data by users) may be investigated, eliminated according to the phenomenon, and then, sensing may be performed based on the remaining tweets. Second, given a collection of tweets, social sensing can be performed utilizing the tweets within the collection assuming that all the tweets are related to the phenomenon, i.e., the sensors are providing information about the phenomenon to be observed. While investigating and searching for the related tweets, the algorithms used in the first approach are more complex than the second one since they require detailed content analysis, e.g.,

sentiment analysis. However, the content analysis may not be a feasible option during the events that require instantaneous social sensing. In this case, we may use the second approach. Fortunately, there is a feature in Twitter called *hashtag*, which is a special keyword used in tweets that enables collecting the tweets with the same *hashtag* word under a single pool, called hashtag timeline. Hence, hashtag provides the natural collection of the tweets related to the phenomenon, similar to a sink in WSNs.

For this approach, we may consider the following example. A Twitter user in NYC who wants to decide to go out or not, and it needs to learn about weather status before making its decision by reading through tweets collected in hashtag timeline with hashtag keyword $\#WeatherInNYC$. Namely, by reading through all related tweets, the user may make a guess about the weather. This social sensing case may be extended to more major events such as a social disaster, where information sensed by user may influence its actions. Therefore, in order to maintain control during a social disaster, the authorities may feel the need of investigating users' perception, i.e., learning, out of collective observations, and design their emergency plan accordingly. The collection set may include tweets, which are either not related to the phenomenon or misleading intentionally (i.e., liar users), nevertheless they are included in the collection since the hashtag attached to them. Overall, the latter model is more complex in sensing a phenomenon, and we believe that it is worthwhile to investigate the reliability of social sensing in Twitter through this aspect since the use of Twitter as a sensing tool is likely to increase during social events, and the gained insights may influence the future social network design for social sensing applications.

The existing studies on Twitter mostly focus on event detection/estimation, and the algorithms mainly depend on machine learning techniques and semantic analysis of tweets [22], [31], [40]. In [41], an earthquake detection system is proposed. Similar to the other studies, by analyzing the content of the tweet, each tweet is converted into a binary sensor value. Then, Kalman and particle filters are used to estimate the location and trajectory of earthquake out of sensory data along with spatial data. In [41], estimation theoretical tools are used on extracted tweets after sentiment analysis. These studies use exact user observations by performing content analysis on tweets and use estimation theoretical tools to design event estimation systems, however, they do not analyze the reliability of social sensing from an estimation theoretical perspective. Different from classical estimation problems, social

sensing includes observations whose existence are not certain, i.e., users may not be sharing their observations all the time or in the same manner [49]. Social sensing has been studied from an estimation theoretical aspect in [49]. Humans are treated as sensors, and a sensing problem is analytically defined. A maximum-likelihood (ML) approach is used to propose a solution for binary sensing problem. However, unlike [49], in our previous study [8], we develop a communication theoretical model of information propagation and social sensing in Twitter, which enables modeling of user observations in Twitter for various user actions and network conditions, and we analyze the accuracy of social sensing mainly based on user behaviors and network characteristics assuming that the estimation is made by collective observations of the users, i.e., social consensus. In this study, we consider continuous sensing problem and investigate the estimation in Twitter through classical estimation and Bayesian estimation. We compare the reliability performance of social sensing with the performances of the fundamental estimators in estimation theory in order to evaluate its performance for different observation models. To that end, by utilizing the information propagation model proposed in [8], we analytically derive the error functions for social sensing (with social consensus) and the error bounds for fundamental estimators in classical and Bayesian settings, then compare the performance of social sensing with these fundamental bounds for varying user behavior and number of available users in Twitter.

The remainder of the chapter is organized as follows. In Section 3.2, we describe the information propagation model in Twitter, and define an observation model. We analytically derive mean square error functions of classical and Bayesian estimators in Section 3.3. In Section 3.4, numerical analysis is performed and the results are presented.

3.2 Observation Model

In this section, we briefly describe the information propagation model presented in [8] to develop observation model for social sensing. Consider a simple scenario as follows. Using a discrete time setting, suppose there exists a major phenomenon θ in the region R at time n. The source of the phenomenon to be observed ($\theta = \theta_0$) is located at (x_0, y_0) and i^{th} user, i.e., U_i resides at the location (x_i, y_i), $\forall\ U_i \in E$ for $i = 1...N$, where E is the set of Twitter users observing the event or interested in the observations. Note that E does not include all Twitter users but only the ones participating in sensing by using the relevant hashtag.

Furthermore, users do not need to be physically around the event region R, they may be following the event in a far location from R.

A wireless sensor network perspective [47] may be used to model the information propagated in Twitter and accumulated in hashtag timeline. Similarly, a user's observation $s_i[n]$ is modeled as

$$s_i[n] = \theta_i[n] + w_i[n] \tag{3.1}$$

where $\theta_i[n]$ is event signal at U_i's location (x_i, y_i) at time n, and $w_i[n]$ is the observation noise. The noise can be considered as user's bias on the source event or it may be due to sensual impairments in users, i.e., visual, hearing impairments. The noise prevents U_i from receiving the actual observation signal $\theta_i[n]$. Instead of sharing its own observation, the user may prefer to share another observation, e.g., U_k's observation, by retweeting its message. Denoting this observation as $t_i[n]$, i.e., retweet signal of U_i,

$$t_i[n] = \theta_k[n-a] + w_k[n-a]. \tag{3.2}$$

In (3.2), $t_i[n]$ may be interpreted as follows. At time n, U_i prefers to tweet U_k's observation shared at time instance $n-a$, i.e., $s_k[n-a]$, where a is a constant defining the time difference between current observation signal and retweet signal. Note that U_i must be following U_k in order to be able to see its tweet in Twitter. Another possible case is that U_i may be following an intermediate user U_m who follows U_k so that U_i retweets a message retweeted by U_m before. Selection of U_k and the tweet may depend on the factors such as the content of the tweet and friendship relations. In this work, we simply assume that U_i retweets a message equally likely among possible retweet set.

In social sensing, the data forms that users represent their observation in Twitter might vary depending on the phenomenon, i.e., it might include text or images if the phenomenon is traffic status in a region [40], recent updates about a social disaster [41], or result of a political election [12], numbers if the phenomenon is the stock price of an asset in a market [39]. In this study, we assume that θ is an unknown deterministic or stochastic scalar parameter, and we use a stochastic processes to model the phenomenon θ. Accordingly, the users share tweets consisting of scalar signals to describe the phenomenon θ.

3.3 Estimation via Twitter

Using the observation signal model presented in Section 3.2, we treat event estimation in Twitter as a parameter estimation problem. As explained in [49], what makes the estimation problem different than the typical ones is that the types of observations with probabilistic existence, and we use a probabilistic approach while modeling the actions of the users. Due to features of Twitter, several factors should be considered in social sensing with Twitter:

- *Reliability of the users*: Although a user includes the hashtag in its tweet, it might not be tweeting about the relevant phenomenon to be observed, hence, it does not provide any helpful information for social sensing. In this work, we assume that all tweets with the relevant hashtag include information about the phenomenon.

- *Geotag Use*: Based on availability of geotag (location information), the users' observations may be correlated with respect to their locations. If a user does not use geotag in its tweet, then its location information is not available. Hence, this might require changes in modeling of users' observations. If U_i does not use geotag in its tweet, we assume that its observation $\theta_i[n]$ is uncorrelated with the phenomenon θ_0 and with other users' observations, i.e., $E[\theta_i\theta_j] = 0$ for $j = 0...N$ and $i \neq j$.

- *Behaviors of the users*: When a user is active on Twitter, there are several actions that it can take including tweet, retweet, search tweets, read tweets, send a direct message, comment and several more. In our study, we only consider tweet and retweet actions since they stand as two main mechanisms to disseminate information via Twitter. In our model, we assume that a user might either tweet or retweet during social sensing. For each user U_i, we define the Bernoulli random variable A_i^{tw} such that $A_i^{tw} = 1$ with probability p_i if U_i decides to tweet, $A_i^{tw} = 0$ with probability $1-p_i$ if U_i retweets another tweet. Here, we assume that all users behave the same during the social sensing, i.e., $p_i = p_j$ for $i, j = 1..N$ and $i \neq j$. Thus, a user's observation to the hashtag timeline becomes

$$X_i[n] = A_i^{tw} s_i[n] + (1 - A_i^{tw}) t_i[n]. \tag{3.3}$$

- *Network Conditions*: From a WSN perspective, there might be impurities in the sensors in Twitter sensor network. For example, a crowded observation region may result in delayed observations from the users' internet devices depending on the environment and observed phenomenon, i.e., network infrastructure may be damaged during natural disaster. In this study, we are mainly interested in user behavior, hence we consider that there is not a network connection problem, i.e., users deliver their observations without a delay.

  With these assumptions on users' observations, in the next section we derive the error functions for classical and Bayesian estimation setting.

3.3.1 Classical Approach

In classical estimation, parameters to be estimated are taken to be unknown deterministic constants [25]. In social sensing, this approach may be used in estimation of deterministic constants, i.e., θ may represent the weight of an animal (given as a motivating example in [10]) or price of a product. While θ is a constant, there exists observation noise, denoted as w_i for each U_i. Depending on the observed phenomenon, various statistical models may be used for the noise. Here, the observation noise is assumed to be constant in time n, and modeled as zero-mean jointly-Gaussian random variables such that

$$E\{w_i\} = 0,\ var\{w_i\} = \sigma_w^2,\ \text{for } i = 1, 2, 3, ...N$$

Assuming that the observation noise is correlated, a spatial exponential covariance function is used for user observation noise, which is a function of the distance between the users $d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ and c_s, spatial correlation parameter that determines the degree of the correlation [46], such that

$$E[w_i w_j] = \sigma_w^2 e^{-d_{i,j}/c_s}. \tag{3.4}$$

The effect of correlation in noise might be considered as the bias resulting from other user's location assuming that users know other users' locations by their tweets. Under this setting, observation signal model in (3.3) can be modified as

$$X_i = A_i^{tw}\,(\theta + w_i) + (1 - A_i^{tw})\,(\theta + w_k)\,, \tag{3.5}$$

where time index n is removed since observed phenomenon is taken to be a constant under classical setting.

Assuming that the user does not have any a priori information about the phenomenon, we model social sensing based on users' social consensus using a weighted average learning rule, in which the weights are assumed to be equally allocated among all the sensors [1]. We derive the mean square error function of the estimator for the instantaneous estimation (social sensing) as

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} X_i[n] \tag{3.6}$$

$$mse(\hat{\theta}) = E[(\hat{\theta}-\theta)^2] \tag{3.7}$$

$$mse(\hat{\theta}) = \frac{\sigma_w^2}{N}+\frac{1}{N^2}\sum_{\substack{i,j=1\\ i\neq j}}^{N}\Big(p^2E[w_iw_j]+(1-p)^2E[w_kw_l]+p(1-p)(E[w_iw_l]+E[w_jw_k])\Big)$$

$$= \frac{\sigma_w^2}{N}+\frac{\sigma_w^2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\Big(p^2e^{-\frac{d_{i,j}}{c_s}}+(1-p)^2e^{-\frac{d_{k,l}}{c_s}}+p(1-p)(e^{-\frac{d_{i,l}}{c_s}}+e^{-\frac{d_{j,k}}{c_s}})\Big). \tag{3.8}$$

In classical estimation, Cramer Rao Lower Bound (CRLB), which defines the minimum variance of unbiased estimators, is the fundamental bound on the performance of unbiased estimators, and it is defined for most of the general observation models in literature [25]. In this study, we use a linear Gaussian observation model, i.e., $\vec{X} = \vec{H}\vec{\theta}+\vec{w}$, while deriving CRLB. Towards this end, we define the following vectors and matrices to form corresponding linear observation model. Here, $\vec{A^{tw}}$ is a $N\times 1$ random vector such that

$$\vec{A^{tw}} = \begin{bmatrix} A_1^{tw} \\ \vdots \\ A_N^{tw} \end{bmatrix}, \tag{3.9}$$

where the elements of $\vec{A^{tw}}$ are simply the random variables A_i^{tw}, for $i = 1..N$, defined in the beginning of Section 3.3. $\vec{w^{tw}}$ and $\vec{w^{rt}}$ are $N\times N$ random matrices such that

$$\vec{w^{tw}} = \begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & 0 & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \dots & w_N \end{bmatrix} \vec{w^{rt}} = \begin{bmatrix} w_p & 0 & \dots & 0 \\ 0 & w_c & 0 & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \dots & w_z \end{bmatrix},$$

where the diagonal elements of $w^{\vec{tw}}$ and $w^{\vec{rt}}$ are simply noise elements belonging to user observations. Note that dummy indices in $w^{\vec{rt}}$, i.e., p, c, z are used for retweet messages since their index values depend on the user's selection at the moment. The linear observation model may be written as

$$\begin{aligned}\vec{X} &= \left(A^{\vec{tw}}+(\vec{1}-A^{\vec{tw}})\right)\theta+w^{\vec{tw}}A^{\vec{tw}}+w^{\vec{rt}}(\vec{1}-A^{\vec{tw}}) \\ &= \vec{1}\theta+w^{\vec{rt}}(\vec{1}-A^{\vec{tw}})+w^{\vec{tw}}A^{\vec{tw}} \\ &= \vec{1}\theta+\vec{w}, \end{aligned} \tag{3.10}$$

where $\vec{1}$ is a $N\times 1$ vector, θ is the unknown parameter, and $\vec{w}$ is

$$\vec{w} = w^{\vec{tw}}A^{\vec{tw}}+w^{\vec{rt}}(1-\vec{A}^{tw})$$

For linear Gaussian observation model, i.e., $X = \vec{H}\vec{\theta}+\vec{w}$, CRLB (mean square error) is given as [25]

$$C_{\hat{\theta}} = \left(H^T C^{-1} H\right)^{-1},$$

and after rearranging the terms in (3.10), $C_{\hat{\theta}}$ reduces to

$$C_{\hat{\theta}} = \left(\vec{1}^T C^{-1}\vec{1}\right)^{-1}, \tag{3.11}$$

where C is a $N\times N$ noise covariance matrix, whose entries, for $i, j = 1..N$ and $i \neq j$, are given as

$$C_{ii} = \sigma_w^2$$

and

$$\begin{aligned} C_{ij} &= p^2E(w_iw_j)+(1-p)^2E(w_kw_l)+p(1-p)E(w_iw_l)+p(1-p)E(w_jw_k) \\ &= \sigma_w^2\left(p^2e^{-\frac{d_{i,j}}{c_s}}+(1-p)^2e^{-\frac{d_{k,l}}{c_s}}+p(1-p)(e^{-\frac{d_{i,l}}{c_s}}+e^{-\frac{d_{j,k}}{c_s}})\right) \end{aligned}$$

In the next section, we consider social sensing from a Bayesian estimation perspective, in which the phenomenon is not a scalar constant, yet it changes with time.

3.3.2 Bayesian Approach

In this setting, a priori parameter PDF is known for the parameter to be observed. For example, the phenomenon to be observed may be modeled as a random process, i.e., stock

price of an asset, temperature distribution over a region over a time period. In this setting, we assume that each user has independent zero mean Gaussian noise $w_i[n]$ with variance σ_w^2, and users observations θ_i, for $\theta = 1..N$, where θ_0 corresponds to source signal, are modeled as zero-mean jointly-Gaussian random variables such that

$$E\{\theta_i[n]\} = 0,\ var\{\theta_i[n]\} = \sigma_\theta^2,\ i = 0, 1, 2, 3, ...N$$

Similar to the Gaussian noise model in classical approach, a separable spatio-temporal exponential covariance model is used for modeling the user observations such that

$$E[\theta_i[n_i]\theta_j[n_j]] = \sigma_\theta^2 e^{-d_{i,j}/c_s} e^{-|\delta|/c_t}, \tag{3.12}$$

where $\delta = (n_i - n_j)/f$, f is the sampling rate, i.e., observation frequency, c_s and c_t are spatial and temporal correlation coefficients [46]. Different correlation functions may be used in the model depending on the type of observed events. Hence, $X_i[n]$ becomes

$$X_i[n] = A_i^{tw}[n]s_i[n] + (1 - A_i^{tw})t_i[n] \tag{3.13}$$

By using (3.6), (3.7) and (3.13), mean square error for weighted average is derived as in (3.14). In (3.14), N is the number of total users participating in social sensing, U_i tweets $s_i[n]$ and retweets $t_i[n] = s_k[n - a_k]$, U_j tweets $s_j[n]$ and retweets $s_l[n - a_l]$, where k, a_k, l and a_l are dummy indices, which are determined by randomly choosing retweet messages among the possible message sets of U_i and U_j.

There are various estimation approaches in Bayesian estimation such as Minimum Mean Square Error (MMSE) estimator, Maximum A Posteriori (MAP) estimator and Linear Minimum Mean Square Error (LMMSE) estimator. When the parameter and observations are jointly Gaussian, these estimators perform the same [25]. Hence, in this work, we use MAP estimator as a reference, which aims to maximize the posterior PDF $p(\theta|\vec{X})$. MAP Bayesian

$$\begin{aligned} mse(\hat{\theta}) = \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \frac{\sigma_\theta^2}{N^2} \Big(& p^2 e^{\frac{-d_{i,j}}{c_s}} + p(1-p)(e^{\frac{-d_{i,l}}{c_s}} e^{\frac{-a_l}{c_t}} + e^{\frac{-d_{k,j}}{c_s}} e^{\frac{-a_k}{c_t}}) + (1-p)^2 (e^{\frac{-d_{k,l}}{c_s}} e^{\frac{-|a_k - a_l|}{c_t}}) \\ & + (1-p)^2 \frac{\sigma_w^2}{\sigma_\theta^2} \delta(k-l)\delta(a_k - a_l) \Big) + \frac{\sigma_\theta^2(N+1) + \sigma_w^2}{N} - \frac{2}{N} \sum_{i=1}^{N} \sigma_\theta^2 \Big(p e^{\frac{-d_{i,0}}{c_s}} + (1-p) e^{\frac{-d_{k,0}}{c_s}} e^{\frac{-a_k}{c_t}} \Big) \end{aligned} \tag{3.14}$$

error for jointly Gaussian parameter and observations is given in [20] as

$$Bmse(\hat{\theta}) = \phi_{\theta\theta} - \phi_{\theta S}\phi_{SS}^{-1}\phi_{\theta S}^{T} \tag{3.15}$$

where $\phi_{\theta\theta}$, $\phi_{\theta S}$ and ϕ_{SS} are 1×1, $1\times N$ and $N\times N$ spatio-temporal covariance matrices respectively. $\phi_{\theta\theta}$ is the variance of θ. The elements of each covariance matrix are defined as (Time indices are dropped for simplicity, i.e, $E(\theta_i[n_i]\theta_j[n_j]) = E(\theta_i\theta_j)$.)

$$\begin{aligned}\phi_{SS}(i,j) &= p^2E(\theta_i\theta_j) + p(1-p)E(\theta_i\theta_l) + p(1-p)E(\theta_j\theta_k) \\ &+ (1-p)^2\Big(E(\theta_k\theta_l) + \sigma_w^2\delta(k-l)\delta(n_k-n_l)\Big),\end{aligned} \tag{3.16}$$

$$\phi_{SS}(i,i) = \sigma_s^2 + \sigma_w^2 \tag{3.17}$$

$$\phi_{\theta S} = \begin{bmatrix} pE[\theta_0\theta_1] + (1-p)E[\theta_0\theta_k] \\ \vdots \\ pE[\theta_0\theta_N] + (1-p)E[\theta_0\theta_z] \end{bmatrix} \tag{3.18}$$

The terms $E[\theta_i\theta_j]$ in (3.16) and (3.18) may be found using (3.12).

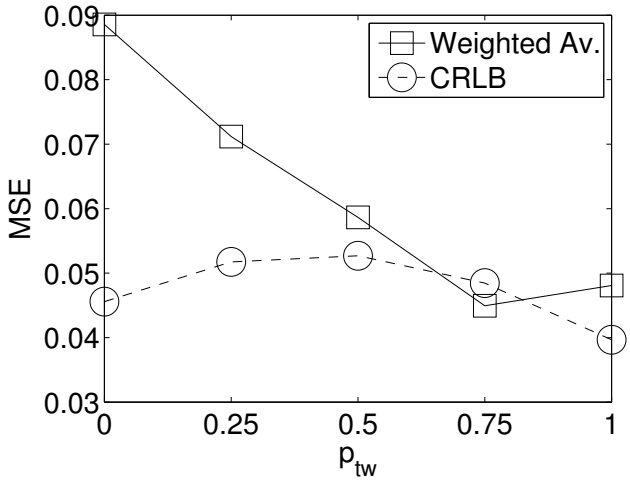
3.4 Numerical Analysis

In this section, we analyze the accuracy performance in mean square error (MSE) sense of social sensing with Twitter and compare it with the performances of the fundamental estimators in classical and Bayesian estimation theory. We use MATLAB in numerical analysis along with an iterative approach, since based on our observation model MSE depends on both directed network structure, i.e., who follows whom, and the locations of the users [8]. Most of the social networks including Twitter have the characteristics of scale free graphs, whose degree distribution is defined by power law, with power exponent α between 2 and 3 [29]. The major feature of these kinds of network graphs is that while a minor number of nodes have large number of connections to other nodes, a major number of nodes has small number of connections with other nodes. In Twitter, this feature represents the fact that while many users have less number of followers in Twitter, a small number of users in the network have high number of followers. For each analysis, we generate random scale free graphs using Complex Network Toolbox [34], then we randomly locate users in 100 x 100 m^2

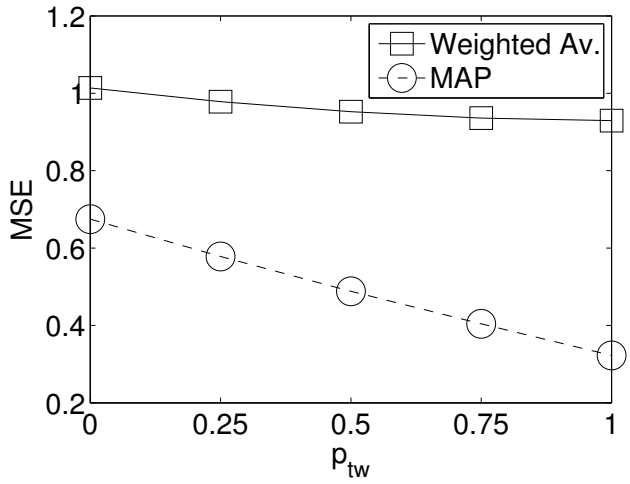
square grid area. Source signal is assumed to be located at the center of grid area. For each analysis, we iterate MSEs over for 1000 different network and deployment configurations, and take the average in order to obtain the final average MSEs. The dummy variables related to retweet signal k, l, a_k and a_u are chosen randomly out of possible retweet message sets of U_i and U_j. We assume that users retweet messages created at time $n-1$ and $n-2$ for simplicity. In the analysis, we form possible retweet message sets by finding all possible messages created by followees of U_i using the network graph. The smallest time frame is taken as a minute and for correlation models, simply $f = 1$ observation/minute, $c_s = 10$ (unitless) and $c_t = 10$ (unitless) are used. Note that we do not have any specific units for MSEs in the results since the event to be observed is not defined explicitly in the model and the MSE may be considered as deviations from the true value under continuous sensing model. During analysis, unless otherwise is stated, we assume that (i) default network size is 1000 users, i.e., $N = 1000$, and (ii) all users behave the same, i.e., $p_i^{tw}[n] = p_{tw} = p$, and (iii) all users use geotag. Throughout the analysis, we will refer to the social sensing error (with social consensus) as WA, fundamental errors in classical estimation as CRLB and in Bayesian estimation as MAP.

3.4.1 Correlated Users

In Fig. 3.1, MSEs in classical and Bayesian case are analyzed for varying p_{tw}. Comparing Fig. 3.1(a) and Fig. 3.1(b), we observe that MSE values improve for increasing p_{tw} in both classical and Bayesian approaches, and the performance of social sensing in classical setting is approximately 90% less erroneous than the Bayesian approach since the parameter is taken to be a simple constant in classical approach, and the randomness of the phenomenon worsens the performance in Bayesian setting. In Fig. 3.1(b), it is seen that in general WA is worse than MAP for varying p_{tw}. As p_{tw} increases, the gap between MAP and WA widens. In Fig. 3.1(a), while WA reaches its minimum MSE for $p_{tw} = 0.75$, CRLB peaks at $p_{tw} = 0.5$, the case when the observations are the most random. In Fig. 3.2, the performances are compared for varying number of people N and varying p_{tw}. The fundamental bounds (CRLB and MAP) improve as N increases as expected. However, MSE for WA in Bayesian setting remains almost constant with N as seen in Fig. 3.2(b) and can not attain the MAP bounds since the increasing number of participants in social



(a) Classical Estimation.



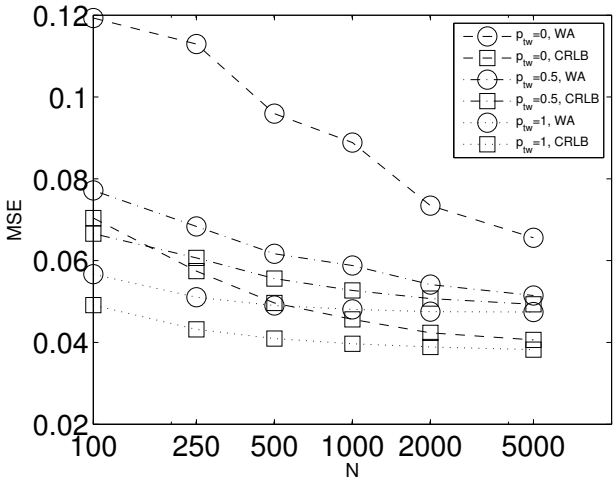
(b) Bayesian Estimation.

Figure 3.1: MSE for varying p_{tw}.

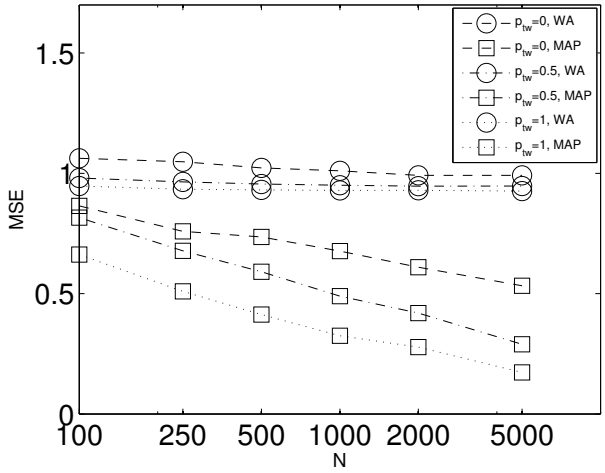
sensing is not likely to able to compensate the randomness of the phenomenon.

3.4.2 α-spatially Correlated Users

In Fig. 3.3, the performances of the estimators for varying α, the ratio of the users using geotag, are analyzed. When α percent of users are α-spatially correlated, it is assumed that first αN of the users, i.e., $i=1..\alpha N$, use geotag, and their observations are correlated with each other and the source phenomenon (θ_0). The rest of the users, i.e., $i=\alpha N+1..N$,



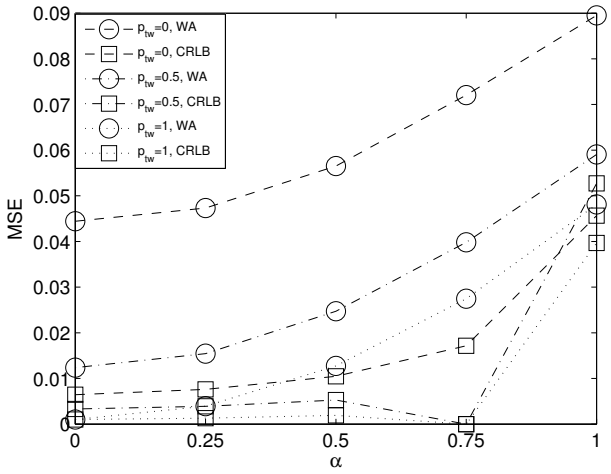
(a) Classical Estimation.



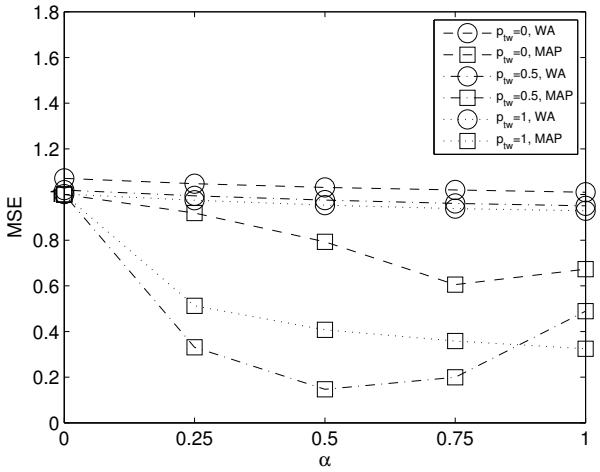
(b) Bayesian Estimation.

Figure 3.2: MSE for varying N.

share uncorrelated observations with θ_0 and the other users. The MSE expressions in (3.8), (3.11), (3.14) and (3.15) given in Section 3.3 are valid for the case when all users use geotag. For α-spatially correlated users, we compute the covariance terms in the MSE expressions accordingly, as described in Section 3.2. In classical setting in Fig. 3.3(a), since users' observation noises are taken to be correlated when geotag is used, its use actually worsens the error (both WA and CRLB). For varying p_{tw} and α, the performance bounds on accuracy (MAP) may be obtained as seen in Fig. 3.3(b) in Bayesian setting. WA does not significantly



(a) Classical Estimation.



(b) Bayesian Estimation.

Figure 3.3: MSE for varying α.

improve for varying p_{tw}. In MAP, we observe that when the observations become either all uncorrelated or correlated, it seems that the bound worsens. This might be due to the fact that correlated observations may introduce redundancy based on our observation model, and MAP takes advantage of correlations up to a limited level, approximately $\alpha = 0.5$. Still, WA can not attain the MAP bound in this analysis. Broadly, tweeting that results in independent observations improves the performance of social sensing. On the other hand, retweeting worsens the performance by introducing redundant observations that result in

noise in the estimation. Accuracy of sensing in classical setting mainly varies with user behavior and number of users. CRLB may be attained for various user behaviors. In Bayesian approach, social consensus results in an adverse performance compared to the fundamental bounds. The performance of WA may slightly improve depending on users' behaviors and users' preferences in geotag use.

Chapter 4

EVENT ESTIMATION ACCURACY OF SOCIAL SENSING WITH FACEBOOK COMMENT THREAD NETWORK

4.1 Introduction

Social sensing, utilization of the information shared in Online Social Networks (OSNs) for sensing a physical phenomenon, is a recently new research field of interest in computer science and associated disciplines. Increased use of OSNs enabled the collection and analysis of data from OSNs (including Twitter, Instagram and Facebook) to understand collective observations on an event of interest. The significance of the collective observations appears in major social events and online marketing strategies. During major social events, people use OSNs to learn about the general situation regarding the event [15, 16]. As an example, the active use of OSNs during 2011 Arab Spring and 2013 Boston Marathon Bombings illustrates the significance of the information dissemination through OSNs [24, 18]. When the global information sources were not available for information retrieval, people followed the course of events by following users' posts in OSNs. Furthermore, fast information spreading capability of OSNs has provided firms to market their products through Online Marketing strategies [11]. The strategies include maintaining a product page in an OSN, in which they share advertorial information about the product and the users share their opinions about the product. The observations collected in these pages may be an important signal indicating the performance of the product on the market. Therefore, understanding the social sensing mechanisms and the accuracy (or reliability) of collective observations in OSNs is an important problem.

Up to now, the research on social sensing has mainly evolved in two directions. First, there are studies that perform event detection for various events with OSNs (mostly Twitter) by sentiment analysis on the collected data [41, 40, 9, 31]. This approach removes the uncertainty in reliability of social sensing by eliminating false or irrelevant user data, however, we are interested in the uncertainty of user observations since social sensing may

be required during instantaneous situations, in which the analysis of data is not feasible and the decision-making authorities (e.g., governments in the case of disasters) may rely on the collective observations. In the second approach, several studies introduce the subject from a more abstract level, in which they define OSN users as *social sensors*, to understand the *social* and *uncertain* nature of social sensing. However, they do not provide theoretical understanding of the process on how user observations are formed and how the consensus is reached [17, 38, 7]. Differently, [28] introduces social sensing from a theoretical level by utilizing social learning and game-theoretic models to investigate the interactions between the users by focusing on decision-making applications. However, the study does not consider the effect of OSN in use during social sensing. In social sensing, OSNs provide the media for generating user observations, hence studying the distinct mechanisms of OSNs is essential in understanding the reliability of social sensing. In this study, we direct our attention to Facebook, which is the most popular OSN (with respect to statistics in [13]), yet whose social sensing capabilities have not been studied from a theoretical level before.

Facebook, currently with 1.3 billion active users (as of June 2014), is a popular free OSN that allows registered users to create profiles, upload photos and videos, send messages and keep in touch with friends, family and colleagues [14]. There exist several studies in literature, which investigate the network characteristics and user activities in Facebook by analyzing collected data [11], [33], [36]. On the other hand, [43] studies the effect of social influence of a Facebook fan page on movie box offices and develops an influence model to forecast box offices for movies, which shows the potential of Facebook as a medium for social sensing.

Facebook has indeed several features that provide a medium for social sensing. For example, *Opinion Polls* are the platforms that allow users to survey the opinions of its friends. Similar to Twitter, Facebook also has *Hashtag Pages*, in which Facebook posts labeled with a specific keyword are collected in a single Facebook page, and the available data may be utilized for social sensing. More importantly, the social sensing mechanism that is most authentic to Facebook is based on the interactions of users through user walls, the Facebook pages in which users either add comments or like the posts that are shared by their friends [33]. The discussion of the opinions through these interactions on Facebook continually yields in a collection of user posts, which is called as *comment thread* [4]. Comment thread

is a medium for social sensing since it facilitates the collection of the opinions of Facebook users on a specific event, which may be estimated using the posts in the comment thread. We name the overall opinion formation process as *Comment Thread Network* (CTN), and in this study, we wish to investigate the social sensing capabilities of Facebook CTN, which, to the best of our knowledge, have not been studied so far.

Social sensing with CTN is radically different from the existing sensing schemes due to several factors, which stem mainly from the *social sensors*, i.e., Facebook users along with the features of Facebook. Observation signals are generated by Facebook users through CTN, and they include noise component due to users' perceptions. Moreover, sharing observation signals in various forms in the comment thread may introduce additional noise. After all, studying social sensing mechanisms and the social sensors in OSNs will foster the research on OSNs, and the gained insights are likely to contribute the effective designs of future Online Social Networks along with social sensing applications in many areas in our lives mainly including disaster management, online marketing and many more to come in future.

In this chapter, we develop an analytical social sensing model for the formation of user observations and social consensus in Facebook CTN from a signal processing perspective, by modeling user observations as *social signals*, and we investigate the accuracy of collective observations for varying user behaviors, relationships and features of Facebook. For the notion of social sensing contains concepts from statistical signal processing along with social learning models in economics, we will use the following group of words *estimate signal-social consensus-collective observations* and *user post-observation signal* interchangeably throughout the study.

The remainder of the chapter is organized as follows. In Section 4.2, we introduce the concept of Facebook CTN, describe the main features of Facebook in social sensing and state our assumptions on the Facebook CTN model. In Section 4.3, we develop an analytical social sensing model and derive the estimation error function. In Section 4.4, numerical analysis is performed and the results are presented.

4.2 Preliminaries on Facebook CTN

Social sensing with Comment Thread Network in Facebook is a sequential process, in which user observations evolve in time, and most comments in comment thread are made in few hours [37]. Therefore, social sensing with CTN may be studied in a limited time window. In a discrete time setting, we limit the time window with the number of participating users in CTN assuming that given an initial post, a single comment is generated at each discrete time step. Briefly, CTN is described as follows. Denoting the present time by n and assuming that the social sensing process ends at time n with N commenting users along with the initial user, who started the comment thread, we examine the course of user actions in CTN from past to present. The illustration of the process is shown in Fig. 4.1. Starting from the initialization of the comment thread, the initial Facebook user U_0 creates a post $p_0(n–N)$ at time $n–N$, which includes its observation about an ongoing event θ. Due to features of Facebook, this post appears in the Facebook pages of the users who are friends with U_0. We may call this group of the people as the *community* C of U_0. After seeing the post, the users in C may want to share their observations on the initial post. At the next time step, first commenting user, U_1 comments on the initial post with $p_{i+1}(n–N+1)$. Similarly, new posts are created sequentially (by the users $U_2, U_3 ... U_N$), and at time n, we obtain the set of posts (or comments) P

$$P = \{p_0(n–N), p_1(n+1–N), ..., p_N(n)\} \tag{4.1}$$

Besides the set of posts P, Facebook releases an additional information about the popularity of the comments, which is the like count information of each post, denoted by $L(p_i)$ or simply L_i. Some users in C may also share their attitudes about the event by liking the relevant user posts in the comment thread. The final set of like count information of the posts, L becomes

$$L = \{L_0, L_1, ..., L_N\}. \tag{4.2}$$

Ultimately, using the available data in the comment thread (i.e., the sets P and L) provided by the community, an estimation $\hat{\theta}$ about θ is made as shown in Fig. 4.2, in which the estimator block represents the social consensus. In the next section, we briefly describe and assess the main user action related features of Facebook in social sensing with CTN, i.e., comment and like actions, in order to comprehend Facebook CTN completely.

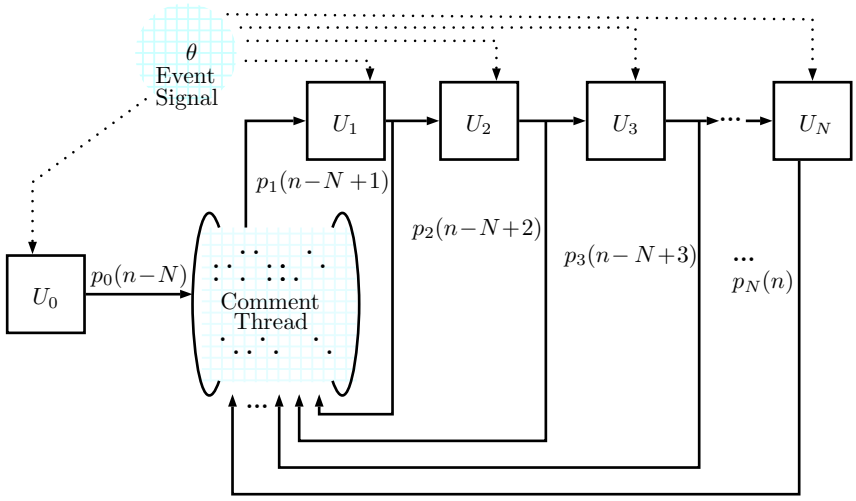


Figure 4.1: Comment Thread Network Model.

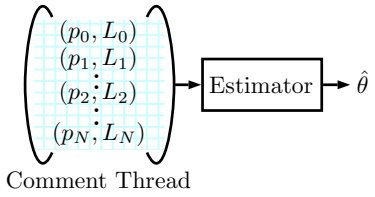


Figure 4.2: Social Consensus Formation in CTN.

4.2.1 User Actions in Facebook CTN.

In Facebook CTN, users generate the observation signals through several ways, i.e., by *commenting* and/or *liking* a given Facebook post.

Comment

By definition, a *comment* is a remark that expresses a personal reaction or attitude about a given fact. In Facebook, given a user's post, the friends of the user may share their opinions about the post via comment, a feature that allows users to add their opinions under the user's post. The opinion reflects their attitudes and recently learned facts about the state of event. Comments (comment posts) in Facebook may include various types of data such as text, image, video or a link to a web page.

Like

Like is a Facebook feature, available for every user post, which allows other users to demonstrate their support on the shared opinion by clicking like button. In other words, it counts the number of likes that a post gets after it is created. The posts belonging to influential users and containing valuable information are likely to get more likes than the other posts, and users on Facebook use like button for highlighting the post among the other posts [11]. Therefore, like count information may be important in defining the estimate signal $\hat{\theta}$ in Facebook CTN.

Before giving the detailed description of Facebook CTN model in Section 4.3, in the next section, we give the assumptions on the user behaviors and CTN model in Facebook.

4.2.2 Model Assumptions

Social sensing with Facebook CTN is a process that evolves in a user's community through comments and likes generated by the user's friends. It is a sequential in the sense that users' posts will be affected from preceding users' posts. In our model, we use a discrete time framework and assume that users post their comments in sequence at fixed time steps. It is also assumed that the user will make a comment based on its current belief and the observation shared by the preceding user. In CTN, we assume that each user in the community is a rational user in the sense that it is aware of the event to be estimated θ, hence it does not lie or share irrelevant observation.

4.3 Comment Thread Network Model

Social sensing with Facebook CTN mainly relies on the event of interest θ to be estimated and a community C, in which users' observations on θ are generated. We start our modeling with the user relationships in Facebook CTN.

Assume that the friendship relationships within the community C (of size N+1) are extracted from the complete social network as a weighted graph, which consists of a set of nodes (Facebook users) and an adjacency matrix W of size $(N+1)\times(N+1)$, whose elements w_{ij} are non-binary values between $[0, 1]$ representing the level of the friendship between U_i and U_j. For higher values of w_{ij}, users U_i and U_j have a closer relationship. In our

modeling, we assume that the relationships are symmetric ($w_{ij} = w_{ji}$), and the values of w_{ij} for $i = j$ are ignored due to their irrelevance in defining friendship relationships.

In general, discussions on social issues such as politics, sport events, natural disasters and product reviews are seen in Facebook comment threads [42]. Therefore, statistical model of the phenomenon θ will heavily depend on the event of interest. For example, if the discussion is on the occurrence of an event (e.g., result of a soccer game), then the event may be modeled as a binary signal. On the other hand, the event of interest may not be about a polarized subject, yet it may be on a subject, in which user observations consist of a continuous range of standpoints. For example, in the estimation of weather temperature, user observations consist of scalar temperature values within a continuous range. In our model, we will simply take phenomenon to be an unknown constant scalar parameter, θ. In the next section, we introduce the social signal model for user observations in Facebook CTN.

4.3.1 User Observations

Main components of social sensing with Facebook CTN are the participating users, who act as sensors in observing and sampling the event of interest. They are social sensors, which perform sensing through their perceptions based on the sensory inputs from outside world. However, the users may introduce noise at different levels of sensing process arising from their perceptions of reality and the methods that they reflect their observations in comment thread. Thus, in this section, we decompose user observation signal into two signals, which are pure observation signal and observation encoding noise, and assess their characteristics.

Pure Observation Signal

A user's pure observation on the event θ is the reflection of event in user's mind through its perception, which may be dependent on the user including its sensory reception, personal characteristics and social status. This fact implies that U_i's observation may not be exactly θ due to variance in its perception and may change with time. To capture the impact of user perceptions in CTN model, we model user observation signals as jointly Gaussian random variables $\theta_i[n]$ for $i = 0...N$ with mean θ and variance σ_θ^2. Since social sensing in CTN occurs within a community and users are connected via friendship relationships, it is

highly possible that users observations are correlated based on their friendship. Friendship distance between two users U_i and U_j is defined by

$$d_{i,j} = 1 - w_{ij}, \tag{4.3}$$

the higher the strength of friendship, the distance between two users will be less, hence their ideas will be more correlated. Considering the correlation resulting from time distance between users' posts and friendship relationship among the users, we assume a stochastic relationship between user observations of U_i and U_j. In a more explicit form, the correlation relationship between $\theta_i[x]$ and $\theta_j[y]$ (at times x and y, respectively) is given as

$$\begin{aligned} E(\theta_i[x]\theta_j[y]) &= cov(\theta_i[x], \theta_j[y]) + \theta^2 \\ &= \sigma_\theta^2 C(d_{i,j}, x-y) + \theta^2, \end{aligned} \tag{4.4}$$

where $C(.)$ is the covariance function that defines the correlation relationships between user observations with respect to friendship and time distance among the posts. In this study, we use a general multivariate exponential covariance function in the form of

$$C(d_{i,j}, x-y) = e^{\frac{-(1-w_{ij})}{\theta_F}} e^{\frac{-\frac{|x-y|}{f}}{\theta_T}}, \tag{4.5}$$

where θ_F and θ_T are correlation coefficients that determine how quickly covariance falls off with respect to friendship and time distance, respectively, and f is the sampling rate of the observed event [30]. Sampling rate is related to how frequently posts are created in Facebook CTN. For simplicity, we assume that posts are generated in each every minute in correspondence with the discrete time framework (the smallest unit of time is a minute), and f rate is set to one sample per one minute.

Observation Encoding Noise

When U_i shares its pure observation $\theta_i[n]$ via a post on Facebook, it has to determine how to encode (or represent) this observation by choosing a proper post type. Facebook allows users to share data in many forms including text, image, video and sound or a web link. The post types may cause different representations of the pure observation signal, which may change the information quality of the post, i.e., users' perceptions and learning on given pure observation signal [35]. The reason is the fact that the representation of a user's

view may be perceived differently by the other users, and the used post type may affect this perception. For example, while a user can alter the content of its observation shared in a text-based post more easily, since the perception of the content heavily depends on the user's writing style, the user's manipulation capabilities are limited in the posts shared as an image or a link since there is no extra room for the user to change these posts in Facebook. Consequently, receiving the pure observation signal in a text-based post is more prone to errors than receiving it in other post types due to both mechanisms of human learning and user characteristics, i.e., the form, which U_i uses to represent its observation in, may introduce additional noise for social sensing. Thus, the use of different post types is likely to change the overall social consensus by altering the observation signals. We may incorporate the effect of different post types by introducing additional noise in users' observations based on the post type they use in CTN, and we will refer to this noise as *observation encoding noise*.

In our model, we limit the available post types on Facebook by only considering text and links to web pages, and we assume that text-based posts are less reliable than link-based posts (including both image and text). Therefore, while forming its comment, the user will introduce additional observation encoding noise on its pure observation signal. Consequently, user U_i's overall observation $s_i[n]$ becomes

$$s_i[n] = a_i v_i^T[n] + b_i v_i^L[n] + \theta_i[n], \tag{4.6}$$

where a_i and b_i are variables with binary values of $\{0, 1\}$, which indicate the use of either a text based or a link based post type by the user, respectively, and $a_i + b_i = 1$ for all $i = 0...N$. In (4.6), $v_i^T[n]$ and $v_i^L[n]$ are the random variables that model the observation encoding noise for each user U_i and for each post type, i.e., text-based and link-based posts, respectively. In our model, we assume that $v_i^T[n]$ and $v_i^L[n]$ are i.i.d Gaussian random variables with zero mean and variances σ_T^2 and σ_L^2, respectively, and we assume that $\sigma_T^2 \geq \sigma_L^2$.

4.3.2 Comment and Like Process

Comment Thread Network starts with the first user U_0's post $p_0(n-N)$

$$p_0[n-N] = s_0[n-N]. \tag{4.7}$$

In the next time step, upcoming user U_1 sees $p_0[n–N]$ and posts its comment $p_1[n–N+1]$, in which it takes into account preceding user's post $p_0[n–N]$ and its observation signal $s_1[n–N+1]$. The user updates its observation with a weighted average social learning rule by allocating weights on each opinion [21]. The weight to be allocated for the preceding post is determined by U_0's friendship level with U_0, i.e., w_{10}. Thus, U_1's post $p_1[n-N+1]$ becomes

$$p_1[n-N+1] = w_{10}p_0[n-N]+(1-w_{10})s_1[n-N+1]. \tag{4.8}$$

Note that since time indices of the posts can be obtained from the information on the sequence of the users in CTN, we may simplify the notation as $P=\{p_0, p_1, ..., p_N\}$, where $p_i=p_i[n–N+i]$, and similarly $s_i = s_i[n–N+i]$.

Using the simplified notation, general form of (4.8) for commenting users reduces to

$$p_{i+1} = w_{(i+1)i}p_i+(1-w_{(i+1)i})s_{i+1}, \text{ for } i = 0...N–1 \tag{4.9}$$

The generation of user posts continues sequentially, as described in Section 4.2, and P is obtained.

In addition to user posts, like count information, which is available as a set $L = \{L_0, L_1, ..., L_N\}$ in Facebook should also be considered while modeling social sensing with CTN. Like is the count information of a Facebook post, which may be considered as a signal indicating the favorability of the opinion shared. Given a user post p_i, like count of p_i, L_i, depends on several factors. First, given two Facebook posts, the post which is posted earlier than the other one, is likely to get more likes since its probability of being seen by more users is higher than the other post. Namely, each post p_i in CTN will appear for different amount of time in the Facebook pages of the users. Second, it is known that the author of the post, who is a more favorable person in the community than the author of the other post, will get more likes. Thus, we may model the number of likes L_i, which post p_i receives, with a Binomial distribution, in which L_i is the number of successes (likes) in a sequence of X (duration of the post) independent like attempts by a single user, each of which results in success with the probability f_i. Assuming that K users attempt to like the post at each time step, we may approximate the total number of likes L_i by its mean

$$L_i = XKf_i, \tag{4.10}$$

where $X = N-i+1$ is the number of attempts (duration of the post), K is the number of Facebook users who attempts to like the post at the given discrete time step, and f_i is the

probability that the post is liked by a user, approximated by the favorability (popularity) of the author of the post, U_k (for $k = 0...N$) as

$$f_k = \frac{\sum_{i=0, i\neq k}^{N} w_{ik}}{\sum_{i,j=0, i\neq j}^{N} w_{ij}}, \tag{4.11}$$

which is simply the normalized average of the weights given by the users in C to U_k. In other words, p_k reflects the user's relationship strength with the community. The higher values of p_k (compared to other users) implies that U_k is a more favorable user in the community.

4.3.3 Facebook CTN Estimator

The impact of each user's observation on the estimate signal (social consensus) $\hat{\theta}$ in CTN depends on the like count information of the posts, since like counts in Facebook demonstrate the level of the support on the post and increase its visibility among the other posts. Given that N users post on p_0, user posts accumulate in comment thread, and ultimately we obtain the collection of data $\{(p_0, L_0), (p_1, L_1), ..., (p_N, L_N)\}$, the elements in each tuple represent the comment signal and like count information of the corresponding user, respectively. The general agreement $\hat{\theta}$ on θ may be obtained from the observations

$$\hat{\theta} = \frac{\sum_{i=0}^{N} L_i p_i}{\sum_{i=0}^{N} L_i}, \tag{4.12}$$

which is the weighted average of user observations with respect to their like counts. By explicitly writing the expression in (4.12) using (4.9) and (4.10), we obtain

$$\hat{\theta} = \frac{\sum_{i=0}^{N} \Big(L_i(1-w_{(i+1)i}) + L_{i+1} w_{(i+1)i} \Big) s_i}{\sum_{i=0}^{N} (N-i+1) K f_i}. \tag{4.13}$$

Since we assume that w_{ij}, L_i, K and p_i are fixed parameters for all users in CTN, we simplify the expression in (4.13) as

$$\hat{\theta} = \frac{\sum_{i=0}^{N} c_i s_i}{M}, \tag{4.14}$$

where

$$c_i = L_i(1-w_{(i+1)i})+L_{i+1}w_{(i+1)i}, \tag{4.15}$$

for $i = 0...N$ and $L_{N+1} = 0$, and

$$M = \sum_{i=0}^{N}(N-i+1)Kf_i. \tag{4.16}$$

In evaluation of the social sensing capabilities of Facebook CTN, we use the estimation error function while analyzing the accuracy of the estimates, defined by mean square error (MSE)

$$D(\hat{\theta}) = E[(\hat{\theta}-\theta)^2]. \tag{4.17}$$

The explicit form of $D(\hat{\theta})$ is given in (4.18).

4.4 Numerical Analysis

In this section, we perform numerical analysis of the estimation error $D(\hat{\theta})$ (denoted as D in Figs. 3-7), using (4.18) for varying user actions and behaviors, Facebook features and network size in MATLAB. Friendship network on Facebook follows power law characteristics similar to general social network characteristics, however, this type of graph contains overall network structure including all users [36]. Since in this study we focus on small communities inside Facebook, we simply use random weighted graphs while modeling the relationships within the community and assume that all users in C are friends with each other. Furthermore, since $D(\hat{\theta})$ depends on the friendship relationships within the community defined by W, the analyses are iterated for various random weighted graphs (10000 times) and $D(\hat{\theta})$ is found by taking average of the results in these runs. Since we do not explicitly define θ

$$D(\hat{\theta}) = \theta^2+\frac{1}{M^2}\sum_{i=0}^{N}\left(a_i^2\sigma_T^2+b_i^2\sigma_L^2+\theta^2+\sigma_\theta^2\right)c_i^2-\frac{2\theta^2}{M}\sum_{i=0}^{N}c_i+\sum_{\substack{i,j=0\\ i\neq j}}^{N}\frac{c_ic_j}{M^2}\left(\theta^2+\sigma_\theta^2(e^{\frac{-(1-w_{ij})}{\theta_F}}e^{\frac{-|i-j|}{\theta_T}})\right) \tag{4.18}$$

in our modeling, $D(\hat{\theta})$ is unitless, and in the analyses, the default parameters in Table 4.1 are used unless otherwise stated.

Table 4.1: Numerical Analysis Parameters

Parameter	**Value**	**Parameter**	**Value**
N	20	θ_T	100
K	10	θ_F	1
θ	0	σ_T	0.5
σ_θ	0.2	σ_L	0.1
f	1 min^{-1}	a_i for $i = 0...N$	1

The analysis results are presented in two sections. First, we investigate the changes in estimation accuracy for varying network size and post characteristics. In the second section, we mainly analyze the effect of user behaviors and relationships on social sensing with Facebook CTN.

4.4.1 Estimation Accuracy Analysis for Varying Network Size and Post Characteristics

The graph in Fig. 4.3 illustrates the changes in $D(\hat{\theta})$ for social sensing with CTN for varying event signal characteristics σ_θ along with varying number of users N. (Note that Facebook CTN has N+1 users in total, but in our analyses we represent it with the total number of commenting users N for simplicity.) In the analysis, the size of community (N) is taken within the range $(5, 50)$, which is the average number of comments in comment thread [4]. The analysis result shows that while for high values of event signal variance ($\sigma_\theta = 1$ and $\sigma_\theta = 5$), the estimator is defective and increasing N does not improve the accuracy of estimation, for relatively low values of signal variance ($\sigma_\theta = 0.2$), estimation accuracy is significantly higher and it improves with increasing number of participants. This trend indicates that CTN may not be suitable in social sensing for all event types, in which user observation signals are highly volatile.

In Fig. 4.4, the effect of variations in post type on estimation accuracy has been investi-

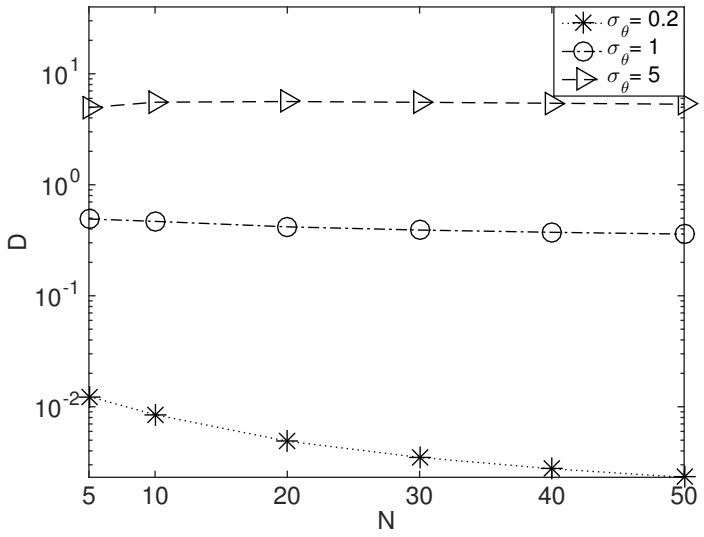


Figure 4.3: $D(\hat{\theta})$ for varying N and σ_θ.

gated. The x-axis values represent the percentage of text input data $(a_0{+}a_1{+}..{+}a_N)/(N{+}1)$ in all users' observations, which consist of text and link-based posts only. The figure depicts that as more text-based posts are included in CTN, the accuracy of social sensing deteriorates by approximately 5 times considering the link-based posts only. The reason is that while sharing observations in CTN, use of data types which are more prone to inclusion of user bias (such as text) is likely to alter the consensus of social sensing adversely.

4.4.2 Estimation Accuracy Analysis for Varying User Behaviors and Relationships

While modeling *like* process, we assume that the most popular users get more likes compared to other users. In the analysis in Fig. 4.5, we compare the accuracy of social sensing for the case, in which we assume that all users are equally liked by their friends, rather than our initial assumption based on popularity-based liking. The graph in Fig. 4.5 indicates that the estimation is less erroneous with equal likes compared to popularity-based likes. Namely, average of all observation signals yields in better performance instead of amplifying the observations of the several users in the community. Furthermore, equally weighting of the observations makes social sensing more robust to more noisy data types.

By following a similar approach to the previous analysis, in the next analysis we in-

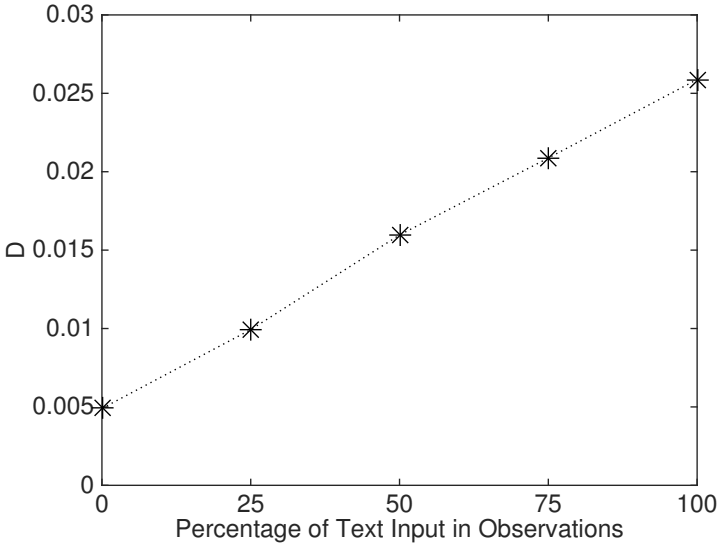


Figure 4.4: $D(\hat{\theta})$ for varying percentage of text-based posts in comment thread.

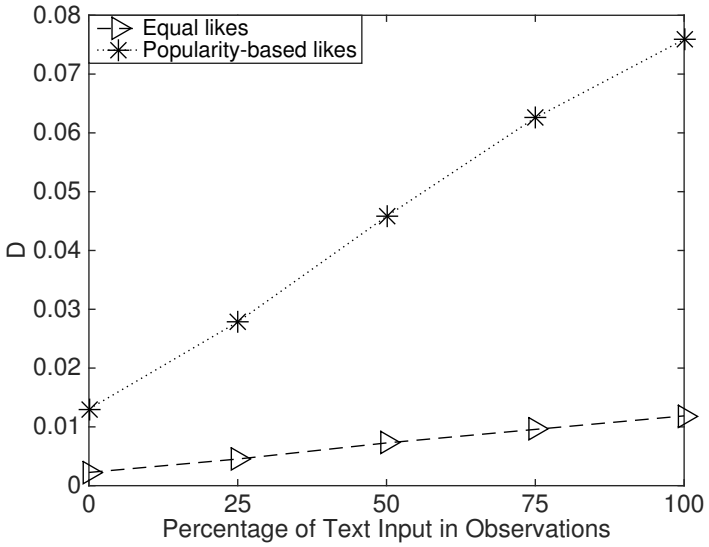


Figure 4.5: $D(\hat{\theta})$ for varying like behaviors.

vestigate the effect of users' opinion weighting behavior while commenting in CTN. In the previous analyses, we assume that users weight preceding comments based on their friendship relationship with the corresponding user. However, sometimes users may choose to allocate more weight their own observations than the preceding user's comment in CTN.

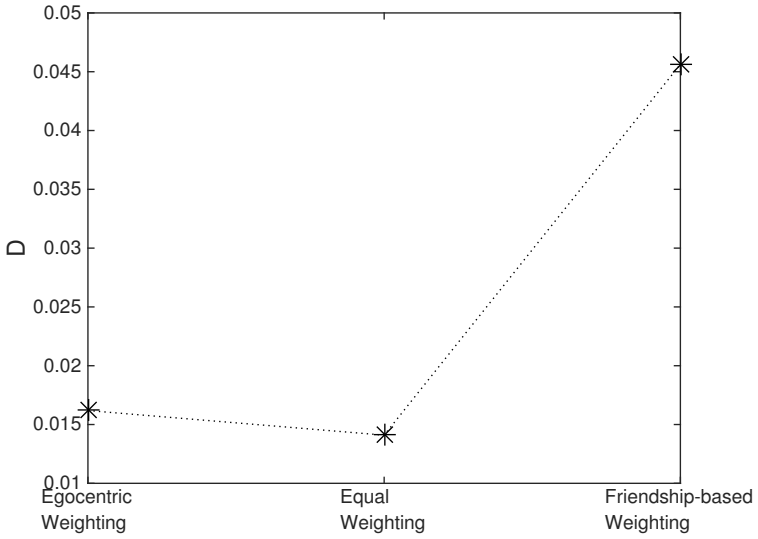


Figure 4.6: $D(\hat{\theta})$ for varying opinion weighting behaviors.

We call this type of behavior as *egocentric* commenting, in which users share only their observations by ignoring the preceding comments. In this scenario, $w_{i,i+1} = 1$ for all $i = 0...N$, however note that the observation correlations and like counts are still computed based on the friendship relationships. In the analysis, we also investigate the case in which the users behave temperately by equally weighting their observations against preceding comment. Fig. 4.6 shows the change in $D(\hat{\theta})$ for varying user behavior. As can be seen, friendship-based weighting results in estimation with a higher value of error (approximately 3 times order) compared to egocentric and equal weighting. These differences in $D(\hat{\theta})$ indicate that unequal weighting of observations via friendship-based weighting while forming social consensus is likely to result in more errors by focusing on few popular individuals' observations when compared to equal weighting.

In the last analysis, we investigate the effect of the correlation between the users in C, which we define using (4.5) for varying θ_F values that change the level of correlation between users' observations. The correlation relationship between users' observations is significant since the accuracy of social sensing in Facebook CTN may change depending on the characteristics of the community. This aspect may especially be important in social sensing applications regarding social events (e.g., political elections). In Fig. 4.7, the change

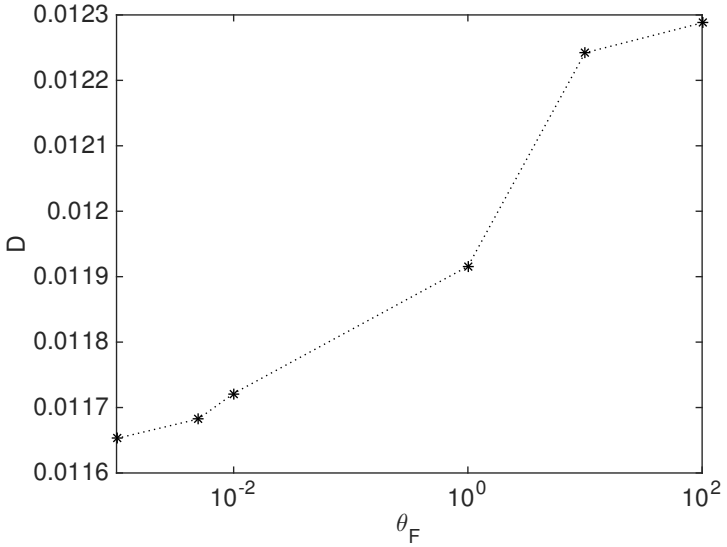


Figure 4.7: $D(\hat{\theta})$ for varying θ_F.

in $D(\hat{\theta})$ is given for varying θ_F. In the analysis, we assume that all observations are equally weighted in order to investigate the effect of correlations more clearly. The analysis shows that estimation error increases by approximately 40% as the correlation between the users increases. In other words, less correlated observations benefit to the accuracy of social sensing, and increasing level of correlation in the community is likely to bias the social consensus. Therefore, the selection of community might be an important factor in social sensing with Facebook CTN.

Chapter 5

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

The frequent use of OSNs during major events recently led to a new application of OSNs called *social sensing*, which is simply the utilization of the information shared in OSNs to make an estimate about an observed (yet unknown) phenomenon. In this thesis, we investigate social sensing capabilities of OSNs for the first time in the literature. To that end, we use a sensor networking approach and introduce the paradigm, Wireless Social Sensor Network (WSSN) and explore the WSSNs in Twitter and Facebook. First, we develop a communication theoretical model for tweet propagation in Twitter, and the accuracy of the estimated signal is explored with mean square error analysis. Later, we investigate social sensing in Twitter from the estimation theoretical perspective. Using a simple observation model by considering the features of Twitter, i.e., tweet and retweet, we analyze the performance (in terms of mean square error) of social sensing, defined as social learning, i.e., social consensus, and compare with fundamental estimators in classical and Bayesian estimation theory for various factors such as user behavior, number of people participating to social sensing and geotag use percentage. It is observed that social consensus is more defective in Bayesian estimation setting than the classical estimation when comparing the measures with the bounds on the fundamental estimators, and social sensing can not attain the fundamental error bounds in Bayesian and classical estimation. The performance of social learning varies with user behavior, however, the change is rather small in Bayesian setting. Last, we study the main social sensing mechanism in Facebook, Comment Thread Network (CTN), which is based on the interactions of users through user walls in Facebook. After seeing their friends' contents about an event, users either add comments or like these posts, and Facebook CTN emerges as a social sensing medium in estimation of an event through social consensus. By modeling user observations as social signals and developing a social sensing model for Facebook CTN from a signal processing perspective, we analyze the reliability of Facebook CTN for varying user behaviors, user relationships, Facebook

features (i.e., post types) and network size. The accuracy depends on the user characteristics and the features of the OSN, since users' perceptions and how they use Facebook may manipulate their observation signals.

In the next sections, we provide the details of the contributions and state future issues regarding social sensing based on WSSNs.

5.1 Contributions

In this section, we summarize the contributions of each chapter and underline the important results.

5.1.1 Social Sensing Model and Analysis for Event Detection and Estimation with Twitter

Twitter is a naturally formed sensor network where user observations are collected, and event estimation may be performed using collected tweets during major events. There are several studies on event detection in Twitter, which are primarily based on semantic analysis of collected tweets. To the best of our knowledge, there exists no analytical study on the accuracy of event detection and estimation of event signal with Twitter using observations collected. In this study, a communication theoretical model is developed for tweet propagation, and the accuracy of the estimated signal is explored with mean square error analysis. The results indicate that, in addition to user observations and network conditions, the accuracy of the estimation is also affected by user behavior patterns. A theoretical communication model for Twitter is developed for the first time and the accuracy of estimate signal formed by collected tweets at hashtag timeline has been investigated. While the accuracy of the estimate signal is mostly dependent on users observation noise and adversely affected by network conditions, i.e., delays or blocks, the results also indicate that the accuracy of the estimation varies with user behavior patterns due to Twitter's features, as well.

5.1.2 Estimation Theoretical Analysis for Social Sensing with Twitter

Social sensing in Twitter is studied from an estimation theoretical perspective using a probabilistic observation model. Defining social sensing as a learning process, the performance of

social sensing is investigated and compared with fundamental accuracy bounds using classical and Bayesian estimation theory. It is observed that social learning is more defective in Bayesian observation setting compared to classical estimation. The performance varies with user behavior, however the change is rather small in Bayesian case. It is likely that when phenomenon is a random process, the estimator, defined as social consensus, is not sufficient to make an accurate estimation and attain the fundamental error bounds. The error level of social sensing and the fundamental bounds depend on the user behavior. Hence, it is crucial to investigate the behavior patterns of users in Twitter to further understand the social sensing capabilities of Twitter.

5.1.3 Event Estimation Accuracy of Social Sensing with Facebook Comment Thread Network

Comment Thread Network is the main mechanism for social sensing in Facebook, which is actively used by Facebook users during the course of major social events for estimating an event of interest. Understanding the reliability of estimation made out of collective user observations in CTN is significant due to applications of social sensing in numerous aspects of human lives. In this chapter, by developing an analytical social sensing model for user observations in CTN from a signal processing perspective, in which we model Facebook users as *social sensors* and their observations as *social signals* by incorporating their perceptions, we analyze the reliability of Facebook CTN for varying user behaviors and relationships, Facebook features and network size. The results indicate that while the characteristics of the event signal affect the accuracy of the social sensing, also the polarized weighting of the observations and use of less reliable post types in CTN during social sensing are important in the sense that they deteriorate the accuracy of the estimate signal, i.e., social consensus. Furthermore, user relationships inside the community may alter the reliability of social sensing implying that the selection of OSN users in social sensing is essential in obtaining an accurate estimation.

5.2 Future Research Directions

In this section, we investigate WSSNs at a lower level, i.e., by studying the formation of user observations and their delivery to the end receiver through the OSN medium. To that

end, we consider an OSN user who wishes to send a message about a phenomenon to its friend and model the information transmission as shown in Fig. 5.1. The user senses the phenomenon, i.e., source signal θ, through its perception, denoted by S. Then, the user encodes its observation by using the available post types in the OSN, e.g., text, image, a video or a web link, and sends its message P. A distorted form of the source signal reaches

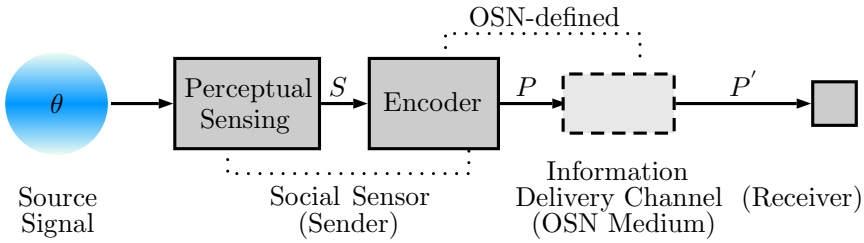


Figure 5.1: Information generation and propagation in the OSN.

to the receiver through the channel, i.e., OSN medium, due to the characteristics of both the sender and also the OSN medium. The sender may distort the source signal by its perception and encoding, and also the OSN medium may alter the characteristics of the message P during its transmission, e.g., other users may comment on the post or the post may be delivered with a delay due to Internet connection. Thus, both the user and OSN medium are likely to alter the characteristics of the observation signal. Next, we discuss the impact of each block in Fig. 5.1 on user observations and outline the open issues.

5.2.1 Source Signal

In social sensing, source signal is the event of interest. The event may be related to a case from various subjects including social disasters, political events, financial markets, sport events and many more. Broadly, we may use various categorizations in defining the source signal. In general, the events can be either *binary* or *continuous* in terms of the value of the unknown parameter. For example, if the unknown parameter is about the realization of an event, e.g., result of a political election, then the event may be modeled as a binary signal. Differently, the unknown parameter may be related to a subject, in which user observations consist of a continuous range of standpoints, e.g., estimation of the price of a product.

Another categorization may depend on the interactions between the source signal and user observations. Some events are independent of user actions, e.g., temperature. We call this kind of events as *static events*. On the contrary, the source signal may change along with the user actions. As the users make an estimate about the source signal and then take actions, source signal changes according to the user actions. Stock price estimation with OSNs can be given as an example. This kind of events may be referred to as *dynamic events* including the interaction between the users and the source signal.

Overall, each event type might have different signal characteristics and may result in different reliability levels in social sensing. Open issues regarding the source signal are as follows.

- To study dynamic events in social sensing, models for the interaction between the OSN users and the source signal should be developed.

- The effect of source signal characteristics for continuous/binary and static/dynamic events should be investigated in social sensing. Different source signals may lead to different reliability levels in social sensing.

5.2.2 Information Source

OSN users provide the observation signals about the source event. They are *social sensors*, which observe and sample the event through their perceptions based on the sensory inputs from outside world, and perform different behavior patterns depending on their state of mind.

Sensor Perceptions

Users observe the source signal through their sensory mechanisms that can sense auditory, visual, tactile, olfactory and gustatory stimuli. The perception, defined as *the process of recognizing and interpreting sensory stimuli*, creates an interpretation of the sensed signal, and the users share these interpreted signals in the OSN. Users' perceptions may alter the observation signal quality. This change is the sensor noise in WSSNs. For instance, users with visual impairments may provide noisier observation signals in estimation of an event observed through visual stimuli.

Sensor Behaviors

Reliability of the social sensors heavily depends on users' behaviors. Social learning theory is useful in defining user behaviors [2]. For example, most users aim to learn the source event in social sensing. Therefore, we may assume that participating users are confronted to share their true observations. However, deviant behavior, i.e., the actions that adversely affect the overall sensing, is frequently observed in OSNs. Some users may alter the performance of the social sensing by proving false information in different levels depending on their presence in the community.

Sensor Samples

OSNs provide different post types for information sharing, e.g., while Twitter mostly allows 140-character limited text messages, Instagram allows images and short duration videos. Moreover, Facebook allows posts including text, image, web link or a video. The data type determines the representation of user observations, i.e., encoding of the user observation signal as shown in Fig. 5.1. Information can be represented in many forms, and the forms may introduce different levels of error while encoding [26].

Social sensors sample the source signal through their perceptions and behaviors, and encode their signals to share them in the OSN. Therefore, the main open issue regarding the social sensors is the characterization of the social sensors.

- OSN users are social sensors with different noise characteristics limited by their perceptive capabilities. Therefore, noise processes included in samples of social sensor nodes should be modeled in social sensing using social learning theories on human perception.

- Social sensors may perform different behavior patterns while sharing their observation signals. These behavior patterns, e.g., deviant behavior, should be investigated and modeled to observe the effect of user behaviors on user observation signals.

- Information-bearing capabilities of the data types may vary for different source signal characteristics. Therefore, understanding the impact of different representations of

information is essential in social sensing. The different representations of observation signals should be modeled and error processes in encoding should be studied.

5.2.3 Information Transmission

User observation signals may deteriorate during their transmission in the OSN. The problems encountered in this stage may lead to delays and even failures in transmission of the user observation signals. The factors influencing the information transmission can be examined under two main lines, i.e., networking infrastructure of the OSN and the wireless channel that the users access to the OSN through Internet.

Network Topologies

Information retrieval and transmission are realized differently based on the networking infrastructures in OSNs. Networking infrastructure determines how information is stored, which is mainly classified as *centralized* and *decentralized* [6]. While centralized OSNs rely on the Internet connection for each information transmission, decentralized OSNs are distributed systems that aim to achieve delay-tolerant information transfer by utilizing the local connections in the OSN. Namely, they are less prone to delays and transmission failures compared to centralized OSNs in the loss of Internet connection.

Wireless Channel

In the context of social sensing, most OSN users are mobile, i.e., they may connect to the OSN via their mobile devices. These devices access to the Internet through wireless mobile networks with different networking architectures, e.g., 3G, 4G and LTE, and each device may support different data rates in wireless communication of user observations. Furthermore, these standards affect the users' Internet experience during social sensing. Thus, the capabilities of mobile network architectures may have an impact on the transmission of user observation signals.

Internet connectivity is an important factor, especially in major events, it may be quite unstable due to excessive access, resulting in bottleneck problems. Understanding of the impact of Internet connectivity is essential as the performance of social sensing relies on the delay levels of the user observations for varying source signal characteristics. The open

issues regarding the information transmission are as follows.

- Information transmission in the OSN server should be modeled and the effect of various networking infrastructures should be investigated. Utilizing these models, analytical bounds of the reliability of social sensing for varying network topologies can be derived and networking topologies for optimum WSSNs can be developed.

- A physical-layer communication model should be developed for information transmission in OSNs. Next, the effect of different mobile networking standards, e.g., 4G, 5G, should be investigated. Thus, these standards may be exploited to enhance the performance of social sensing and the reliable rate of communication between the users.

5.2.4 The OSN in Use

The OSN is the platform that defines the users and how they are connected in the WSSN. OSNs have many features that provide distinct functionalities in information dissemination. For example, most OSNs give users the option of sharing extra information, e.g., location, in addition to the content of the post. Moreover, side information about the user post, e.g., favorability of a user's post, may be extracted from the mechanisms in some OSNs. Namely, users may use *like/favorite* actions in OSNs to show their support for the user's post. The features of OSNs may be leveraged for reliable social sensing.

- User relationships may be useful in modeling user observation signals. For various source signals, friendship levels of the users may determine the relationship between the user observation signals in the network, e.g., closer users tend to have similar opinions on political issues. The effect of user relationships on user observation signals should be investigated. Thus, friendship-based correlation models for source signal can be developed. The models may be exploited to select social sensor nodes in constrained social sensing problems with user limits.

- Mobile OSN users may exhibit distinct mobility patterns in the event region during social sensing. These mobility patterns may impact both the quality of network connectivity and also the user observation signal characteristics. Using the location

information available in some OSNs, users' mobility patterns should be investigated. Moreover, spatial correlation between user observations can be utilized for event estimation algorithms in social sensing.

- Side information can be leveraged to improve the performance of WSSNs. For example, favorite count in some OSNs may be a signal indicating the favorability or reliability of the post. To that end, like/favorite mechanisms in OSNs should be investigated and their use in social sensing must be statistically studied with real user data.

5.2.5 Event Estimation

Event estimation, estimating the specific characteristics of source signal using collected user observations, is the ultimate goal of social sensing. As we discussed in the previous sections, the reliability of the estimation depends on the source signal, the OSN users and the features of the OSN in use. Hence, open issues in event estimation is about understanding the social sensing and communication capabilities of OSNs by studying social sensing models that incorporate the factors influencing the reliability of social sensing. With the gained insight, event detection/estimation algorithms can be developed. We identify key open issues as follows.

- Estimation theoretical bounds on the reliable level of event estimation/detection with social sensing should be derived for varying source signal characteristics, social sensor characteristics and OSNs to understand social sensing capabilities of different OSNs.

- Information theoretical analysis should also be performed to derive the input-output mutual information, in which the output is the estimate signal in social sensing. Next, input-output mutual information can be utilized for investigating the social sensing capabilities of OSNs.

- To develop application-based event estimation/detection algorithms, constrained social sensing problems should be studied utilizing WSSN models. Following constraints can be considered, e.g., the number of the participating users in social sensing, time interval, the type of the OSN and the source signal.

- By understanding the social sensing capabilities of WSSNs and constrained social sensing problems, we may gain insights about which components in an OSN makes it to superior to other OSNs in social sensing. With the extraction of these components, design metrics for future OSNs with advanced social sensing capabilities can be determined and new OSNs can be designed.

BIBLIOGRAPHY

[1] D. Acemoglu and A. Ozdaglar, "Opinion dynamics and learning in social networks," *Dynamic Games and Applications*, vol. 1, no. 1, pp. 3–49, 2011.

[2] R. L. Akers, and J. G. J. Wesley, "Social Learning Theory," in *In 21st Century Criminology: A Reference Handbook*, edited by J. Mitchell Miller. Thousand Oaks, CA: SAGE Publications, Inc., 2009, pp. 323-332.

[3] O. Aulov and M. Halem, "Human sensor networks for improved modeling of natural disasters," *in Proc. IEEE*, vol. 100, no. 10, pp. 2812–2823, Oct 2012.

[4] L. Backstrom, Lars, J. Kleinberg, Jon, L. Lee, and C. Danescu-Niculescu-Mizil, "Characterizing and curating conversation threads: expansion, focus, volume, re-entry," in *Proc. ACM WSDM'2013*, Feb. 2013, pp. 13–22.

[5] D. M. Boyd, N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, 2007, pp. 210-230.

[6] A. Datta, S. Buchegger, L. Vu, T. Strufe, K. Rzadea, "Decentralized Online Social Networks," in *Handbook of Social Network Technologies and Applications*, edited by B. Furht. Springer US, 2010, pp. 349-378.

[7] A. Campbell, S. Eisenman, N. Lane, E. Miluzzo, R. Peterson, H. Lu, X. Zheng, M. Musolesi, K. Fodor, and G.-S. Ahn, "The rise of people-centric sensing," *Internet Computing, IEEE*, vol. 12, no. 4, pp. 12–21, July 2008.

[8] K. Cepni and O. Akan, "Social sensing model and analysis for event detection and estimation with twitter," in *Proc. IEEE CAMAD'2014*, Dec 2014, pp. 31–35.

[9] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. Ebert, and T. Ertl, "Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition," in *Proc. IEEE VAST'2012*, Oct 2012, pp. 143–152.

[10] C. Chamley, A. Scaglione, and L. Li, "Models for the diffusion of beliefs in social networks: An overview," *Signal Processing Magazine, IEEE*, vol. 30, no. 3, pp. 16–29, May 2013.

[11] I. P. Cvijikj, E. D. Spiegler, and F. Michahelles, "The Effect of Post Type, Category and Posting Day on User Interaction Level on Facebook," in *Proc. IEEE PASSAT'2011 and IEEE SocialCom'2011*, Oct. 2011, pp. 810–813.

[12] J. Digrazia, K. McKelvey, J. Bollen, and F. Rojas, "More tweets, more votes: social media as a quantitative indicator of political behavior," *PloS one*, vol. 8, no. 11, p. e79449, 2013.

[13] M. Duggan, N.B. Ellison, C. Lampe, A. Lenhart and M. Madden. "Social Media Update 2014," Pew Research Center, January 2015. Available at: http://www.pewinternet.org/2015/01/09/social-media-update-2014/

[14] Facebook Statistics. `http://newsroom.fb.com/company-info/`

[15] H. Gao, G. Barbier, and R. Goolsby, "Harnessing the crowdsourcing power of social media for disaster relief," *IEEE Intelligent Systems*, vol. 26, no. 3, pp. 10–14, 2011.

[16] M. Greer and J. Ngo, "Personal emergency preparedness plan (pepp) facebook app: Using cloud computing, mobile technology, and social networking services to decompress traditional channels of communication during emergencies and disasters," in *Proc. ACM CHI'2011*, May 2011.

[17] B. Guo, Z. Yu, D. Zhang, and X. Zhou, "Cross-community sensing and mining," *Communications Magazine, IEEE*, vol. 52, no. 8, pp. 144–152, Aug 2014.

[18] A. Gupta, and H. Lamba, and P. Kumaraguru, "$1.00 per RT #BostonMarathon #PrayForBoston: Analyzing fake content on Twitter," in *Proc. eCRS'2013*, 2013, pp. 1–12.

[19] J. Heidemann, M. Klier, F. Probst, "Online social networks: A survey of a global phenomenon," *Computer Networks*, vol. 56, no. 18, 2012, pp. 3866-3878,

[20] C. W. Helstrom, *Elements of Signal Detection and Estimation.* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1995.

[21] C. Chamley, *Rational Herds: Economic Models of Social Learning.* Cambridge University Press, 2004.

[22] A. Iyengar, T. Finin, and A. Joshi, "Content-based prediction of temporal boundaries for events in twitter," in *Proc. PASSAT'2011*, Oct 2011, pp. 186–191.

[23] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: Understanding microblogging usage and communities," in *Proc. 9th WebKDD and 1st SNA-KDD 2007*, pp. 56–65, 2007.

[24] A. Kavanaugh, S. Yang, S. Sheetz, L. Li, and E. Fox, "Microblogging in crisis situations: Mass protests in iran, tunisia, egypt," *in Proc. ACM. CHI'2011*, 2011.

[25] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory.* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[26] Paul A. Kolers and William E. Smythe, "Symbol manipulation: Alternatives to the computational view of mind," *Journal of Verbal Learning and Verbal Behavior*, vol. 23, no. 3, 1984, pp. 289-314.

[27] A. Kongthon, C. Haruechaiyasak, J. Pailai, and S. Kongyoung, "The role of twitter during a natural disaster: Case study of 2011 thai flood," in *Proc. PICMET'2012*, July 2012, pp. 2227–2232.

[28] V. Krishnamurthy and H. Poor, "A tutorial on interactive sensing in social networks," *Computational Social Systems, IEEE Transactions on*, vol. 1, no. 1, pp. 3–21, March 2014.

[29] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" *in Proc. WWW'2010*, pp. 591–600, 2010.

[30] *Linear Models for Multivariate, Time Series, and Spatial Data.* Springer-Verlag New York, 1991.

[31] R. Li, K. H. Lei, R. Khadiwala, and K.-C. Chang, "Tedas: A twitter-based event detection and analysis system," in *Proc. ICDE'2012*, April 2012, pp. 1273–1276.

[32] M. Mendoza, B. Poblete, and C. Castillo. "Twitter under crisis: Can we trust what we RT?," in *Proc. SOMA'2010*, ACM, New York, NY, USA, 2010, pp. 71–79.

[33] A. Muangngeon and S. Erjongmanee, "Analysis of facebook activity usage through network and human perspectives," in *Proc. IEEE KST'2015*, Jan. 2015, pp. 13–18.

[34] L. Muchnik, "Complex networks package for matlab (version 1.6)," http://levmuchnik.net/, 2013.

[35] "Multimodal Learning Through Media, Cisco Systems," Tech. Rep., 2008.

[36] K. Nguyen and D. A. Tran, "An analysis of activities in Facebook," in *Proc. IEEE CCNC'2011*, Jan. 2011, pp. 388–392.

[37] "Local news in a digital age," Pew Research Center, Tech. Rep., 2015.

[38] A. Rosi, M. Mamei, F. Zambonelli, S. Dobson, G. Stevenson, and J. Ye, "Social sensors and pervasive services: Approaches and perspectives," in *Proc. IEEE PERCOM Workshops'2011*, March 2011, pp. 525–530.

[39] E. J. Ruiz, V. Hristidis, C. Castillo, A. Gionis, and A. Jaimes, "Correlating financial time series with micro-blogging activity," in *Proc. WSDM '12.* New York, NY, USA: ACM, 2012, pp. 513–522.

[40] T. Sakaki, Y. Matsuo, T. Yanagihara, N. Chandrasiri, and K. Nawa, "Real-time event extraction for driving information from social sensors," in *Proc. CYBER'2012*, May 2012, pp. 221–226.

[41] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *Proc. WWW'2010*, pp. 851–860, 2010.

[42] "Sandy is year's No. 2 topic on Facebook," `http://edition.cnn.com/2012/10/30/tech/social-media/sandy-facebook-chatter/.` Last accessed on April 5, 2015.

[43] W.-H. Tang, M.-Y. Yeh, and A. J. Lee, "Information diffusion among users on Facebook fan pages over time: Its impact on movie box office," in *Proc. IEEE DSAA'2014*, Oct. 2014, pp. 340–346.

[44] D. Thom, H. Bosch, S. Koch, M. Worner, and T. Ertl, "Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages," *in Proc. IEEE PacificVis'2012*, pp. 41–48, Feb 2012.

[45] Twitter Statistics. `https://about.twitter.com/company`

[46] M. C. Vuran and O. Akan, "Spatio-temporal characteristics of point and field sources in wireless sensor networks," in *Proc. ICC '06*, vol. 1, June 2006, pp. 234–239.

[47] M. C. Vuran, O. B. Akan, and I. F. Akyildiz, "Spatio-temporal correlation: Theory and applications for wireless sensor networks," *Comput. Netw.*, vol. 45, no. 3, pp. 245–259, Jun. 2004.

[48] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher, "On truth discovery in social sensing: a maximum likelihood estimation approach," in *Proc. IPSN'2012*, 2012, ACM, New York, NY, USA, pp. 233–244.

[49] D. Wang, M. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. Aggarwal, R. Ganti, X. Wang, P. Mohapatra, B. Szymanski, and H. Le, "Using humans as sensors: An estimation-theoretic perspective," in *Proc. IPSN'2014*, April 2014, pp. 35–46.

VITA

KARDELEN ÇEPNİ was born in Sinop, on October 17, 1990. She received the B.S. degree in Electrical and Electronics Engineering from Koç University, Istanbul, Turkey in 2013. She received her M.S. degree in Electrical and Electronics Engineering under the supervision of Prof. Özgür B. Akan from Koç University, İstanbul, Turkey in 2015. Her research interests include social networks, social sensing and estimation theory.