Mining Social Signals in Cyber-Human Systems: Collective Behavior, Personal Health, and Modeling Methods

---

A Dissertation

Presented to

the Faculty of the School of Engineering and Applied Science

University of Virginia

In Partial Fulfillment

of the requirements for the Degree

Doctor of Philosophy (Systems and Information Engineering)

by

Congyu Wu

May 2019

Approval Sheet

This Dissertation is submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Systems and Information Engineering)

Congyu Wu

This Dissertation has been read and approved by the Examining Committee:

William T. Scherer, Committee Chair

Matthew S. Gerber, Advisor

Laura E. Barnes

Gerard P. Learmonth

Adam Slez

Accepted for the School of Engineering and Applied Science:

Craig H. Benson, Dean, School of Engineering and Applied Science

May 2019

Abstract

With increasingly advanced pervasive technology many applications centered on assessing human behavior and health stand to benefit from new data and analytics. Among the new data captured by smart technology, *social signals*, the stimuli exchanged via modes of online and offline social interactions, are promising yet under-exploited source of information to help understand and infer human outcomes. This dissertation is focused on the data mining methodologies that transform social signals data available from smart devices in daily use into human serving insights. Specifically, I focus on three major components: (1) using Twitter data and protest participation theory to forecast daily civil unrest activities during the Arab Spring, through which I demonstrate the value of theoretical underpinnings in mining online social signals for macro-level, collective behavior prediction; (2) using smartphone-based physical proximity data to improve cognitive stress recognition through two novel feature engineering methods that are applicable to generic social signals for micro-level, personal outcome inference, and; (3) the theoretical connections between inverse reinforcement learning and relational event model in discovering group social interaction dynamics, through which I broaden the scope of modeling methods for characterizing sequence of social signals in cyber-human systems. I then propose future research directions incorporating and integrating multiple sources of human-centered sensing data to contribute to aspects of personal well-being and collective good.

To my dearest grandparents, Guirong and Weilan

Acknowledgments

I have always dreamed of, but at the same time dreaded, writing the acknowledgments; because one must have accomplished something substantial to be paying tribute to other people's involvement. If I have remotely done so over the past seven years of my life, I owe it to all of you who have guided, supported, and believed in me. I thank my parents, who have been through all the ups and downs with me with their unconditional love; it is ultimately because of you that I have the courage to brave all the challenges in life. I thank my advisor Professor Matthew S. Gerber, who introduced me to the field of cyber-human systems and shaped me into a better data miner, technical writer, and scientist; it is the skills I gained working with you that have equipped me for a brighter future. I thank my committee chair Professor William T. Scherer, for being a fantastic mentor to me over the years; I could almost count on feeling enlightened and just plain better after a conversation with you. I thank the wonderful faculty members and colleagues at University of Virginia who taught me classes or with whom I worked on my research; the knowledge and wisdom you instilled in me will go long ways in my future career. I thank my friends, for all the fun times and camaraderie; it is my honor and pleasure to have navigated my mid-20s alongside you. I realize I cannot name all of you who made a positive difference on my academic and personal life during my PhD; but please know that I would not have become myself today without you.

Contents

List of Tables

List of Figures

Introduction

1.1 Social Signals in Cyber-Human Systems

Humans are social animals, constantly interacting with one another in various ways. Throughout our daily lives, we enter physical spaces occupied by other people, with whom we maintain different degrees of physical proximity. We exchange eye contact and facial expressions. We engage in conversations that are of different length, content, sentiment, and purpose. The individuals and groups of individuals we interact with are of different social significance to us, whether it is their social status, professional role, and familiarity levels. We are likely to have a smartphone or an internet-enabled computer, with which we call people, text people, email people, and receive phone calls, text messages, and emails. We use social media often, on which we communicate with people by commenting on their postings, sending them direct messages, or non-verbally, just browsing the content other people created, or registering a follow/friend request. The users we interact with are again of different significance to us, with some being public figures or organizations while some being our close friends and family. Regardless of the medium, form, and content, the cues we receive from other people in our social environment constitute **social signals**, which are events that potentially elicit cognitive or behavioral responses within us and fundamentally influence our perception, behavior, performance, and health. Therefore, the ways in which we interact with other people, such as where, when, how often, with whom, under what context, of what content, all affect and reveal various aspects of our individual and collective well-being.

Social signals are a pervasive phenomenon that researchers have attempted to define and approach in many disciplines. Biologists and ecologists have studied social signals among animals, manifested as chemical cues such as pheromones [2] and visual cues such as extravagant ornament in sexual selection [3]. Social signal processing [4] has been an emerging

research area where computer scientists and electrical engineers use sensors to detect and process human non-verbal behavioral cues, such as vocal patterns and body language, in order to understand the "unconscious channels of communication between people" [5]. Some others define social signals in a social media context: within the search engine optimization (SEO) community, social signals are annotations (e.g., likes, shares, votes) made to webpages by social media users that are visible to search engines, which contribute to the pages' search ranking [6]; social signals can also refer to the collective textual information circulated on social media [7]. An effort to conceptually define and categorize comprehensive social signals has been made [8]. I unify social signals under the notion of social stimuli, emphasizing the potential effects social signals have on human behavior and health outcomes.

We need to quantify social signals in order to understand their effects on human outcomes. Traditionally, survey method was used to solicit measurements of respondents' social interaction patterns. Researchers have also analyzed social signals from recordings of human social interaction scenarios in laboratories or controlled settings such as meeting rooms [4, 5]. As information technology becomes increasingly integrated into people's daily life, opportunities arise for capturing social signals data in natural, ecologically valid settings. The prevalence of social media allows real-time, high-resolution, online social signals data to be recorded on the population level. Smartphones, carried by many people as a companionship technology, are equipped with sensors such as Bluetooth radio and microphone that can measure users' physical proximity with other devices and voice patterns during conversations. The fusion of technology and human constitutes **cyber-human systems** (CHS), where internet, sensors, and computers are not only reflecting how we live our lives but also creating new functions and reality to improve our lives. Therefore, research into cyber-human systems aims to discover how smart, wearable, pervasive technologies can be engineered to enhance human capabilities and well-being. Endorsed by the findings from behavioral science literature on the effects of social stimuli, enabled by the growing amount of social signals data captured in in-situ cyber-human systems, encouraged by the benefit we can potentially reap from deeper

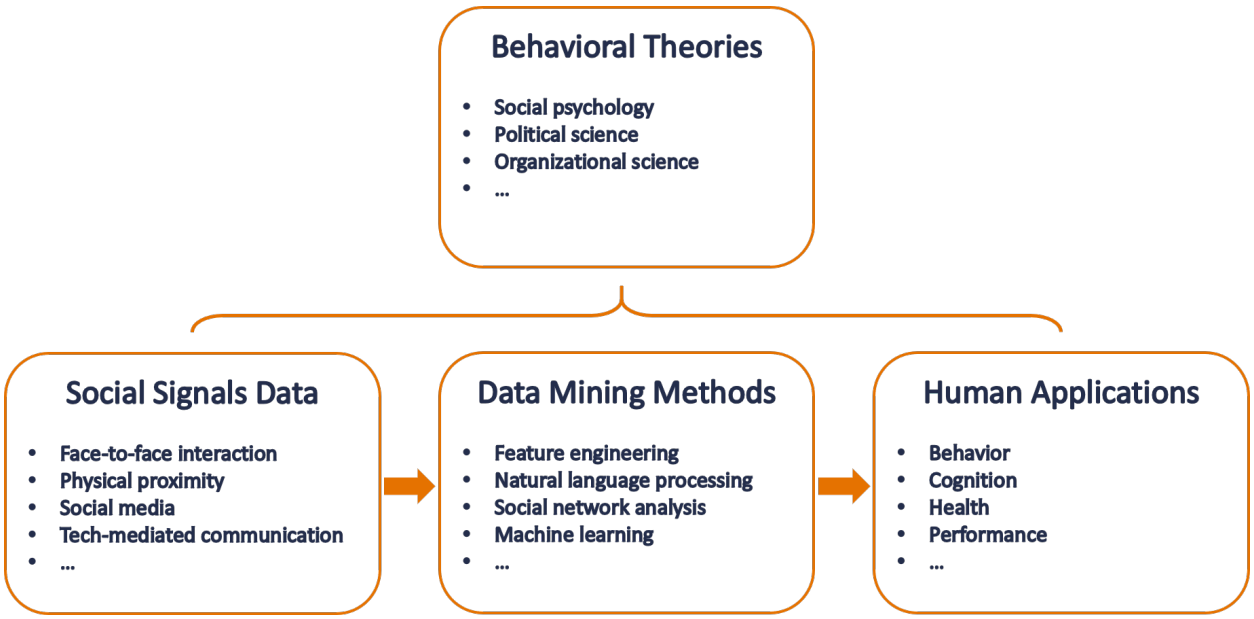


Figure 1.1: Data-Method-Application-under-Theory diagram: through this pipeline the mission of Cyber-Human Systems research is fulfilled.

understandings and sharper predictions of human behavior and health, I propose **mining social signals in cyber-human systems** as the focus of this dissertation, addressing challenges in the process of transforming social signals data into human serving insights.

I outline the landscape of research questions and tasks for mining social signals in cyber-human systems as a Data-Method-Application-under-Theory framework, shown in Figure 1.1. First, on the human side, human-centered applications are what drive the entire research pipeline, therefore we ask the question: what aspects of human well-being and capability should we better understand and improve? The answer often originates from clinical needs and life-style objectives. Second, on the technology side, data enables all research and solution, therefore we ask the questions: how do we measure social signals using technology? What technologies, hardware and software, allow the capture of social signals ranging from online communication to physical co-presence, and from face-to-face interactions to behavioral cues? How can we develop new technologies that afford more accurate, portable, and discrete measurement of social signals? Answering these questions often inspires new tasks in electrical and computer engineering. Third, to bridge the gap between technology and human, we

ask the question: how do we process and extract insights from social signals data? We often need to resort to quantitative methods from a variety of disciplines (e.g., statistics, machine learning, network science, signal processing) and continue to broaden the scope and deepen the depth of these methods to build better solutions for modeling and sense-making of social signals data, which typically falls within data science and systems engineering. Finally, findings from domain research such as social and behavioral sciences provide rich theoretical underpinnings that not only justify the entire analytical approach but also point to new opportunities in coupling pervasive technology with human outcomes. The direction of the arrows between components of Figure 1.1 indicates a working pipeline mapping from technology to human; however, it does not necessarily reflect an intellectual order, for that innovations can be identified within either component and inspire advancement in the others, which naturally facilitates interdisciplinary collaboration. In the rest of this chapter, I discuss in further details the data (Section 1.2), methods (Section 1.3), and applications (Section 1.4) of mining social signals in cyber-human systems, followed by Section 1.5 where I outline the challenges addressed in the rest of this dissertation.

1.2 Data: online and offline

Depending on whether the social signals are received through a physical space or a virtual channel, they can generally be divided into two categories: **online** and **offline**. Phone calls, emails, text messages, online messages, other people's social media postings are examples of online or cyber social signals; whereas physical proximity, face-to-face conversations, and body language are examples of offline or physical social signals. Each type of social signal entails a specific measuring instrument and reflects a particular level of technological development. Phone call records have long been utilized to understand human social networks while in-person interactions were recorded from study participants in laboratory environments. However, it is not until the recent proliferation of smartphone technology and internet-based services have we been able to take advantage of social media and mobile sensor data. **Social media** has been one of the hottest research topics in the last decade: social media sites have

been popular data sources of online social signals for researchers to draw upon to tackle wide ranges of analytical tasks and validate existing and new social and behavioral science research questions. Almost during the same time, **mobile sensing** has been offering unprecedented tools for measuring personal behavioral information (e.g., location, proximity, activity, audio, app usage), covering social signals in the physical space that were previously extremely difficult to obtain in daily life settings. As social media is usually accessible from smart devices where mobile sensors are embedded, the two modes of human-centered sensing [9] can not only be housed within the same hardware, but also considered jointly to design and implement new human-centered applications [10, 11].

Online social signals data comes from users creating content and interacting with one another via social media services. The number of users and their level of activity determine the scale of the social system and the volume of data we can observe. Twitter and Facebook are popular among choices. On Twitter, users receive social signals passively through exposure to tweets composed by other users he or she chooses to follow; and actively take part in sending and receiving social signals through actions such as following, @-mentioning, re-tweeting, and replying to a tweet. On Facebook, users also passively receive other user's content (which has more freedom in terms of format than Twitter) and have multiple actions available to interact with other users such as add friend, follow, direct message, reply, and share. Different social media platforms allow different formats of content to be posted and accommodate different social actions, which are usually annotated by timestamps and location tags, especially with more and more people accessing social media services from GPS-activated smartphones. When usage of a social media service is widespread within a population, textual content created, behaviors exhibited, and networks emerged yield interesting large scale patterns [12].

Unlike social media, each different type of passively sensed physical social signal data entails a different sensor with its own physical properties, hardware requirements, and limitations, which makes each type of physical social signal data better suited for particular occasions and scenarios [13]. Although video cameras are widely used in laboratory settings

to record pose, posture, and facial features during social actions [14], to measure social signals in the physical space in a natural, daily life setting, we are confined to the devices people carry and use throughout the day. Due to relative ease to be built into smartphones and wearable devices, Bluetooth radio, RFID (radio frequency identification), and infrared sensor are three popular sensor choices in existing work for offline social signals. Bluetooth radio is able to detect other Bluetooth enabled devices within a detection range of 10 meters [15]. Bluetooth encounters are considered evidence of physical proximity between devices [16] and often used as proxy signals for in-person interaction [17]. Research has shown strong correlation between Bluetooth encounters and real-life social interaction events [18] as well as social relationships [19]. Due to its prevalent availability in smartphones and wearables, Bluetooth encounter data is collected in most human sensing studies [20–24]. RFID is an alternative choice for physical proximity data and can be tuned to detect other RFID tags in close ranges (1 to 1.5 meters). Compared to Bluetooth radio, RFID transmitters are available in far less smartphones and require additional wireless devices such as the sociometric badges [25]; moreover, RFID requires a base station to which the contact data can be sent. As a result we see more studies conducted within shared public spaces (e.g., office) where residents wear the RFID tags while they were occupying the space [23, 26]. Unlike Bluetooth and RFID, infrared transmissions detect alignment of sight in addition to physical proximity. For an infrared sensor to detect another, the two sensors must be within a certain short distance (e.g., 1 meter) and angle (e.g., 15 degrees) from each other. As such, infrared is considered to be able to provide more accurate evidence of face-to-face interactions between people wearing a infrared-enabled badge on their chest. Common to all three sensors is a transmission rate, at which a device sends a signal into the environment in search for other devices; if another device is detected, an encounter is registered. The number of repeated detection between two devices usually indicates the length of proximity or interaction.

1.3 Methods: features and prediction

Inference of human outcomes based on social signals data involves a **feature extraction** step where raw data gets transformed into predictor variables, and a **predictive modeling** step that outputs predicted outcome values, to which supervised learning models and techniques are directly applicable. I identify four major aspects of information from social signals in cyber-human systems that warrant feature extraction.

(1) Content. We are able to observe and record the content carried by many social signals: when a user receives a message on social media, the words, sentences, and symbols form the content; when two people engage in a conversation, the sound of their voice forms the content; when someone expresses their discontent through a facial expression, the muscle movements involved form the content. Different feature engineering techniques are employed to process textual, auditory, or motor-sensory data. For textual content, natural language processing tools are especially useful: a lexical approach seeks to quantify words based on their frequency, relevance, and psychological connotation (e.g., Linguistic Inquiry and Word Count [27]) whereas as a thematic approach focuses on the co-occurrence of words and aims to discover composition of thematic clusters (e.g., Latent Dirichlet Allocation [28]) within textual content. For auditory and motor-sensory content, transformation, segmentation, and classification are often performed on raw sensor signals to extract useful annotations such as as voice vs. non-voice status from audio recordings [29] and affective states from facial expressions [30].

(2) Context. The context of a social signal refers to the environmental, behavioral, and cognitive state the recipient is in. Time, location, concurrent activity, and psychological status are all examples of context, and may be strongly indicative of the nature and effect of the social signals situated therein. Thanks to mobile sensing technology, contextual data is becoming increasingly available and should be represented in the feature extraction process.

(3) Network. Networks emerge over time as social signals transmit among individuals [26]. One may quantify the patterns of these networks through structural features established

in graph theory literature (e.g., degree, density, transitivity) and temporal features describing the time-related properties of the social signals (e.g., duration, frequency, entropy). For the same group of individuals, we may have observations of multiple networks formed by different types of social signals (e.g., social media messages and face-to-face interaction). We may aggregate these networks at different temporal resolutions to reflect the evolution of network patterns over different time frames. We may also choose to focus on different subsets of these networks, ranging from ego-centric to global when the unit of analysis varies. Moreover, statistical network modeling (e.g., Exponential Random Graph Models [31]) are formal inferential methods aimed at learning underlying factors that result in particular network structures.

(4) Sequence. Patterns exist within the temporal order of social signals received by an individual or circulated within a group of individuals. These patterns which may reveal behavioral routines and preferences, and are correlated with health and performance outcomes. Relational Event model [32] is a useful modeling framework to mine dependencies between events that take place over time and has been predominantly applied on team communication dynamics discovery problems [33]. Alternatively, Inverse Reinforcement Learning [34] has been used to extract behavioral preferences from human routine activities [35], which may be useful for mining sequence of social signals.

1.4 Applications: micro- and macro-level

The ultimate motivation for mining social signals in cyber-human systems is to help estimate and forecast human outcomes that are critical to personal and societal well-being. An important distinction within the applications lies in the scale of the unit of analysis, that is whether the outcome of interest is associated with an individual or a population. The scale of the human outcome in question stays on a micro-macro spectrum with one end being an individual, the other end being a population, and teams and organizations occupying the middle. On a micro level, personal social interaction patterns can be harnessed to inform smart health and behavior change applications. On a macro level, social media platforms can

Table 1.1: Major applications of mining online social signals

Human outcome	Scale	Features	Predictive Model
Box office revenue [40]	Collective	Volume of tweets referring to a movie per hour; sentiment polarity	Linear regression
Stock market index [41]	Collective	Values of mood polarity (Opinion Finder) and mood states (Google Profile of Mood States)	Self-organizing fuzzy neural network
Patient intake [42]	Collective	Volume of tweets containing H1N1 influenza-related terms in a region	Support vector regression
Crime rates [43]	Collective	Proportions of topic clusters detected in recent tweets at the location	Logistic regression
Election result [44]	Collective	Volume of tweets mentioning a candidate political party; LIWC sentiment scores	Proportion comparison
Citations [45]	Collective	Social media impact metrics of tweets containing links to a JMIR article	Linear regression
Political alignment [46]	Personal	Unigrams and hashtags in tweets created by a user; user cluster membership based on retweeting and mentioning	Support vector machine
Depression [47]	Personal	Behavioral attributes relating to social engagement, emotion, language and linguistic styles, ego network, and mentions of antidepressant medications	Support vector machine

be highly useful to predict emergent crowd behaviors and harvest crowdsourced knowledge [36, 37]. On a group or organizational level (or mezzo if one will), social interactions sensed within workplace teams, college classes, and hospital wards are crucial for understanding performance [24, 38, 39].

We identify a large group of works tackling with predicting collective behaviors using social media data, thanks to its broad user base often covering significant populations of cities and regions. Particularly, Twitter data has been fueling population-level predictive analytics since early 2000s, proving useful in applications spanning economics, commerce, politics,

and health. Table 1.1 shows a representative sample of such research. Features extracted are often volumes of certain content (terms, topics, messages) that is qualified as relevant. Findings typically discover that features extracted from social media data satisfactorily serve as independent predictors of real world response variables or significantly enhance the performance of traditional predictive models after being incorporated. Facebook data, on the other hand, proves more suitable for investigating personal outcomes such as personality, preferences, and mental health [48–50]. Another type of online platforms, Massive Open Online Courses (MOOC), where users browse course materials, complete required tasks, and interact with one another, are gaining popularity as a testbed for online social signals mining to understand users' learning performance and experience [51, 52].

Compared to online social signals, offline social signals captured by smartphone and wearable sensors in the physical space are more powerful in terms of detecting personal behavioral patterns due to its high-resolution recording of daily life. Physical proximity events detected among a group of people by Bluetooth radio or RFID have been used as primary evidence to infer nature of interpersonal relationship [19, 53], type of social occasion [18], and to inform strategies for infectious disease control [54]. A body of social sensing studies have shown strong correlation between the physical social signals an individual receives and emits and their personal outcomes, which I empirically categorize into four classes: (1) health outcomes, including acute physical symptoms such as common cold and influenza [55, 56], mental health such as affect and stress [57–60], and physiological indices such as body mass index (BMI) [61]; (2) cognitive outcomes, such as politics- and health- related opinions [62, 63]; (3) behavioral outcomes, such as dietary habit [56], place visit pattern [64], and physical activity [65]; (4) performance outcomes, covering academic performance [24, 66–68], workplace performance [23, 38], as well as effectiveness of communication at occasions such as job interviews [69]. As evident in current literature, physical social signals data are mostly used for personal or group level inference; applications on a population level are still very limited due to difficulty of sensor data collection on a larger scale.

1.5 Challenges addressed in this dissertation

In this section I identify three challenges within mining social signals in cyber-human systems that I help address in this dissertation.

First, in the past decade there has been wide-spread enthusiasm in the computer science community for macro-level predictive analytics using social media data. There is never a shortage of predictive tasks as one could choose a response variable of interest, train predictive models with social media data as input, and predict with a certain level of performance. However, a missing part is often the social and behavioral theories that govern the process. What do we learn? What about social media drives the predictive power? How can we utilize the findings for future decision support purposes? These questions are often overlooked. As such, the first challenge I address is in the identification of **theoretical underpinning** to a collective behavior prediction problem using online social signals data. In Chapter 2, I look into the domain of political science, specifically the early warning of civil unrest in politically unstable areas [70]. I take on forecasting daily fluctuation of civil unrest activity during the Arab Spring revolution using Twitter and automated political event data. Mapping protest participation theory [1] to textual features, I discover driving factors of predictive power in online social activity and real-time political event for daily civil unrest activity. With this effort I demonstrate a process of feature engineering and predictive modeling driven by social science theory and its value.

Second, compared to online social signals, currently we have a relative shortage of physical social signals data in terms of both variety (i.e., sensing technologies) and quantity. Existing works reviewed in Section 1.4 targeting micro-level outcomes are more or less based upon the notion of dynamic homophily [63], that is an individual's states (e.g., health, behavior, psychology, performance) are similar to, thus can be more accurately inferred if we incorporate, the states of the individual's close social neighbors. However, existing research is lacking data mining methods to utilize an individual's smart device based in-situ physical social signal data without their social contacts's ground truth data available, which is normally the reality.

To address this challenge, I make several innovations in **feature engineering** methods for interpersonal proximity data for mental health inference. In Section 3.3 of Chapter 3, I tap into the graphic aspects of Bluetooth encounter network and advance the state-of-the-art of Bluetooth feature engineering to improve real-time cognitive stress recognition. In Section 3.4 of the same chapter, I borrow insights from natural language processing and propose and validate a novel bag-of-words approach to representing raw Bluetooth encounter signals for smart mental health inference.

Lastly, regardless of the space of social signals (i.e., online vs. offline), current efforts in mining social signals data are largely focused on application-driven and data-based feature extraction but are lacking official **modeling theories** that are applicable to the more generic data structures such as network and event sequence. In Chapter 4, I look to expand the current scope of modeling theories that are suitable for learning behavioral insights from raw social signals data. Specifically, I identify inverse reinforcement learning [34] as a fitting candidate for modeling social interactions and, for the first time in both machine learning and network science community, draw theoretical connections between inverse reinforcement learning and relational event models [32], advancing both the theory and the practice of social signals mining and provide novel grounds for interventions for group communication dynamics and beyond.

This dissertation is an effort towards the advancement of cyber-human systems research, with a special focus on social signals. This dissertation covers applications, addressing two critical human outcomes, one collective, the other one personal, in the domains of political science and mental health respectively. This dissertation also covers data, as it encompasses two types of sources, spanning online and offline social signal measurements. This dissertation is about theory as well, from the social and behavioral theories guiding the cyber-human systems research pipeline (Figure 1.1), to the theoretical modeling frameworks that offer novel data mining methods for social signals.

Mining Social Signals for Macro-level Applications

Cyber social signals are relatively easy to measure and collect due to universal usage of social media. When people log onto a social media site everyday to follow people and have themselves followed, to send and receive direct messages, to like other people's statuses, and to post pictures others can see, social signals are generated and circulated among large groups of people in the cyberspace. These cyber social signals are digital traces that are potentially telltale of people's real-time behavioral and psychological states. As social media platforms are typically owned and served by commercial companies who store every piece of information ever created by users, cyber social signals are faithfully gathered and available for research, development, and marketing purposes. Collecting cyber social signals data usually entails acquiring bulk data from these commercial "gathering places", without needing to seek out information actively from individual users or bear the brunt of direct privacy-related negotiations. As a result, we have access to a relatively high quantity of fine-grained, information rich cyber social signals data.

Empowered by the availability of —and relative ease of collecting— cyber social signals data among large groups of users, there has been great interest in using cyber social signals data to help predict collective, macro-level human outcomes, such as election results, crime rates, and social movements. Researchers correlate these collective outcomes, which are associated with the population that reside in a region, with cyber social signals originating within the same population. Response variables in many application domains have been targeted and many machine learning techniques have been implemented to forecast or detect these response variables. With prediction performance a high priority, we sometimes claim victory without sufficient knowledge of where the predictive power comes from within cyber

social signals data. However, only when this question is answered can we truly understand the implications of online social behavior on offline collective outcomes and begin to social media based intervention methods to enhance social good. As such, what I want to address is how to make use of existing theories in social science literature and test them on new, big data offered by technology and create new theories and knowledge; and in the meantime improve feature engineering and prediction performance.

In this chapter, we situate this task in the political science domain, specifically a political crisis early warning problem, using social media data to predict development of civil unrest activities in the near future. Activists have used social media during modern civil uprisings, and researchers have found that the generated content is predictive of offline protest activity. However, questions remain regarding the drivers of this predictive power. We begin by deriving predictor variables for individuals' protest decisions from the literature on protest participation theory and then test these variables on the case of Twitter and the 2011 Egyptian revolution. We find significant positive correlations between the volume of future-protest descriptions on Twitter and protest onsets. We do not find significant correlations between such onsets and the preceding volume of political expressions, individuals' access to news, and connections with political activists. These results locate the predictive power of social media in its function as a protest advertisement and organization mechanism. We then build predictive models using future-protest descriptions and compare these models with baselines informed by daily event counts from the Global Database of Events, Location, and Tone (GDELT). Inspection of the significant variables in our GDELT models reveals that an increased military presence may be predictive of protest onsets in major cities. In sum, this work highlights the ways in which online activism shapes offline behavior during civil uprisings [71].

2.1 Introduction

From national political upheavals such as the protests in Thailand, Spain, Turkey, and Brazil, to revolutionary waves such as the Arab Spring and the Occupy movements that

reverberated across national borders, the 2010s have seen many parts of the world engaging in contentious politics to oppose authoritarianism and demand social change. During these civil uprisings, episodes of mass protests, riots, and even civil wars erupted sporadically, compromising civil and military operations in the affected regions. Decision makers struggled to anticipate the evolution of civil unrest and to initiate effective preparation and mitigation efforts. These difficulties motivate research into predicting near-term spikes of civil unrest. Forecasts of whether a region will transition to a significantly increased chaotic state within the near future could assist government and non-government organizations with planning and implementing response efforts that are more accurate, timely, and cost-effective [72, 73].

The above civil uprisings share a common characteristic: they were mediated by social media (e.g., Twitter). Evidence indicates that activists and other participants used social media during civil unrest to express political opinions, converse with fellow citizens, and organize future events [74–77], all of which entail the exchange of social signals in the cyberspace. Research —predominantly from computer science— suggests a predictive relationship between social media use and offline activities [78–82]. Political science studies [83–85] have investigated the correlations between an individual's engagement in offline political participation and his or her online activist behavior. The extant literature in these fields indicates the following gap: predictive analyses have not systematically identified the underlying mechanisms that drive the predictive power of social media content, whereas political science studies have focused on inferring static correlations between protest decisions and personal and behavioral attributes, neglecting the ways in which online activism dynamically affects collective offline activism. Thus, the relationship between social media usage and offline, near-term protest activity remains an open question.

We seek to bridge this gap. We first identify variables from political science literature on protest participation theory [1, 84, 86] that impact an individual's decision to participate in protests. Conventionally, these variables are measured using questionnaires. In this chapter we (1) measure these variables within tweets composed by Cairo users during the early months

of the Arab Spring, (2) examine the correlations between these variables and the occurrence of protest onset in Cairo, Egypt, and (3) use the validated variables to predict protest onsets. Our approach entails the manual curation of ground-truth data describing protest events, followed by hypothesis testing and predictive modeling involving baseline models informed by the Global Database of Events, Location, and Tone (GDELT) [87]. Thus, work presented in this chapter deepens our understanding of the role played by online social networks in modern civil uprisings.

2.2 Related Work

2.2.1 Political conflict prediction

Works on political conflict prediction have explored a host of different explanatory variables, response variables, and machine learning techniques. Predicted phenomena include political instability and dyadic international relations in future time periods. Example response variables include high or low levels of conflict in the next week [88], two powers existing in a state of cooperation or conflict [89], and whether an intra-state political upheaval will begin in the upcoming years [90]. Prediction lead-time in these studies ranges from one month to five years. Explanatory (predictor) variables fall into two broad categories: socioeconomic features and historical political events. Socioeconomic features, including regime type, GDP, demographics, state policy, and those of neighboring states [91], capture changes in the socioeconomic landscape of a region that might lead to political crisis. Historical event records, containing the occurrence of political events leading up to the present time, have been used to extract patterns of social interaction within one or between multiple countries. Methods used include predominantly logistic regression, as well as hidden Markov models, time series analysis, and agent-based simulation [92].

In the studies reviewed above, the predicted phenomena are of low temporal resolution and the prediction lead-times are longer than one month. Such methods do not address questions such as "Will there be a protest onset in [CITY] in the next three days?". This higher

temporal resolution and more granular unit of analysis motivate extracting predictors from online social media, where activist content is produced rapidly and is potentially predictive of offline crowd behavior. Researchers have predicted significant protests in the next three days using the volume of event mentions automatically extracted from social and news media [79]. Others have conducted a similar task using Twitter data and discovered that distributions over user-centric meta-information (e.g., number of followers and followees) are more predictive than content-based features such as topic proportions [81]. Others have found that the amount of inequality in hashtag distributions on Twitter on a given day is predictive of the next day's intensity of protest activity [78]. Others have focused on extracting descriptions of future protest events from social and news media content to generate alerts for potential events with specific time, location, and actor information [80]. Apart from reports of prediction performance and accounts of salient variables, a theoretical explanation of what makes social media content predictive of real world events is lacking. This motivates our integration of protest participation theory in hypothesis testing and predictive modeling.

2.2.2 Key roles of social media

Research indicates that social media played an important role during the Arab Spring uprisings [74], where its impact was partially attributed to its function as an information dissemination platform [75]. For example, active flows of revolution-related information were detected between multiple types of Twitter users including activists, bloggers, and journalists [76]. Social media provides a space for collective dissent to be articulated [77], political debates to be shaped [74] and mass protests to be organized [75]. From the perspective of resource mobilization theory, social media facilitated large-scale mobilization [93] so that the revolution happened "sooner rather than later" [75]. Moreover, evidence suggests that real-world civil unrest events often succeed spikes of online revolutionary conversations [74, 82]. As the vehicle of revolutionary conversations, opinions, and sentiment during the Arab Spring, we find locally created social media content a suitable source from which to measure features of public behaviors.

2.2.3 Political event databases

A central task in political conflict prediction is defining variables that quantify social and political events and then coding these variables within available information sources for subsequent statistical analysis. A coded event is a tuple with structured elements *who*, *did what*, *to whom*, *where* and *when* (sometimes augmented with a Goldstein score, an index of the event's impact on stability). In past work, coding an event has entailed the assignment of these labels to news articles according to a set of coding rules such as WEIS (World Event/Interaction Survey) [94] and CAMEO (Conflict and Mediation Event Observations) [95]. Coding can be performed manually by reading news articles, which is the case for WEIS, COPDAB (Conflict and Peace Data Bank) [96], ACLED (Armed Conflict Location & Event Data Project) [97], and SCAD (Social Conflict Analysis Database) [98]. The coding process is automated in systems such as TABARI (Textual Analysis By Augmented Replacement Instructions) [99], which produced the GDELT (Global Database of Events, Language, and Tone) database. GDELT is a database of political events of the form "COUNTRY A invaded COUNTRY B" that are extracted from news articles published by global news media. Publicly accessible, free-of-charge, and updated daily, GDELT serves as a candidate source for predictor variables and has been used in several studies [80, 100].

2.2.4 Protest participation theory

A number of social science theories, such as relative deprivation theory [101], political opportunity theory [102], and resource mobilization theory [103], have been proposed to account for the causal mechanisms of social movements and revolutions. These theories help explain the long-term development of revolutions, but not the near-term evolution of protests during times of political upheaval. Consider the prediction of political stability within a country. Such theories are suitable for answering questions such as "How likely is it that a revolution will take place in COUNTRY X within the next three years?". However, these theories are ill-suited to questions such as "How likely is it that a mass demonstration will

take place in CITY X within the next three days?".

Marx [104] sees revolution as an inevitable outcome of the class conflict between the exploited proletarians and the exploiting bourgeois. Tocqueville [105] attributes revolutions to people's boosted will to rebel when experiencing relaxed pressure from the authority after an endured period of oppression. Into the 20th century, many theories were proposed: (1) relative deprivation theory [101], or the so-called "J-curve" theory, arguing that it is the gap between people's expected level of well-being and what people end up experiencing, rather than the absolute level of life conditions, that causes revolution; (2) political opportunity theory [102], arguing that revolutions are a result of people's will to make changes to the society combined with people's perceived opportunity to successfully make a difference; (3) resource mobilization theory [103], which emphasizes the necessity of social organizations that provide resources such as funds, supporters, and press coverage for a revolutionary movement to happen.

On the other hand, protest participation theory aims to explain individuals' decisions to participate in protest events [86]. Variables described in protest participation theory fall into three broad categories: biographical availability, political engagement, and structural availability [1]; or —similarly— personal characteristics, political attitudes, and group effects [86]. Biographical availability indicates the absence of constraints (e.g., marriage, children, or employment) that would increase the cost of protest participation. Political engagement has several dimensions, including one's interest and knowledge in politics as well as one's political efficacy, the belief that one can make a difference politically. Structural availability refers to an individual's "presence in an interpersonal network that facilitates recruitment to activism" [1]. The most salient manifestation of such presence is being asked to or knowing to participate in a protest event. The structure of protest participation theory can be summarized as *can participate*, *wants to participate*, and *has been asked to* or *knows to participate* [106], corresponding to the three categories of variables, respectively. Individuals' propensity to participate collectively determines the occurrence and scale of civil unrest events.

Table 2.1: Significant variables in protest participation theory [1]

Category	Variable
Biographical availability	Being young
	Without children
	Educated
Political engagement	Knowledgeable in politics
	Interested in politics
	Being liberal
Structural availability	Asked/knows to participate
	Affiliated with social organizations
	Having civic skills

Table 2.1 provides a list of variables associated with each category in protest participation theory. Note that, although these variables can vary over time, some do so at faster rates than others, especially when aggregated for a population. For example, the level of interest in politics within a population can change overnight as a result of a provocative event, whereas a population's measured age distribution is likely to remain steady for years, leaving the variable "being young" unaffected over a short time frame. Given the real-time nature of our prediction task, we require the predictors in our models to track the rapid evolution of civil unrest. We identify four candidate variables from protest participation theory: knowledgeable in politics, interested in politics, having been asked to participate, and affiliated with social organizations, and we use these variables to drive the design of predictors in our statistical models. "Affiliated with social organizations" is included because although official membership in an organization does not often change overnight, an internal affiliation (e.g., an individual's political alignment with certain social actors) can.

2.3 Hypotheses

We define a **protest onset** to be the commencement of one or more successive days of gatherings that are politically motivated in the same way. Given this definition, we seek to test the following hypotheses, that during a period of political upheaval:

- **Hypothesis 2.1**: A protest onset is more likely to happen in the near future if we observe an increase in the collective level of political knowledge in the cyberspace.

- **Hypothesis 2.2**: A protest onset is more likely to happen in the near future if we observe an increase in the collective level of political interest in the cyberspace.

- **Hypothesis 2.3**: A protest onset is more likely to happen in the near future if we observe an increase in the collective level of knowledge of a future protest in the cyberspace.

- **Hypothesis 2.4**: A protest onset is more likely to happen in the near future if we observe an increase in the collective level of affiliation with political activists in the cyberspace.

We hypothesize that the timing of such onsets will correlate strongly with increases in one or more of these collective levels. Factors originating from outside of the social system in question, such as intervention from a foreign power, may directly or indirectly impact the development of civil unrest and thus be predictive of future events. We will focus on endogenous characteristics of a population derived from protest participation theory, leaving such exogenous factors to future work.

To formalize the concepts "increase in the collective level" and "near future", we define two parameters: *base period* and *prediction horizon*. The base period (denoted B) is a period of time immediately preceding the current time step; we compare the current value of a predictor with the average value over the base period to measure the increase. The prediction horizon (denoted H) is a period of time immediately following the current time step (i.e., "near future"); a predicted protest onset will be considered correct if a protest onset does in fact occur within the prediction horizon. Both parameters take positive integer values. In this article, we will explore different combinations of $B \in \{1, 2, 3\}$ and $H \in \{1, 2, 3\}$ in our modeling process.

Table 2.2: Manually curated timeline of Cairo protest onsets 12/1/2010-4/1/2011

Date	Day	Main Venue	Scale*	Theme/Demand	Narrative
12/12/2010	Sunday	Supreme Court building	2	Protest against the cheating in recent election	Hundreds of opposition activists protested Sunday over what they said were bogus elections that had produced an illegitimate parliament
1/2/2011	Sunday	Shubra, downtown Cairo	2	Denounce Saturday's church bombing in Alexandria and show solidarity with the Egyptian Coptic minority	Some 500 Muslim and Coptic activists, politicians and other civil society leaders protested
1/25/2011	Tuesday	Tahrir Square	3	"Day of Revolt": urge Hosni Mubarak to stepdown	Thousands of people took part in rare anti-government protests
2/18/2011	Friday	Tahrir Square	5	Celebrate the first week after the downfall of Mubarak and pay tribute to the people who died in the uprising	Hundreds of thousands of protesters returned to Tahrir Square in a mass rally
2/25/2011	Friday	Tahrir Square	3	Urge the new military government to purge the cabinet of ministers appointed by Mubarak	Thousands of protesters have gathered once again in Cairo's Tahrir Square
3/4/2011	Friday	Tahrir Square	4	Show determination of Ahmed Shafiq's stepdown and then celebrate the appointment of Essam Sharaf as new prime minister	Ten thousand people gathered as the newly appointed prime minister Essam Sharaf spoke to them asking for support and help
3/7/2011	Monday	Multiple locations	3	Against the burning of the church in the province of Helwan	Coptic Christians held protests in different areas of Cairo to demand better treatment and an end to what they perceive as discrimination in Egypt
3/22/2011	Tuesday	Interior Ministry	3	Demand better pay and conditions	About 3,000 police protested outside Interior Ministry building
3/25/2011	Friday	Central Cairo	2	Demand more reform	Hundreds of people gathered Friday in central Cairo in the familiar tableau of chants and slogans demanding reform
4/1/2011	Friday	Tahrir Square	3	Demand the ruling military council move faster to dismantle lingering aspects of the old regime "save the revolution"	Thousands of demonstrators filled Tahrir Square on Friday for the largest protest in weeks

*Number of participants in the power of 10

2.4 Data

This section describes data acquisition for our response variable, protest participation theory variables, and GDELT variables. The response variable is a binary variable indicating whether a protest onset occurred in Cairo, Egypt on a particular day of the Arab Spring. Protest participation theory variables have been introduced in Section 2.2.4 and formalized in Section 2.3 and are the main focus of this research. We seek to measure protest participation theory variables automatically within Twitter data. GDELT variables will serve as predictor variables in our baseline models, against which we compare the predictive performance of the protest participation theory variables. The temporal scope of our data is the 121-day period from 1 December 2010 to 31 March 2011, covering the early months of the Arab Spring.

2.4.1 Ground truth

We considered two alternatives for establishing the ground truth of our response variable: (1) utilizing automatically curated event databases like GDELT, and (2) manually examining news articles. Each alternative aims to produce a vector of ground-truth labels indicating whether a protest onset occurred in Cairo during each day, and each alternative is grounded in news articles that are written by humans. The advantage of GDELT is convenience, as it is automatically extracted from news articles. Researchers have used GDELT as a proxy for ground truth by aggregating daily counts of GDELT protest events on a three-day basis and labeling statistical outliers as positive instances [81]. In addition to concerns related to errors introduced by GDELT's automatic coding process, we find GDELT incompatible with our definition of protest onset. For example, sustained daily GDELT records of protest events might result from continued coverage of a single protest. This is a single onset by our definition, but GDELT does not make such distinctions. Manual examination of news articles can help differentiate such patterns. This manual analysis requires considerable human effort, but we believe the resulting ground truth will be better aligned with our stated response variable. As such, we decided to obtain Cairo protest onset data through manual examination

Table 2.3: Predictors of Protest Onsets

Theoretical Variable	Implemented Predictor	Notation (index)
Knowledgeable in politics	Daily count of Cairo tweets @-mentioning popular news media	NumTweetsNews (1)
	Daily count of Cairo Twitter users @-mentioning popular news media	NumUsersNews (2)
Interested in politics	Daily count of Cairo political tweets	NumTweetsPolitics (3)
	Daily count of Cairo Twitter users who authored at least one political tweet	NumUsersPolitics (4)
Having been asked or presented information to participate	Daily count of Cairo tweets that present future protest information	NumTweetsProtests (5)
	Daily count of Cairo Twitter users @-mentioning the authors of future protest tweets	NumUsersProtests (6)
Presence of ties in an activist network	Daily count of Cairo tweets @-mentioning salient Egyptian political activists	NumTweetsActivists (7)
	Daily count of Cairo Twitter users @-mentioning salient Egyptian political activists	NumUsersActivists (8)

of online news articles reporting the daily progress of the Egyptian revolution.

We considered the 121-day period from 1 December 2010 to 31 March 2011. Associated with this time span is a vector of 121 boolean indicators, which are 1 if a protest onset occurred during the given day and 0 otherwise. To determine the boolean value of each day, we examined news articles published online by global news media. For each day within the period, we queried Google News (http://news.google.com) with the date in question and *Cairo* as the query term. From the query results, we examined articles containing descriptions of protest activity in Cairo on that day. When descriptions of a politically charged gathering occurred in at least one news article, we documented the location, theme (motivation or demand of the gathering), scale (number of participants by magnitudes of 10), a short narrative of the gathering, and the URL of the news article. Then for each day within the period, we determined whether the day was a protest onset day based on our definition: a day was labeled 1 if protest activity was reported on that

day and the immediately preceding day has no evidence of protest activity with the same political motivation or at all. For example, on 25 January 2011, "Thousands of protesters spilled into the streets of Cairo on Tuesday, an unprecedented display of anti-government rage..."[1]. On 25 February 2011, "Thousands of Egyptians have returned to Cairo's Tahrir Square to mark one month since the start of their uprising which toppled President Hosni Mubarak"[2]. In total, 10 protest onsets were identified, and 9 of them were within the 121-day period, with one more on 1 April 2011 during which "Thousands of demonstrators gathered in Cairo's famous Tahrir Square on Friday"[3]. Our response data is listed in Table 2.2, with a more detailed spreadsheet showing curation effort downloadable at https://www.dropbox.com/s/ssp3rsh3e7983z1/cairo_protest_onsets_data.xlsx?dl=1.

2.4.2 Twitter

In Section 2.2.4, we discussed biographical, political, and social structure variables that might explain the occurrence of protest onsets. We measured these variables within Twitter data generated by Cairo Twitter users from 1 December 2010 to 31 March 2011. GPS coordinate information is the most accurate way to ascertain the location of a tweet; however, such metadata was rarely attached to tweets posted within Egypt in the 2010-2011 time period; instead, we identified a user as residing within Cairo if the user had the English city name "Cairo" in the *bio-location* field of their user profile. We purchased this subset of tweets (3,661,036) from Gnip. In addition to tweet text and timestamp, the dataset also contains user profiles of all users. Section 2.5 will discuss in detail how we extracted predictors from this Twitter data.

2.4.3 GDELT

We used GDELT data to build baseline models for comparison with our Twitter-driven models. For every detected event within a news source, GDELT assigns an event type based on the CAMEO coding scheme. The CAMEO coding scheme categorizes an action into one of

[1] http://news.blogs.cnn.com/2011/01/25/thousands-protest-in-rare-cairo-mass-uprising/
[2] http://www.bbc.com/news/world-middle-east-12583189
[3] http://www.nytimes.com/2011/04/02/world/middleeast/02egypt.html?_r=4&

the 20 root event types and assigns additional labels if finer distinctions are detected. The 20 root event types are: (1) make public statement; (2) appeal; (3) express intent to cooperate; (4) consult; (5) engage in diplomatic cooperation; (6) engage in material cooperation; (7) provide aid; (8) yield; (9) investigate; (10) demand; (11) disapprove; (12) reject; (13) threaten; (14) protest; (15) exhibit military posture; (16) reduce relations; (17) coerce; (18) assault; (19) fight; (20) engage in unconventional mass violence. Additional digits specify subcategories under a root event type. For example, code 03 will be assigned to an "express intent to cooperate" root event type; code 033 will be assigned to an "express intent to provide not specified material aid" event; and code 0333 will be assigned to an "express intent to provide humanitarian aid" event.

We extracted from GDELT counts of events that occurred in Egypt aggregated by day and by root event type within our study's time period. We did this by querying the database for all events tagged with an *ActionGeoCode* of "EG", which is the Federal Information Processing Standard (FIPS) code for Egypt. This process was automated by the R package *GDELTtools*. Note that the daily event counts are not directly used as predictor values for the baseline models; we will discuss baseline model feature design in Section 2.2.1.

2.5 Feature Engineering

This section presents the design of features that capture the protest participation hypotheses laid out in Section 2.3. We also explain how these features are measured using the Twitter data described in Section 2.4.2. Table 2.3 lists the theoretical variables and associated predictors derived from our Twitter data. In the following paragraphs we describe the four groups of features in detail.

First, we used the volume of user mentions that target popular news media Twitter accounts to approximate the collective level of political knowledge among Cairo users. On Twitter, user mentions (also known as @-mentions) are used to retweet or reply to another user's tweets, or to engage another user in a conversation. We observed that the vast majority of tweets @-mentioning a news media account are retweets of news posted by said news media

account, for example: "RT @AJEnglish: Egypt's protest dispersed by force: Army uses batons to break up demonstrations in Cairo demanding purging..."[4]. The rest are predominantly tweets discussing the @-mentioned news media, for example: "@AJEnglish You really have the corrupt #Egyptian government running scared. BRAVO! If only more news sources were as good as you. #Jan25"[5]. In previous work, political knowledge has been measured by asking individuals whether they read a daily news article [1]. We broaden this notion to encompass access to information sources (e.g., Twitter accounts) that frequently disseminate political stories. We reason that @-mentioning the Twitter account of popular news agencies is an online analog of accessing a traditional newspaper, since the user's @-mention was likely precipitated by content distributed by the news agency.

Second, one's interest in politics is reflected by the extent to which one engages in political discussions. Although an interest in politics does not necessarily generate political discussion, the volume of political tweets one creates may proxy for the user's level of political interest. Therefore, we approximated the collective level of political interest on each day by measuring the volume of political tweets created by Cairo users on that day. We will explain the selection of political tweets later in this section.

Third, we observe that protest advertisements and exhortations to participate often contain basic information about an upcoming protest event and the mention of a future day. The volume of such tweets indicates awareness and social pressure to engage in an upcoming protest.

Last, we mapped the notion "ties in an activist network" to the volume of @-mentions that target salient political activists on Twitter. The @-mention is a primary means whereby users interact with each other; therefore the frequency of @-mentioning a political activist represents the degree of affiliation. Note that so far we have used the word "volume" to describe quantity (e.g., the volume of @-mentioning certain users); concretely, this quantity

[4]Tweet link deprecated, original news article: https://www.aljazeera.com/news/middleeast/2011/02/2011226221957428.html

[5]http://twitter.com/lollybrubs/statuses/33213052178403329

Table 2.4: Query terms for political tweets selection.

Terms without Arabic equivalent	
jan25, 25jan	dostor2011[4]
feb17	amndawla[5]
mar19	newegypt
17feb	freeegypt
19mar	egyrevolt

Terms in both English and Arabic	
tahrir ميدان التحرير	essam[1] عصام
shafiq, shafi2, shafik[2] شفيق	sharaf[1] شرف
mubarak مبارك	council المجلس
army الجيش	suleiman[3] سليمان
vote التصويت	military العسكري
protest مظاهرة المتظاهرين المتظاهرون	police الشرطة شرطة
revolution ثورة	constitution الدستور دستور الدستورية
liberation التحرير	wael[6] وائل
amendment التعديلات للتعديلات	ghonim[6] غنيم
referendum الاستفتاء	minister الوزراء
security أمن	government حكومة
thugs البلطجية	baradei[7] البرادعي

[1]Essam Sharaf, Prime Minister of Egypt (3 March 2011 - 7 December 2011)
[2]Ahmed Shafiq, Prime Minister of Egypt (18 September 2002 - 28 January 2011)
[3]Omar Suleiman, Vice President of Egypt (29 January 2011 - 11 February 2011)
[4]Egyptian Constitutional Declaration of 2011
[5]The Egyptian State Security Investigation Service
[6]Wael Ghonim, Egyptian Internet activist and computer engineer
[7]Mohamed ElBaradei, Egyptian law scholar and diplomat.

could be represented by both the number of tweets and the number of users that satisfy a constraint (e.g., the number of tweets @-mentioning certain users vs. the number of users that have created tweets @-mentioning certain users). We did not have reasons to choose one over the other so we created both as measurements of the same concept and left the selection to up to the statistical models.

To measure the predictors described above, we first selected Cairo political tweets from our data set, as these tweets and associated users are the basis for all of our predictor

measurements. For example, our identification of popular news media and salient political activists (which we collectively call political Twitter users) is based on the activity of users who generated political tweets. To select political tweets, we adopted a keyword match method with query expansion. Since we did not know exactly how to characterize Cairo users' online political speech, we began with a list of four terms, *jan25*[6], *tahrir*[7], *mubarak*[8], and *amndawla*[9], which were the four most frequently used hashtags of clear political nature for the Egyptian protests. We selected the tweets containing these four hashtags, aggregated the selected tweets by day, and built a corpus of 121 tweet documents each of which was the concatenation of tweets on that day. We then calculated the average term-frequency inverse document-frequency (TF-IDF) score for each word in the corpus and ranked them in descending order. Most of the top-ranked terms, both in English and Arabic, are related to politics. From the top-ranked 300 words in the list, we manually selected words representing (1) entities of political nature such as revolution, government, and vote; (2) names of important political actors such as suleiman and ghonim; (3) dates and locations of revolutionary events such as tahrir, jan25, and feb17; and (4) political slogans such as newegypt and freeegypt. Note that among these political terms there are some English words and Arabic words that share the same meaning; for those words that were only in one language (English or Arabic), we obtained the translation (when applicable) in the other language using Google Translate and added the translated words to the list. In this way, we curated a list of 67 query terms (listed in Table 2.4). We used this list of query terms to select political tweets from our full collection of tweets, producing 946,988 Cairo political tweets (roughly 26% of all Cairo tweets). We selected Cairo tweets that contain at least one of the query terms when not used as a username being @-mentioned (e.g., "@ghonim" is not considered as containing the query term "ghonim"). We then measured our predictors using these political tweets.

[6]25 January 2011, the official start date of the Egyptian Revolution of 2011

[7]Tahrir Square, a major public town square in Downtown Cairo that was the location and focus for political demonstrations in Cairo

[8]Hosni Mubarak, the then-president of Egypt in 2011 Spring

[9]The Egyptian State Security Investigations Service, which was the highest national internal security authority of Egypt

Table 2.5: Most mentioned news media by political tweets

Twitter account	Real name
Shorouk_News	Al-Shorouk
AlMasryAlYoum_A	Al-Masry Al-Youm, "the Egyptian Today"
DostorNews	Al-Dostor, "the Constitution"
RassdNews	Rassd News Network
AlArabiya_Ar/alarabiya_ar	Al Arabiya Arabic
ONtveg	ONTV Egypt
MasrawyFans	Masrawy
Youtube	Youtube
eahram	Al-Ahram
AJArabic	Al Jazeera Arabic

Table 2.6: Most mentioned political activists by political tweets

Twitter account	Real name
Ghonim	Wael Ghonim
alaa	Alaa Abdel Fattah
ElBaradei	Mohamad ElBaradei
waelabbas	Wael Abbas
3arabawy	Hossam el-Hamalawy
Sandmonkey	Mahmoud Salem
monasosh	Mona Seif
wael	Wael Khalil
Zeinobia	Zeinobia
gamaleid	Gamal Eid

To measure the daily volume of @-mentioning popular news media (predictors 1 and 2) and @-mentioning salient political activists (predictors 6 and 7), we identified the key news media and political activist Twitter accounts active during the revolution. To this end, we extracted a list of all @-mentioning instances from the collection of Cairo political tweets. We then aggregated the list by mentioned user and sorted it in descending order of the number of times (i.e., tweets) each user was mentioned. From the highest ranking user down, we examined each user's profile on Twitter and manually identified the 10 most mentioned news media accounts and the 10 most mentioned political activist accounts. A complete list of

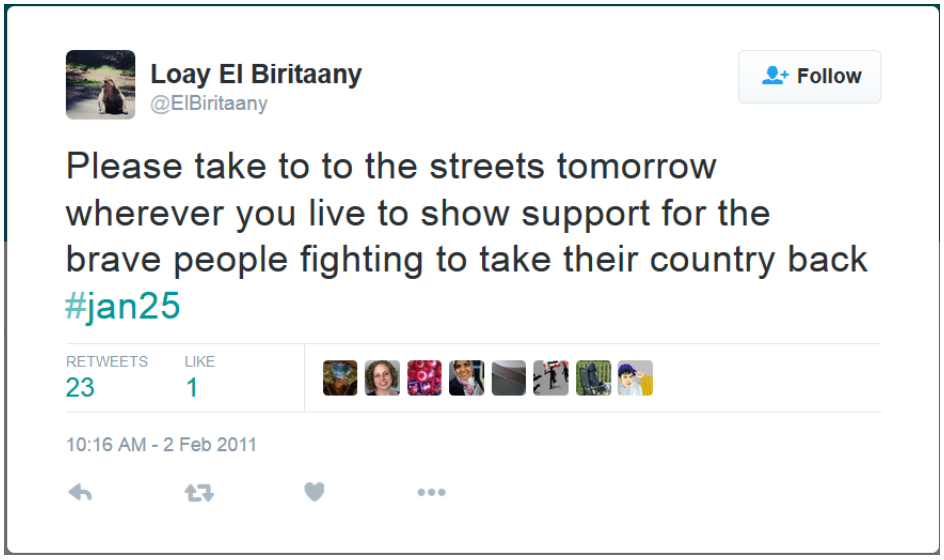


Figure 2.1: Example of protest advertising tweets.

these news media and political activists are shown in Table 2.6. Tweets that @-mentioned these news media and political activists accounts were used to measure predictors 1, 2, 6, and 7.

We adopted a keyword matching method to measure the number of tweets that present future-protest information (predictor 5). From Cairo political tweets we selected those that simultaneously contain protest related words and future-oriented words and measured their daily volume. We used the following English words related to protests: *tahrir*, *marching*, *demonstration*, *protest*, their Arabic translations, and the English phrase *to the streets*. Future-oriented words were English and Arabic words and phrases for *tomorrow*, and combinations of *this* and *next* with any one of the days of the week (e.g., this Friday, next Tuesday, etc.). Figure 2.1 shows an example protest-advertising tweet.

2.6 Experiments and Results

2.6.1 Hypothesis testing

To test our hypotheses, for a given day, we modeled the probability of a protest onset occurring in Cairo within the next H days (denoted as O_H, $H \in \{1, 2, 3\}$) given the feature vector $\boldsymbol{X}_B$ ($B \in \{1, 2, 3\}$) as

$$Pr(O_H|\boldsymbol{X}_B) = f_H(\boldsymbol{X}_B) = \frac{1}{1+\exp[-(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{X}_B)]}, \tag{2.1}$$

where the feature vector $\boldsymbol{X}_B$ was obtained through two processes. First, to quantify the "increase over recent level" in the hypotheses formulated in Section 2.3, we calculated the quotients of the current value of each of the eight predictors in Table 2.3 divided by the predictors' average values over the past B days. Thus, under each level of B, we created eight features. We then implemented a stepwise variance inflation factor (VIF) selection process on these features to resolve multicollinearity, which is a concern due to the similarity between the pairs of features that originate from the same theoretical variable (e.g., predictor 1 and 2 in Table 2.3). The process is carried out as follows: for each feature created in the first process, (1) we calculate VIF and discard the feature with the highest VIF; (2) calculate VIF for the remaining features again and discard the feature with the highest VIF; (3) repeat (1) and (2) until all existing features have a VIF $\leqslant 5$ (a conventional cutoff). As a result of the two processes, we obtained a feature vector under each $B \in \{1, 2, 3\}$. We denote these feature vectors as $\boldsymbol{X}_B$. Variables that remain after the stepwise VIF selection process are shown in Table 2.7.

Given a vector of predictor variables, VIF of the j-*th* predictor variable is defined as

$$VIF_j = \frac{1}{1 - R_j^2}, \tag{2.2}$$

where

$$R_j^2 = 1 - \frac{\Sigma_i (x_{ij} - f(\mathbf{x}_{i\bar{j}}))^2}{\Sigma_i (x_{ij} - \overline{x_{ij}})^2}, \tag{2.3}$$

and where $f(\mathbf{x}_{i\bar{j}})$ is the value fitted by a linear regression model with the j-*th* predictor variable as response and all other variables as predictor variables at the i-*th* observation. Thus, VIF quantifies the extent to which a predictor is explained by other predictors (i.e., the level of multicollinearity).

Table 2.7: Feature vectors $\boldsymbol{X}_B$ after stepwise VIF selection

Features	$\boldsymbol{X}_B$		
	$B = 1$	$B = 2$	$B = 3$
NumTweetsNews	✓		
NumUsersNews			
NumTweetsPolitics	✓	✓	✓
NumUsersPolitics			
NumTweetsProtests	✓	✓	✓
NumUsersProtests	✓	✓	✓
NumTweetsActivists		✓	✓
NumUsersActivists			

Table 2.8: Variables in $f_H(\boldsymbol{X'}_{B,H})$ under each configuration of base period and prediction horizon. Variables with a p-value smaller than 0.05 are shown in ***bold italic***. Variable names are followed by their coefficients (β) and p-values (p).

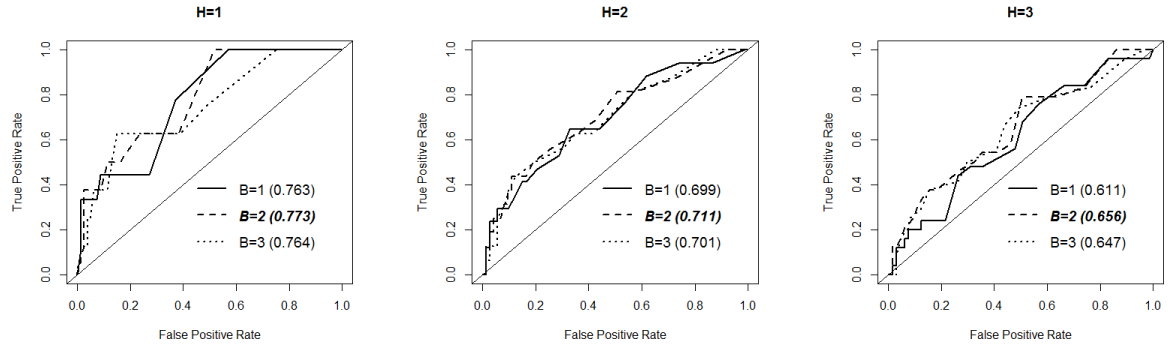
Base Period	Predicting Horizon	$\boldsymbol{X'}_{B,H}$	β	p
B = 1	H = 1	***NumTweetsProtests***	0.422	0.002*
		NumTweetsPolitics	-4.76	0.107
		NumTweetsNews	1.016	0.165
	H = 2	***NumTweetsProtests***	0.399	0.005*
		NumTweetsPolitics	-1.045	0.286
	H = 3	***NumTweetsProtests***	0.208	0.042*
B = 2	H = 1	***NumTweetsProtests***	0.596	0.002*
	H = 2	***NumTweetsProtests***	0.438	0.008*
	H = 3	***NumTweetsProtests***	0.259	0.05*
B = 3	H = 1	***NumTweetsProtests***	1.046	0.001*
		NumTweetsPolitics	-3.415	0.099
		NumUsersProtests	-0.009	0.148
	H = 2	***NumTweetsProtests***	0.673	0.006*
		NumTweetsPolitics	-1.446	0.232
		NumUsersProtests	-0.008	0.294
	H = 3	NumTweetsProtests	0.289	0.116
		NumUsersProtests	-0.007	0.494

Having fitted $f_H(\boldsymbol{X}_B)$ for each configuration of B and H, we used backward stepwise selection [107] to select the best features based on the Akaike information criterion (AIC). We denote the feature vector selected by stepwise regression under base period B and prediction horizon H as $\boldsymbol{X'}_{B,H}$ and the corresponding logistic regression that models $Pr(O_H|\boldsymbol{X'}_{B,H})$ as $f_H(\boldsymbol{X'}_{B,H})$. The variables, coefficients, and p-values in each $f_H(\boldsymbol{X'}_{B,H})$ are shown in Table 2.8. Significant variables are shown in bold italic.

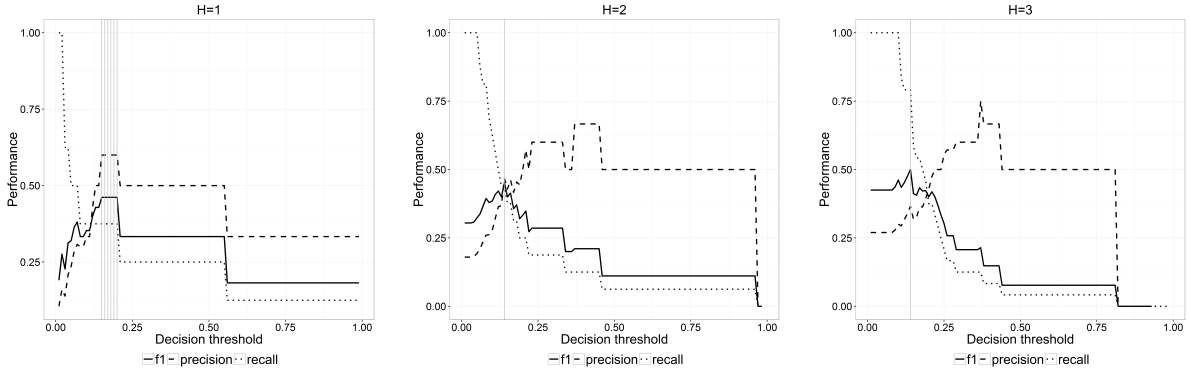
Evidently, the number of tweets presenting future-protest information (*NumTweetsProtests*) is the most significant positive factor, with its p-value smaller than 0.05 under all configurations of H and B except when $H = B = 3$. This suggests that an increase in the volume of information about future protest events on social media often accompanies increased likelihood of protest onset within a few days. To give a numeric example, when a 10% increase in the amount of protest-advertising tweets over the past three days (i.e., $B = 3$) is observed on a day, the odds of a protest onset occurring in the next day (i.e., $H = 1$) increases by $\exp(0.1 * 1.046) - 1 = 11.0\%$. This finding supports hypothesis H2.3: an increase over the recent collective level of knowledge of a future protest is correlated with the occurrence of a protest onset in the near future. None of the other features proved significant at a significance level of 0.05 in any of the predicting schemes. As such, we do not have evidence to support hypothesis H2.1, H2.2, and H2.4. The lack of evidence to support H2.1 corroborates the findings of [84] that using social media for news seeking purposes is not significantly correlated with an individual's decision to protest. The work of [84] also indicates that an individual who expresses political opinions on social media is more likely to protest; however, an increased level of non-protest-specific political discussion on Twitter failed to significantly translate into offline protest onsets in the near future, as H2.2 was not supported in our results.

2.6.2 Predictive modeling

In addition to hypothesis testing, we also used the variables to make predictions of future protest onsets and examined their predictive power. Since *NumTweetsProtests* was the only



(a) Protest Advertisement models (b) Protest Advertisement models (c) Protest Advertisement models



(d) Best thresholds: 0.15-0.2 (e) Best threshold: 0.14 (f) Best threshold: 0.14

Figure 2.2: ROC curves for the protest advertisement models (panels (a), (b), (c)) under each prediction horizon $H \in \{1, 2, 3\}$ and base period $B \in \{1, 2, 3\}$; the corresponding AUC values are shown in parentheses in the legend, with the highest value shown in ***bold italic***. Precision-recall trade-off plots for the best performing main models under each H are shown in panels (d), (e), and (f).

statistically significant variable, it was the only predictor used in our models. We situated our model on a day within the 121-day period and made predictions of protest onsets after that day based on all data available up to that day. Specifically, for each (B, H) where $B \in \{1, 2, 3\}$ and $H \in \{1, 2, 3\}$, we treated the first 30 observations as the initial training data (containing two protest onsets) and we trained a logistic regression model using one predictor *NumTweetsProtests*, which we call the protest advertisement model —denoted $f_H(NumTweetsProtests_B)$— to predict the outcome of the 31st observation. We then moved one day forward by adding day 31 to the training set, retraining the model on days 1-31, and making a prediction for day 32. This reflects a natural setup in which a daily prediction is

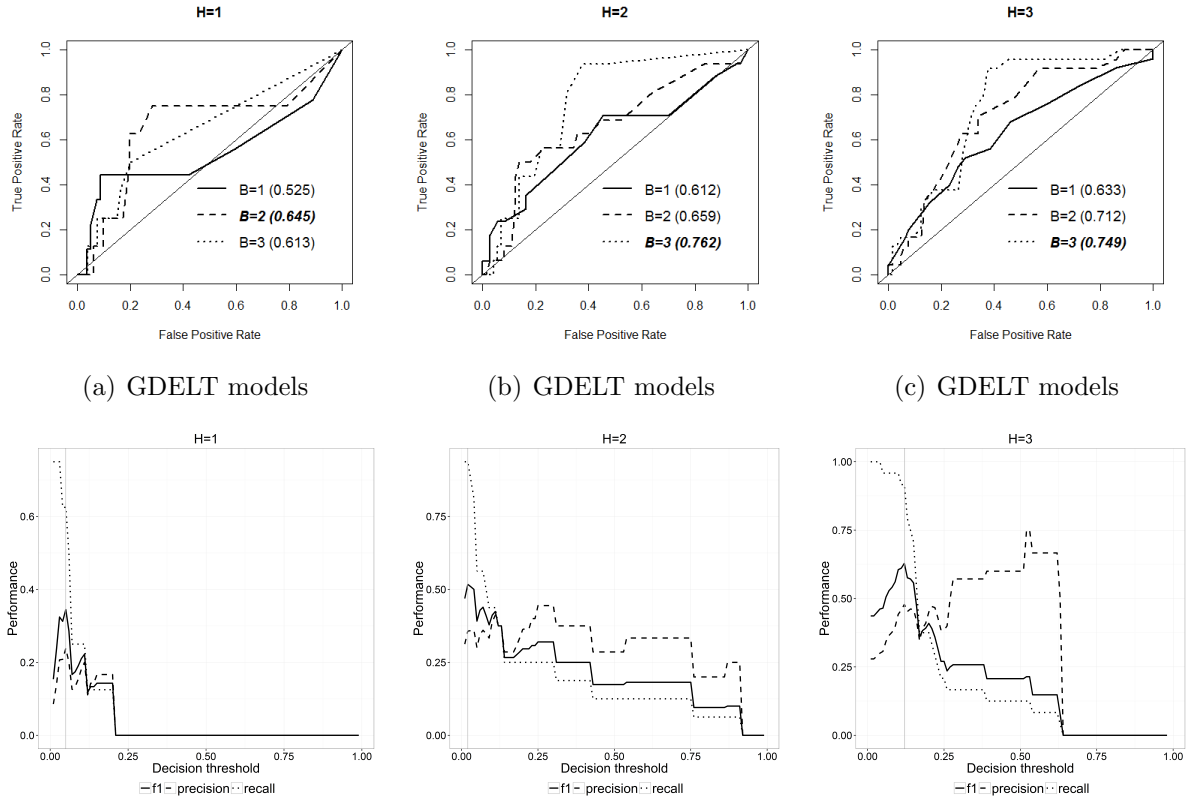
made using all available historical data. We used the first 30 days as the initial training period to obtain at least one positive observation. Predictions are compared with the protest onset data introduced in Section 2.4.1. Panels (a), (b), and (c) of Figure 2.2 show the Receiver Operating Characteristic (ROC) curve and the area under ROC curve (AUC) of each model trained and tested according to the experimental setup described above. The ROC curves are grouped by prediction horizon H and the associated base period B. The AUC is highlighted in bold italic where a model achieved the highest AUC. We denote the best performing protest advertisement model under a prediction horizon H as $f_H(NumTweetsProtests_{\hat{B}})$, where $\hat{B}$ is the B at which $f_H(NumTweetsProtests_B)$ achieves the best performance under a prediction horizon H. We see that $\hat{B} = 2$ under all three levels of H. This suggests the existence of an optimal choice of base period. Judging from the highest AUC score, $f_H(NumTweetsProtests_{\hat{B}})$ performs better when the prediction horizon H is shorter, which is an intuitive pattern also observed in [79].

Although ROC curves and the AUC values indicate model prediction performance, one does not specify a decision threshold for the predicted probabilities; rather, ROC curves are plotted by traversing all decision thresholds ranging from 0 to 1. Deciding the decision threshold to use when applying the model in practice is not a trivial problem. To investigate this, we plotted in panels (d), (e), and (f) of Figure 2.2 a precision-recall plot for our best performing models under each prediction horizon H. These figures show the trade-off between precision and recall and locate the best decision threshold for each of our best performing models. We defined the optimal threshold as the location of highest F1 score, calculated as the harmonic mean of the precision and the recall: $F1 = 2 * (precision^{-1} + recall^{-1})^{-1}$. We discovered that for the three prediction horizons $H = 1, 2, 3$, the best thresholds were 0.15-0.2, 0.14, and 0.14, respectively.

Next, under each prediction horizon, we compared the prediction performance of $f_H(Num\ TweetsProtests_{\hat{B}})$ with that of baseline models trained with GDELT features. As discussed in Section 2.4.3, we obtained the daily counts of 20 types of events in Egypt from GDELT

from 1 December 2010 to 31 March 2011. For each of the 20 event types, we computed the ratio between a day's event count and the average event count over the past B ($B \in \{1, 2, 3\}$) days (to match the "increase over recent level") and used it as a GDELT feature. Then we used stepwise VIF selection and AIC-based backward stepwise logistic regression (as described in Section 2.6.1) to choose the best set of GDELT features under each configuration of B and H. The selected GDELT features along with their coefficients and p-values are shown in Table 2.9. We will discuss further details about this table in Section 2.7. Finally, we trained and tested GDELT models using the selected best GDELT features following the same experimental setting described in Section 2.6.1, retaining the models that achieved the best prediction performance (AUC) under each prediction horizon.

Panels (a), (b), and (c) of Figure 2.3 show the ROC curves and corresponding AUC values of our selected GDELT models. Panels (d), (e), and (f) show the precision-recall trade-offs and the best decision thresholds for the best performing GDELT models under each configuration of prediction horizon H. It is interesting to observe that, unlike the previous models, the best performance of the GDELT models when $H = 2$ and $H = 3$ is remarkably better than that achieved when $H = 1$, which is a pattern contrary to that of the protest advertisement models. Furthermore, the performance competition between the previous models and the GDELT models is contingent upon the prediction horizon chosen (shown in Figure 2.4). When $H = 1$, our best performing protest advertisement model outperforms the best GDELT model by $(0.773 - 0.645)/0.645 = 19.8\%$; whereas when $H = 2$ and $H = 3$, the best GDELT models outperform the previous models by $(0.762 - 0.711)/0.711 = 7.2\%$ and $(0.749 - 0.656)/0.656 = 14.2\%$, respectively. This suggests that the predictor *NumTweetsProtests* is better able to predict protest onsets that take place on the next day whereas the GDELT features reveal more about protest onsets further into the future.



(a) GDELT models (b) GDELT models (c) GDELT models

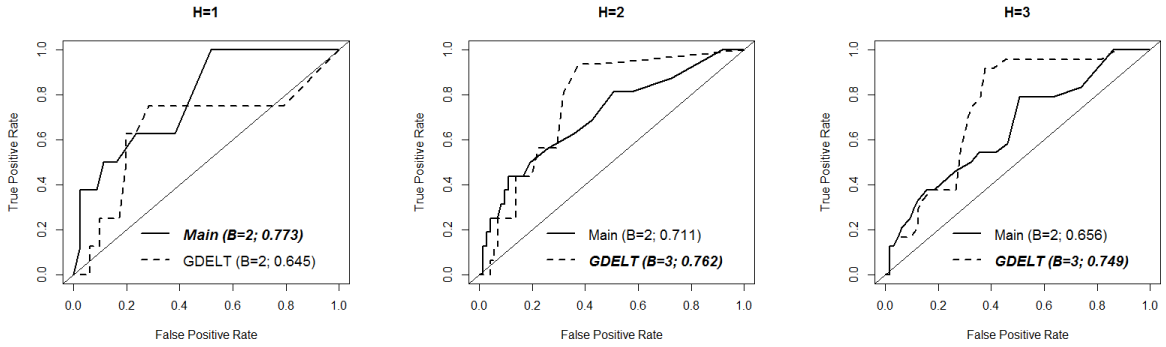
(d) Best threshold: 0.05 (e) Best threshold: 0.02 (f) Best threshold: 0.12

Figure 2.3: ROC curves for our GDELT models (panels (a), (b), (c)) under each prediction horizon $H \in \{1, 2, 3\}$ and base period $B \in \{1, 2, 3\}$; the corresponding AUC values are shown in parentheses in the legend, with the highest value shown in ***bold italic***. Precision-recall trade-off plots for the best performing GDELT models under each H are shown in panels (d), (e), and (f).

2.6.3 GDELT models

In the previous section we observed the superiority of the GDELT models under prediction horizons $H = 2$ and $H = 3$. A closer look at Table 2.9 indicates where the predictive power comes from. We observe two patterns, neither of which has been discussed in recent GDELT-based research (e.g., [80]).

First, an increase of military presence in Egypt appears to be an important predictor of upcoming protest onsets in its capital, Cairo, especially when predicting 2 or 3 days into the future. Out of the 20 GDELT event types, three are military-oriented: *exhibit military posture*,



(a) (b) (c)

Figure 2.4: Comparing the performance of the best protest advertisement models and the best GDELT models under each prediction horizon $H \in \{1, 2, 3\}$. Winners are shown in ***bold italic*** in the legend. In the parentheses are the base period B and AUC value of the models.

fight, and *engage in unconventional mass violence*. According to [108], *exhibit military posture* includes mobilizing and increasing police power, armed forces, and cyber forces; *fight* refers to the use of conventional armed forces, including imposing blockade, occupying territory, using light, heavy, and aerial weapons, as well as violating ceasefires; *engage in unconventional mass violence* includes mass expulsion, mass killing, ethnic cleansing, and using weapons of mass destruction. All of the three military-related event types are significant and positive in the selected GDELT models: *fight* being the most significant variable when $H = 2$ and $H = 3$, *exhibit military posture* when $H = 1$ and $H = 2$, and *engage in unconventional mass violence* when $H = 1$.

Second, an increase of protest events decreases the likelihood of an upcoming onset. Variable *Protest* is significantly negatively correlated under all three configurations of the prediction horizons and only when $B = 3$. This phenomenon is consistent with reasoning that when a series of related protests is gaining momentum, the onset of new, unrelated protests is lowered.

2.7 Discussion

In previous sections we evaluated model performance using ROC, AUC, precision, recall, and F1 score. However, we did not examine the predicted series themselves. We now

Table 2.9: Variables (the increase of the count of a GDELT type of events over the past B days) the selected GDELT models under each configuration of base period and prediction horizon, their coefficients (β), and p-values (p). An empty cell indicates the absence of the corresponding variable in the selected GDELT model; the coefficient and p-value of a significant variable (at a confidence level of 0.95) are shown in ***bold italic***.

GDELT event type	B=1, H=1		B=2, H=1		B=3, H=1	
	β	p	β	p	β	p
make public statement						
appeal						
express intent to cooperate						
consult						
engage in diplomatic cooperation					***-2.868***	***0.034****
engage in material cooperation	***0.435***	***0.046****			1.049	0.092.
provide aid	-1.144	0.077.	-1.500	0.063.		
yield						
investigate						
demand						
disapprove						
reject						
threaten			***0.763***	***0.018****	***1.078***	***0.018****
protest	-0.319	0.203	-0.754	0.052.	***-1.264***	***0.020****
exhibit military posture	***0.530***	***0.011****	***0.706***	***0.011****	***0.983***	***0.007****
reduce relations						
coerce						
assault	0.191	0.105				
fight						
engage in unconventional mass violence	***0.647***	***0.037****	***0.776***	***0.016****	***0.725***	***0.036****

GDELT event type	B=1, H=2		B=2, H=2		B=3, H=2	
	β	p	β	p	β	p
make public statement						
appeal						
express intent to cooperate			-0.816	0.127		
consult						
engage in diplomatic cooperation	-0.811	0.183			***-3.374***	***0.005****
engage in material cooperation						
provide aid						
yield					1.090	0.107
investigate					0.564	0.070.
demand						
disapprove						
reject						
threaten						
protest					***-1.653***	***0.004****
exhibit military posture					***0.633***	***0.025****
reduce relations			-0.376	0.270	-0.292	0.334
coerce						
assault	0.177	0.081.				
fight			***0.620***	***0.022****	***2.110***	***0.000****
engage in unconventional mass violence						

GDELT event type	B=1, H=3		B=2, H=3		B=3, H=3	
	β	p	β	p	β	p
make public statement						
appeal						
express intent to cooperate						
consult						
engage in diplomatic cooperation						
engage in material cooperation						
provide aid						
yield						
investigate						
demand						
disapprove						
reject						
threaten						
protest					***-0.838***	***0.027****
exhibit military posture					0.338	0.111
reduce relations			-0.517	0.134	-0.470	0.133
coerce						
assault						
fight			***0.454***	***0.010****	***1.027***	***0.002****
engage in unconventional mass violence						

collate the ground-truth labels and the predictions made by our best performing protest advertisement models and GDELT models and show how the two series line up with each

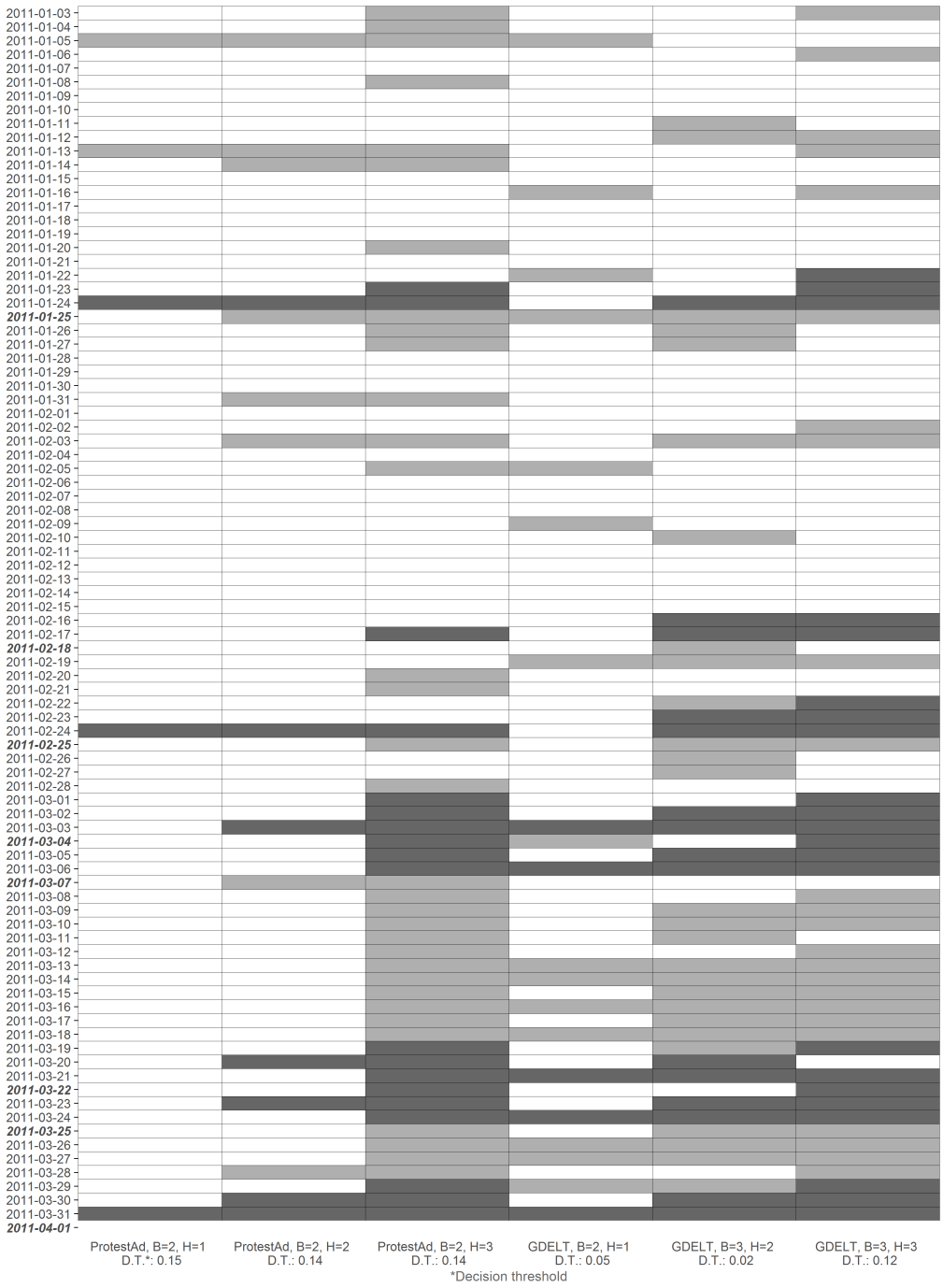


Figure 2.5: Six predicted series (best protest advertisement model and best GDELT model under three prediction horizons) lined up with the actual protest onsets. Dark grey tiles represent positive predictions made that detected an actual protest onset. Light grey tiles indicate positive predictions that failed to detect an onset. White tiles indicate that a negative prediction was made on the corresponding day. On the left, the dates of the protest onset days are highlighted in ***bold italic***.

Table 2.10: Detection precision and detection recall

Performance metric / Model	ProtestAd B=2, H=1 D.T.*: 0.15	ProtestAd B=2, H=2 D.T.: 0.14	ProtestAd B=2, H=3 D.T.: 0.14	GDELT B=2, H=1 D.T.: 0.05	GDELT B=3, H=2 D.T.: 0.02	GDELT B=3, H=3 D.T.: 0.12
Number of positive predictions that detect an onset	3	7	19	5	15	22
Total number of positive predictions	5	15	52	20	42	46
Proportion of successful positive predictions (detection precision)	0.6	0.47	0.37	0.25	0.36	0.48
Number of onsets detected	3	6	8	5	8	8
Total number of onsets within investigated period	8	8	8	8	8	8
Proportion of detected onsets (detection recall)	0.375	0.75	1	0.625	1	1

*Decision threshold

other. Figure 2.5 visualizes the outcomes predicted by our 6 best performing models (3 protest advertisement models, 3 GDELT) using their optimal thresholds. In each of the 6 columns, a dark grey tile represents a positive prediction made that detected an actual protest onset. We define detection as the case where a protest onset happens within H days of a prediction being made, where H is the prediction horizon of the predictive model. A light grey tile indicates a positive prediction that failed to detect an onset and a white tile indicates that a negative prediction was made on the corresponding day. On the left, the dates of the protest onset days are highlighted in bold italic.

With Figure 2.5 we can evaluate the performance of predictive models in a practical way that complements the above evaluations. Early detection of a protest onset is important to decision makers; however, successful detection is a looser requirement than correct prediction.

Consider a predictive model with a prediction horizon of 3 days and the prediction made on each of the three days before an actual protest onset. Each of the three predictions made needs to be positive to be correct; however, only one of them needs to be positive to successfully detect the upcoming protest onset. To quantify detection-based performance, we defined and calculated two measures: *detection precision*, which refers to the proportion of successful positive predictions out of all positive predictions made; and *detection recall*, which refers to the proportion of detected onsets out of all onsets in the investigated period. Their values are shown in Table 2.10. We see that when $H = 3$, both the best protest advertisement model and the best GDELT model are able to detect all 8 onsets but a higher proportion (0.48 vs 0.37) of the positive predictions makes the GDELT model a superior choice. However when $H = 1$ or $H = 2$, there exists a trade-off between detection precision and detection recall that leaves neither model dominant.

2.8 Concluding Remarks

This chapter presented a daily forecasting approach for protest onsets in Cairo, Egypt during the early months of the Arab Spring. We implemented a process of feature design guided by protest participation theory for daily prediction of civil unrest. We supported hypothesis H2.3 upon observation of significant positive correlation between an upcoming protest onset (Table 2.8) and an increase in the volume of expressions describing future protests on Twitter, a manifestation of structural availability in protest participation theory. On the other hand, we did not find support for hypotheses H2.1, H2.2, and H2.4: the correlations between (1) future protest onsets and (2) preceding increases in the number of political tweets, number of political Twitter users, and volume of @-mentioning news media and political activists were not significant. We built two groups of predictive models: protest advertisement models using the validated predictor *NumTweetsProtests* and GDELT models using daily counts of GDELT events. We conducted prediction using the selected models and GDELT models and discovered that when predicting the next day's outcome ($H = 1$), the best protest advertisement model outperformed the best GDELT model whereas for longer

prediction horizons ($H = 2$ and $H = 3$), the dominance is reversed. This indicates that the increase in the volume of expression describing future protests is better suited to predict only the immediately following day. Further inspection of the significant variables in our GDELT models reveals that an increase of military presence in a country may be predictive of upcoming protest onsets in its major cities.

This work helps to bridge the gap between social science theory and civil unrest forecasting by building predictive models using features tested in relevant social science literature. In the realm of social science research, hypotheses are usually tested within the sample and theories are proposed based on the results of hypothesis testing; train-test separation and prediction are not conventionally conducted. On the other hand, predictive modeling research tends to incorporate predictors from various sources when building models, often without a theoretical basis for feature engineering. This work also allows us to partially uncover the driver of predictive power of social media when predicting offline protest activity. We find support for H2.3, that individuals utilized social media to advertise future protest that ultimately materialized. On the other hand, the other 3 hypotheses remained unsupported: other politically relevant user activities (e.g., @-mentioning salient activists) were uncorrelated with protest onsets. Finally, this work also provides guidance for decision makers who wish to alter (discourage or encourage) the progression of a protest through early detection. As demonstrated in Table 2.10, our best performing models were able to detect all 8 protest onsets in the investigated period.

The broader lesson we learn from this work is the value of integrating social science theories and findings into social computing tasks, especially for those where the computation part is usually emphasized while the interpretation remains overlooked. Without grounding feature engineering in relevant theories, one faces the risks of unsound premises for their predictive modeling tasks and not understanding the underlying mechanisms with the human problem at hand, which is especially important if, in addition to predicting a human social outcome with high accuracy, we are interested in intervening with the social systems as

well (e.g., reduce crime). We should look into relevant social science literature and identify theoretical frameworks that could be mapped into measurements taken through mobile sensing technology. We should beware of tailoring feature engineering to specific data itself or solely relying on automatic feature representation techniques. Reflecting the nature of fusing social and computing sciences, theoretically driven feature engineering proves beneficial to not only predictive modeling in social computing tasks but also the creation of new knowledge in cyber-human systems. "To facilitate the design of social-technical systems and enhance their performance, social computing must learn from sociology and anthropology and integrate psychological and organizational theories." [109]

We focused on a macro-level outcome —civil unrest activity— fitting the theme of social computing where content and behavior generated by a crowd are used to understand and potentially bring about positive changes to the crowd. However, the outcome or response variable needs not be associated with a population; there are many micro-level, personal outcomes worth addressing that could utilize cyber social signals whose data mining work flow is demonstrated in this chapter. Mental health are among the most important personal outcomes, with existing works addressing the detection of stress [110], loneliness [111], and suicidal ideation [112] from online speech and posting patterns. Current theaters of social computing research are mainly online platforms, however, when we switch our focus from society to individuals, social computing can be enhanced by social signals in the physical space as well. Besides online platforms, a major proportion of social interactions take place offline and the characteristics of our in-person interactions with other people (where, when, how often, with whom, of what content, under what contextual activity) affect and reveal various aspects of our well-being. Sensors embedded in smartphones and wearable devices are increasingly able to capture these physical social signals. In the next chapter, we delve into discussions of social signals in the physical space and their measurement and mining for smart mental health applications.

Mining Social Signals for Micro-level Applications

In this chapter we investigate ways in which social signals in the physical space captured through mobile sensing technology can be utilized to improve personalized mental health tracking and potentially intervention, representative of micro-level human applications. The term "physical" refers to "in-person", in a sense that is the opposite of "online" or "virtual". As such, what constitutes a physical social signal would be one that is received within one's in-person social environment, which consists of physical proximity and face-to-face interactions with other people. Many micro-level outcomes such as adoption of innovation and political opinions are affected by such social environment; among them, health, both physiological and mental, is at the very center of human well-being and an area to which pervasive sensing technology has especially been seeking to contribute. The main goal achieved by this chapter is to, on a technical level, advance feature engineering methodology using smartphone proximity sensor data to improve performance of automated stress recognition, and; on a theoretical level, demonstrate that physical social signals detected by smart wearable devices provide valuable evidence (in addition to the non-social signals) to better understand and predict micro-level human behavior and health.

3.1 Introduction

Medical research has found strong connections between cognitive stress and physical and mental health [113, 114]: chronic stress experience is found to increase risks of not only respiratory infection, immunodeficiency, diabetes, and atherosclerosis, but also neurosis, depression, and schizophrenia. With stressful experiences common for students at school [115], for employees at work [116], and with family at home [117], stress remains a barrier that prevents people from living well and reaching their health and lifestyle goals, especially

among younger populations [118]. Study shows that between 37-84 percent of college students with elevated stress symptoms in the US never receive help [119]. To worsen the situation, some people are simply unaware of their needs for stress relief while others will not seek help until chronic stress develops into other mental disorders. As such, effective stress management should be placed high priority within personal health care.

Although therapeutic techniques such as autogenic training have been clinically validated [120] to combat diagnosed stress, a primary challenge in stress management is obtaining stress level measurements in an accurate, timely, and unobtrusive fashion. With many people reluctant to seek help with elevated stress symptoms [119], timely treatment is especially challenging. Clinicians have developed multiple survey inventories such as the Perceived Stress Scale [121] to solicit subjective stress measurements from patients. These inventories, although validated and reliable, are practitioner administered procedures that can be expensive and time intensive, and they are typically retrospective and suffer from various degrees of recall bias [122] (e.g., erroneous memory about the stressor). Computerized ecological momentary assessment (EMA) surveys have been deployed via mobile technology to assess momentary stress levels [123], thus reducing recall bias and increasing ecological validity. However, the burden of stress self-report lowers EMA compliance [124].

Recent improvements in and uptake of mobile sensing technology present opportunities to use passively collected data from smartphone-embedded sensors for objective and unobtrusive stress inference. This approach mitigates cognitive biases and allows for continuous and unobtrusive data collection, thus enabling just-in-time interventions [125, 126]. Recent studies have used passively collected data from smartphone-embedded sensors to infer and predict mental health outcomes [127–129]. While most existing work focuses on characterizing a subject's personal behavior (e.g., place visit, physical activity level), information on one's in-person social environment —physical proximity and face-to-face interactions with other people— has not been adequately utilized in mental health tracking. With (1) mobile sensing data, specifically proximity triggered Bluetooth encounter data serving as proxy to these

social signals and having been used to reconstruct social networks among groups of individuals [18, 19, 130], and (2) multifaceted effects of social interaction on personal stress discovered in psychology studies [131–133] suggesting the value of incorporating physical social signal patterns in stress inference, questions remain whether, to what extent, and how Bluetooth encounter data can be used to predict stress and thus enhance real time mental health tracking.

I address this question through a feature engineering approach. A Bluetooth encounter data point typically comprises three elements: a detecting device ID, a detected device ID, and a timestamp at which the proximity detection occurred. Depending on the data available, one may have at hand (1) Bluetooth encounter data among a group of smartphone users who have relatively close social ties (e.g., classmates, residents of the same apartment complex), or (2) Bluetooth encounter data from individual smartphone users who do not have significant social interactions with one another. From the perspective of an individual subject, in the former case we would have substantial information about how the individual's social contacts themselves encounter their own social contacts whereas in the latter case we would not. In terms of the networks that emerge from these Bluetooth encounters, in the former case we could construct a relatively well-connected network with few singletons whereas in the latter case we would observe multiple, disconnected egocentric networks. In reality the latter case places a more relaxed requirement on data quality. In this chapter, we present two feature engineering methodologies to tackle the two cases, in Sections 3.3 and 3.4 respectively: for the better connected Bluetooth encounter data, we resort to social network analysis to systematically extract key metrics from the Bluetooth encounter networks [134]. For the more egocentric data, we borrow insights from Natural Language Processing and propose a vector space approach to achieve effective feature representation [135]. For both approaches we conduct correlation analysis and predictive modeling to investigate the theoretical implications and validate the predictive power of physical social signals incorporated in stress recognition models.

3.2 Related Work

3.2.1 Automated stress sensing

The physiological and behavioral covariates of human stress responses have been extensively studied in recent years, with a body of research focusing on physiological and motor-sensory indicators. Ranabir et al. [136] identified hormonal changes associated with stressful experiences. Sun et al. [137] used electrocardiogram, galvanic skin response, and accelerometer data to detect stress in sitting, standing, and walking activities. During sessions of human-computer interaction, human stress levels are correlated with gaze and click patterns [138], touch intensity and duration [139], as well as physiological measures including blood volume pulse, galvanic skin response, pupil diameter, and skin temperature [140]. In real-time driving scenarios, the driver's stress level can be detected from facial expression features [141], galvanic skin response, and photoplethysmography [142]. Other research has focused on detecting stress from human speech [143, 144]. Sharma et al. [145] provide a comprehensive review of the physiological and motor-sensory indicators used for stress recognition.

The studies above were conducted in laboratory settings using special purpose equipment to collect data. Such measurements are impractical in natural settings outside of the laboratory. The need for timely, accurate, and unobtrusive mental health management motivates automated mental health inference and prediction using data passively collected from mobile phones and wearable devices that are widely owned and carried around. Sano et al. [146] showed preliminary success using wearable motion sensors and mobile phone usage data to classify stressed versus non-stressed individuals. Maxhuni et al. [147] used phone-embedded accelerometer readings before and after phone calls to detect stress levels of office workers. Bogomolov et al. [148] predicted an individual's daily stress levels using a range of predictors including personality type, weather conditions, SMS, phone call, and Bluetooth features. A similar study was conducted by Gjoreski et al. [149] targeting momentary stress

detection using multiple data sources available from a smartphone, such as accelerometer, audio sensor, GPS, WiFi access, call log, and light sensor. However, despite rich evidence showing the stress-modulating effect of social stimuli [131–133], existing work on automated stress recognition has only made limited use of information regarding an individual's social interaction patterns.

3.2.2 Bluetooth encounters

Passive detection of social interaction in natural settings remains a challenge. As far as current mobile sensing technology goes, Bluetooth is a popular choice for capturing in-person social interactions in uncontrolled settings. Reasons are the following. First, Bluetooth encounters are triggered by physical proximity between two Bluetooth enabled devices with an expected detection range of 10 meters [15]. As many in-person social interactions require physical proximity, researchers have found that Bluetooth encounter data contains valuable information about, thus can serve as proxy signals for, an individual's social connections. Moreover, Bluetooth sensors are widely available in smartphones and functional in both indoor and outdoor environments [18], thus providing major advantages over other encounter detecting devices such as radio-frequency identification (RFID; as used in this study [150]) or infrared sensors (IR; as used in the Sociometric badges [66]) which require external hardware and making potential findings using Bluetooth data more easily applicable to real-world smart health applications.

As such, Bluetooth data is collected in multiple large scale mobile sensing data collection projects around the world [20–22]. Eagle et al. [19] used Bluetooth encounter data to infer friendship networks. Do et al. [18] proposed a generative probabilistic model to extract latent human interaction types based on Bluetooth encounters. Zheng et al. [130] adopted an egocentric, unsupervised learning approach to learn an individual's social circles (e.g., family vs colleague) using Bluetooth encounter data. Yan et al. [151] focused on classifying the context (e.g., in a meeting or at lunch) of Bluetooth encounters and clustering users based on their encounter patterns. Madan et al. [56] used Bluetooth proximity to measure

exposure to peers that are overweight or have unhealthy dietary habits and inactive lifestyles and discovered its influence on personal weight changes. In each of these studies, Bluetooth encounter data drives inferences regarding subjects' social connections.

Despite the growing body of research on sense-making of Bluetooth encounters, little has been done to correlate Bluetooth encounters with mental health outcomes. Boonstra et al. [152] demonstrated the feasibility of collecting Bluetooth encounter data for depression recognition but offered no further findings on the relations between Bluetooth encounters and depression. A set of nine Bluetooth features covering encounter counts, entropy, and inter-encounter times were explored in a daily stress estimation problem [148] but the extent to which the proposed Bluetooth features enhance performance was not assessed quantitatively. To the best of our knowledge, current research lacks systematic (1) feature engineering from Bluetooth encounter data and (2) evaluation of Bluetooth encounters as stress predictors. We seek to address these limitations in this chapter.

3.2.3 Mining physical social signals

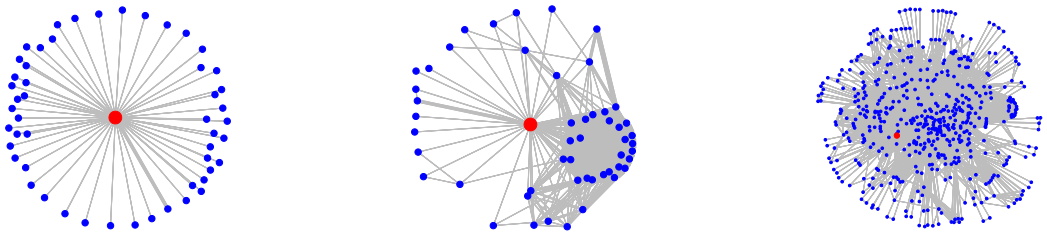
From indoor co-location to face-to-face conversation, various types of physical social signals have been the target of mathematical and statistical modeling to understand various aspects of human behavior. Some works [153] [154] focus on the movement pattern of people engaging in face-to-face interactions in confined, indoor spaces (e.g., reception at a conference). They typically adopt an agent-based approach, characterize movements through random walk models with actor characteristics specified that govern movement micro-dynamics, and seek to reproduce the pattern observed with agent-based models. This approach is useful in studying the epidemic spreading of infectious diseases [155] and corresponding intervention strategies. Some works [38, 156] focus on the structure of the static networks emerging from physical proximity and face-to-face interactions accumulated over a period of time. Network graphical measures such as density and transitivity are the backbone of this approach and a typical application of these works is discover the correlation between structural attributes to employee performance at workplace, partly due to the fact that many datasets [23] were collected from

office employees wearing sensor badges. Still another approach is statistical network modeling, which attempts to explain the formation of particular network structures by estimating the effects of underlying factors of a group of individuals. Exponential-family Random Graph Model is the central method and has been adapted and enhanced to model conversation networks [157] with data collected by face-to-face interaction or proximity sensors. We will discuss in more detail this approach and potential ways to enrich its modeling methods in Chapter 4.

3.3 Bluetooth Encounter Networks

In this section we focus on Bluetooth encounter networks, addressing the case where we have Bluetooth encounter data for a relatively close-knit group of social actors. First, we conduct systematic feature engineering on Bluetooth encounter data based upon network analysis as well as social and temporal commonalities. Then, we identify four experimental settings for momentary stress recognition, differentiating stress estimation versus stress forecasting in practical applications. Finally, we evaluate the proposed features in correlation analyses and predictive modeling, achieving significant improvement in goodness-of-fit and prediction performance in multiple evaluation settings.

Key contributions of this section are twofold. First, to the best of our knowledge, this work is the first to evaluate the value of Bluetooth encounter network data in individualized real-time stress recognition. Second, we propose novel features designed from Bluetooth signals that not only demonstrate a significant relationship with stress outcomes but also provide sociological and psychological insights into the implications of social network on mental health. These features will be applicable to many electronic and mobile health tasks monitoring other types of mental health status such as loneliness, depression, and social anxiety.



(a) Ego (b) Local (c) Global

Figure 3.1: Illustration of the three network scales from which structural attributes are extracted. Each red node represents a subject and blue nodes represent other devices; grey weighted edges represent Bluetooth encounters and their volume. Given all encounter events accumulated over a period of time, "ego" includes only the subject, its neighbors, and edges between the subject and a neighbor; "local" includes the identical set of devices as "ego" but also includes encounter events between pairs of neighbors; "global" includes all devices and all encounters within.

3.3.1 Feature engineering

I consider three aspects of an individual's social interaction network when conducting feature engineering using Bluetooth encounter data. The first aspect is structure. Some people have more social contacts than others; some people prefer to distribute their social interaction more evenly among their social contacts whereas others like to focus on a small subset. Then come the characteristics of an encounter event, which can be measured by the time, location, and the content of verbal and non-verbal expressions involved. These characteristics usually require additional sensors to capture (e.g., GPS and audio sensors). Moreover, an individual's social interaction experience can also be affected by the nature of his or her social contacts. Capturing these aspects of a subject's social interaction network, we propose three classes of features, namely **structural** attributes, **edge** attributes, and **neighbor** attributes, and ground their measurement in Bluetooth encounter data. We will define the baseline features following the discussion of my proposed Bluetooth features.

Table 3.1: Structural features used to describe the topology of the Bluetooth encounter network surrounding a subject

Feature	Description
ego_deg	Number of neighbors (devices encountered)
ego_avgWeight	Average weight (number of encounters) over all neighbors
ego_giniWeight	Gini inequality of weight distribution over all neighbors
loc_den	Density of the local network
loc_avgWeight	Average weight over all edges in the local network (sum of the number of Bluetooth encounters over the number of distinct edges)
loc_trans	Global transitivity of the local network
loc_avgCompSize	Mean group size (a group is defined as a connected component after ties with the subject node are removed)
loc_giniCompSize	Gini inequality of group size
glo_avgDegNb	Average degree of the neighbors
glo_giniDegNb	Degree inequality of the neighbors
glo_avgBetwNb	Mean betweenness centrality of the neighbors
glo_giniBetwNb	Gini inequality of betweenness centrality among neighbors
glo_avgTransNb	Mean local transitivity of the neighbors
glo_giniTransNb	Inequality of local transitivity of the neighbors
glo_betw	Betweenness centrality of the subject node in the global network
glo_trans	Local transitivity of the subject node in the global network

3.3.1.1 Structural Attributes

I create 16 features to characterize the topology of the Bluetooth encounter network surrounding a subject formed over a time window Δt. A complete list with definitions is provided in Table 3.1. The suffix of a feature name indicates the scale of the social interaction network the feature is extracted from. *ego* is an alias for egocentric network, which indicates the network containing a subject and its neighbors (devices encountered) and Bluetooth encounter events (edges) only between a subject and a neighbor. *loc* refers to a local network, which refers to the network containing a subject and its neighbors and edges between a subject and a neighbor or between two neighbors. Lastly, *glo* indicates a "global" network encompassing all nodes detected through Bluetooth encounters given a period of time. Figure 3.1 demonstrates the Bluetooth encounter network on the three network scales surrounding a subject (red node); grey weighted edges represent Bluetooth encounters and their volume. With the structural features we pay special attention to three network metrics: (1) degree, the number of edges connecting a subject with another, approximating level of social activeness; (2) betweenness centrality, defined by the proportion of shortest paths in a network that go through a vertex [158] to describe how central in the Bluetooth encounter network a subject is, and; (3) transitivity, which as a global measure is defined by the proportion of closed connected triples (i.e., triangles) out of all connected triples in a network and as a local measure the proportion of closed connected triples connected to a vertex out of all connected triples centered on the vertex [159]. Transitivity quantifies the propensity for a network to exhibit (global) and a subject to be present in (local) triangular relations, which is an indicator of community forming.

3.3.1.2 Edge Attributes

Edge attributes are the characteristics of the encounter events. The only native edge attribute in Bluetooth encounter data is the timestamp of each encounter event (see Table 3.5). One can obtain other edge attributes providing availability of corresponding sensor data (e.g., the geographic location of a Bluetooth encounter with GPS data). As this paper

Table 3.2: Social commonality features

Feature	Description
usual	The proportion of the subject's past encounters with a neighbor out of the subject's past encounters with any device
usual_nb	The proportion of the subject's past encounters with a neighbor out of the neighbor's past encounters with any device
shared	The proportion of shared social contacts between the subject and a neighbor out of past social contacts of the subject
shared_nb	The proportion of shared social contacts between the subject and a neighbor out of past social contacts of the neighbor

focuses on Bluetooth data exclusively, we will use a daily epoch feature available from the timestamps, as our only edge attribute in this study. We define the daily epoch feature as a categorical variable with 4 levels *morning* (6am-noon), *afternoon* (noon-6pm), *evening* (6pm-midnight), *night* (midnight-6am), marking section of the day where the current time (the temporal upper bound for Bluetooth features extraction) resides.

3.3.1.3 Neighbor Attributes

Named neighbor attributes, this class of features aim to characterize the social nature of the devices a subject has encountered. Further divided into three categories detailed below, each feature in this class is associated with a particular neighbor of a given subject.

a) Social commonality: An encounter between two individuals tends to have different significance to their respective social lives. For example, meeting with a familiar friend that one usually spends long time with would likely incur different emotional responses than meeting with a less familiar acquaintance. Another example concerns the interactions between two pairs of individuals: if one pair have a large group of mutual friends and the other have none, the nature and content of their respective interactions are likely to be very different. Materializing these notions on familiarity and overlapping social circles, we design four social commonality features as listed in Table 3.2.

To compute these features, we first choose the $7 \times 24 = 168$ hours leading up to $t - \Delta t$ as a base period to extract past social contacts. For a subject and for each neighbor encountered

Table 3.3: Temporal commonality features

Feature	Description
tempCom	The degree to which the times when the subject encounters a neighbor agree with the usual times the subject encounters any device
tempCom_nb	The degree to which the times when the subject encounters a neighbor agree with the usual times the neighbor encounters any device

within $t - \Delta t$ to t we follow the following procedure: (1) compile a set of device IDs the subject and the neighbor encountered respectively during the 168-hour base period, producing a set L_s for the subject and L_n for the neighbor; note that L_s would include the neighbor and L_n would include the subject; (2) compute feature *shared* as $|L_s \cap L_n|/|L_s|$ and feature *shared_nb* as $|L_s \cap L_n|/|L_n|$; (3) for each device ID in L_s and L_n count the number of times (frequency) Bluetooth encounter events were recorded during the base period, producing a frequency vector F_s over L_s for the subject and F_n over L_n for the neighbor; (4) suppose during the 168-hour period the subject and the neighbor encountered F_{sn} times, compute feature *usual* as $F_{sn}/\sum F_s$ and *usual_nb* as $F_{sn}/\sum F_n$. Note that all four social commonality features have a maximum of 1 and a minimum of 0.

b) Temporal commonality: Similar to the concept of social commonality, the extent to which an encounter happened at a usual time for the subject and for the neighbor could also be a telling sign of the nature of an encounter and potentially its effect on mental health. We construct two features describing such temporal commonality, *tempCom* and *tempCom_nb*, with definitions given in Table 3.3. A high temporal commonality indicates that the encounter transpired at a usual time that an individual (the subject or the neighbor) tends to be "encounterable" whereas a low value would reflect that an individual may be venturing more out of his or her usual schedule for the encounter to be happening. Combining temporal commonality and social commonality we can get a read on the social nature of a Bluetooth encounter between two devices.

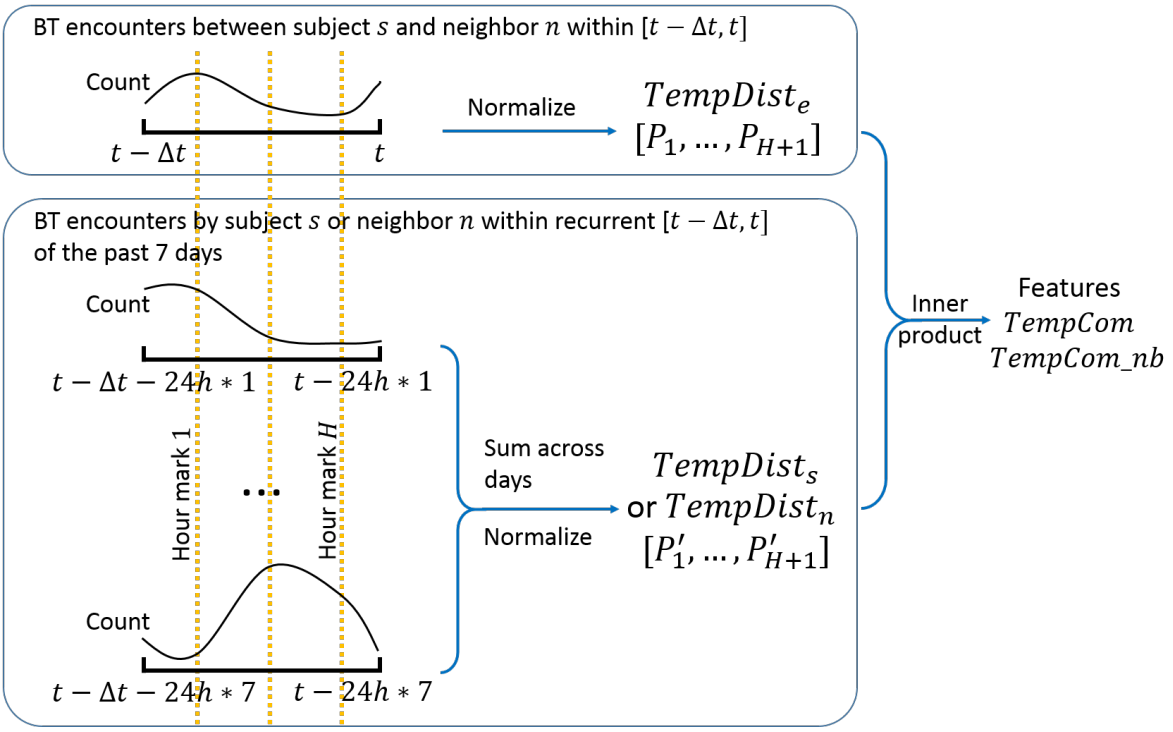


Figure 3.2: Illustration of the computing process for temporal commonality features

The process for computing temporal commonality is illustrated in 3.2 and explained as follows. We choose again the $7 \times 24 = 168$ hours leading up to $t - \Delta t$ as a base period to extract evidence on usual encounter times.we(1) break down the encounter events by their timestamps into hourly blocks, compute event counts that fall within each block, and create a probabilistic distribution vector $TempDist_e$ over the blocks; (2) compile all encounter events involving the subject (regardless of who the neighbor is) detected during the same time period $t - \Delta t$ to t but on different days within the base period (which covers seven days) and create a similar event count distribution vector $TempDist_s$ using these past encounter events over the same hourly blocks as $TempDist_e$; (3) repeat step (2) for all encounter events involving the neighbor and create an analogous $TempDist_n$. The three $TempDist$ vectors each sums up to 1 and all have the same support. Finally, the value feature *tempCom* is the inner product of $TempDist_e$ and $TempDist_s$ while the value of feature *tempCom_nb* is the inner product of $TempDist_e$ and $TempDist_n$, thus capturing the agreement between the

Table 3.4: Seven semantic labels of GPS locations

Label	Criterion
Home	The place a subject spends the most time 12-6am
Education	University buildings
Food	Dining and food vending establishments
Health	Healthcare facilities and gymnasiums
Transit	Bus stops and stations, parking lots, airports
Religion	Places of worship (e.g., churches)
Other	Places not categorized above (e.g., post offices)

distributions. Both temporal commonality features have a maximum of 1 and a minimum of 0.

c) Network structure: As the structural features described in Section 3.3.1.1 can be used to characterize the interaction network structure surrounding any device in an Bluetooth encounter network, we build the same 16 features for each neighbor a subject encounters within Δt.

Note that each neighbor attribute is associated with one neighbor of a subject's; and since a subject tends to encounter multiple devices during a given time period we compute four aggregation statistics mean, standard deviation, maximum, and minimum values of all neighbor attributes and use them all as features. This way, we designed in total 16 (structural attributes) + 3 (edge attributes: three dummy variables created for the four-level categorical variable that is daily epoch) + 22 (neighbor attributes) $\times$ 4 (aggregation statistics) = 107 Bluetooth features (later referred to as main features) that we will use for my further analysis.

3.3.1.4 Baseline Features

I also build features measured from other standard mobile sensors as baseline in order to evaluate the additional predictive power of our Bluetooth features when combined with them . Our baseline features cover four groups: (1) semantic location features: a 7-level categorical variable indicating one of seven place types a subject is at at current time t (denoted "_now" in later references) and their proportions within a feature extraction window Δt (denoted

"_past"); to learn place types, we performed clustering of GPS coordinate traces [160], then obtain semantic annotations from Google Map API (https://developers.google.com/maps/), and manually created seven place types as shown in Table 3.4; (2) activity level features: a 4-level categorical variable indicating the stationary/walking/running/unknown status of a subject at current time t and their proportions within Δt; (3) sound status features: a 3-level categorical variable indicating the silence/voice/noise status detected by a subject's mobile device at current time t and their proportions within Δt, and; (4) sleep quality: self-reported sleep quality of the previous night (5-scale ordinal value ranging from poorest to best); sleep quality is included as a baseline feature due to its proved positive correlation with next-day stress [161]. After encoding dummy variables we create $[(7-1)+(4-1)+(3-1)] \times 2 + 1 = 23$ baseline features.

3.3.2 Hypotheses

Below, we hypothesize relationships between Bluetooth encounters and personal stress outcomes. Hypotheses 1 and 2 concern the statistical relationships between Bluetooth encounter networks and stress outcomes, whereas hypotheses 3 and 4 concern predictive power. We further differentiate between the estimation of current stress levels and the forecasting of future levels. In addition to Bluetooth encounter data, we use data from GPS, accelerometer, audio sensors as well as sleep quality self-reports to drive baseline models for comparison (Section 3.3.4).

- **Hypothesis 3.1**: a subject's physical proximity based social environment, as measured through Bluetooth encounter networks, is more correlated with his or her *current* stress outcomes compared with prior probabilities of such outcomes.
- **Hypothesis 3.2**: a subject's physical proximity based social environment, as measured through Bluetooth encounter networks, is more correlated with his or her *future* stress outcomes compared with prior probabilities of such outcomes.
- **Hypothesis 3.3**: a subject's recent Bluetooth encounter data together with baseline variables can estimate his or her *current* stress outcomes more accurately than only

using the baseline variables.

- **Hypothesis 3.4**: a subject's recent Bluetooth encounter data together with baseline variables can forecast his or her *future* stress outcomes more accurately than only using the baseline variables.

3.3.3 Data

I use the StudentLife dataset [24] to test the hypotheses stated above within experiment designs to be described in Section 3.3.4. The StudentLife dataset records Bluetooth encounters among 49 college student participants over 66 days (27 March 2013 through 31 May 2013).

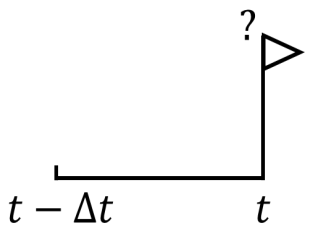
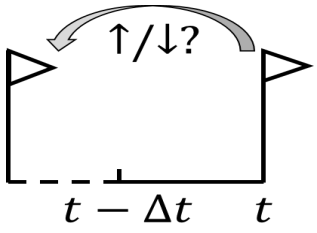
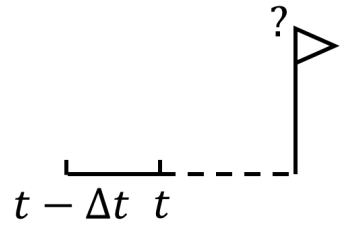
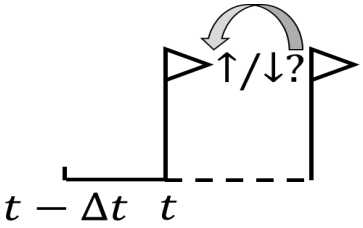
Table 3.5: A sample of Bluetooth encounter data

Timestamp	Observer ID	Observed ID
2013-03-29 00:00:55	u56	E4:CE:8F:73:CE:65
2013-03-29 00:01:34	u39	u26
2013-03-29 00:02:42	u08	04:0C:CE:EB:14:7E
2013-03-29 00:02:42	u08	04:0C:CE:EC:5C:21

Table 3.5 shows a sample of the Bluetooth encounter data, each row of which consists of a timestamp and two device identifiers. A study participant's device has an ID starting with "u" followed by a two-digit number whereas the ID of a non-participant is a 17-character string. The observer IDs are only from participant's devices whereas the observed IDs can be devices of both participants and non-participants. We add to the dataset an encounter event between two non-participant devices when both of them are found to be observed by a study participant's device at the same timestamp. For example, in Table 3.5, noting that the third and fourth row share the same timestamp ("2013-03-29 00:02:42") and the same observer ID ("u08") but different observed IDs ("04:0C:CE:EB:14:7E" and "04:0C:CE:EC:5C:21"), we add another row with "04:0C:CE:EB:14:7E" and "04:0C:CE:EC:5C:21" being the two encountering device identifiers.

Stress level data is obtained through ecological momentary assessment surveys deployed on the subjects' smartphones multiple times a day at random times. The survey comprises a

Table 3.6: Four experimental settings for momentary stress recognition. t represents the current time and Δt represents the feature extraction window. Flags represent momentary self-reports of stress levels. Value/Change indicates whether the predicted outcome is stress level itself or the increase and decrease therein. Diagnostic/Prognostic indicates whether the predicted outcome temporally parallels or succeeds Bluetooth encounter observations

	Value	Change
Diagnostic		
Prognostic		

question text "Right now, I am..." and 5 response options "feeling great", "feeling good", "a little stressed", "definitely stressed", and "stressed out". We convert ordinal outcome values to binary ones. Concretely, for value diagnosis and value prognosis, we assign self-reported stress "feeling great" and "feeling good" as negative (0) and "a little stressed", "stressed", and "stressed out" as positive (1); for change diagnosis and change prognosis, we treat increased and not increased stress level as positive and negative respectively. By doing so we simplify modeling and attain larger sample size associated with each level of the response variable.

3.3.4 Experiments

3.3.4.1 Experimental Design

I identify two key distinctions in stress recognition experiments: (1) the goal being diagnostic (estimating current outcomes, which is the primary focus of existing stress recognition work) versus prognostic (predicting future outcomes); and (2) the response variable being the value of a mental health status (e.g., stress level) versus its change (e.g., increase/decrease in stress level). In practice, the diagnostic setting will inform reactive interventions that aim to

mitigate negative consequences of the present outcome, whereas the prognostic setting will inform proactive interventions that aim to mitigate negative consequences of future outcomes. The four resulting designs are illustrated in Table 3.6 and further defined below. Each of the four settings also entails the setting of a window size parameter Δt, a period of time preceding a stress self report from which we extract features.

- **Value Diagnosis**: to estimate stress level at current time t given mobile sensing data available by t: "What is the stress level of a user now?"
- **Change Diagnosis**: to estimate at current time t whether the current stress level has increased compared to the previous known level given mobile sensing data available by t: "Has a user's stress level increased?"
- **Value Prognosis**: to forecast at current time t the next stress level given mobile sensing data available by t: "What will the next stress level of a user be?"
- **Change Prognosis**: to forecast at current time t whether the next stress level will increase compared to the current level given mobile sensing data available by t: "Will a user's stress level increase?"

3.3.4.2 Correlation Analysis

The objective of this step is to evaluate how well my Bluetooth features account for the variance within an individual's momentary stress outcomes in the four experiment settings introduced in Section 3.3.4.1. The result from this analysis will allow us to test Hypotheses 1 and 2 and discover salient features that are responsible for the correlation.

To achieve this, we use two model selection methods (1) backward stepwise logistic regression with AIC (Akaike Information Criterion, defined as $2k - 2lnL$, where k is the number of estimated parameters and L is model likelihood) as the selection criterion and (2) 10-fold cross validated logistic regression with LASSO regularization (least absolute shrinkage and selection operator [162]) and model likelihood as the selection criterion on the following feature groupings: (1) a null model; (2) baseline features as defined in Section 3.3.1.4; (3) Bluetooth features as described in Sections 3.3.1.1, 3.3.1.2, and 3.3.1.3, and lastly;

(4) our Bluetooth features combined with baseline features. We choose Δt to be 6 hours as it represents the length of one entire daily epoch (e.g., morning or afternoon). Response variables in each setting are as explained in Section 3.3.3.

I then conduct bootstrapping to break down potential dependency between observations and further prove the statistical relationships between my Bluetooth features and personal stress outcomes. Concretely, we draw 100 bootstrap samples on the original dataset and repeat stepwise and LASSO model selection on all four feature groupings. Then we compare the resulting AIC's using Student's t-tests.

3.3.4.3 Predictive Modeling

The objective of this step is to evaluate the predictive power of my proposed Bluetooth features and test Hypotheses 3.3 and 3.4. We experiment with two different evaluation settings as discussed below.

a) Across-subject, leave-one-subject-out (LOSO): In this setting we set aside one subject's data, train models on data pooled together from all other subjects' data, and evaluate the performance on the one subject's data that was set aside. For each subject, we obtain a sequence of predicted outcome values and a corresponding AUC (area under ROC curve) score. This experiment is applicable in practical cases where models can be built from available, existing subjects' data and performance on a new, unseen subject is desired.

b) Within-subject, leave-one-observation-out (LOOO): In this experiment, for each subject, we set aside each observation of his or hers, train models on the remaining data that belongs to the same subject, and evaluate on the one observation that was set aside. For each subject, we also obtain a sequence of predicted outcome values and a corresponding AUC score. This experiment is applicable in cases where only a user's own historical data is available for insights regarding said user's future outcome values.

The critical difference between the across-subject LOSO and within-subject LOOO evaluation settings is that in the former, to predict an observation of a subject, only his or her peers' data and none of his or her own data is used; whereas in the latter, the situation is

Table 3.7: AIC scores of selected models using backward stepwise logistic regression on different feature groupings

Stepwise	Value Diagnosis	Change Diagnosis	Value Prognosis	Change Prognosis
NULL	1,733	1,873	1,704	1,715
BASE	1,691	1,841	1,678	1,714
MAIN	1,656	1,828	1,626	1,701
BASE + MAIN	1,618	1,825	1,603	1,706

reversed: to predict an observation of a subject, only his or her own data and none of his or her peers' data is used. Note that in both experiments, we pool data from the entire available time period and thus ignore the temporal sequence of observations; this is justified by (1) the focus on non-temporal effect (own data versus peers' data) in these two experiments and (2) limited data size (1,702 observations in total covering 49 subjects over a 66-day period).

As a preprocessing step, we discard (1) the first week's data as the social and temporal commonality features require a preceding seven days (168 hours) to serve as a base period, and; (2) for all prediction settings except change diagnosis, which uses differenced feature values as predictors, observations that have a zero *ego_deg* as it indicates zero Bluetooth encounters. For all predictive experiments in this section, we choose Δt as 6 hours and use a random forest learner with 3000 trees grown, 20 predictors randomly selected at each split, until reaching node size 5. We also explore the sensitivity of prediction performance to varied Δt sizes.

3.3.5 Results

3.3.5.1 Correlation Analysis Result

We list the AIC scores of selected models using stepwise and LASSO methods in Table 3.7 and 3.8 respectively, both of which show that models using selected Bluetooth features (MAIN) achieve higher goodness-of-fit than null models when explaining momentary personal stress outcomes. Overall, stepwise logistic regression achieves lower AIC scores than LASSO as a feature selection method. Among the four experimental settings, the goodness-of-

Table 3.8: AIC scores of selected models using LASSO logistic regression on different feature groupings

LASSO	Value Diagnosis	Change Diagnosis	Value Prognosis	Change Prognosis
NULL	1,733	1,873	1,704	1,715
BASE	1,662	1,853	1,689	1,713
MAIN	1,653	1,853	1,653	1,713
BASE + MAIN	1,626	1,850	1,652	1,714

fit improvement is more pronounced in value diagnosis and value prognosis. Under all experimental settings, baseline features (BASE) and our Bluetooth features each outperformed the null models. Best performances are obtained when models are selected from Bluetooth and baseline features combined (BASE + MAIN), dominating baseline and main features individually, under all experimental settings except change prognosis, where using all features together seems to harm goodness-of-fit.

Mean and standard deviation of each group of bootstrapped AIC values are shown in Figure 3.3 and clearly with strong confidence (p-value < 0.001) we achieve significantly better goodness-of-fit with selected Bluetooth features compared to null models, with either feature selection method. We successfully confirm Hypotheses 3.1 and 3.2 that a subject's Bluetooth encounter networks correlates with his or her current and future stress outcomes significantly better than null models.

Finally we look into the important Bluetooth features that have driven the performance demonstrated above. As expected, LASSO and stepwise logistic regression selected moderately different but overlapping sets of features and LASSO resulted in more parsimonious models. We focus on the features in the BASE + MAIN models that are both significant in the stepwise models and selected by LASSO method under each experimental setting, for they are most likely important features. The selected features, the sign of their effect, and their p-values in the corresponding stepwise model are listed in Table 3.9, organized by the category (BASE and MAIN) and sorted in descending order of significance. We make several interesting

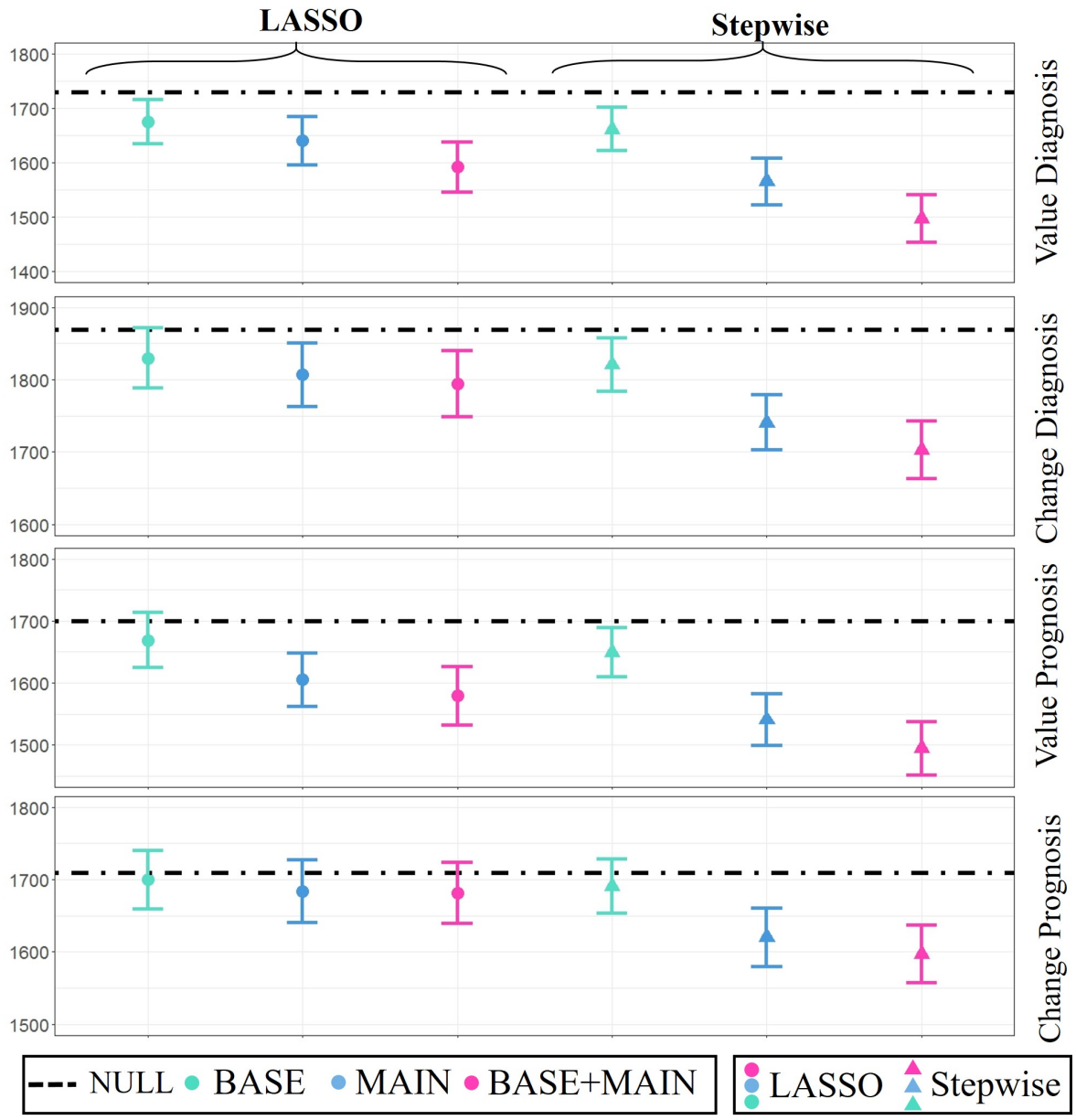


Figure 3.3: Bootstrap result: mean (round and triangular dots) and standard deviation (upper and lower error bars) of AIC scores achieved by LASSO and stepwise logistic regression on each bootstrap sample compared against the corresponding null models (dashed horizontal lines).

observations on the correlations between our Bluetooth features and stress outcomes. First, *usual_max* is the number one significant and negative predictor associated with both present and future level of personal stress (value diagnosis and value prognosis). This suggests that

Table 3.9: Important features in the MAIN + BASE models selected by both stepwise and LASSO logistic regression under each experimental setting; following each feature is the sign of their effect and p-value in the corresponding stepwise models

Value Diagnosis					
BASE			MAIN		
sleepQ	−	0.000	usual_max	−	0.000
walking_past	−	0.002	ego_deg	+	0.000
stationary_now	+	0.003	shared_nb_mean	−	0.004
home_now	−	0.006	shared_mean	+	0.004
health_now	−	0.007	tempCom_max	+	0.005
transit_now	−	0.011			
Change Diagnosis					
BASE			MAIN		
religion_now	−	0.000	loc_trans_mean	−	0.000
noise_now	+	0.003	usual_sd	+	0.001
noise_past	−	0.046	ego_avgWeight_min	−	0.008
Value Prognosis					
BASE			MAIN		
stationary_now	+	0.008	usual_max	−	0.000
transit_past	−	0.009	loc_avgWeight_mean	−	0.000
religion_now	−	0.009	evening	−	0.002
edu_now	+	0.016	tempCom_mean	−	0.005
noise_past	−	0.030	loc_giniCompSize_mean	+	0.012
health_now	−	0.035	usual_nb_min	−	0.022
Change Prognosis					
BASE			MAIN		
			shared_nb_max	−	0.019

interaction with a familiar and regular social contact has a potentially lasting stress relieving effect ("protective effect under stress" [131]); in fact, it was confirmed that "the presence of a friend in a stressful situation reduces stress more than the presence of a stranger" [132], which constitutes the social buffering theory [163]. Second, an increased value of *usual_sd* appears to be positively correlated with increased stress level in the change diagnosis setting.

This indicates that interacting with people of different familiarity levels may comprise a stressful experience. Such effect is mentioned in [60, 148] and agrees with the role strain theory [164]. The significance of feature *loc_trans_mean* in change diagnosis indicates that increased transitivity within a subject's local encounter network, which can be a result of real life scenarios like gatherings and group study sessions, is likely to accompany decreasing personal stress. Last, for change prognosis only one feature made the list, indicating it is less well-modeled by our Bluetooth features than other settings; *shared_nb_max*, representing the highest proportion of shared social circle with a subject among his or her encounter neighbors, exhibited a similar effect as *usual_max* which is also a social commonality feature.

As for baseline features, sleep quality of the previous night is found to have a significant correlation with low current stress level, which confirms findings in [161]. Among semantic location features, *religion_now* shows strong correlation with low or lowered stress level, suggesting mental health benefit of worship places. The contrast between the positive correlation of feature *stationary_now* and the negative correlation of features like *walking_past* and *transit_past* in both value diagnosis and value prognosis in both suggest the benefit of traveling and not staying still in stress management. These interpretations provide motivation and evidence for future studies and smart health applications to further investigate and utilize their effect.

3.3.5.2 Predictive Modeling Result

Table 3.10 shows the mean and standard deviation of area under ROC (AUC) values obtained for all subjects in the across-subject LOSO and within-subject LOOO experiments using different groupings of features. One evident pattern we found is that regardless of feature groups and experimental settings, prediction performance is significantly higher under within-subject LOOO than across-subject LOSO. This observation suggests that historical Bluetooth encounters are more predictive of stress than encounter networks measured from other individuals. This can be explained by the fact that an encounter of the same particular characteristics can impact different individuals differently, due to personal differences in the

Table 3.10: Prediction performance (Area under ROC curve) achieved by different feature groupings under the across-subject leave-one-subject-out and the within-subject leave-one-observation-out evaluation settings

Across-subject LOSO								
AUC	Value Diagnosis		Change Diagnosis		Value Prognosis		Change Prognosis	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
BASE	0.639	0.123	0.626	0.111	0.628	0.136	0.639	0.127
MAIN	0.653	0.117	0.603	0.122	0.609	0.110	0.636	0.135
BASE+MAIN	0.660	0.120	0.602	0.110	0.611	0.122	0.624	0.136
Within-subject LOOO								
AUC	Value Diagnosis		Change Diagnosis		Value Prognosis		Change Prognosis	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
BASE	0.671	0.124	0.675	0.137	0.697	0.154	0.657	0.133
MAIN	0.712	0.158	0.672	0.162	0.666	0.143	0.663	0.136
BASE+MAIN	0.726	0.151	0.676	0.159	0.670	0.152	0.662	0.136

reaction to affective stimuli [165]. As such, in real world applications, caution should be taken applying existing models to new subjects. Calibration with personal data may be important when incorporating Bluetooth based features in stress recognition tasks.

My Bluetooth features prove predictive: all MAIN models achieved AUC values significantly greater than the random guess baseline of 0.5. However, when comparing performance achieved by our Bluetooth features combined with baseline features (BASE + MAIN) versus by baseline features only (BASE), we discovered mixed results. For value diagnosis, incorporating our Bluetooth features evidently improves prediction performance than baseline, confirming Hypothesis 3: mean AUC rose from 0.6706 to 0.7262 in the within-subject LOOO experiment (p-value 0.0564 in a one-tailed t-test). However, our results do not support Hypothesis 3.4 as performance change is minimal for the prognosis settings. Moreover, value diagnosis enjoys the highest AUC regardless of the prediction approach, regardless of feature groups. The relatively stronger performance under value diagnosis indicates more promising value of our proposed features in stress estimation applications, compared to forecast-oriented

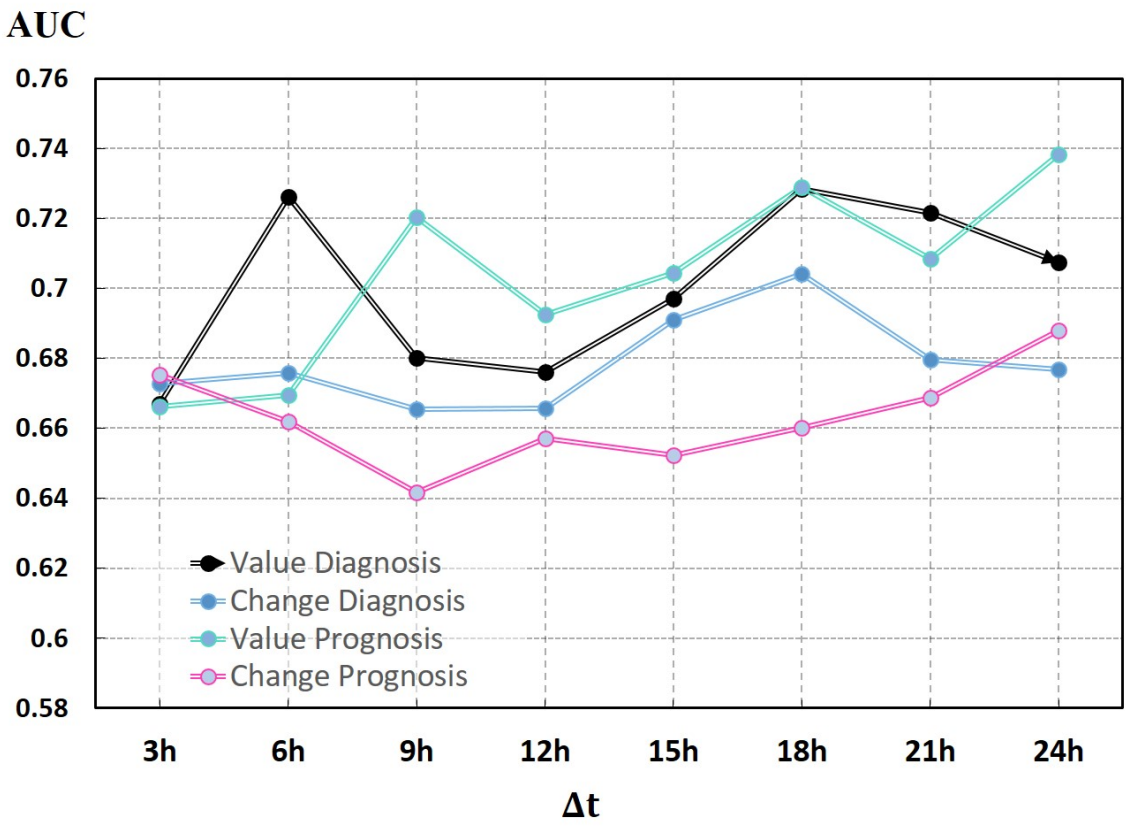


Figure 3.4: Sensitivity of prediction performance to the choosing of Δt, in the within-subject, leave-one-observation-out experiments using BASE + MAIN features. Shown in figure are mean AUC values for each feature extraction window size and prediction setting.

settings.

Our motivation to perform stress recognition using passively sensed proximity network data is rooted in the effect of social interaction on personal stress level studied in psychology research [133]. However, questions remain regarding the inertia and decay of the effect over time, which governs the extent to which past social interaction episodes affect current stress level and in turn concerns the choosing of feature extraction window in our predictive modeling tasks. We attempt to answer this question empirically by exploring the sensitivity of prediction performance to the size of feature extraction window Δt. Specifically, we repeat our within-subject leave-one-observation-out prediction experiments using MAIN + BASE features built with different sizes of Δt, namely $3h$, $9h$, $12h$, $15h$, $18h$, $21h$, and $24h$; and compare the resulting performances with that obtained with $\Delta t = 6h$ as shown in Table 3.10.

We plot the results in Figure 3.4.

The pattern shown in Figure 3.4 indicates that prediction performance does vary with different feature extraction window sizes but does not exhibit a monotonous or unimodal trend. For all experimental settings except change prognosis, we observe a bimodal shape in the trend of prediction performance: performance is relatively low at $\Delta t = 3h$; there seems to be a local maximum at $6h$ (value diagnosis and change diagnosis) or 9h (value prognosis) early on and a later local maximum that comes quite unanimously at $18h$. Although, for value prognosis prediction performance reached global maximum with $\Delta t = 24h$, unlike value diagnosis and change diagnosis for which performance reached maximum at $\Delta t = 18h$. In contrast, performance under change prognosis appears to have a V-shaped trend with $24h$ being the optimal Δt. From our observations it is difficult to draw a general conclusion on the optimal choice of feature extraction window but we recommend examination of various window sizes up to greater than 15 hours and be cautious using shorter-than-6-hour window sizes.

3.3.6 Conclusion

In this chapter, I addressed the need for accurate, timely, and unobtrusive stress monitoring technology through passively sensed Bluetooth encounter networks. Bluetooth is a widely available mobile sensor that can detect an individual's social environment based on physical proximity. Systematic feature engineering and examination of predictive value using Bluetooth encounter data is lacking in existing literature. We have investigated Bluetooth encounter data in terms of structural attributes, edge attributes, and neighbor attributes, and built features accordingly incorporating measures from social network analysis and concepts of social and temporal commonalities. Our correlation analysis, involving Bluetooth encounters among 49 student subjects over 66 days, suggests that our features extracted from Bluetooth encounter data account for the variance of stress outcomes significantly more effectively than null models. In doing so, we supported our Hypotheses 1 and 2, confirmed existing findings (e.g., sleep quality as a predictor of next day stress), and called attention to salient features

(e.g., *usual_max*) for consideration in real world stress recognition applications as well as their interpretations in social psychology.

Moreover, we tested the predictability of momentary stress levels from Bluetooth encounter data using random forest under four different prediction settings (value diagnosis, change diagnosis, value prognosis, change prognosis) and two evaluation settings (within-subject leave-one-observation-out and across-subject leave-one-subject-out). The results presented in the previous sections provide preliminary but promising evidence for the performance boosting effect of our Bluetooth features in momentary stress recognition applications. When incorporated with baseline features built with data collected from other standard mobile sensors and sleep quality self-reports, our Bluetooth features achieved performance improvement ($0.726 - 0.671 = 0.055$ in AUC) for value diagnosis but did not for other experimental settings. As such, we supported Hypothesis 3.3 and did not support Hypothesis 3.4. Further evidence is needed to evaluate the utility of our Bluetooth features in stress forecasting tasks. In practice, we recommend incorporation of our proposed features with those extracted from other mobile data sources (e.g., GPS, accelerometer, phone usage) as input for predictive modeling to enhance stress recognition performance.

The work presented in this chapter has several limitations. As the study subjects in the StudentLife dataset are all college students, the generalizability of our results to other demographic groups awaits further evidence. Moreover, as in many mental health monitoring tasks, ground truth curation is subject to the availability of self-report. Cognitive biases and low self-report compliance from the subjects could negatively impact the validity of response variable data, especially for change-based and forecast-oriented experimental settings. Finally, the data we assume available encompasses a relatively close-knit group of individuals who share significant social relations, as is the case in the StudentLife dataset and other group sensing scenarios; however, when this is not the case, we would only have a set of isolated, egocentric encounter "networks" as input. We propose and validate a feature engineering method addressing this very issue in the following section.

Future work may involve further evaluation of our proposed features in prediction tasks targeting other health outcomes (e.g., affect recognition). Another important task is to incorporate comprehensive contextual information obtained from other data sources such as GPS and accelerometer to more effectively and concisely represent social interaction. Evidence of the value Bluetooth encounter data provides for personal stress estimation presented in this paper motivates the incorporation of passively sensed in-person social network data with other sensor data sources in automated mental health monitoring.

3.4 Vector Space Model for Bluetooth Encounters

In this section, we propose a vector space representation of Bluetooth encounter data for mental health inference tasks, without needing information from any other individual's devices than the subject's own. Vector space modeling originated in the field of information retrieval [166] as an approach to quantifying the content of textual documents. Each document in such a model is a real-valued vector with its elements representing words in the overall vocabulary. The resulting space of documents supports comparison, similarity measurement, and prediction of outcomes such as thematic category.

I draw an analogy between a word in a vector space model of documents and a Bluetooth encounter between a subject's device and another device in natural environment. Over time the subject's device encounters multiple other devices, forming a collection of Bluetooth encounters that are analogous to the text of a document. In broad terms, the resulting encounters proxy for the narrative of a subject's social interactions. The aim of this section is to study the relationship between this narrative and mental health outcomes. Concretely, we convert Bluetooth encounters (time and device identifier tuples) into spatiotemporal tokens and treat each token as a separate dimension in a feature space. Our approach provides a fine grained representation of a subject's Bluetooth encounters, and we hypothesize statistical correlations between locations in the vector space and mental health outcomes.

When applied to mobile sensing data, the vector space model transforms raw sensor signals into tokens, and distinct tokens become orthogonal dimensions. Several studies have

applied this approach to represent mobile sensing data and extract behavioral patterns. Eagle et al. created a vector space over time and locations for each subject [167]. They divided each day into 24 hourly bins, categorized each GPS location as *home*, *work*, *other*, or *no-signal*, and considered each hour-location pair as a binary-valued feature. Thus, an individual's daily movements are represented in a 24×4-length vector. The authors then applied principal component analysis to reduce the dimensionality of the vector space, discovering salient components among the hour-location tokens that correspond with elements of an individual's daily routine. Do et al. [18] represented the proximity events among a group of co-workers using a vector space with each dimension being a distinct user-pair/time-of-the-day token, and then applied latent Dirichlet allocation (LDA) to the vector space to extract semantically coherent clusters. The clusters identified meaningful social communities within the group of co-workers. The goal of vector space modeling in these studies has been to understand human behavior through dimensionality reduction (whether PCA or LDA). In contrast, in this section I use vector space representations of Bluetooth encounters as a basis for predicting mental health outcomes.

To the best of my knowledge, this study is the first to apply vector space feature representations to Bluetooth encounter data for mental health inference. We propose and evaluate (1) Bluetooth encounter token designs including different combinations of temporal and spatial information; and (2) feature value weighting schemes such as binary value, term frequency (TF), and term frequency inverse document frequency (TFIDF) based on prediction performance in stress recognition tasks. We also compare our vector space features with two baseline feature groups: the Bluetooth network features described in Section 3.3 and vector space features of non-Bluetooth-encounter mobile sensing data similar to those constructed in [167]. Finally, we propose future work with topic modeling and word embedding methods on our vector space model to discover meaningful clusters of behavioral patterns. With this study we hope to inspire further discussion and research on bag-of-words approaches to representing mobile sensing signals for more effective mental health inference and management.

3.4.1 Feature engineering

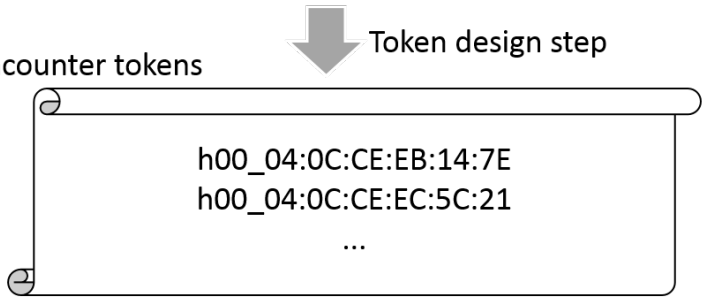
3.4.1.1 Bluetooth Encounter Vector Space Model

Within a Bluetooth data stream, an encounter takes the form of a device identifier with a timestamp. Representing these encounters in a vector space entails treating each distinct type of Bluetooth encounter — a distinct device identifier at a distinct time — as a dimension. The rationale behind this is that each Bluetooth encounter may represent a meaningful real-life proximity or interaction event that has implications for a mental health outcome. Concretely, we codify such information as follows. First, we bin all timestamps by hour (e.g., "2013-03-29 00:02:42" would be binned to "hour00") such that Bluetooth encounters that occurred within the same hour are given the same temporal label. Second, we concatenate the hour block information (e.g., "hour00") with the encountered device identifier to create a Bluetooth encounter token, which takes the format of "[hour block]-[device ID]". Alternatively, one may also choose to keep only the device identifier in the tokens and leave out the time, resulting in tokens with only the device identifier string itself. This treatment will further reduce the size of the resulting feature space. Figure 3.5 illustrates this process.

We assign values to the vector space features of Bluetooth encounters following the conventions of text-based vector space modeling. Specifically, we compute a binary token value, a token frequency (TF) value, and a token frequency inverse period frequency (TFIPF) value for each Bluetooth encounter in the vector space. Given multiple periods of time (e.g., multiple days) during which collections (documents) of Bluetooth encounters are detected by a subject's device, the binary token value is defined as 1 when a token appears in a document regardless of frequency and 0 otherwise; token frequency is defined as the number of times a token appears within a particular period of time; and token frequency inverse period frequency is defined as token frequency multiplied by the natural logarithm of the ratio of the number of collections where a particular token appears to the total number of periods. These three sets of feature values will be used for further analyses and compared.

Bluetooth encounter data

Timestamp	Device ID 1 (self)	Device ID 2 (other)
2013-03-29 00:02:42	u08	04:0C:CE:EB:14:7E
2013-03-29 00:02:42	u08	04:0C:CE:EC:5C:21
...	...	...



Vector space features

u08	h00_04:0C:CE:EB:14:7E	h00_04:0C:CE:EC:5C:21	...
Time period 1	1	1	...
Time period 2	...	...	...
...	...	...	...

Figure 3.5: Illustration of the vector space construction procedure. The upper table shows a sample of raw Bluetooth encounter data. The lower table shows the representation of time periods (analogous to documents) as vectors of time-device token frequencies (analogous to term frequencies).

3.4.1.2 Token Augmentation with GPS Data

The encountered device and the encounter time are two pieces of information native to the Bluetooth data stream. One can also incorporate information from external sensors, when available, to enrich the vector space of Bluetooth encounters. In this paper we explore the augmentation of Bluetooth encounter tokens with GPS data and evaluate the resulting predictive value. Ideally, at the time of any Bluetooth encounter, a subject should have a location that is detectable by a GPS sensor; therefore, one can concatenate the GPS information with the [hour block]-[device ID] tokens to create richer tokens representing a Bluetooth encounter event, with one downside being an enlarged feature space. We encode

GPS information through coordinate anonymization, which truncates significant digits of a GPS coordinate to map a raw coordinate (a point) to the rectangular area it belongs to such that no finer geographic information is retained. GPS coordinate anonymization is typically performed to protect user privacy in data transmission. In vector space construction it serves a purpose similar to binning timestamps.

3.4.1.3 Bluetooth Encounter Network Features

Creating a vector space model for Bluetooth encounters does not require any information other than what is available from the subject's device. However, in some cases other sensor data are provided by encountered devices. For example, when we have mobile sensing data from all students who enroll in the same class in college, we will usually observe a Bluetooth encounter network encompassing multiple individuals as nodes and encounters between them as edges. This enriched network provides topological attributes of the local encounter network surrounding a subject as well as information on the social nature of a subject's encountered devices (e.g., are they familiar or fortuitous encounters, how many common encountered devices two encountering devices share). We previously proposed and evaluated three groups of Bluetooth encounter network features, namely structural, edge, and neighbor attributes in Section 3.3.1. These encounter network features will be used as a baseline for comparison with vector space features in the present paper. Intuitively, Bluetooth encounter network features take advantage of information about a subject's encountered "neighbors" in addition to the subject itself; therefore, we hypothesize these network features will produce higher performance than the vector space features, which only utilize the subject's encounter patterns and thus assume no knowledge of the behaviors of the encountered devices.

3.4.1.4 Baseline Vector Space Features

To measure the predictive value of Bluetooth encounter vector space features, we construct non-Bluetooth tokens containing only temporal and spatial information as a baseline feature group. We divide a day into 24 hourly bins, treat a subject's anonymized GPS locations as places, and create distinct hour-place pairs as separate dimensions in a vector space. For

example, if a subject is detected to be at place A at 2:36pm, then feature/hour-place token [hour 14]-[place A] will be created and assigned value 1, otherwise 0. We will compare the performance of these spatiotemporal features alone versus their combination with Bluetooth encounter features in stress recognition tasks.

3.4.2 Hypotheses

The following hypotheses examine the predictive value of our Bluetooth encounter vector space features compared to non-Bluetooth, baseline features (Hypothesis 3.5) and Bluetooth encounter network features (Hypothesis 3.8) targeting personal stress outcomes. The hypotheses also investigate the implications of vector space design (Hypothesis 3.6) and feature value weighting scheme (Hypothesis 3.7) for prediction performance.

- **Hypothesis 3.5**: Binary, device-only vector space features of Bluetooth encounters, when combined with baseline vector space features, predict personal stress levels more accurately than baseline vector space features alone.

- **Hypothesis 3.6**: As the vector space token design of Bluetooth encounters changes from device-only to device-time to device-time-location, prediction performance targeting personal stress levels will increase.

- **Hypothesis 3.7**: As the vector space feature weighting scheme changes from binary value to TF to TFIPF, prediction performance targeting personal stress levels will increase.

- **Hypothesis 3.8**: Vector space features of Bluetooth encounters predict personal stress levels less effectively than the Bluetooth encounter network features [134].

3.4.3 Data

We use two mobile sensing datasets to test our hypotheses: (1) the StudentLife [24] dataset and the Friends & Family [21] dataset. As introduced in Section 3.3.3, the StudentLife dataset was collected from 49 Dartmouth college students who enrolled in the same class and contains Bluetooth encounters scanned every 10 minutes among them over two months. The Friends

& Family dataset was collected from young faculty members and their spouses totalling 117 people who lived in the same residential complex at a major research university in North America and contains Bluetooth encounters scanned every 5 minutes among them over 10 months. Bluetooth encounters in both datasets take the format shown in Figure 3.5. GPS data are available in both datasets although they are affine-transformed in the Friends & Family dataset.

Ground truth data on personal stress level is obtained through mobile phone surveys in both datasets but with different temporal resolutions. In the StudentLife dataset, stress level measurements are obtained through ecological momentary assessment (EMA) surveys deployed on the participants' smartphones multiple times per day at random times. The survey consists of a question with text "Right now, I am..." and 5 response options "feeling great", "feeling good", "a little stressed", "definitely stressed", and "stressed out". We consider the first two as non-stressed and the latter three as stressed. In the Friends & Family dataset, stress surveys are deployed at the end of each day soliciting an assessment of a participant's perceived stress level on the past day. The question reads "On a scale of 1 to 7, how stressed were you on [day] (with 1 being very unstressed, 4 being neither stressed nor unstressed, and 7 being very stressed)?" and as the question text suggests, participants are asked to choose among 7 options with 5,6,7 being a stressed response and 1,2,3,4 being a non-stressed response.

3.4.4 Experiments

To test Hypotheses 3.5-8, we perform predictive modeling targeting a categorical stressed/non-stressed response variable as discussed in Section 3.4.3 using vector space features constructed following the descriptions in Section 3.4.1. For the StudentLife dataset, we treat Bluetooth encounters detected within the 18-hour window leading up to each stress level self-report as a document. For the Friends & Family dataset we treat the entire day corresponding to each end-of-day self-report as the time period associated with each document. We experiment with naive Bayes, support vector machine, and random forest classifiers. Random forest yielded

the best results and we will only report performance by random forest in the next section. To evaluate the prediction results we adopt a subject specific leave-one-out cross validation setup where area under the ROC curve (AUC) is computed for each subject in each dataset.

3.4.5 Results

As shown in Table 3.11, incorporating Bluetooth encounter vector space features (binary valued, device-only tokens) improved prediction performance compared to spatiotemporal baseline features alone for both the momentary and the daily stress recognition tasks. This result supports Hypothesis 3.5 and our proposed vector space representation of Bluetooth encounter data for mental health inference.

Listed in Table 3.12 are the average AUC scores achieved by the three vector space token designs and the three feature value weighting schemes with the two datasets. The best performing group is the device-time-GPS tokens with TFIPF feature values for the momentary recognition task using the StudentLife dataset (AUC = 0.714) whereas the best performing group for the daily recognition task using Friends & Family dataset is the simpler binary valued device-only token features (AUC = 0.664). Our results do not support Hypothesis 3.6 or 3.7 as more sophisticated token design and feature value scheme did not result in better performance.

Comparing the best performance achieved across the three token designs, we found that device-time tokens performed poorer in both stress recognition tasks than device-only and device-time-GPS tokens. Our results were inconclusive regarding the relative performance of the latter two designs. This suggests that the identity of one's Bluetooth encounters drives the predictive power and the timestamp of the encounters introduces more noise to predictive modeling than the additional information it provides. In other words, whether one is in proximity of a particular device may matter more than when such proximity events occur, as far as the effect on one's stress level is concerned. This result also indicates the value of geographic information in mental health inference, as the addition of GPS information in token design improved the performance of device-time tokens.

Table 3.11: Binary valued, device-only vector space features of Bluetooth encounters enhance stress recognition prediction performance (AUC).

Features	StudentLife (momentary)	Friends&Family (daily)
Hour-GPS	0.688	0.649
Hour-GPS + Bluetooth	0.714	0.662

Table 3.12: Predictive performance (AUC) with each vector space token design, feature value scheme, and dataset; the best value for each token design is bolded.

Token design	Weighting	StudentLife (momentary)	Friends&Family (daily)
Device-only	Binary	0.689	**0.664**
	TF	0.705	0.642
	TFIPF	**0.711**	0.654
Device-hour	Binary	0.683	0.634
	TF	0.688	0.636
	TFIPF	**0.709**	**0.643**
Device-hour-GPS	Binary	0.709	**0.654**
	TF	0.683	0.644
	TFIPF	**0.714**	0.649

Table 3.13: Predictive performance (AUC) comparison of vector space and network features of Bluetooth encounters.

Features	StudentLife (momentary)	Friends&Family (daily)
Best Vector Space	0.714	0.664
Network	0.735	0.631
Network + Best Vector Space	0.760	0.654

Comparing the best performance achieved with the three feature value weighting schemes, we found that TF is consistently the worst choice in both tasks with each token design. We expected that TFIPF would perform better than TF as in many text-based problems; however, TF did not outperform binary values in our tasks. Between results obtained with the two datasets, a shared pattern is that device-time tokens with TF feature values appear to be the worst choice when modeling personal stress outcomes, regardless of its temporal scale (momentary versus daily).

Table 3.13 compares the prediction performance of our best vector space features with Bluetooth network features as well as the two groups combined. For the momentary problem, Bluetooth network features performed better than vector space features, supporting Hypothesis 3.8 that information about behaviors of the devices (and by extension their users) encountered by a subject carries information regarding the subject's mental health status. When the two feature spaces are combined, they achieved better performance than each alone. For the daily task, the comparison between vector space and network features is reversed. We suspect that the differences in these comparisons results from the different degree of closeness of the social relationships between study participants in the two datasets. In the StudentLife study, participants are undergraduate students enrolled in the same class who likely have integrated social lives. We observed many non-participant Bluetooth encounters shared by these subjects. In contrast, in the Friends & Family study, the tie between participants is that they live in the same apartment complex, which might not include daily socialization events. Although Bluetooth encounter features can be computed regardless of the closeness of the network, the characteristics of a subject's Bluetooth encounter network are likely more indicative of health outcomes when the Bluetooth encounter network reflects proximity events with a higher proportion of a subject's social contacts. Testing this hypothesis is beyond the scope of this paper; however, this finding (1) validates the utility of vector space representation of Bluetooth signals especially when we do not have access to proximity network data of a subject's close social contacts (as in the case of Friends & Family dataset); and (2) motivates smart health researchers and practitioners to incorporate Bluetooth encounter network features (Section 3.3.1) with vector space features proposed in this paper when data are available for a relatively close-knit group.

3.4.6 Conclusion

This study proposed a vector space representation of Bluetooth encounter data for mental health inference and measured predictive utility in stress recognition tasks with two public datasets. Our results support Hypothesis 3.5 regarding the value of vector

space representations of Bluetooth encounter data. Our results do not support Hypothesis 3.6 or 3.7, revealing implications of token design and feature value scheme for prediction performance. Lastly, our results partially support Hypothesis 3.8, indicating that closeness of social relationships could be a factor in Bluetooth encounter network features' predictive advantage over vector space features.

We envision several directions of future research, as a vector space representation of Bluetooth encounters opens the door to several additional methods. Topic modeling might be used to extract clusters of proximity events that convey social insights. In particular, we anticipate the applicability of structural topic modeling (STM) [168], which allows the effect of a document-level covariate on the topic composition to be modeled. A fluctuating mental health status associated with a period of time from which a collection of Bluetooth encounters took place naturally serves as a document-level covariate. How such covariates (e.g., high or low stress levels) correlate with topical composition of Bluetooth encounters would be an interesting avenue of exploration. Moreover, word embedding methods [169] might be used to compute a weight vector for each token in a vector space, with higher degree of similarity between weight vectors indicating higher likelihood of token co-occurrence. Such weights may provide a useful representation for understanding human behavior and predicting health outcomes. Finally, the applicability of vector space representation extends to many mobile sensors beyond Bluetooth encounters (e.g., accelerometer and audio). Future discussion and research on bag-of-words approaches to representing raw mobile sensing signals for intelligent health inference would further validate and extend our preliminary results.

3.5 Discussion

A primary limitation we currently face is in the existing sensing technologies to capture genuine in-person social interactions in an accurate, discreet, and privacy preserving fashion. Existing sensing technologies differ in their physical properties and thus the theoretical construct they may proxy: Bluetooth and RFID both measure physical proximity (although through different mechanisms) whereas infrared sensors are triggered by alignment of sight

and thus can better capture face-to-face interactions between two equipped individuals. Then comes different requirements of hardware, which further determine their applicability in different real-life settings: RFID and infrared sensors require special purpose devices (e.g., sensor badge and base station) which are out of the comfort zone of daily wearing for most users and render large-scale ubiquitous usage unfeasible; Bluetooth radio has advantage over the other technologies in this regard.

Bluetooth encounters, triggered by physical proximity, is not necessarily produced by an in-person interaction [170]. Because of this, the correlation discovered and predictability achieved in this chapter should be regarded as evidence for as well as rationale for incorporating our proposed features in future stress recognition tasks, as opposed to for causal relations between certain characteristics of in-person social interaction and mental health. To accomplish the latter purpose we would need more advanced sensing technology and face-to-face interaction discovery algorithms as well as causal experimental designs that are out of the scope of this study but will be the focus of future work.

The work presented in this chapter is based on Bluetooth encounter data; however, one should note that the methods are broadly applicable to physical social signals, agnostic to the measuring technology. The core notion we study is a timestamped link between the subject in question and his or her social contacts, which can manifest as proximity-based encounter, face-to-face conversation, or online interaction. In fact, the feature engineering methodology described in this chapter has been incorporated in recent studies addressing loneliness issues through the lens of online conversation dynamics [111].

3.6 Concluding Remarks

In this chapter we have created two major feature engineering methods with Bluetooth encounter data collected from smartphone users in uncontrolled settings to improve automated cognitive stress recognition: a network analysis approach extracting topological and social insights from the Bluetooth encounter network in which a subject is situated, and; a vector space approach encoding Bluetooth encounter events captured from a subject's personal device

as individual spatio-temporal features. Mining physical social signals to help understand fluctuations in personal mental health is motivated by social psychology research that predates wide adoption of mobile sensing technology; but proves a beneficial component to incorporate in future sensing-based smart health solutions.

Future work should primarily focus on developing new hardware and algorithms for more effective social interaction detection that are better suited for discreet daily usage. An important source of information to harness is user voice pattern, which has not been sufficiently utilized partly due to privacy concerns and partly due to limitations of existing wearable devices. Improvement in detection may best be achieved by combining different sensor data streams and through fusion methods. Moreover, my current work in stress recognition has been based on undiagnosed, "healthy" cohorts. I anticipate the next steps to zero in on specific health conditions and behavioral scenarios, such as depression and autism, to investigate the effect of physical social signals on symptoms and fulfill targeted clinical needs. Finally, the methods proposed and validated in this chapter are generalizable to other micro-level human applications targeting affective, cognitive, and health outcomes and I recommend their incorporation in future tasks.

Out-scoping Modeling Methods for Mining Social Signals in CHS

Over the recent decade, we have witnessed a surge in data collection and analytics effort aimed at understanding the relations between pervasive sensing signals and critical human outcomes to enhance human well-being and capability. Among these signals are social ones, which are the central construct of this dissertation, measuring and proxying for social stimuli people receive from their online and offline social environment. Various sources of social signals data have been collected, ranging from social media platforms to smartphone embedded sensors. Various feature engineering and machine learning techniques have also been extensively explored and validated. However, besides application-driven and data-driven feature engineering, which was the main perspective of Chapters 2 and 3, questions remain what theoretical frameworks and models are useful to extract insights from social signals in cyber-human systems otherwise unobtainable.

In this chapter, I seek to identify theoretical frameworks suitable for modeling social signals in cyber-human systems and broaden the scope of modeling methods currently in use in existing literature. I —for the first time— discover a connection between two theoretical frameworks and modeling approaches, namely relational event model (REM) and inverse reinforcement learning (IRL), on their mathematical machinery to characterize and predict directed social interactions. The quantitative characteristics learned by both modeling methods may serve as sole or additional evidence to support predictive tasks targeting outcomes such as team performance or satisfaction. Moreover, I prove that maximum entropy inverse reinforcement learning [171], when applied to a group social interactions modeling problem, is equivalent to a relational event model. The demonstrated connections, commonalities, and uniquenesses introduce fresh perspective and tools for social network

mining, identifies novel applications for reinforcement learning algorithms, and inaugurates research opportunities at the nexus of social network analysis and machine learning.

4.1 Introduction

Network is an effective abstraction of virtually infinite physical and social phenomena. Whenever one can define a set of entities (i.e., nodes) and find some relationships (i.e., edges) between the entities, one has at hand a network. Network is especially a natural fit for social signals in cyber-human systems, with individual technology users as nodes, and different types of social signals that can be unified under the notion of edges and characterized by different edge attributes such as time, location, modality (online, offline), and content (textual, verbal): in the cyberspace, the edges are multimedia information circulated among users whereas in the physical space the edges are certain modes of in-person interaction.

As such, I argue that the art of mining social signals of cyber-human systems stands to benefit from the rich theories and models in network science. In addition to well established groups of graphic metrics (some of which are discussed in Section 3.3), statistical network modeling is a major advantage network science research could provide to mining cyber-human system social signals especially in group social interaction scenarios. Researchers have typically resorted to a family of exponential random graph models [31] (ERGM) to understand the network *structure* emerging from group social dynamics. The central question ERGM answers is why certain edges exist in a network whereas others do not; and the model expresses the likelihood of a particular tie existing in a network as a transformed linear combination of features that characterize said tie. ERG family models have been studied predominantly in the context of sociology and with survey and panel data, and whose value in the context of cyber-human systems and pervasive sensing data awaits further investigation.

Not only the network structure, but the temporal *sequence* of edges matters as well, especially when we are increasingly able to obtain fine-grained, time-stamped social actions (such as phone call, SMS, email, and face-to-face conversations) between mobile technology users. Among the statistical network models, **relational event model** (REM) [32] was

proposed to specifically model the order of time-stamped or time-ordered dyadic actions in a social network. The premise of REM is that the dynamics of dyadic actions in a social system follow a set of intrinsic but latent characteristics the system possesses: for example, if a group of social actors value the reciprocation of a dyadic action one would see much back-and-forth between the entities (e.g., physical attacks that justify retaliation) whereas if the group enjoy the relay of a dyadic action one would tend to see the direction of the action forming a chain along the entities (e.g., secrets getting circulated in a group). Thus, given an observed recent history of dyadic actions among entities in a social system, and a particular set of intrinsic characteristics, some dyadic actions are more likely to follow than others. By fitting REM to dyadic action data, one can learn the effects of the hypothesized characteristics that are most probable to have resulted in the observed dynamics of social actions. These learned effects of the characteristics constitute valuable insights into how a social system functions.

Now we look to the field of artificial intelligence. There lies another theoretical framework, namely reinforcement learning (RL), that theorizes that the behavior of an intelligent agent is driven by accumulating higher reward when interacting with its environment. Comparing the state the agent is in before and after a behavior is committed, the behaviors that result in high rewards would be favored and those that result in lower rewards would be avoided in the future. Reinforcement learning assumes knowledge of such rewards and aims to find an optimal behavioral strategy, or policy, that would achieve highest reward accumulated over time. Its inverse form, **inverse reinforcement learning** (IRL) on the other hand, observes the behavior of an intelligent agent and seeks to solve for the reward that must have driven the behaviors observed. In this sense, we speculate a theoretical connection between IRL and REM, as both approaches observe a trajectory of actions, both assume the existence of particular underlying characteristics that are supposed to have given rise to the observed patterns, and both try to find their values as a way to explain observed behaviors. If confirmed, such connections will show the applicability of inverse reinforcement learning algorithms to social signals mining problems in group interaction settings and thus broaden

the scope of existing modeling methods.

4.2 Related Work

4.2.1 Exponential random graph model

To begin to understand social networks, a primary question to answer is why some edges exist while others do not among a group of nodes. Exponential random graph model (ERGM) [172] answers this question by theorizing that the existence of edges in an observed network is due to the underlying phenomenon represented by the network preferring certain configurations of network topology and nodal attributes, and offers an inferential solution. For example, if the social process within a group of entities favors homophily, then we would expect to observe more edges forming between nodes that share similar traits than those who do not. In the setup of ERGM, we observe a network y that consists of N nodes and is represented by an adjacency matrix with each element $y_{ij} = 0, i, j \in \{1, ..., N\}$ if there is an edge between nodes i and j and 0 otherwise. The observed network y is a realization of a random graph Y over the same set of N nodes with edges pending and the probability of the observed network y as:

$$P(Y = y|\theta) = \frac{\exp[\theta^\top s(y)]}{\sum_{y' \in Y} \exp[\theta^\top s(y')]} \tag{4.1}$$

where y' is any possible realization (sample of edges) of random graph Y, $s(y)$ is a feature vector quantifying the network topology and nodal attributes that influence edge existence, and θ is the weight vector associated with the features. Through Equation 4.1 one can derive the logit (log odds) of each edge y_{ij} given all other edges unchanged:

$$logit(y_{ij} = 1) = \log[\frac{P(Y = y^+|\theta)}{P(Y = y^-|\theta)}] = \theta^\top [s(y^+) - s(y^-)] \tag{4.2}$$

where y^+ and y^- are the network realization with and without edge y_{ij}, respectively, with the rest of the network identical to the observed y.

The core of ERGM is that it expressed the probability of an observed network as proportional to the exponential of a weighted sum of the features $s(y)$. Then, not unlike a logistic regression model, ERGM is fitted to real network data to infer the weights θ, through which one learns the inherent characteristics (hypothesized by the researcher) that have shaped a network into the pattern it appears. ERGM is most suitable to model static networks that are like snapshots of group interactions and relations. Therefore, the traditional theater of applications of ERG models is in sociology research to understand patterns and driving factors of social phenomena such as friendship [173], co-authorship [174], and inter-organizational relations [175, 176].

Not only the existence of edges in a static network is of interest, but why some edges come into existence while some others disappear over time in a temporal or dynamic network is also highly important and relevant in human behavior. Several different but related models (Table 4.1) have been devised based on ERGM to explain such edge formation (and dissolution in some cases). We identify three key aspects of distinction. First, the **directed** or **undirected** property of edges in a social network, requiring different modeling techniques; for example, liking someone is a directed tie as the affection is not necessarily reciprocated, whereas being in a relationship with someone is an undirected tie as the pair involved have to be in agreement to be with one another. Second, a dynamic network model may adopt a **tie-oriented** or **actor-oriented** perspective. A tie-oriented approach makes no assumptions about the decision making process of individual actors and weighs all legal edges in a social network based on their likelihood to exist or happen next; whereas an actor-oriented approach explicitly models the desirability by an individual actor to alter (i.e., form, terminate, strengthen, weaken) its edges linking with other actors. Finally, although all dynamic network models by definition require network data with temporal information ingrained, whether one knows when a change in network edges happens leads to the distinction between **panel** and **timestamped** network

Table 4.1: Summary of major exponential-family random graph models for networks

Model	Modeling Perspective	Input Data
Stochastic actor-oriented model [177]	Actor-oriented	Panel
Relational event model [32]	Tie-oriented	Timestamped
Temporal exponential random graph model [178]	Tie-oriented	Panel
Dynamic network actor model [179]	Actor-oriented	Timestamped

as model input. Panel data, on which longitudinal social network research typically relied on before mobile technology became advanced and widely available, is obtained like snapshots of the same group at different points in time; as such, changes of edges can happen between two consecutive panel collections and their exact timing is nowhere to be known. On the contrary, timestamped data records the exact time of every event that takes place within a social network. This distinction is not only about the applicable input data, but also has implications on the type of relations modeled: a timestamped event usually constitutes a discrete, sometimes ephemeral action such as sending an email or terminating relationship with someone; however, panel data can accommodate more "status-like" relations which are difficult to pinpoint exact time of changing but can be measured at a given time.

4.2.2 Relational event model

Due to its capability of taking full advantage of time-stamped (or time-ordered) activity data, relational event model (REM) is especially useful for modeling directed social interactions that are monitored and recorded in cyber-human systems. Full details can be found in its original paper [32] and we review the key building blocks here. First and foremost, a dyadic action or "relational event" $a \in \mathbb{A}$, where $\mathbb{A}$ is the set of all legal actions, is fully characterized by five elements: (1) a sender $s(a) \in S$, where S is the set of senders; (2) a receiver $r(a) \in R$, where R is the set of receivers; (3) an action type $c(a) \in C$, where C is the set of action types; (4) a timestamp $\tau(a)$, and; (5) descriptive covariate(s) X_a, which may or may not be applicable. At a given time t, we have observed A_t, a time-stamped (thus time-ordered) history of M dyadic actions $a_1, a_2, ..., a_i, ..., a_M$ (a_M is the most recent) and we denote a_0 as

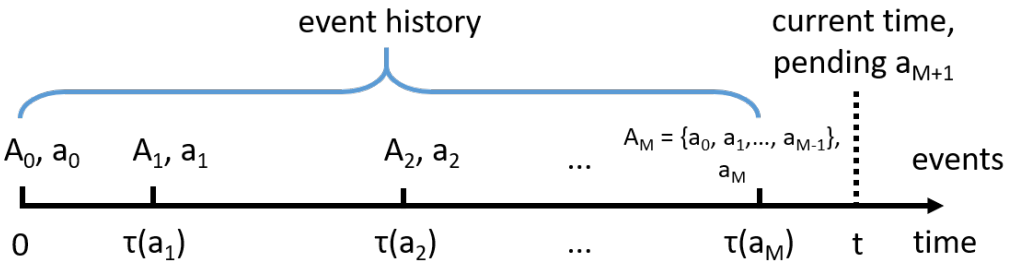


Figure 4.1: Illustration of timestamped relational event history, the input data for REM

a place-holder for the beginning point of the observed sequence with its timestamp $\tau(a_0) = 0$. The event history is illustrated in Figure 4.1.

To specify the probability density of such an event history, REM resorts to survival and hazard functions [180]. This is a fitting approach because an event history clearly consists of multiple eventless periods between discrete dyadic actions scattered over time. A survival function $S_a(t)$ expresses the probability of an action a not happening for duration $0{\sim}t$ while a hazard function $h_a(t) = \frac{\partial[1-S_a(t)]}{\partial t}/S_a(t)$ quantifies the propensity that an action a did occur when some action was to happen. As such, the probability of an event history can be written as a product of multiple survival functions and hazard functions as Equation 4.3.

$$p(A_t) = \prod_{i=1}^{M}[h_{a_i}(\tau(a_i)) \prod_{a\in\mathbb{A}} S_a(\tau(a_i) - \tau(a_{i-1}))] \times \prod_{a\in\mathbb{A}} S_a(t - \tau(a_M)) \tag{4.3}$$

$$h_{a_i}(\tau(a_i)) = \lambda(a_i, A_{\tau(a_{i-1})}, X_{a_i}) \tag{4.4}$$

$$= \exp\theta^T u(a_i, A_{\tau(a_{i-1})}, X_{a_i}) \tag{4.5}$$

Based on the intuition that actions with different senders, receivers, types, preceding events, and other descriptive covariates should have different likelihood of happening, REM

further specifies the hazard function as a rate function (Equation 4.4), which is further expressed (Equation 4.5) as the exponential of the weighted sum of a vector of *sufficient statistics*, $u(a, A_t, X_a)$, featurizing an action based on its sender, receiver, type, event history, and descriptive covariates. θ is a coefficient vector associated with the sufficient statistics quantifying their effects. The rate function governs the distribution of probability associated with each action given the past event history. In cases where the temporal order of the events is available but exact timestamps are not, the probability of an event history realizing can be equivalently written as Equation 4.6 (proof available [32]), which relaxes the requirement for timestamped data.

$$p(A_t) = \prod_{i=1}^{M} [\frac{\lambda(a_i, A_{\tau(a_{i-1})}, X_{a_i})}{\sum_{a' \in \mathbb{A}} \lambda(a', A_{\tau(a_{i-1})}, X_{a'})}] \tag{4.6}$$

By expressing the hazard function in terms of linear combinations of sufficient statistics, REM allows not only straightforward parameterization of certain realizations of event trajectory but also straightforward inference procedure through likelihood-based methods (e.g., maximum likelihood estimation or MLE) to learn the values of coefficients θ that best fit real data. We find applications of REM in diverse domains to understand social behavior such as team processes [181][182], friendship [183], zoology [184][185], education [186], and health care [187]. The group interaction dynamics learned by REM are anticipated to be useful for individual and group outcome prediction such as task performance [182].

4.2.3 Inverse reinforcement learning

The life of all living creatures naturally involves observing changing environment and acting upon it in a way favorable for their survival and prosperity. In artificial intelligence applications where we train a computer to complete human tasks (e.g., playing chess), we also require it to be able to make "good moves" given a situation (e.g., positions of self and opponent pieces) that are conducive to favorable outcomes (e.g., winning). As such, questions

Table 4.2: Summary of major inverse reinforcement learning algorithms

Algorithm	Strategy
Linear programming IRL [34]	Maximize minimum difference in feature expectation with linear constraint
Apprenticeship learning IRL [188]	Maximize minimum difference in feature expectation with quadratic constraint
Maximum entropy IRL [171]	Maximize likelihood of state-action trajectory under principle of maximum entropy

arise as to how a human or machine should choose actions based on the observed environment, which can change as a result of previous actions taken, to achieve particular goals over time.

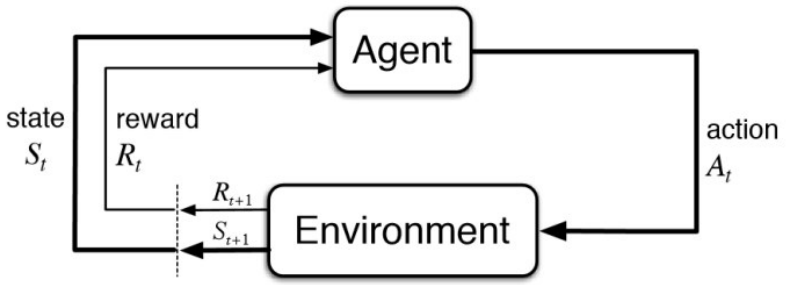


Figure 4.2: Reinforcement learning diagram

Reinforcement learning is a theoretical framework aimed at officially answering these questions. In its basic setup, an agent finds itself in a *state* s at each time step t, which characterizes the context in which the agent is situated, and takes an *action* a, which alters the context and thus brings the agent to a new state s'. Both s and s' belong in a *state space* S, the set of all possible states, and all possible actions form an *action space* A. State transition is governed by transition probabilities $P_a(s, s')$, specifying the probability that the agent will arrive in state s' after taking action a while in state s. Upon arriving in the new state s', the agent receives a *reward* (or *reward function*) $R(s)$, which quantifies the desirability of being in the new state; a reward value is associated with each state in the state space S. The agent interacts with the environment by repeating the state-action cycle and receiving and accumulating reward in the meantime. The way the agent behaves follows a *policy* $\pi(s) = P(a|s)$, governing the probability with which the agent is to take an action

$a \in A$ when in state s. A deterministic policy specifies one action to take given a state whereas a stochastic one provides a probabilistic distribution over multiple actions. Following different policies, an agent will take actions and visit states differently, thus achieving different amount of accumulated reward over time. A *value* (or *value function*) $V^{\pi}(s_0) = \sum_{t=0}^{\infty} \gamma^t R(s_t)$ is defined for each state as the discounted accumulated reward when the agent starts from state s_0 and behaves onward following policy π. γ is a discount factor converting a reward received later on into its current value, indicating a preference for long- or near-term reward seeking. The core problem of reinforcement learning is to find an optimal policy π^* for the agent that maximizes its accumulated reward regardless of starting state, given knowledge of the states, actions, and rewards. This optimization problem can be solved by dynamic programming algorithms and has extremely broad applications in artificial intelligence.

$$\pi^* = argmax_{\pi} V^{\pi}(s_0) = argmax_{\pi} \sum_{t=0}^{\infty} \gamma^t R(s_t), \forall s_0 \tag{4.7}$$

Reinforcement learning assumes knowledge of reward and aims to solve for an optimal policy to serve as a behavioral guidance for the agent; for example, to direct a robot to successfully navigate a labyrinth or win a chess game against human opponents. However, in some behavioral cases (e.g., driving a car), we don't know or can't define the reward straightforwardly; instead, we know the state and action spaces, observe an example agent's behavior, and want to find the underlying rewards the agent must be seeking to be behaving the way it does, which can be further used to train other agents or simply to understand the example agent's behavioral patterns. This problem calls for inverse reinforcement learning, where we observe a trajectory of state-action trajectory ζ (as illustrated in Figure 4.3) and aim to solve for the rewards. In the process of solving for the rewards, policies can be computed to generate agents that behave similarly to the demonstrated behavior of the example agent.

$$R(s) = \theta^\top f_s \tag{4.8}$$

We identify several major IRL algorithms in Table 4.2; all of them take demonstrated or "expert" state-action trajectories as input data, assume knowledge of (or the ability to empirically estimate) transition probabilities, and aim to solve for the most fitting rewards and the corresponding policies. A key approach to solving for rewards that is common across these IRL algorithms is to represent state s as a feature vector $f(s)$ and theorize the corresponding reward $R(s)$ as a linear combination of the features with the weights θ indicating the importance of a feature. This further entails: (1) in maximum entropy IRL [171], the decomposition of the reward (i.e., non-discounted accumulated value) of a given trajectory ζ into a weighted sum of *feature counts* $\mathbf{f}_\zeta$:

$$V_\zeta = \sum_{s_i \in \zeta} R(s_i) = \sum_{s_i \in \zeta} \theta^\top f_{s_i} = \theta^\top \mathbf{f}_\zeta \tag{4.9}$$

and (2) in Linear Programming and Apprenticeship Learning IRL [34][188], the decomposition of the expected value of a policy π (i.e., of all possible trajectories realized under policy π) into a weighted sum of *feature expectations* $\mu(\pi)$:

$$\begin{aligned} E[V^\pi] &= E_\pi[\sum_{t=0}^{\infty} \gamma^t R(s_t)] = E_\pi[\sum_{t=0}^{\infty} \gamma^t \theta^\top f_{s_t}] \\ &= \theta^\top E_\pi[\sum_{t=0}^{\infty} \gamma^t f_{s_t}] = \theta^\top \mu(\pi) \end{aligned} \tag{4.10}$$

with the same set of weights θ as in Equation 4.8. IRL algorithms then seek to find the θ that will behavior that is as close as possible to that of the demonstrated state-action trajectories.

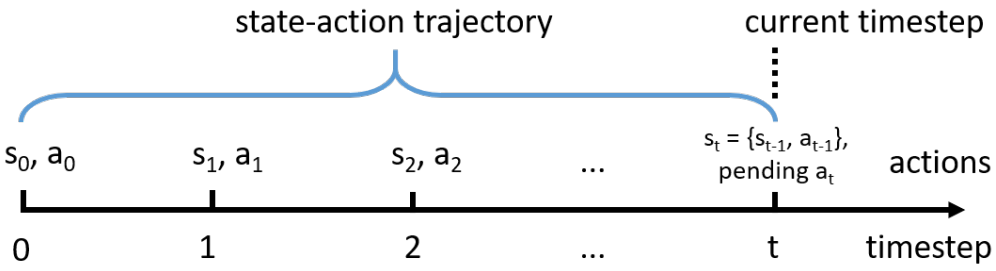


Figure 4.3: Illustration of state-action trajectory, the input data for IRL

Key differences exist in the optimization strategy and procedure (see Table 4.2). On one hand, Linear Programming IRL [34] and Apprenticeship Learning IRL [188] are both based on feature expectation; the core problem is to find a policy π^* whose feature expectation is as close as possible to that of the example agent's state-action trajectories. Maximum entropy IRL [171], on the other hand, utilizes principle of maximum entropy to express the probability of any trajectory to be proportional to the exponential of the feature counts of the trajectory:

$$P(\zeta|\theta) = \frac{\exp \theta^\top \mathbf{f}_\zeta}{Z(\theta)} = \frac{\exp \theta^\top \mathbf{f}_\zeta}{\sum_{\zeta'} \exp \theta^\top \mathbf{f}_{\zeta'}} \tag{4.11}$$

and then maximize this probability over θ. We find applications of IRL in many imitation learning applications; however, using IRL to solve social interaction modeling problems is still a largely unbeaten path. One study [189] exists on modeling group interaction dynamics using Markov decision process (MDP) to solve for the reward associated with different interaction states. Another study identifies the theoretical connections and equivalence between IRL [190] and general adversarial network (GAN), a class of unsupervised machine learning algorithm. In the next sections we (1) examine the theoretical connections between REM and IRL in general and (2) prove that when applied to a group interactions modeling problem with the group rather than an individual actor treated as the agent, maximum entropy IRL is in fact equivalent to REM.

Table 4.3: Equivalent elements of relational event model and inverse reinforcement learning frameworks

REM	IRL
Group with N actors	Agent
Event history $A_i = A_{\tau(a_{i-1})}$	State s_i
Relational event/dyadic action a_i	Action a_i
Newly realized event history $A_{i+1} = \{a_i, A_{\tau(a_{i-1})}\}$	New state s_{i+1}
Trajectory of M realized histories $\{A_1, ..., A_M\}$	State trajectory $\zeta_{(T)} = \{s_1, ..., s_T\}$
Sufficient statistic $u(A_i)$	State feature $f(s_i)$
Coefficient θ	Reward weight θ
Rate/hazard function $\lambda(A_i) = \exp[\theta^\top u(A_i)]$	Reward function $R(s_i) = \theta^\top f_{s_i}$

4.3 Theoretical Connections between REM and IRL

Relational event model and inverse reinforcement learning originated independently from two different fields; however, both posit *actions* as a central concept and theorize that the tendency of a future action depends on a current situation resulting from recently realized history. Such similarity inspires us to examine parts of REM and IRL that are analogous. In Table 4.3 we list elements of the two theoretical frameworks that we find fundamentally equivalent to one another.

First, the social network or actor group studied in REM is equivalent to an RL agent. Although it is reasonable and somewhat more intuitive to conceive an individual sender in REM as an RL agent (on which we will elaborate in Section 4.5.1), the group in REM as a whole is equivalent to one agent because (1) under the setting of REM it is the group as a whole rather than individual senders that are the observer of their own event history, which encompasses behaviors of all senders, and; (2) the goal of REM is to infer underlying group-level, rather than sender-specific, characteristics that drive group social interaction dynamics.

Next, we find directly matching notions in REM for *action* and *state* in RL. Straightforwardly, a relational event or dyadic action a in REM maps to the action a in RL. Action space $\mathbb{A}$ in REM is determined by the number of senders $|S|$, the number of receivers $|R|$,

and the number of action types $|C|$. Assuming all dyadic actions in a group are legal and no self-directed actions are permitted, the action space size would be $N(N-1) \times |C|$, where N is the total number of nodes in the group. In REM, the notion of an RL *state* is effectively fulfilled by a group's past relational event history, as it "creates the context for present action, forming differential propensities for relational events to occur" [32]. A newly taken dyadic action a_i directly updates the past event history $A_\tau(a_{i-1})$ of the group at the time a is taken, as the newly taken dyadic action becomes the most recent event in event history, and thus places the group in a new state. The new state s' is simply a concatenation of the old state and the action taken upon it: $A_\tau(a_i) = \{a_i, A_\tau(a_{i-1})\}$. In other words, REM is intended for a Markov decision process with *history* as *state*. This property assumes group interaction process to be (1) a deterministic MDP, as once a new event happens, the event with 100% probability gets placed on top of past event history, and (2) an MDP with non-revisitable states, as event history always grows (assuming infinite memory) and once a dyadic action is realized, none other could happen. As IRL algorithms require state-action trajectories as input, and we find the nature of past event history in REM equivalent to an RL state, we find that the input data for REM is also suitable for IRL.

The core of both REM and IRL lies in the mechanism to determine what actions are more likely or favorable for the agent to take given a context, whether it is an event history or a state. On the highest level, both methods quantify the desirability of a realized state by a function: in REM it is the rate function λ while in IRL it is the reward function r. Given a current state, actions that can transition to states with a higher rate function value in REM or reward function value in RL are more likely chosen than others. Zooming in, both REM and IRL decompose the "desirability" function as a linear combination of features that describe a state. In REM they are the sufficient statistics $u(A)$ while in IRL they are the features $f(s)$. Both these variables are the *de facto* descriptors of a context (analogous to the explanatory variables or predictors in the supervised learning sense) and need to be pre-specified by the researcher. Lastly, the parameters θ associated with the features are the

output of both REM and IRL and the vehicle for learned insights. Inferred θ values in both cases can be used as input for further behavioral modeling such as behavioral clustering [191] and performance prediction [192].

4.4 Equivalence between REM and Maximum Entropy IRL

The analogy discussed between REM and IRL theories and valid regardless of the actual IRL algorithm chosen; however, the optimization procedure to find the most fitting rewards differs across different IRL algorithms. While the feature expectation based IRL algorithms [34] [188] use a distinct procedure to drive the learner's policy close to the expert's, we find that maximum entropy IRL [171] and REM share the same strategy that is to maximize the likelihood of the demonstrated event trajectory. Moreover, note the striking commonality in using the exponential function to express (1) rate function λ in REM based on survival analysis (Equation 4.4), and; (2) trajectory probability $P(\zeta)$ in maximum entropy IRL, following its namesake, principle of maximum entropy. Below we prove that when a series of directed social interactions in a group are formulated as an MDP (as discussed in Section 4.3), maximum entropy IRL optimizes the exact same objective function as does REM. Suppose the demonstrated trajectory ζ contains T states $\{s_1, s_2, ..., s_T\}$ and $\zeta_{(T-1)}$ denotes a possible trajectory of length $T-1$.

$$Z(\theta) = \sum_{\zeta'} \exp \theta^\top \mathbf{f}_{\zeta'} \tag{4.12}$$

$$= \sum_{\zeta'_{(T-1)}} \sum_{s'_T} \exp\left(\theta^\top \mathbf{f}_{\zeta'_{(T-1)}} + \theta^\top f_{s'_T}\right) \tag{4.13}$$

$$= \sum_{s'_1} \sum_{s'_2} \cdots \sum_{s'_T} \exp \sum_{i=1}^{T} \theta^\top f_{s'_i} \tag{4.14}$$

$$= \prod_{i=1}^{T} \sum_{s'_i} \exp \theta^\top f_{s'_i} \tag{4.15}$$

Plug $Z(\theta)$ into the likelihood function:

$$P(\zeta|\theta, T) = \frac{1}{Z(\theta)} \exp \sum_{s_j \in \zeta} \theta^\top f_{s_j} \tag{4.16}$$

$$= \frac{1}{Z(\theta)} \prod_{i=1}^{T} \exp \theta^\top f_{s_i} \tag{4.17}$$

$$= \frac{\prod_{i=1}^{T} \exp \theta^\top f_{s_i}}{\prod_{i=1}^{T} \sum_{s'_i} \exp \theta^\top f_{s'_i}} \tag{4.18}$$

$$= \prod_{i=1}^{T} \frac{\exp \theta^\top f_{s_i}}{\sum_{s'_i} \exp \theta^\top f_{s'_i}} \tag{4.19}$$

$$= \prod_{i=1}^{T} \frac{\exp[\theta^\top u(a_i, A_{\tau(a_{i-1})})]}{\sum_{a'_i} \exp[\theta^\top u(a'_i, A_{\tau(a_{i-1})})]}, \tag{4.20}$$

which is the same as Equation 4.6. The last step (from Equation 4.19 to 4.20) holds because at time step i the state s_i the group transitions into is solely determined by the action a_i taken at that step, based on the setup of REM. Therefore we have shown that maximum entropy IRL's equivalence to REM in the time-ordered case.

$$\nabla \log P(\zeta|\theta) = \mathbf{f}_\zeta - \sum_{\zeta'} P(\zeta'|\theta)\mathbf{f}_{\zeta'} = \mathbf{f}_\zeta - \sum_{s_i} D_{s_i}\mathbf{f}_{s_i} \tag{4.21}$$

Despite the same objective function, REM seeks to directly maximize likelihood using MLE or Bayesian methods [32] whereas maximum entropy IRL adopts a special optimization procedure. It expresses the gradient of trajectory log-likelihood in terms of the expected *state visitation frequency* D_{s_i} (Equation 4.21), a vector of expected frequency of visiting each state in a trajectory, and finds its value (thus then the gradient value) through an iterative algorithm (Algorithm 1 [171]). This procedure is valid because the reward of a trajectory (non-discounted value) $\mathbf{f}_\zeta$ eventually depends on the proportion (as opposed to the order) of states it visits along the way and the reward of each state f_{s_i}. The iterative algorithm contains a backward pass and a forward pass. The backward pass is a *value iteration* procedure to return a policy $P(a|s)$ that chooses action with probabilities proportional to the non-discounted value of the next state, which will generate trajectories that possess probabilities stated in Equation 4.11. The forward pass then uses this policy to simulate state visitation and find the expected frequency to plug back into gradient calculation.

The optimization procedure (reward weights $\rightarrow$ policy $\rightarrow$ state visitation frequency $\rightarrow$ gradient) used by maximum entropy IRL, however, is not compatible with REM for the following reasons. First, due to the fact that the states in a group interaction MDP in REM comprise previously realized action history at every point of the process, they are non-revisitable, making "expected state visitation frequency" inapplicable. Moreover, the purpose of the backward pass in Ziebart's Algorithm 1 is to solve for a policy given reward weights θ that behaves according to the theorized probability (Equation 4.11); this procedure is unnecessary in REM as the probability of action given state (i.e., policy) in REM is already specified as the rate function thanks to its survival analysis setup and readily available.

4.5 Discussion

4.5.1 Agent identity

What constitutes an *agent* is the first design choice when formulating an MDP for IRL. The perspective of REM is a bird-eye view of a whole group of N actors: legal actions between *all* pairs of actors are considered and ranked by their propensity to be taken. Learned coefficients of the sufficient statistics also represent the interaction dynamics of the whole group as opposed to individual actors. Therefore, the counterpart IRL problem should consider the group as the agent as discussed in Section 4.3. However, IRL is a more generic approach that can naturally treat an individual actor in the group as agent and other actors and their behaviors as environment if we construct the MDP accordingly. This way, the modeling perspective becomes truly egocentric. Compared to REM, IRL affords higher freedom in choosing what subset of a social system to be the agent and what subset to be the environment, contingent upon the problem at hand. We anticipate IRL to be useful for understanding group interactions in the international politics domain (in which previous work focused on network science approaches [193]) where behaviors of individual countries may be of higher interest than universal dynamics of multiple countries.

This inspires the question: can we make REM suitable for individual modeling in a group interactions setting? In the current REM (and all the models listed in Table 4.1 for that matter), we model every action that takes place in the group, meaning that the process is eventless between two adjacent actions. Therefore, it is reasonable for REM to assume piecewise constant hazard for each action within the group. However, if we only focus on the behavior of one actor in the group, actions that involve other group members would likely happen between two adjacent actions taken by the actor in question, making the piecewise constant hazard assumption unreasonable as the actions between other actors are likely to have an impact on how the actor we focus on behaves (i.e., expediting or delaying its certain actions). As such, if we were to build a variant of REM to learn individual specific insights, we

would expect to resort to accelerated failure time models [194] to make the hazard contingent upon other actors' actions.

4.5.2 State space

As shown in Equation 4.6, the entire event history is retained to fit an REM. However, the sufficient statistics of REM require memory of different lengths of history, often much shorter than the entire history since t_0. For example, to calculate the statistic *reciprocity*, one would only need information about the most recent event, and if the sender and the recipient of the current social action correspond to the recipient and the sender of the most recent event respectively, the feature is assigned value 1 and all otherwise 0; whereas in cases like feature *inertia*, one would need to look up the current action in the entire history and calculate the frequency. This difference does not fundamentally affect REM modeling; however, it reflects greatly in the construction of state space when we formulate an MDP for IRL. REM amounts to an IRL problem with a state space consisting of slices of non-revisitable, potentially-realizable event histories of all possible lengths, which may result in extremely large state spaces and create an issue for using feature expectation computation based on state visitation frequency (SVF). However, if we truncate the slices of event history populating the state space to a fixed number of recent actions, the states become revisitable and the state space considerably smaller. Seeing that, depending on the sufficient statistics/features we choose, and frankly out of the intuition that older events may not matter as much as more recent ones, we may not need to retain the entire event history all the time; the truncation operation may be a viable, even advisable choice when building IRL procedure for group interaction modeling problems.

In an MDP with recently realized action histories as states, the size of state space is eventually determined by the number of past actions retained as part of a state and the number of possible actions, the latter of which is in turn determined by the number of action-capable actors and event types. As the size of state space increases at an exponential rate as do the number of possible actions and the number of recent actions, over-large state spaces

are a pressing issue to resolve and a thinly veiled curse of dimensionality. Countermeasure strategies targeting over-large state spaces are of primary interest and mainly fall into two categories: (1) state aggregation through clustering actors, actions, and state features to directly reduce state space size; (2) value function approximation based on seen states to generalize to unseen states.

4.5.3 Model assumptions

Besides agent identity and state space, several model assumptions also show distinctions of IRL that speaks to its flexibity compared to REM. First, REM explicitly assumes the absence of forward looking [32] whereas IRL is naturally equipped with the discount factor γ to handle future states and how they are reflected in the current decision making. REM essentially amounts to IRL models with $\gamma = 0$.

Also, IRL and REM have different assumptions on the stochasticity of action choosing. REM has one single rule of choosing actions: given a past event history, a rate function value is computed for each possible next action and the probability of an action being chosen as the next action is modeled to be proportional to its rate function value. In this way an action with a higher rate function value has a higher probability to be chosen. This operation amounts to the Thompson sampling, or probability matching strategy [195] in the multi-armed bandit problem. In IRL, multiple action choosing strategies have been proposed. A popular one is ϵ-greedy strategy, where an agent chooses the action with highest reward (action with highest valued rate function in REM) with probability $1 - \epsilon$ and randomly chooses among all possible actions with probability ϵ. We anticipate that different action choosing strategies, reflected in a probabilistic format, can be incorporated into REM as well. We suspect that such incorporation is beneficial in behavioral modeling as different real-world social systems fit different assumptions of action choosing strategies to different degrees. This could become an interesting direction of future research.

However, compared to IRL, REM makes up its lower flexibility with its ability to also model timestamped event sequence, thanks to its survival analysis mechanism (concretely,

Equation 4.3). It is not within the current machinery of IRL to take advantage of continuous timestamps as the states are treated as discrete. The philosophical trade-off between REM and IRL is that REM is a more "ad-hoc", special purpose model and theoretical framework without the horns and whistles of IRL whereas IRL serves a wider range of purposes but may feel somewhat unnatural dealing with specific issues when applied to group interaction modeling problems.

4.6 Concluding Remarks

In this chapter, we have (1) approached social network modeling problem using inverse reinforcement learning from the perspective of the whole group, and (2) discovered, explained, and proved the theoretical connection between relational event model and inverse reinforcement learning in modeling group dyadic actions. Both theories take the viewpoint of an intelligent actor or group of actors, posit a desirability quantity for any particular state the actor(s) are in, choose actions based on the desirability of the potential state they lead the actor(s) onto, express the desirability in terms of a linear combination of state features, and aim to find the value of those features. Not only are the theories of REM and IRL analogous, we found that REM shares the objective function of maximum entropy IRL when applied to a group interactions MDP. Last but not least, We identified the differences in the optimization procedure before REM and maximum entropy IRL and empirically tested the results of an maximum entropy IRL inspired optimization procedure for REM and showed their equivalence.

By identifying the connections between IRL and REM, we demonstrate the applicability of IRL algorithms to social signals mining tasks, straightforwardly for group interaction dynamics. Both modeling methods can be used to extract quantitative characteristics (i.e., weights or coefficients) from social interaction episodes detected in cyber-human systems, thus creating novel features for relevant predictive tasks. One can consider the participants of direct social interactions with an individual as forming a group that is signature of the individual in question, and join the group dynamics features learned by REM or IRL with other existing features to predict personal outcomes. Further, due to technical differences,

IRL is suitable for specific cases of data and applications (e.g., treating a single individual in group interaction as the agent as discussed in Section 4.5.1) that may be difficult for REM to accommodate, thus providing unique utility.

The applicability of REM and IRL goes beyond social signals, as both essentially are theoretical frameworks of **event sequence**. Social interaction dynamics can be considered as sequences of social action, and so can many tasks humans undertake. As a generalization of the REM framework, it has been used to extract sequential dependencies in participants' daily activities collected in the American Time Use Survey (ATUS) data [196]. On the IRL side, we found existing application in discovering underlying patterns in people's routine driving behavior [35]. With growing amount of smart sensing data both in daily life and in workplaces such as manufacturing plant where task procedures are complex, I foresee great research opportunities in the application of REM and IRL for behavioral modeling.

Consider team processes research in organizational science, whose central hypothesis is that aspects of team interaction dynamics affect team performance. Existing research typically tracks group dynamics change through measurements of individual members and aggregation metrics of such measurements [197]; however, insights on the underlying factors that drive the formation of group dynamics structures are still largely missing. A promising next step in team processes research may be to utilize REM and IRL to automate and streamline behavioral learning in order to "distinguish effective groups from ineffective ones" [181]. Consider on MOOCs (Massive Open Online Courses), students' behavioral sequence —how they navigate through a course's materials and resources— is likely telltale of their learning performance. With MOOC students' entire usage history stored, REM and IRL may be the promising analytical tools to understand learner's usage patterns in order to ultimately identify their weaknesses and strengths. Now that we can fuse statistical network models and inverse reinforcement learning and use them to quantify driving factors of behavioral sequences, what can we do with the learned coefficients/reward weights? The answer may reside in their correlation with group and personal performance.

Summary of Contributions and Future Work

5.1 Summary of Contributions

5.1.1 Mechanism how online activism shapes collective behavior

In Chapter 2, I conducted a predictive task targeting mass protest onsets with a daily granularity using Twitter and GDELT data. Protest participation theory driven Twitter features outperform baseline GDELT features when forecasting with a one-day lead time but not with longer periods into the future. The predictive power of Twitter speech originates from the volume of protest advertising tweets, representing a collective level of knowledge of future protest events. By which I locate the predictive power of social media in its role as an organization and mobilization tool. Among GDELT features, event counts indicating stronger military presence within the nation in question, such as exhibiting military posture, fighting, and unconventional mass violence, are important predictors of an upcoming protest onset. I proposed new general-purpose performance metrics *detection precision* and *detection recall* for future event prediction tasks with a sliding window lead-time setting; and my best performing models were able to detect all the protest onsets during the period of study, whereas detection precision remains unsatisfactory. In all, I showed that using online social signals data to forecast daily fluctuation in collective offline behavior is feasible and promising.

The broader lesson learned is the value of the theoretical underpinning to guide feature engineering: in addition to prediction performance gain from theory-driven feature design, I highlighted the mechanism which online activism shapes offline behavior during civil uprisings, which had not been confirmed in other works and serve as crucial insights for decision support.

5.1.2 Physical proximity features that improve stress prediction

In Section 3.3 of Chapter 3, I conducted systematic feature engineering with Bluetooth encounter network, and systematic evaluation of Bluetooth encounter network features as stress predictors, both of which had been missing in the literature. I created 107 Bluetooth network features based on its structural, edge, and neighbor attributes, among which the social and temporal commonality features proved especially useful. I identified four problem setups, namely value diagnosis, change diagnosis, value prognosis, and change prognosis, for sensing-based automated mental health recognition and prediction based on the temporal alignment between input data and symptom as well as response variable being a value or a trend. The Bluetooth encounter network features improved personal stress recognition performance in the value diagnosis setup, justifying the incorporation of Bluetooth encounter data widely collected by personal smartphones and wearable devices into real-time mental health tracking solutions. A comparison between two different training and evaluation schemes, namely within-subject vs. across-subject leave-one-out cross validation, indicates the importance of personalization to improve inference performance. Besides predictive performance, correlation analyses confirmed several Bluetooth encounter network variables that are consistent with select theories in social psychology. The broader impact of this work is the wide applicability of the feature engineering techniques which can be useful to mining other social signals data as well [111].

In Section 3.4, I proposed and validated a vector space model approach to representing Bluetooth encounter events, which is especially suitable for less-data-demanding cases where only a subject's own sensing data is available. I encoded Bluetooth encounters detected by personal smart devices over a period of time into spatio-temporal tokens that represent each distinct proximity encounter scenario, assign feature values based on practice from natural language processing, and use the resulting feature vector to predict with mental health outcomes associated with the corresponding time period. I improved stress recognition performance with Bluetooth encounter vector space features compared to baseline features,

and prediction performance varies based on the vector space token design (i.e., device-only, device-time, and device-time-location) and feature assignment schemes (i.e., binary value, token frequency, token frequency-inverse period frequency). The broader impact is that this bag-of-words approach to representing Bluetooth encounter features is applicable to other types of sensing signals as well and is conducive to many other advanced data mining techniques such as word embedding and deep neural networks. My vector space model work paves the way for future research that aims at representing raw human-sensing signals in automatic and effective ways.

A comparison (Table 3.13) between these two feature engineering methods of Bluetooth encounter data showed that when we have access to the data a group of people sharing close-knit social relationship, Bluetooth encounter network features are the better choice when inferring personal mental health for individuals within the group; otherwise, vector space features would be the preferred choice.

5.1.3 A connection between network science and machine learning

In Chapter 4, I looked into *network* as a unifying abstraction of social signals and for the first time discovered a theoretical connection between inverse reinforcement learning from machine learning and relational event models from network science. Used to model group social interaction dynamics, relational event models share the same logic as inverse reinforcement learning as agent deciding the desirability of the next social action based on interaction history realized so far and similar building blocks that map directly to one another such as REM's rate function and IRL's reward function as shown in Table 4.3. I also proved the mathematical equivalence between Maximum Entropy IRL [171] and REM. I discussed distinctions in modeling details of the two approaches, which inspire research in both fields to design more sophisticated models to serve relevant needs. This discovery connecting machine learning and network science modeling theories and methods provided ground work for future applications of inverse reinforcement learning as a modeling method for social signals in cyber-human systems.

5.2 Summary of Future Work

The limitation in existing in-situ sensing instruments for physical social signals calls for development and fusion of sensing capabilities to more effectively measure and understand in-person social signals. For example, discovering changes in voice patterns with smartphone and wearable sensors (e.g., time and duration of conversations, volume and tone of voice, noise level, without content ever recorded due to privacy concerns) and their correlation with behavioral context and health status can greatly complement existing methods (social media, proximity sensors) to gain insights into an individuals socialization patterns and serve as valuable evidence for personalized health monitoring and adaptive intervention. Moreover, online and offline social signals have largely been studied in separate applications in current literature and I plan to explore ways how they can be fused to create new social sensing functionalities [198], especially the merging of online social activity and in-person interaction record for personal sensing.

Two major methodological topics have emerged from work discussed in this dissertation, which I plan to further develop with new data and applications. The first is bag-of-words approach with human-centered sensing signals. My work on vector space representation of Bluetooth encounter signals indicates the applicability of natural language processing techniques on human-centered sensing data to improve predictive performance targeting health statuses. I plan to further investigate the value of multiple text mining methods (e.g., bag-of-words representation, topic modeling, word embedding) with multiple types of mobile sensing data (e.g., location, physical activity, proximity) in better understanding human behavioral components and their connection with psychological outcomes. The second is mining human behavioral sequences using relational event models and inverse reinforcement learning. Although dyadic social actions data in technology-mediated communications was used to demonstrate the modeling framework of the two methods, their applicability extends beyond to any behavioral sequence. I find these two application cases especially interesting: (a) mobility traces where place visit, transportation means, and route choices reflect an

individuals lifestyle and environment preference, and; (b) user behavior on online education platforms (MOOC), where learners execute a sequence of actions to navigate through a course, which may have strong correlation with their learning goals and performance. This line of research will not only help better understand the underlying mechanism of human behavior but also provide novel grounds for behavior change intervention.

I contributed to mental health sensing and intervention in this dissertation through mining physical social signals; yet many challenges remain in the field. First, in addition to smart health work focused on general undiagnosed cohorts, I anticipate my next steps to zero in on specific health conditions and behavioral issues to serve targeted clinical applications, such as alleviating poor sleep quality, drinking problems, and caring for persons with special needs. Second, in current literature, we see, on one hand, smartphone sensing cohort studies with participants in natural daily-life conditions where mental health ground truth is obtained through ecological momentary assessment (EMA); and on the other hand, wearable sensing research [199] in controlled settings becoming increasingly accurate at detecting stress responses using physiological measures from on-wrist wearable devices. I believe an important next step in mental health sensing is to conduct quality in-situ cohort studies tracking both smartphone and wearable sensing data as well as EMA, in order to (a) discover correlations between anomalies in physiological measures from wearable sensors with behavioral patterns detected by smartphones, and (b) maximize effectiveness of ground truth acquisition by collating EMA responses with physiological measures and eventually reduce dependence on EMA, achieving multi-platform mental health sensing.

I plan to broaden the scope of human applications beyond mental health. Ubiquitous computing literature has seen relatively standalone efforts in detecting cognitive states (alertness, boredom, circadian rhythm) through smartphone sensing. I believe there is great value in synthesizing these tasks and applying them to improve students time management and study experience (e.g., Csikszentmihalyis flow [200]). We should investigate, in ecologically valid manners, the covariation and predictability of the different cognitive constructs, the

correlation between cognitive states and behavioral traces captured by smartphone sensors, as well as their implications on the productivity of study sessions and eventually academic performance. Real time measurements and findings should be made accessible to users in a dashboard fashion similar to the screen time functionality in the latest iOS versions. On a population level, I am intrigued by the inter-dependencies between urban environment features, local business performance, and human mobility patterns. Specifically, how peoples mobility patterns (transportation and route choice, place visit sequences) are correlated with the performance of local businesses and affected by urban environment features. With increasingly available human mobility data through mobile sensing and location-based social media, I want to leverage these data to inform business management and urban planning.

Bibliography

[1] Alan Schussman and Sarah A Soule. Process and protest: Accounting for individual protest participation. *Social forces*, 84(2):1083–1108, 2005.

[2] Shao-Zhong Zhang, Eric Block, Lawrence C Katz, et al. Encoding social signals in the mouse main olfactory bulb. *Nature*, 434(7032):470, 2005.

[3] Gary R Bortolotti, Francois Mougeot, Jesus Martinez-Padilla, Lucy MI Webster, and Stuart B Piertney. Physiological stress mediates the honesty of social signals. *PLoS One*, 4(3):e4983, 2009.

[4] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759, 2009.

[5] Alex Pentland. *Honest signals: how they shape our world*. MIT press, 2010.

[6] John B Killoran. How to use search engine optimization techniques to increase website visibility. *IEEE Transactions on Professional Communication*, 56(1):50–66, 2013.

[7] Amit Sheth. Citizen sensing, social signals, and enriching human experience. *IEEE Internet Computing*, 13(4):87–92, 2009.

[8] Isabella Poggi and Francesca D'Errico. Cognitive modelling of human social signals. In *SSPW@ MM*, pages 21–26, 2010.

[9] Mani Srivastava, Tarek Abdelzaher, and Boleslaw Szymanski. Human-centric sensing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1958):176–197, 2012.

[10] Emiliano Miluzzo, Nicholas D Lane, Kristóf Fodor, Ronald Peterson, Hong Lu, Mirco Musolesi, Shane B Eisenman, Xiao Zheng, and Andrew T Campbell. Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application. In *Proceedings of the 6th ACM conference on Embedded network sensor systems*, pages 337–350. ACM, 2008.

[11] Bin Guo, Chao Chen, Daqing Zhang, Zhiwen Yu, and Alvin Chin. Mobile crowd sensing and computing: when participatory sensing meets participatory social media. *IEEE Communications Magazine*, 54(2):131–137, 2016.

[12] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. AcM, 2010.

[13] Michele Starnini, Bruno Lepri, Andrea Baronchelli, Alain Barrat, Ciro Cattuto, and Romualdo Pastor-Satorras. Robust modeling of human contact networks across different scales and proximity-sensing techniques. In *International Conference on Social Informatics*, pages 536–551. Springer, 2017.

[14] Marianne Schmid Mast, Daniel Gatica-Perez, Denise Frauendorfer, Laurent Nguyen, and Tanzeem Choudhury. Social sensing for psychology: Automated interpersonal behavior assessment. *Current Directions in Psychological Science*, 24(2):154–160, 2015.

[15] Daniel Chaffin, Ralph Heidl, John R Hollenbeck, Michael Howe, Andrew Yu, Clay Voorhees, and Roger Calantone. The promise and perils of wearable sensors in organizational research. *Organizational Research Methods*, 20(1):3–31, 2017.

[16] Shu Liu, Yingxin Jiang, and Aaron Striegel. Face-to-face proximity estimation using bluetooth on smartphones. *IEEE Transactions on Mobile Computing*, 13(4):811–823, 2014.

[17] Claudio Martella, Matthew Dobson, Aart Van Halteren, and Maarten Van Steen. From proximity sensing to spatio-temporal social graphs. In *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on*, pages 78–87. IEEE, 2014.

[18] Trinh Minh Do and Daniel Gatica-Perez. Human interaction discovery in smartphone proximity networks. *Personal and Ubiquitous Computing*, 17(3):413–431, 2013.

[19] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

[20] Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. Measuring large-scale social networks with high resolution. *PloS one*, 9(4):e95978, 2014.

[21] Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. Social fmri: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 7(6):643–659, 2011.

[22] Niko Kiukkonen, Jan Blom, Olivier Dousse, Daniel Gatica-Perez, and Juha Laurila. Towards rich mobile phone datasets: Lausanne data collection campaign. *Proc. ICPS, Berlin*, 2010.

[23] Daniel Olguín Olguín, Benjamin N Waber, BN Taemie Kim, Akshay Mohan, Koji Ara, and Alex Pentland. Sensible organizations: Technology and methodology for automatically measuring organizational behavior. Institute of Electrical and Electronics Engineers, 2008.

[24] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. Studentlife: assessing

mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 3–14. ACM, 2014.

[25] Taemie Kim, Erin McFee, Daniel Olguin Olguin, Ben Waber, and Alex Sandy Pentland. Sociometric badges: Using sensor technology to capture new forms of collaboration. *Journal of Organizational Behavior*, 33(3):412–427, 2012.

[26] Lorenzo Isella, Juliette Stehlé, Alain Barrat, Ciro Cattuto, Jean-François Pinton, and Wouter Van den Broeck. What's in a crowd? analysis of face-to-face behavioral networks. *Journal of theoretical biology*, 271(1):166–180, 2011.

[27] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[28] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[29] Danny Wyatt, Tanzeem Choudhury, and Jeff Bilmes. Conversation detection and speaker segmentation in privacy-sensitive situated speech data. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[30] Irfan A. Essa and Alex Paul Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 19(7):757–763, 1997.

[31] Tom AB Snijders. Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40, 2002.

[32] Carter T Butts. A relational event framework for social action. *Sociological Methodology*, pages 155–200, 2008.

[33] Andrew Pilny, Alex Yahja, Marshall Scott Poole, and Melissa Dobosh. A dynamic social network experiment with multi-team systems. In *Big Data and Cloud Computing (BdCloud), 2014 IEEE Fourth International Conference on*, pages 587–593. IEEE, 2014.

[34] Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, pages 663–670, 2000.

[35] Nikola Banovic, Tofi Buzali, Fanny Chevalier, Jennifer Mankoff, and Anind K Dey. Modeling and understanding human routine behavior. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 248–260. ACM, 2016.

[36] Huiji Gao, Geoffrey Barbier, and Rebecca Goolsby. Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, 26(3):10–14, 2011.

[37] Zheng Xu, Yunhuai Liu, Neil Yen, Lin Mei, Xiangfeng Luo, Xiao Wei, and Chuanping Hu. Crowdsourcing based description of urban emergency events using social media big data. *IEEE Transactions on Cloud Computing*, 2016.

[38] Lynn Wu, Benjamin Waber, Sinan Aral, Erik Brynjolfsson, and Alex Pentland. Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an it configuration task. 2008.

[39] Lorenzo Isella, Mariateresa Romano, Alain Barrat, Ciro Cattuto, Vittoria Colizza, Wouter Van den Broeck, Francesco Gesualdo, Elisabetta Pandolfi, Lucilla Ravà, Caterina Rizzo, et al. Close encounters in a pediatric ward: measuring face-to-face proximity and mixing patterns with wearable sensors. *PloS one*, 6(2):e17144, 2011.

[40] Sitaram Asur and Bernardo A Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 492–499. IEEE Computer Society, 2010.

[41] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[42] Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467, 2011.

[43] Matthew S Gerber. Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125, 2014.

[44] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*, 2010.

[45] Gunther Eysenbach. Can tweets predict citations? metrics of social impact based on twitter and correlation with traditional metrics of scientific impact. *Journal of medical Internet research*, 13(4), 2011.

[46] Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. Predicting the political alignment of twitter users. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 192–199. IEEE, 2011.

[47] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*, 2013.

[48] Adriana M Manago, Tamara Taylor, and Patricia M Greenfield. Me and my 400 friends: The anatomy of college students' facebook networks, their communication patterns, and well-being. *Developmental psychology*, 48(2):369, 2012.

[49] Ethan Kross, Philippe Verduyn, Emre Demiralp, Jiyoung Park, David Seungjae Lee, Natalie Lin, Holly Shablack, John Jonides, and Oscar Ybarra. Facebook use predicts declines in subjective well-being in young adults. *PloS one*, 8(8):e69841, 2013.

[50] Robert E Wilson, Samuel D Gosling, and Lindsay T Graham. A review of facebook research in the social sciences. *Perspectives on psychological science*, 7(3):203–220, 2012.

[51] Srećko Joksimović, Areti Manataki, Dragan Gašević, Shane Dawson, Vitomir Kovanović, and Ines Friss De Kereki. Translating network position into performance: importance of centrality in different network configurations. In *Proceedings of the sixth international conference on learning analytics & knowledge*, pages 314–323. ACM, 2016.

[52] Collin F Lynch, Tiffany Barnes, Jennifer Albert, and Michael Eagle. Graph-based educational data mining (g-edm 2015). *CEUR-WS*, 2015.

[53] Jun-Ki Min, Jason Wiese, Jason I Hong, and John Zimmerman. Mining smartphone data to classify life-facets of social relationships. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 285–294. ACM, 2013.

[54] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean-François Pinton, Marco Quaggiotto, Wouter Van den Broeck, Corinne Régis, Bruno Lina, et al. High-resolution measurements of face-to-face contact patterns in a primary school. *PloS one*, 6(8):e23176, 2011.

[55] Zhaoyang Zhang, Ken CK Lee, Honggang Wang, Dong Xuan, and Hua Fang. Epidemic control based on fused body sensed and social network information. In *2012 32nd International Conference on Distributed Computing Systems Workshops*, pages 285–290. IEEE, 2012.

[56] Anmol Madan, Manuel Cebrian, David Lazer, and Alex Pentland. Social sensing for epidemiological behavior change. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 291–300. ACM, 2010.

[57] Mark C Pachucki, Emily J Ozer, Alain Barrat, and Ciro Cattuto. Mental health and social networks in early adolescence: a dynamic study of objectively-measured social interaction behaviors. *Social science & medicine*, 125:40–50, 2015.

[58] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Sandy Pentland. Pervasive stress recognition for sustainable living. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, pages 345–350. IEEE, 2014.

[59] Sai T Moturu, Inas Khayal, Nadav Aharony, Wei Pan, and Alex Pentland. Using social sensing to understand the links between sleep, mood, and sociability. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 208–214. IEEE, 2011.

[60] Aamena Alshamsi, Fabio Pianesi, Bruno Lepri, Alex Pentland, and Iyad Rahwan. Network diversity and affect dynamics: The role of personality traits. *PloS one*, 11(4):e0152358, 2016.

[61] Rahman O Oloritun, Taha BMJ Ouarda, Sai Moturu, Anmol Madan, Alex Sandy Pentland, and Inas Khayal. Change in bmi accurately predicted by social exposure to acquaintances. *PloS one*, 8(11):e79238, 2013.

[62] Anmol Madan, Katayoun Farrahi, Daniel Gatica-Perez, and Alex Sandy Pentland. Pervasive sensing to model political opinions in face-to-face networks. In *International Conference on Pervasive Computing*, pages 214–231. Springer, 2011.

[63] Anmol Madan, Manuel Cebrian, Sai Moturu, Katayoun Farrahi, et al. Sensing the" health state" of a community. *IEEE Pervasive Computing*, 11(4):36–45, 2012.

[64] Wen Dong, Bruno Lepri, and Alex Sandy Pentland. Modeling the co-evolution of behaviors and social relationships using mobile phone data. In *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia*, pages 134–143. ACM, 2011.

[65] Nicholas D Lane, Li Pengyu, Lin Zhou, and Feng Zhao. Connecting personal-scale sensing and networked community behavior to infer human activities. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 595–606. ACM, 2014.

[66] Tanzeem Choudhury and Alex Pentland. Sensing and modeling human networks using the sociometer. In *null*, page 216. IEEE, 2003.

[67] Valentin Kassarnig, Enys Mones, Andreas Bjerre-Nielsen, Piotr Sapiezynski, David Dreyer Lassen, and Sune Lehmann. Academic performance and behavioral patterns. *EPJ Data Science*, 7:1–16, 2018.

[68] Huaxiu Yao, Min Nie, Han Su, Hu Xia, and Defu Lian. Predicting academic performance via semi-supervised learning with constructed campus social network. In *International Conference on Database Systems for Advanced Applications*, pages 597–609. Springer, 2017.

[69] Denise Frauendorfer, Marianne Schmid Mast, Laurent Nguyen, and Daniel Gatica-Perez. Nonverbal social sensing in action: Unobtrusive recording and extracting of nonverbal behavior in social interactions illustrated with a research example. *Journal of Nonverbal Behavior*, 38(2):231–245, 2014.

[70] Sean P O'brien. Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review*, 12(1):87–104, 2010.

[71] Congyu Wu and Matthew S Gerber. Forecasting civil unrest using social media and protest participation theory. *IEEE Transactions on Computational Social Systems*, 5(1):82–94, 2018.

[72] Ryan Compton, Craig Lee, Tsai-Ching Lu, Lalindra De Silva, and Michael Macy. Detecting future social unrest in unprocessed twitter data: "emerging phenomena and big data". In *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference On*, pages 56–60. IEEE, 2013.

[73] Feng Chen, Jaime Arredondo, Rupinder Paul Khandpur, Chang-Tien Lu, David Mares, Dipak Gupta, and Naren Ramakrishnan. Spatial surrogates to forecast social mobilization and civil unrests. In *Position Paper in CCC Workshop on "From GPS and Virtual Globes to Spatial Computing-2012*, 2012.

[74] Philip N Howard, Aiden Duffy, Deen Freelon, Muzammil Hussain, Will Mari, and Marwa Mazaid. Opening closed regimes: what was the role of social media during the arab spring? 2011.

[75] Ekaterina Stepanova. The role of information communication technologies in the "arab spring". *Ponars Eurasia*, 15:1–6, 2011.

[76] Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, et al. The arab spring— the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International journal of communication*, 5:31, 2011.

[77] Miriyam Aouragh and Anne Alexander. The arab spring— the egyptian experience: Sense and nonsense of the internet revolution. *International Journal of Communication*, 5:15, 2011.

[78] Zachary C Steinert-Threlkeld, Delia Mocanu, Alessandro Vespignani, and James Fowler. Online social networks and offline protest. *EPJ Data Science*, 4(1):1–9, 2015.

[79] Nathan Kallus. Predicting crowd behavior with big public data. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 625–630. International World Wide Web Conferences Steering Committee, 2014.

[80] Naren Ramakrishnan, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Khandpur, Parang Saraf, Wei Wang, Jose Cadena, Anil Vullikanti, Gizem Korkmaz, et al. 'beating the news' with embers: forecasting civil unrest using open source indicators. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1799–1808. ACM, 2014.

[81] Benedikt Boecking, Margeret Hall, and Jeff Schneider. Predicting events surrounding the egyptian revolution of 2011 using learning algorithms on micro blog data.

[82] Marco T Bastos, Dan Mercea, and Arthur Charpentier. Tents, tweets, and events: The interplay between ongoing protests and social media. *Journal of Communication*, 65(2):320–350, 2015.

[83] Henrik Serup Christensen. Political activities on the internet: Slacktivism or political participation by other means? *First Monday*, 16(2), 2011.

[84] Sebastián Valenzuela. Unpacking the use of social media for protest behavior the roles of information, opinion expression, and activism. *American Behavioral Scientist*, 57(7):920–942, 2013.

[85] Meredith Conroy, Jessica T Feezell, and Mario Guerrero. Facebook and political engagement: A study of online political group membership and offline political engagement. *Computers in Human Behavior*, 28(5):1535–1546, 2012.

[86] Sidney Verba, Kay Lehman Schlozman, Henry E Brady, and Henry E Brady. *Voice and equality: Civic voluntarism in American politics*, volume 4. Cambridge Univ Press, 1995.

[87] Kalev Leetaru and Philip A Schrodt. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA Annual Convention*, volume 2, page 4, 2013.

[88] Philip A Schrodt. Forecasting conflict in the balkans using hidden markov models. In *Programming for Peace*, pages 161–184. Springer, 2006.

[89] Patrick T Brandt, John R Freeman, and Philip A Schrodt. Real time, time series forecasting of inter-and intra-state political conflict. *Conflict Management and Peace Science*, 28(1):41–64, 2011.

[90] Jack A Goldstone, Robert H Bates, David L Epstein, Ted Robert Gurr, Michael B Lustik, Monty G Marshall, Jay Ulfelder, and Mark Woodward. A global model for forecasting political instability. *American Journal of Political Science*, 54(1):190–208, 2010.

[91] Michael D Ward, Nils W Metternich, Cassy L Dorff, Max Gallop, Florian M Hollenbach, Anna Schultz, and Simon Weschle. Learning from the past and stepping into the future: Toward a new generation of conflict prediction. *International Studies Review*, 15(4):473–490, 2013.

[92] Ian Lustick. Ps-i: A user-friendly agent-based modeling platform for testing theories of political identity and political stability. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.

[93] Nahed Eltantawy and Julie B Wiest. The arab spring— social media in the egyptian revolution: Reconsidering resource mobilization theory. *International Journal of Communication*, 5:18, 2011.

[94] Charles McClelland. World event/interaction survey, 1966-1978. *WEIS Codebook ICPSR*, 5211, 1978.

[95] Deborah J Gerner, Philip A Schrodt, Omür Yilmaz, and Rajaa Abu-Jabr. Conflict and mediation event observations (cameo): A new event data framework for the analysis of foreign policy interactions. *International Studies Association, New Orleans*, 2002.

[96] Edward E Azar. The conflict and peace data bank (copdab) project. *Journal of Conflict Resolution*, 24(1):143–152, 1980.

[97] Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. Introducing acled: An armed conflict location and event dataset special data feature. *Journal of peace Research*, 47(5):651–660, 2010.

[98] Idean Salehyan, Cullen S Hendrix, Jesse Hamner, Christina Case, Christopher Linebarger, Emily Stull, and Jennifer Williams. Social conflict in africa: A new database. *International Interactions*, 38(4):503–511, 2012.

[99] Philip A Schrodt. Tabari: Textual analysis by augmented replacement instructions. *Dept. of Political Science, University of Kansas, Blake Hall, Version 0.7. 3B3*, pages 1–137, 2009.

[100] James E Yonamine. Predicting future levels of violence in afghanistan districts using gdelt. *Unpublished Manuscript*, 2013.

[101] James C Davies. Toward a theory of revolution. *American sociological review*, pages 5–19, 1962.

[102] Doug McAdam. *Political process and the development of black insurgency, 1930-1970*. University of Chicago Press, 2010.

[103] John D McCarthy and Mayer N Zald. Resource mobilization and social movements: A partial theory. *American journal of sociology*, pages 1212–1241, 1977.

[104] Karl Marx and Friedrich Engels. *The communist manifesto*. Penguin, 2002.

[105] Alexis De Tocqueville. *The Old Regime and the Revolution, Volume II: Notes on the French Revolution and Napoleon*, volume 2. University of Chicago Press, 2001.

[106] Jeroen Van Laer. Why people protest. *PhD Universiteit Antwerpen*, 2011.

[107] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009.

[108] Philip A Schrodt. Cameo: Conflict and mediation event observations event and actor codebook. *Pennsylvania State University*, 2012.

[109] Fei-Yue Wang, Kathleen M Carley, Daniel Zeng, and Wenji Mao. Social computing: From social informatics to social intelligence. *IEEE Intelligent systems*, 22(2):79–83, 2007.

[110] Huijie Lin, Jia Jia, Jiezhong Qiu, Yongfeng Zhang, Guangyao Shen, Lexing Xie, Jie Tang, Ling Feng, and Tat-Seng Chua. Detecting stress based on social interactions in social networks. *IEEE Transactions on Knowledge and Data Engineering*, 2017.

[111] Sanjana Mendu, Mehdi Boukhechba, Anna Baglione, Sonia Baee, Congyu Wu, and Laura Barnes. Socialtext: A framework for understanding the relationship between digital communication patterns and mental health.

[112] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110. ACM, 2016.

[113] Sheldon Cohen, David AJ Tyrrell, and Andrew P Smith. Psychological stress and susceptibility to the common cold. *New England journal of medicine*, 325(9):606–612, 1991.

[114] Mohd Razali Salleh. Life event, stress and illness. *The Malaysian Journal of Medical Sciences: MJMS*, 15(4):9, 2008.

[115] Ranjita Misra and Michelle McKean. College students' academic stress and its relation to their anxiety, time management, and leisure satisfaction. *American Journal of Health Studies*, 16(1):41, 2000.

[116] Thomas W Colligan and Eileen M Higgins. Workplace stress: Etiology and consequences. *Journal of Workplace Behavioral Health*, 21(2):89–97, 2006.

[117] Hamilton I McCubbin, Constance B Joy, A Elizabeth Cauble, Joan K Comeau, Joan M Patterson, and Richard H Needle. Family stress and coping: A decade review. *Journal of Marriage and the Family*, pages 855–871, 1980.

[118] Norman B Anderson, Cynthia D Belar, Steven J Breckler, Katherine C Nordal, David W Ballard, Lynn F Bufka, and K Wiggins. Stress in america: Paying with our health. *American Psychological Association*, 2015.

[119] Daniel Eisenberg, Ezra Golberstein, and Sarah E Gollust. Help-seeking and access to mental health care in a university student population. *Medical Care*, 45(7):594–601, 2007.

[120] Varvogli Liza. Stress management techniques: evidence-based procedures that reduce stress and promote health. *Health Science Journal*, 5(2), 2011.

[121] Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. A global measure of perceived stress. *Journal of Health and Social Behavior*, pages 385–396, 1983.

[122] Aishwarya Goyal, Shailendra Singh, Dharam Vir, and Dwarka Pershad. Automation of stress recognition using subjective or objective measures. *Psychological Studies*, pages 1–17, 2016.

[123] Silvia Serino, Pietro Cipresso, Gennaro Tartarisco, Giovanni Baldus, Daniele Corda, Giovanni Pioggia, Andrea Gaggioli, and Giuseppe Riva. Computerized experience-sampling approach for realtime assessment of stress. *EAI Endorsed Transactions on Ambient Systems*, 1(2):1–8, 2013.

[124] Saul Shiffman, Arthur A Stone, and Michael R Hufford. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4:1–32, 2008.

[125] Michelle Nicole Burns, Mark Begale, Jennifer Duffecy, Darren Gergle, Chris J Karr, Emily Giangrande, and David C Mohr. Harnessing context sensing to develop a mobile intervention for depression. *Journal of Medical Internet Research*, 13(3), 2011.

[126] Inbal Nahum-Shani, Eric B Hekler, and Donna Spruijt-Metz. Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. *Health Psychology*, 34(S):1209, 2015.

[127] David C Mohr, Mi Zhang, and Stephen M Schueller. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology*, 13:23–47, 2017.

[128] Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, pages 389–402. ACM, 2013.

[129] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research*, 17(7), 2015.

[130] Jiangchuan Zheng and Lionel M Ni. An unsupervised learning approach to social circles detection in ego bluetooth proximity network. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 721–724. ACM, 2013.

[131] Everett W Bovard. The effects of social stimuli on the response to stress. *Psychological Review*, 66(5):267, 1959.

[132] Stanley Kissel. Stress-reducing properties of social stimuli. *Journal of Personality and Social Psychology*, 2(3):378, 1965.

[133] A Courtney DeVries, Erica R Glasper, and Courtney E Detillion. Social modulation of stress responses. *Physiology & Behavior*, 79(3):399–407, 2003.

[134] Congyu Wu, Mehdi Boukhechba, Lihua Cai, Laura E Barnes, and Matthew S Gerber. Improving momentary stress measurement and prediction with bluetooth encounter networks. *Smart Health*, 2018.

[135] Congyu Wu, Mehdi Boukhechba, Lihua Cai, Laura E Barnes, and Matthew S Gerber. Vector space representation of bluetooth encounters for mental health inference. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 1691–1699. ACM, 2018.

[136] Salam Ranabir and K Reetu. Stress and hormones. *Indian Journal of Endocrinology and Metabolism*, 15(1):18, 2011.

[137] Feng-Tso Sun, Cynthia Kuo, Heng-Tze Cheng, Senaka Buthpitiya, Patricia Collins, and Martin Griss. Activity-aware mental stress detection using physiological sensors. In *International Conference on Mobile Computing, Applications, and Services*, pages 282–301. Springer, 2010.

[138] Michael Xuelin Huang, Jiajia Li, Grace Ngai, and Hong Va Leong. Stressclick: Sensing stress from gaze-click patterns. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 1395–1404. ACM, 2016.

[139] Davide Carneiro, José Carlos Castillo, Paulo Novais, Antonio FernáNdez-Caballero, and José Neves. Multimodal behavioral analysis for non-invasive stress detection. *Expert Systems with Applications*, 39(18):13376–13389, 2012.

[140] Jing Zhai and Armando Barreto. Stress recognition using non-invasive technology. In *FLAIRS Conference*, pages 395–401, 2006.

[141] M Paschero, G Del Vescovo, L Benucci, A Rizzi, M Santello, G Fabbri, and FM Frattale Mascioli. A real time classifier for emotion and stress recognition in a vehicle driver. In *Industrial Electronics (ISIE), 2012 IEEE International Symposium on*, pages 1690–1695. IEEE, 2012.

[142] Rajiv Ranjan Singh, Sailesh Conjeti, and Rahul Banerjee. A comparative evaluation of neural network classifiers for stress level analysis of automotive drivers using physiological signals. *Biomedical Signal Processing and Control*, 8(6):740–754, 2013.

[143] John HL Hansen and Sanjay Patil. Speech under stress: Analysis, modeling and recognition. In *Speaker Classification I*, pages 108–137. Springer, 2007.

[144] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T Chittaranjan, Andrew T Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 351–360. ACM, 2012.

[145] Nandita Sharma and Tom Gedeon. Objective measures, sensors and computational techniques for stress recognition and classification: A survey. *Computer Methods and Programs in Biomedicine*, 108(3):1287–1301, 2012.

[146] Akane Sano and Rosalind W Picard. Stress recognition using wearable sensors and mobile phones. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 671–676. IEEE, 2013.

[147] Alban Maxhuni, Pablo Hernandez-Leal, Eduardo F Morales, L Enrique Sucar, Venet Osmani, Angelica Muńoz-Meléndez, and Oscar Mayora. Using intermediate models and knowledge learning to improve stress prediction. In *Applications for Future Internet*, pages 140–151. Springer, 2017.

[148] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Sandy Pentland. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 477–486. ACM, 2014.

[149] Martin Gjoreski, Hristijan Gjoreski, Mitja Lutrek, and Matja Gams. Automatic detection of perceived stress in campus students using smartphones. In *Intelligent Environments (IE), 2015 International Conference on*, pages 132–135. IEEE, 2015.

[150] Martin Atzmueller, Lisa Thiele, Gerd Stumme, and Simone Kauffeld. Analyzing group interaction on networks of face-to-face proximity using wearable sensors. In *Future IoT Technologies (Future IoT), 2018 IEEE International Conference on*, pages 1–10. IEEE, 2018.

[151] Zhixian Yan, Jun Yang, and Emmanuel Munguia Tapia. Smartphone bluetooth based social sensing. In *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, pages 95–98. ACM, 2013.

[152] Tjeerd W Boonstra, Mark E Larsen, Samuel Townsend, and Helen Christensen. Validation of a smartphone app to map social networks of proximity. *PloS one*, 12(12):e0189877, 2017.

[153] Michele Starnini, Andrea Baronchelli, and Romualdo Pastor-Satorras. Modeling human dynamics of face-to-face interaction networks. *Physical review letters*, 110(16):168701, 2013.

[154] Yi-Qing Zhang, Jing Cui, Shu-Min Zhang, Qi Zhang, and Xiang Li. Modelling temporal networks of human face-to-face contacts with public activity and individual reachability. *The European Physical Journal B*, 89(2):26, 2016.

[155] Michele Starnini, Anna Machens, Ciro Cattuto, Alain Barrat, and Romualdo Pastor-Satorras. Immunization strategies for epidemic processes in time-varying contact networks. *Journal of theoretical biology*, 337:89–100, 2013.

[156] Jun-ichiro Watanabe, Nozomu Ishibashi, and Kazuo Yano. Exploring relationship between face-to-face interaction and team performance using wearable sensor badges. *PloS one*, 9(12):e114681, 2014.

[157] Danny Wyatt, Tanzeem Choudhury, and Jeff A Bilmes. Discovering long range properties of social networks with multi-valued time-inhomogeneous models. In *AAAI*, 2010.

[158] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[159] Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. The architecture of complex weighted networks: Measurements and models. In *Large Scale Structure And Dynamics Of Complex Networks: From Information Technology to Finance and Natural Science*, pages 67–92. World Scientific, 2007.

[160] Jong Hee Kang, William Welbourne, Benjamin Stewart, and Gaetano Borriello. Extracting places from traces of locations. *ACM SIGMOBILE Mobile Computing and Communications Review*, 9(3):58, July 2005.

[161] S Taylor, N Jaques, E Nosakhare, A Sano, EB Klerman, and RW Picard. Importance of sleep data in predicting next-day stress, happiness, and health in college students. *Journal of Sleep and Sleep Disorders Research*, 40(suppl_1):A294–A295, 2017.

[162] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[163] Takefumi Kikusui, James T Winslow, and Yuji Mori. Social buffering: relief from stress and anxiety. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 361(1476):2215–2228, 2006.

[164] William J Goode. A theory of role strain. *American Sociological Review*, pages 483–496, 1960.

[165] Randy J Larsen and Timothy Ketelaar. Personality and susceptibility to positive and negative emotional states. *Journal of Personality and Social Psychology*, 61(1):132, 1991.

[166] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[167] Nathan Eagle and Alex Sandy Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.

[168] Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, et al. The structural topic model and applied social science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*, 2013.

[169] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[170] Mathieu Génois and Alain Barrat. Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Science*, 7(1):11, 2018.

[171] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

[172] Tom AB Snijders, Philippa E Pattison, Garry L Robins, and Mark S Handcock. New specifications for exponential random graph models. *Sociological methodology*, 36(1):99–153, 2006.

[173] Andreas Wimmer and Kevin Lewis. Beyond and below racial homophily: Erg models of a friendship network documented on facebook. *American Journal of Sociology*, 116(2):583–642, 2010.

[174] Neha Gondal. The local and global structure of knowledge production in an emergent research field: An exponential random graph analysis. *Social Networks*, 33(1):20–30, 2011.

[175] Alessandro Lomi and Francesca Pallotti. Relational collaboration among spatial multipoint competitors. *Social networks*, 34(1):101–111, 2012.

[176] Olaf N Rank, Garry L Robins, and Philippa E Pattison. Structural logic of intraorganizational networks. *Organization Science*, 21(3):745–764, 2010.

[177] Tom AB Snijders. Stochastic actor-oriented models for network change. *Journal of mathematical sociology*, 21(1-2):149–172, 1996.

[178] Steve Hanneke, Wenjie Fu, Eric P Xing, et al. Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4:585–605, 2010.

[179] Christoph Stadtfeld, James Hollway, and Per Block. Dynamic network actor models: Investigating coordination ties through time. *Sociological Methodology*, 47(1):1–40, 2017.

[180] Rupert G Miller Jr. *Survival analysis*, volume 66. John Wiley & Sons, 2011.

[181] Roger Th AJ Leenders, Noshir S Contractor, and Leslie A DeChurch. Once upon a time: Understanding team processes as relational event networks. *Organizational Psychology Review*, 6(1):92–115, 2016.

[182] Andrew Pilny, Aaron Schecter, Marshall Scott Poole, and Noshir Contractor. An illustration of the relational event model to analyze group interaction processes. *Group Dynamics: Theory, Research, and Practice*, 20(3):181, 2016.

[183] Brooke Foucault Welles, Anthony Vashevko, Nick Bennett, and Noshir Contractor. Dynamic models of communication in an online friendship network. *Communication Methods and Measures*, 8(4):223–243, 2014.

[184] Mark Tranmer, Christopher Steven Marcum, F Blake Morton, Darren P Croft, and Selvino R de Kort. Using the relational event model (rem) to investigate the temporal dynamics of animal social networks. *Animal behaviour*, 101:99–105, 2015.

[185] KP Patison, E Quintane, DL Swain, G Robins, and P Pattison. Time is of the essence: an application of a relational event model for animal social networks. *Behavioral ecology and sociobiology*, 69(5):841–855, 2015.

[186] Duy Vu, Philippa Pattison, and Garry Robins. Relational event models for social learning in moocs. *Social Networks*, 43:121–135, 2015.

[187] Duy Vu, Alessandro Lomi, Daniele Mascia, and Francesca Pallotti. Relational event models for longitudinal network data with an application to interhospital patient transfers. *Statistics in medicine*, 36(14):2265–2287, 2017.

[188] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1. ACM, 2004.

[189] Gabriel Murray. Markov reward models for analyzing group interaction. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 336–340. ACM, 2017.

[190] Chelsea Finn, Paul Christiano, Pieter Abbeel, and Sergey Levine. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *arXiv preprint arXiv:1611.03852*, 2016.

[191] Kamwoo Lee, Mark Rucker, William T Scherer, Peter A Beling, Matthew S Gerber, and Hyojung Kang. Agent-based model construction using inverse reinforcement learning. In *Simulation Conference (WSC), 2017 Winter*, pages 1264–1275. IEEE, 2017.

[192] Alessandro Lomi, Tom AB Snijders, Christian EG Steglich, and Vanina Jasmine Torló. Why are some more peer than others? evidence from a longitudinal study of social networks and individual academic performance. *Social Science Research*, 40(6):1506–1520, 2011.

[193] Emilie M Hafner-Burton, Miles Kahler, and Alexander H Montgomery. Network analysis for international relations. *International Organization*, 63(3):559–592, 2009.

[194] Lee-Jen Wei. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879, 1992.

[195] Steven L Scott. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

[196] Christopher Steven Marcum and Carter T Butts. Constructing and modifying sequence statistics for relevent using informr in . *Journal of statistical software*, 64(5):1, 2015.

[197] Marcial Losada. The complex dynamics of high performance teams. *Mathematical and computer modelling*, 30(9-10):179–192, 1999.

[198] Yu Liu, Xi Liu, Song Gao, Li Gong, Chaogui Kang, Ye Zhi, Guanghua Chi, and Li Shi. Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers*, 105(3):512–530, 2015.

[199] Martin Gjoreski, Hristijan Gjoreski, Mitja Luštrek, and Matjaž Gams. Continuous stress detection using a wrist device: in laboratory and real life. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pages 1185–1193. ACM, 2016.

[200] Mihaly Csikszentmihalyi. *Finding flow: The psychology of engagement with everyday life.* Basic Books, 1997.