Event Analytics on Social Media: Challenges and Solutions

by

Yuheng Hu

A Dissertation Presented in Partial Fulfillment
of the Requirement for the Degree
Doctor of Philosophy

Approved November 2014 by the
Graduate Supervisory Committee:

Subbarao Kambhampati, Chair
Eric Horvitz
John Krumm
Huan Liu
Hari Sundaram

ARIZONA STATE UNIVERSITY

December 2014

ABSTRACT

Social media platforms such as Twitter, Facebook, and blogs have emerged as valuable – in fact, the *de facto* – virtual town halls for people to discover, report, share and communicate with others about various types of events. These events range from widely-known events such as the U.S Presidential debate to smaller scale, local events such as a local Halloween block party. During these events, we often witness a large amount of commentary contributed by crowds on social media. This burst of social media responses surges with the "*second-screen*" behavior and greatly *enriches* the user experience when interacting with the event and people's awareness of an event. Monitoring and analyzing this rich and continuous flow of user-generated content can yield unprecedentedly valuable information about the event, since these responses usually offer far more rich and powerful views about the event that mainstream news simply could not achieve. Despite these benefits, social media also tends to be noisy, chaotic, and overwhelming, posing challenges to users in seeking and distilling high quality content from that noise.

In this dissertation, I explore ways to leverage social media as a source of information and analyze events based on their social media responses collectively. I develop, implement and evaluate *EventRadar*, an event analysis toolbox which is able to identify, enrich, and characterize events using the massive amounts of social media responses. EventRadar contains three automated, scalable tools to handle three core event analysis tasks: Event Characterization, Event Recognition, and Event Enrichment. More specifically, I develop ET-LDA, a Bayesian model and SocSent, a matrix factorization framework for handling the Event Characterization task, i.e., modeling characterizing an event in terms of its topics and its audience's response behavior (via ET-LDA), and the sentiments regarding its topics (via SocSent). I also develop DeMa, an unsupervised event detection algorithm for handling the Event

Recognition task, i.e., detecting trending events from a stream of noisy social media posts. Last, I develop CrowdX, a spatial crowdsourcing system for handling the Event Enrichment task, i.e., gathering additional first hand information (e.g., photos) from the field to enrich the given event's context.

Enabled by EventRadar, it is more feasible to uncover patterns that have not been explored previously and re-validating existing social theories with new evidence. As a result, I am able to gain deep insights into how people respond to the event that they are engaged in. The results reveal several key insights into people's various responding behavior over the event's timeline such the topical context of people's tweets does not always correlate with the timeline of the event. In addition, I also explore the factors that affect a person's engagement with real-world events on Twitter and find that people engage in an event because they are interested in the topics pertaining to that event; and while engaging, their engagement is largely affected by their friends' behavior.

To my parents for their love and support

ACKNOWLEDGEMENTS

I am grateful to many friends and colleagues for their enthusiastic support of my PhD. journey. First and foremost, I must express my deepest gratitude to my advisor and mentor, Dr. Subbarao (Rao) Kambhampati, for his advice and support. Rao is the most wonderful advisor a graduate student can hope to have. His excellent advising skills combining patience, guidance and intellectual freedom made my Ph.D experience both exciting and productive. I have been impressed by Rao's ability to understand immediately the technical details and values of my work. I have also been moved by Rao's efforts to make my work and presentations solid, clean, and easy to understand. I leave ASU having learned many valuable things from him about research and life. Rao, I cannot thank you enough.

Thank you to my committee members, Dr. Eric Horvitz, Dr. John Krumm, Dr. Huan Liu and Dr. Hari Sundaram for helpful suggestions and for holding me to a high standard. I want to thank Eric for broadening my perspectives on research and I enjoyed our conversations. Thank you to John for many discussions about space and time, and their applications in social media. I also cannot forget how much encouragement you gave me. Thank you for everything. I want to thank Dr. Huan Liu for his comments and questions. Last, I must thank Dr. Hari Sundaram. His insightful feedback and comments always provided me many new angles to rethink about my research. Thank you to him for inspiring me.

I cannot accomplish this dissertation without the help, advice, guidance from several mentors during my internships, Ajita John from Avaya Labs and Shelly Farnham from Microsoft Research. Ajita, thank you for bringing me to the research of social computing, and for your advice in personal and professional arenas. Shelly, thank you for sharing with me a sociologist's perspectives of social computing and for teaching me many qualitative and quantitative research skills.

I have been working closely with many colleagues from ASU and outside ASU during my Ph.D. I want to thank Kartik Talamadupula, Sushovan De, Lydia Manikonda, Tuan Anh Nguyen, Tony Yu Zhang, Yilin Wang from ASU, Yi Chen from NJIT, Hankz Hankui Zhuo from Sun Yat-sen University, Shelly Farnham, Andres Monroy-Hernandez, Eric Horvitz, John Krumm from Microsoft Research, Ajita John, Doree Duncan Seligmann from Avaya Labs, Fei Wang from IBM Research, Yu-Ru Lin from University of Pittsburgh, Vagelis Hristidis from UC Riverside, and Panagiotis Ipeirotis from New York University. Thank you for the countless discussions and technical collaborations which have led to a few results, and to a greater number of new questions that undoubtedly will continue to shape my research agenda in the future.

Over the years, many friends and colleagues have provided me very helpful suggestions, insightful comments, support and encouragement during my Ph.D study. I am thankful to Kartik Talamadupula, Sushovan De, Lydia Manikonda, Raju Balakrishnan, Tathagata Chakraborty, Tuan Anh Nguyen, J Benton, Will Cushing and Tony Yu Zhang from the Yochan group, Xia Hu, Jiliang Tang, and Huiji Gao from the DMML lab. Mor Naaman from Cornell, Emre Kiciman from Microsoft Research, Justin Cranshaw from CMU, David McDonald from University of Washington, Munmun De Choudhury from Georgia Tech, Jalal Mahmud from IBM Research, and Iris Howley from CMU.

Last, my parents raised me so that I can do my best. They supported me in my decision to stay in school for so long. Thank you so much to them for all they have given me. The best thing I did while at ASU was to meet Nan. Thank you for being such a good listener and my partner in crime over the years.

TABLE OF CONTENTS

Page

APPENDIX

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

INTRODUCTION

Social media platforms such as Twitter, Facebook, and blogs have emerged as valuable – in fact, the *de facto* – virtual town halls for people to discover, report, share and communicate with others about various types of events. These events range from widely-known events (e.g., the U.S Presidential debate, the Superbowl, the Haiti earthquake) to smaller scale, local events (e.g., a gas leak at 5th street, a local parade, a local Halloween block party). During these events, we often witness a large amount of commentary contributed by crowds on social media. For example, there were more than 1,500,000 tweets about the first U.S Presidential debate in September 2012 in only one and a half hours. Likewise, over 10,000 tweets were posted within the first 15 minutes of a series of shootings in downtown Seattle on May 23, 2012.

This burst of social media responses surges with the "*second-screen*" behavior [1] , and greatly *enriches* an event context as well as the user experience when interacting with the event. For example, during the 2010 Haiti earthquake, enabled by the connectivity and ubiquity of social media, people engaged with that event by writing, sharing, reposting, commenting and curating tens of thousands of stories, thoughts, videos and photos of the disaster on Twitter and blogs from various perspectives (Yates and Paquette, 2011). These responses can be seen as a dynamic source of information enabling individuals, corporations, and government organizations to stay

[1] The first screen usually refers to the traditional medium like TV, where the live events and/or linear content (such content progresses without any navigational control for the viewer) is rendered. Second screen on the other hand refers to an additional portable device (usually tablet and smartphone) that allows a content consumer to interact with the content in different ways (e.g., tweeting about the event on Twitter) while it is being displayed on TV in parallel. More references about First vs. Second screen can be found in (Google, 2012; Times, 2012; eMarketer, 2011).

informed of “what is happening now”, “who is involved”, “where is it happening”, “when did it happen” and so on.

Monitoring and analyzing this rich and continuous flow of user-generated content/responses can yield unprecedentedly valuable information about the event, since these responses usually offer far more rich and powerful views about the event that mainstream news simply could not achieve during and following the event. For instance, people are likely to be interested in getting advice, opinions, or updates on news or events (Java *et al.*, 2007; Krishnamurthy *et al.*, 2008; Zhao and Rosson, 2009). Companies are increasingly using social media to advertise products, brands and services; to build and maintain reputations; and to analyze users’ sentiment regarding their products (or those of their competitors), especially during and following their press conferences. Such (mined) deep knowledge from such a collection of responses – known as collective intelligence – is crucial to many applications such as decision making and business intelligence, as highlighted by James Surowiecki in his landmark book (Surowiecki, 2005) and studied by other researchers (Pak and Paroubek, 2010; Liu *et al.*, 2012; Jiang *et al.*, 2011).

Moreover, social media also *enriches* people’s awareness of an event. Due to the immediacy and rapidity of social media, news events are often reported and spread on Twitter or Facebook ahead of traditional news media. For example, the news of both the 2012 Aurora shootings in Colorado and the 2012 Empire State Building shooting in New York City were reported by social media users minutes earlier than by traditional news outlets (Fillion, 2012; Lotan, 2012). This also results in the transformation of the established paradigm of – a government or news media which inform its citizens – to a *citizen body* that not only informs its government and news media, but also the world.

Despite these benefits, social media also tends to be noisy, chaotic, and overwhelming, posing challenges to users in seeking and distilling high quality content from that noise. It should be no surprise that regardless of the popularity of social media as a source of event information, people are still using television and newspapers (among other traditional sources) as their main channels for acquiring information about the events that they care about(Rosenstiel *et al.*, 2011).

In this dissertation, we explore ways to leverage social media as a source of information and analyze events based on their social media responses collectively. Here, events are defined as real-world occurrences that unfold over space and time (Troncy *et al.*, 2010; Xie *et al.*, 2008; Yang *et al.*, 1998; Allan, 2002). We aim to develop a set of computational tools that is able to automatically identify, enrich, and characterize events using the massive amounts of social media responses. Enabled by these tools, another important goal of this dissertation is to gain deep insights into *how* people respond to events on social media and *why* they do so.

As a result, the technical contributions of the dissertation is that I develop, implement and evaluate *EventRadar*, a toolbox for event analysis using Twitter responses. Unlike traditional human-based or primitive tools (e.g., (Shamma *et al.*, 2009; Livne *et al.*, 2011), see Chapter 2 for more discussions), *EventRadar* contains three automated, scalable tools that leverage the connectivity and rapidity of social media to handle three core event analysis tasks (see Figure 1.1 for the structure of EventRadar). These tools facilitate novel applications in *computational journalism, event playback, storytelling, and civic engagement* (see the usage scenario of EventRadar in journalism in Section 1.1).

These tools are:

- ET-LDA (Hu *et al.*, 2012c), a Bayesian model and SocSent(Hu *et al.*, 2013c), a matrix factorization framework for handling the **Event Characterization** task, i.e., modeling characterizing an event in terms of its topics and its audience's response behavior (via ET-LDA), and the sentiments regarding its topics (via SocSent).

- DeMa(Hu *et al.*, 2013b), an unsupervised event detection algorithm for handling the **Event Recognition** task, i.e., detecting trending events from a stream of noisy social media posts.
- CrowdX (Hu *et al.*, 2014), a spatial crowdsourcing system for handling the **Event Enrichment** task, i.e., gathering additional first hand information (e.g., photos) from the field to enrich a given event's context.

Moreover, based on the computational power of the EventRadar toolbox, we are able to make scientific contributions in this dissertation by uncovering insights that have not been explored previously and re-validating existing social theories with new evidence. For example, we gain deep insight into how people's Twitter responding behavior evolves before, during and after an event, and how this behavior changes with respect to different topics of the events (Hu *et al.*, 2012a). In addition, inspired by prior theoretical constructs that bridge social science, linguistics, and computer mediated communication, we identify and exploit multiple important factors that could potentially affect a person's presence and degree of Twitter engagement with various real-world events (Hu and Farnham, 2014).

Before going into the details of the technical and scientific contributions of this dissertation, next, we describe a motivating use-case scenario for EventRadar in order to provide a concrete context for the functionalities that EventRadar will support. Although this particular scenario focuses on a disaster event, EventRadar can be easily adapted to handle other kinds of events as well.

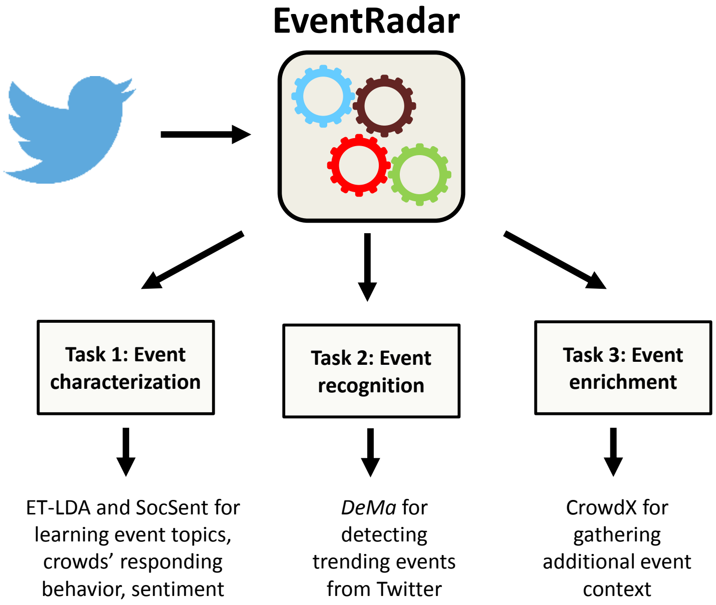

EventRadar
Task 1: Event characterization
Task 2: Event recognition
Task 3: Event enrichment
ET-LDA and SocSent for learning event topics, crowds' responding behavior, sentiment
DeMa for detecting trending events from Twitter
CrowdX for gathering additional event context

Figure 1.1: The Architecture of EventRadar

1.1 Motivating Use-case Scenario for EventRadar

Consider the recent 2014 Hong Kong protest when activists in Hong Kong protested outside the Hong Kong Government headquarters and occupied several major city intersections. In such fluid scenario, a journalist would benefit significantly from a system that is able to monitor, detect, and analyze social media responses in order to provide up-to-date information relevant to the event. We will use this scenario to illustrate the contributions of the various components of our EventRadar toolbox.

EventRadar consistently monitors social media streams on Twitter (since Twitter is a major platform used to spread event information (Kwak et al.*, 2010)) posted from and/or about Hong Kong. Hence, right after the event starts trending on Twitter (i.e., many people suddenly start to talk about the event on Twitter – see the definition of "trending" in Chapter 5), the "trending signal" is immediately recognized and captured by the* DeMa *event detection algorithm.* DeMa *also fetches all event-related tweets* *and summarizes them using a standard text summarization algorithm. Last,* DeMa *computes a novelty score for this event based on how interesting it is compared to previously detected trending events from the Hong Kong metropolitan area.*

Once the novelty score passes a pre-defined threshold, an alert is issued to the journalist. He then starts to review the summarization of the detected event to determine whether it is news worthy. he realizes that the event is high news worthy. Therefore, he decides to conduct further investigation himself before writing his report. More interestingly, he finds the event (protest) is happening at multiple spots in Hong Kong at the same time

Unfortunately, the journalist is at one event spot which is far away from the other event spots' locations. It is also very likely that the event at those spots will be over before the journalist arrives (note that the event will be much less worthy of coverage once it ends). Therefore, he decides to use the CrowdX system to handle his journalistic investigation. As a result, he sends a task request to CrowdX with his budget limit, the descriptions of the event spots and the task (e.g., take photos and videos at each event spot), and time constraints for finishing the task. CrowdX then automatically outsources his request (with monetary incentives) to a group of workers who are close to the event spots and have availability, interests, time, and skills to finish their assigned task. After the workers finish their assignments, the collected event information is sent back to the journalist.

With the additional information collected from both Twitter and the event spots, the journalist can do several in-depth event analyses and use the mined insights to support his report on the Hong Kong protest. For example, in order to learn how the topics (of people's responses to the event) evolved over the event's timeline and how people responded to the event, he uses ET-LDA which automatically segments the entire event progress and aligns the event-related tweets to different segments based on *their topical associations. Moreover, the journalist is also interested in characterizing people's sentiment regarding the event, especially during different developments of the event. Therefore, he uses* SocSent *to automatically learn the latent sentiment of people based on their tweets regarding the entire event, or the different parts of the event.*

Finally, based on the event detected by DeMa*, the additional first-hand information returned by CrowdX and the mined knowledge from ET-LDA and* SocSent*, the journalist obtains enough relevant material and is confident that he can write an insightful report about the event.*

1.2 Technical Contributions

Our contribution in this dissertation is to realize scenarios such as the above, by developing tools that target Event Recognition, Enrichment, Characterization, and Sensemaking. These tools are: (i) ET-LDA, an effective method for aligning an event and its social media responses, (ii) SocSent, a tool for aggregating public receptions on social media regarding the topics/segments of the event, (iii) DeMa, an effective event detector which detects trending events on Twitter, (iv) CrowdX, a spatial crowdsourcing platform for gathering first hand event information. Next, we briefly review these tools below.

1.2.1 Detecting Trending Events with DeMa

As mentioned in the scenario, we need an approach to detect unspecified (i.e., unplanned) trending events (since journalists normally are aware of planned large events quite well), and their associated Twitter responses when they break out on social media. Besides, the approach needs to handle any events that are trending, rather than focusing on a particular type of event as in previous research, e.g., earthquakes (Sakaki *et al.*, 2010). In this dissertation, we propose and develop DeMa, an unsupervised online event detector which identifies each trending event – and its associated Twitter responses, from a stream of noisy Twitter messages (Hu *et al.*, 2013b). DeMa takes a two-step approach: First, it identifies trending features (terms) from Twitter posts using a trends indicator; later, it clusters the topically-related trending features to form event-related clusters (each cluster represents a trending event). More specifically, in order to identify trending features from a substantial volume of Twitter posts, we are inspired by the definition of "trending" from the literature of computational finance (Taylor and Allen, 1992; Blume *et al.*, 1994) and apply it to the Twitter context. By this definition, we develop our trends indicator using EMA (exponential moving average), MACD (Moving Average Convergence Divergence), and MACD histograms to characterize the time series of every feature in a Twitter stream and capture the trending features. Since not every trending feature is novel (e.g., "good morning" will be trending between 8:00 AM to 10:00 AM but everyday), we compute a novelty score for each feature based on the normalized difference between the feature's current time series and the the feature's past time series. After that, we group those novel trending features into topically-related event-clusters using the shared nearest neighborhood (SNN) clustering algorithm, where the distance between two neighbors is defined based on their topical similarities. With the help of the DeMa algorithm, regular social media users and journalists can easily learn about any trending events (and their associated tweets) on Twitter within any given time period.

1.2.2 Gathering Event Information with CrowdX

Next, in order to help the journalists collect more event information from the field (we call it the Event Enrichment task), we consider the crowdsourcing approach.

Unlike traditional crowdsourcing schemes which focus mostly on online tasks, we require our crowdsourcing approach to handle spatial tasks (i.e., tasks associated with locations). Moreover, the crowdsourcing system should take into account the cost of hiring workers, the dynamic utility of the tasks and the uncertainty of the workers accepting and finishing their assignments. To this end, we propose and develop CrowdX (Hu *et al.*, 2014). It harnesses programmatic access to workers like traditional crowdsourcing. More importantly, it constructs 1) a model to assess the non-uniform utility of each task assignment, 2) a model to estimate the expected cost of each worker, and 3) a probabilistic model to quantify the worker's uncertainty. More specifically, we develop a utility-theoretic approach which considers workers and event spots jointly and selects worker-spot assignments that simultaneously maximize the overall utility, and achieve low cost within the given budget. The utility of each worker-spot assignment is assessed based on two properties: representativeness (an ideal event spot should cover the most representative and diverse aspects of the event at the same time) and quality (an ideal worker should complete her task with high quality). Moreover, we develop a Bayesian model to quantify each worker's uncertainty by predicting the likelihood of her accepting a task request and finishing the task, based on several features. Based on this, we also develop a cost function that takes into account different cost metrics and estimates the expected cost for each work-spot assignment. Finally, balancing utility of user-spot selection and assignment with the need to achieve a probabilistic budget-aware cost efficiently can be formalized as a discrete optimization problem. Exploiting the concept of submodularity, we develop a greedy algorithm which is guaranteed to provide a near-optimal solution for this hard problem. With the help of the CrowdX system, the journalists can easily gather rich and first-hand event information from the field by hiring a group of workers with a (controllable) hiring cost.

1.2.3 Joint Event-Tweets Analysis with ET-LDA

Next, a key challenge in the in-depth event analysis (i.e., the Event Characterizing task) is to establish the alignment between an event and its associated Twitter responses. To address this problem, we propose and develop ET-LDA (joint Event and Tweets LDA) (Hu *et al.*, 2012a,c, 2011), a hierarchical Bayesian model. It aligns an event and its associated Twitter messages (tweets) through joint modeling of the inter-dependency, i.e., topical influences between them. In order to characterize such influences, we first propose that rather than enforcing tweets to be correlated only with the topics of the event that occur within time-windows around the tweets's timestamps, they should correlate to any topic in the event. Next, we assume that that a tweet can be composed in multiple ways to respond to the event, depending on the topic influences from the event. For example, a person may comment directly on a specific topic/part of the event, so, her tweets would be deeply influenced by that topic. In contrast, she could comment broadly about the event. In this case, the tweets would be more influenced by the general topics/themes of the event. ET-LDA models exactly these two distinct tweeting behaviors of the crowd in a generative process. We deem the tweets specific tweets if they are generated in the first way, since their content refers to the specific topics of the event. To determine what these topics are about and which topic a tweet refers to, ET-LDA splits the event into several sequential segments in which a particular set of topics is covered. On the other hand, we deem the tweets general tweets if they are generated in the second way, because their topics stay steady on the general topics across the event rather than being affected by its varying context. Learning in ET-LDA is proven to be intractable (Griffiths and Steyvers, 2004). To resolve this, we derive an approximate inference procedure and estimate parameters using collapsed Gibbs sampling. Enabled by ET-

LDA, the journalists are able to conduct complex in-depth analysis such as learning the topic transitions about the event and inferring people's responding behavior over the event's timeline.

1.2.4 Aggregated Twitter Sentiment Analysis with SocSent

Once we have the alignment in place, we would like to support aggregate in-depth sentiment analysis to characterize the events. In particular, we are interested in how different segments of an event have been received by social media in terms of the sentiment of the crowd in response to various parts and topics of the event. To this end, we propose and develop SocSent, a flexible factorization framework, that learns aggregated sentiments of both (a) segments and (b) topics of an event they elicited on Twitter (Hu *et al.*, 2013c). SocSent advances the state-of-the-art sentiment analysis approaches in several aspects: (1) SocSent requires very little or no manually-labeled tweet sentiment as training data, (2) SocSent has the ability to relate Twitter sentiments to segments and topics of the event, and (3) SocSent has domain knowledge about contextual cues in the tweets. Specifically, SocSent seeks a new representation of the Twitter sentiment and its correlations to the event by factorizing an input tweet-term matrix into four factors corresponding to tweets-segment, segment-topic, topic-sentiment and sentiment-words. The ET-LDA alignment can be seen as providing the initial information ("prior knowledge") on the tweet-segment and segment-topic factors. Similarly, the availability of labeled tweets can be used to constrain the product of tweet-segment, segment-topic and topic-sentiment matrices. Finally, the sentiment lexicon is used to regulate the sentiment-words matrix. We pose this factorization as an optimization problem where, in addition to minimizing the reconstruction error, we also require that the factors respect the prior knowledge to the extent as much as possible. We derive a set of multiplicative update rules that efficiently solve this optimization problem. As a result of this factorization, the journalists can readily determine whether people appreciate the segments or topics of the event or dislike them.

1.3 Scientific Contributions

1.3.1 Understanding Tweeting Behavior w.r.t Event Progress

Enabled by the ET-LDA tool, we are able to gain deep insights into how people respond to the event (that they are engaged in) over the event's timeline. Our results reveal several key insights regarding people's behavior: (1) We discover that the crowd's tweeting behavior varies significantly with the timeline of the event. Often, more specific tweets are witnessed during the event and less are found before or after the event. This indicates that the crowd's responses tend to be general and steady before the event and after the event, while during the event, they are more specific and episodic since the crowd is more engaged in the specific topics of the event. (2) We also discover that people show different levels of engagement in different kinds of events. For example, we find a greater level of engagement (the total number of tweets and the percentage of specific tweets) in domestic events as opposed to international issues. (3) Our final finding is that the topical context of the tweets does not always correlate with the timeline of the event. We have seen that a segment in the event can be referred to by specific and episodic tweets at any time irrespective of whether that segment has already occurred, or is occurring currently or will occur later on. This finding is significant in light of the fact that current approaches such as (Shamma *et al.*, 2009) focus on correlating tweets to the event based on their timestamps.

1.3.2 Understanding People's Twitter Engagement in Events

We are also interested in understanding of factors that affect a person's engagement with real-world events on Twitter (Hu and Farnham, 2014). To address this question, we operationalize a person's event-related engagement on Twitter using three specific user behaviors: 1) posting, 2) retweeting and 3) replying to a tweet about a real-world event. Inspired by prior theoretical constructs that bridge social science, linguistics, and computer mediated communication, we collect factors that could potentially affect the person's Twitter engagement in a real-world event from five broad categories: (i) Twitter activities, (ii) Tweets' topics, (iii) Twitter user types, (iv) Geolocation, and (v) Social network structure. We map these dimensions into 17 numeric predictive variables manifested on Twitter, spanning the volume of tweets produced, burstiness of tweets, frequency of retweets and so on. We construct statistical models to assess the relative contributions of these variables towards predicting the presence of a person's Twitter engagement and the degree of that Twitter engagement in the real-wold events. As a result, we discover that 1) People engage in an event because they are interested in the topics pertaining to that event, 2) People tend to engage more in an event due to their friends (following list) are also posting tweets about it, and 3) The event locations do not affect people's engagement.

1.4 Broader Impact of this work

In addition to the main contributions described above, techniques presented in this dissertation also have some broader impacts to various other areas. Below, we highlight some of those impacts.

Event Analytics on Social Media and Applications in Journalism

Technology is rapidly shifting the ways in which information about news and events is gathered, processed, and disseminated. Computational Journalism is the application of computing to the activities of journalism including information gathering, organization and sensemaking, communication and presentation, and dissemination and public response to news information, all while upholding the core values of journalism such as accuracy and verifiability (Diakopoulos *et al.*, 2010; Cohen *et al.*, 2011b; Anderson, 2013). In recent years, some of the core areas of computing such as databases, information retrieval, and information visualization, are already playing important roles in driving many changes as news organizations re-adjust to the digital era (Cohen *et al.*, 2011a; Diakopoulos *et al.*, 2012; Flaounas *et al.*, 2013). While Computational Journalism is unlikely to replace real journalists, it does enable and augment human journalists through computing. Therefore, we believe that the transfer and use of computing technology in news and journalism can be accelerated, and our work presented here can have a direct impact regarding computational journalism, especially on information gathering, organization and sensemaking.

As we mentioned earlier, the first step in Journalism is to gather relevant information about news and events. Thus far, this has been done mainly based on tips (for breaking events) and/or journalistic investigation. While such ways still work quite well, social networking systems (e.g. Facebook), social awareness streams (e.g. Twitter), location-based social networks (e.g., Foursquare) have explicitly connected the "what", the "who", the "where", and the "when" of reporting. Besides, with the ubiquity and immediacy of social media, news events often are reported on Twitter or Facebook ahead of traditional news media. In addition, social media has also become one of the few sources of local news – and life-saving information – where traditional media is sometimes censored by governments or even criminal organizations. These advantages make social media an ideal information source for journalists to gather more information to learn new stories and/or augment their stories.

However, social media is known to be noisy, chaotic, and overwhelming; hence leveraging it for information gathering for journalists is never easy. Our EventRadar toolbox readily handles these challenges. For example, DeMa automatically discovers, extracts, and summarizes relevant information about trending events on Twitter.

On the other hand, the adoption of cheap and ubiquitous devices (e.g., smartphones) with photo and video capability has already had a substantial impact on how stories are reported, both in the mainstream media and through citizen journalism (see examples in (Gilbert *et al.*, 2010; Das *et al.*, 2010)). The existing way of gathering this kind of information (e.g., photos, videos) requires 1) people upload information voluntarily to the new outlets' websites (e.g., CNN has a website "http://ireport.cnn.com/" for grassroots to upload information about an event/story), or 2) journalists have to spend a lot of time scanning social media sites to gather relevant information. Despite its effectiveness, such an information gathering process is passive and time consuming which will be a concern when events are fast-paced and dynamic. The CrowdX system of EventRadar addresses this common problem in journalism in terms of fast and proactive outsourcing the information gathering tasks to a group of people with monetary rewards.

Next, for information organization and sensemaking, there is already a field where computers have already had a significant impact, namely though Computer Assisted Reporting (CAR) (Reavy, 2001; Garrison, 1998). CAR tools are usually generic in the sense that they are widely applicable to different stories with tools that are designed to interact with geographic data, temporal data, or network data. While many CAR tools succeed in enabling journalists to organize their information, not much effort has been spent on to information sensemaking. However, given massive amounts of information gathered about the event, it is important to convert it into human-readable knowledge. Indeed, there has been recent interest in making sense of social media, however, as we mentioned earlier that the most existing work is either human-based or primitive. Our EventRadar toolbox can helps journalists in the sense that both ET-LDA and SocSent tools are able to convert signals about the world (extracted from Twitter messages) into aggregate and derivative insights (e.g. the sentiment, topics, human behavior with respect to an event).

The work that will be presented in the next few chapters has resulted in several publications at top conferences (see References). The DeMa work also won a best paper nomination at ACM CHI 2013. Besides, the work has been featured in various press including ABC news, [2] PBS, [3] The Seattle Times, [4] FastCompany [5] Computer Magazine, [6] Neowin, [7] ASU news, [8] and so on.

1.5 Dissertation Overview

Next, I will first present event categorization and illustrate specific challenges EventRadar needs to address in Chapter 2. After that, we present technical details for solving the Event Recognition task in Chapter 5. In Chapter 6, I will describe the CrowdX system for handling the Event Enrichment task with the help from a group of workers. Next, for addressing the Event Characterization task, I will discuss the connection between the event and its Twitter responses for modeling the event topics and people's tweeting behavior in Chapter 3 and for inferring the event sentiment in

[2] http://abcnews.go.com/Lifestyle/instagram-vast-favorite-themes-surprisingly-study-shows/story?id=26551891

[3] http://www.azpbs.org/video/horizon.php?vidId=2331

[4] http://blogs.seattletimes.com/microsoftpri0/2013/08/22/microsoft-launches-whooly-localized-twitter-project/

[5] http://www.fastcoexist.com/3017494/a-hyperlocal-twitter-filter-that-gets-neighbors-talking-to-each-other

[6] http://computermagazine.com/2013/08/22/microsoft-research-launches-whooly-localized-twitter-project/

[7] http://www.neowin.net/news/microsoft-research-launches-beta-of-twitter-aggregator-whooly

[8] https://asunews.asu.edu/20130920-kambhampati-twitter-analysis/

Chapter 4. Last, in Chapter 7, I will investigate various factors that affect people's engagement in different types of events for handling the Event Sensemaking task. Finally, I will summarize this work and discuss future research on event analytics using social media in Chapter 8. Readers unfamiliar with probability theory, or belief networks may wish to review Appendix A to understand better Chapters 3 and 4.

Chapter 2

BACKGROUND

To understand events that exist on social media and guide the development of the EventRadar toolbox more properly, in this chapter we define and categorize a large set of events on the social media platform Twitter based on several key distinguishing features. After that, we highlight the challenges EventRadar toolbox has to address in order to realize the motivating scenario mentioned in Chapter 1.

2.1 Event Definition and Categorization

The definition of *event* has received substantial attention across various academic areas, from philosophy (Zalta and Abramsky, 2003) to psychology (Zacks and Tversky, 2001), over the years. An event is often defined as an abstract concept (Zalta and Abramsky, 2003), or with respect to its manifestation in a specific domain (e.g., textual news (Allan, 2002), video and audio clips (Westermann and Jain, 2007), social media (Becker *et al.*, 2011)). However, even within a specific domain, researchers often debate over what an event really is (Gerner *et al.*, 1994), or agree on (and use) a definition that is admittedly problematic and incomplete (Makkonen, 2003). In this dissertation, instead of providing yet another definition of an event, we define specific type of events *in the context of* Twitter (note that events on other social media platforms can be defined in a similar way), based on the traditional event definition (Troncy *et al.*, 2010; Xie *et al.*, 2008; Yang *et al.*, 1998; Allan, 2002). We provide the following definitions:

Definition 1 (Event on Twitter): *An event $\mathcal{E}$ is a real-world occurrence with (1) an associated time period $\mathcal{T}_{\mathcal{E}}$, and (2) a stream of corresponding Twitter messages*

$\mathcal{M}_{\mathcal{E}}$ about $\mathcal{E}$ and published during $\mathcal{T}_{\mathcal{E}}$.

Next, following the event categorization in (Becker *et al.*, 2011), we classify the events in Twitter into two categories – *planned events* and *unplanned events* – depending on whether the context of an event (e.g., topic or hashtags on Twitter, time, location) is available to our toolbox.

Definition 2 (Planned event): *A planned event $\mathcal{PE}$ is an event in Twitter with pre-known corresponding event context information consisting both (a) topic or hashtags of the event, and (b) time, at which $\mathcal{PE}$ is planned to occur.*

Based on this, we further define:

Definition 2.1 (Planned transcribable event): *A planned transcribable event $\mathcal{PE}+$ is an event in Twitter with pre-known corresponding event context information consisting at least of (a) topic or hashtags, (b) time, at which $\mathcal{PE}+$ is planned to occur, and (c) transcripts or captions of $\mathcal{PE}+$.*

The main motivation here is that the transcribable events are often associated with external documents (e.g., transcripts) which can be leveraged to provide additional fine-grained event analysis. Next, in contrast to the planned event, we have no information about the unplanned events in the Twitter stream. To characterize and define such events, we have to rely on other signals that could indicate their presence in Twitter. To this end, we focus on a class of events that we refer to as *trending* events, which are events that exhibit temporal burst patterns (see Chapter 5 for more definitions on trending).

Definition 3 (Unplanned trending event): *An unplanned trending event $\mathcal{UE}$ is an event in Twitter with one or more features (e.g., terms) of the corresponding Twitter messages $\mathcal{M}_{\mathcal{UE}}$ exhibiting bursty patterns during the event's time period $\mathcal{T}_{\mathcal{UE}}$.*

In addition to the event definition and categorization (See Table 2.1 for event examples), it is worth noting that in this dissertation, we consider all Twitter messages

$\mathcal{M}$ that are relevant to an event $\mathcal{E}$ as being corresponding to (or associated with) $\mathcal{E}$, regardless of whether $\mathcal{M}$ was posted before, during or after the event. In other words, the time period $\mathcal{T}$ covers the actual start and end time of $\mathcal{E}$.

Types	**Events**
Planned Non-transcribable Event	Occupy wall street in New York City 2012 Homecoming party at Arizona State University
Planned transcribable Event	2012 U.S Presidential debate 2013 Superbowl
Unplanned trending event	2013 Boston Bombings Shootings in downtown Seattle on May 23, 2013

Table 2.1: Examples Of Different Types Of Events.

2.2 Limitations of Existing Work and Challenges

Next, we illustrate specific challenges that our toolbox EventRadar needs to address in order to perform three tasks: Event Recognition, Event Enrichment and Event Characterization. As we will see later, although some tasks are relatively easy to solve for certain types of events based on their context (Section 2.1), most tasks across all types of events raise several technical challenges.

First, for the unplanned trending events we have no information about whether or not they will happen, when they will happen and what the associated Twitter responses will be about if they happen. Therefore, rather than focusing on the tasks, the most critical task for our toolbox for this type of events is Event Recognition, i.e., for from a stream of Twitter messages. And once the event content is identified, the other tasks can be addressed by using similar techniques that are designed for planned events. It is well known that social media like Twitter tends to be noisy, chaotic and overwhelming, and in the meantime most unplanned trending events go viral very fast on social media. Therefore, we need effective approaches to address challenges such as *how to indicate the presence of an unplanned trending events $\mathcal{UE}$ and associated Twitter responses $\mathcal{M}_{\mathcal{UE}}$ in real-time from a noisy Twitter messages.*

Next, for planned events such as the 2012 U.S Presidential debate and the Occupy wall street in New York City, the Event Recognition task can be performed in a straightforward way due to the fact that these events's context – topics, hashtags and time are often readily available (e.g., the MSNBC news posted both the event's start and end time and official hashtages for the first 2012 U.S Presidential debate nearly 2 week prior to the event). Therefore, the main focus of EventRadar for this type of events is on the rest two tasks: Event Enrichment and Event Characterization. Developing automated analytical approaches here however involves overcoming significant technical challenges. Specifically, given a planned transcribable event $\mathcal{PE}+$, its associated Twitter responses $\mathcal{M}_{\mathcal{PE}+}$ and an event transcript $\mathcal{D}_{\mathcal{PE}+}$, for the task of Event Characterization, we need approaches to address challenges like *how to segment $\mathcal{PE}+$ (based on $\mathcal{D}_{\mathcal{PE}+}$) into smaller yet meaningful parts where each part covers a set of topics. How to align the event $\mathcal{PE}+$ (based on $\mathcal{D}_{\mathcal{PE}+}$) with Twitter messages $\mathcal{M}_{\mathcal{PE}+}$ and infer which parts of the event the responses refer to.* Furthermore, based on this alignment, *how to characterize and classify the aggregate sentiments of segments and topics of $\mathcal{M}_{\mathcal{PE}+}$ that elicited on $\mathcal{M}_{\mathcal{PE}+}$*. Besides, *how to capture and model the effect and influence of $\mathcal{PE}+$ on its Twitter responses $\mathcal{M}_{\mathcal{PE}+}$*, and *how to infer whether or not a Twitter message is deeply influenced by a particular segment of $\mathcal{PE}+$.*

Last, we are interested in enriching the event context for planned (or detected) events by gathering first-hand texts, photos and videos from the field. To conduct this Event Enrichment task, EventRadar relies on a crowdsourced solution via calling a group of people near/on the event spots. However, developing this crowdsourcing scheme is non trivial and it involves in solving several challenges like *given a limit budget, how to select a group of people (workers) and assign them to a set of event spots*, *how to define and assess the utility of each task assignment*, *how to estimate the expected cost of each worker*, and *how to quantify the workers' uncertainty*.

While the tasks of Event Recognition, Event Enrichment and Event Characterization have received considerable attention in recent years from different communities such as information retrieval (Allan, 2002), data mining (Ritter *et al.*, 2012), social computing (Weng and Lee, 2011), natural language processing (Tanev *et al.*, 2008), directly applying these existing solutions to address aforementioned challenges have several major drawbacks. Specifically, for Event Characterization, most existing solutions (such as (Diakopoulos and Shamma, 2010a)) ignore the influence of an event on its Twitter responses and treat them in isolation (which obviously are interdependent as discussed in Chapter 3). Indeed, some recent approaches (Shamma *et al.*, 2010) tend to model them jointly but rely on strong assumptions. For instance, they assume Twitter messages to be correlated *only* with the topics of the event that occur within time-windows around the messages's timestamps, which is not valid based on our preliminary studies (Chapter 3). Next, for Event Recognition, most state-of-the-art algorithms (Kleinberg, 2003; He and Parker, 2010) focus on very specific types of events (e.g., sport events (Nichols *et al.*, 2012)), since they need to be trained using domain-specific features. Although there have been some exceptions, e.g., unsupervised approaches that work cross domains (e.g., (Nichols *et al.*, 2012; Weng and Lee, 2011)), they can neither identify when the detected event started trending nor infer the novelty the event is. Last, for Event Enrichment, the retrieval of rich information about an event using crowdsourcing brings up issues of both spatial task assignment and quality assurance. There has been some work on spatial crowdsoucring lately

(Kazemi and Shahabi, 2012; Kazemi *et al.*, 2013; Alt *et al.*, 2010; Väätäjä *et al.*, 2011). However, most existing solutions are inapplicable to our scenario due to three reasons. First, they can work only with self-incentivised workers. Second, they assume the crowdsourced task has a uniform utility regardless of its location or the dispatched worker. Last, these studies assume that workers will always accept and complete the task requests which however is unlikely in reality. Next,

The weaknesses discussed in the foregoing paragraphs motivated the need for an automated, in-depth event analytics toolbox to support 1) event detecting a trending event from a stream of noisy tweets (Event Recognition), 2) crowdsourcing spatial tasks to a group people to cover the event (Event Enrichment), and 3) segmenting and aligning the event and its responses, and inferring sentiment of the event and event segments (Event Characterization). In the rest of this dissertation, we present more details about our toolbox.

Chapter 3

JOINT MODELING OF EVENTS AND TWEETS FOR EVENT CHARACTERIZATION

As mentioned in the introduction, social media platforms such as Twitter have become the *de facto* platform for crowds to share perspectives and commentaries during various events such as the Superbowl, the U.S. Presidential and Primary debates, the last episode of a TV drama series, etc. Such burst of social media information, on the one hand, enriches the user experience for the live event. On the other hand, it poses tremendous challenges for attempts to analyze and extract sense from the tweets, which is critical to applications for computational advertising, community detection, journalistic investigation, storytelling, playback of events, etc. How can we identify *what these tweets were about*? And *did these tweets refer to specific parts of the event and if so, what parts?* Furthermore, *what was the nature and magnitude of the influence of the event over the tweeting behavior of the crowd?*

In this chapter, we answer these questions by devising a computational method geared toward extracting sense from the crowd's tweets in the context of the public events that they are in response to. Therefore, the focus in this chapter is quite different from the literature of sensemaking of tweets in that the existing techniques tend to focus on either tweets in isolation from the context of the event or their usage patterns, e.g., volumes of tweets, networks of audience members and their tag relations, etc (see Section 3.1 for more discussions on the related works).

More specifically, given an event and an associated large-scale collection of tweets, here we focus on two fundamental problems in analyzing and making sense of them, namely, extracting the topics covered in the event and tweets, and segmenting the event into topically coherent segments. While both topical modeling and event segmentation have received considerable attention in recent years, they have been mainly viewed as separate problems and studied in isolation. For example, there have been significant efforts on developing Bayesian models to discover the patterns that reflect the underlying topics from the document (Blei *et al.*, 2003; Griffiths *et al.*, 2004; Wang and McCallum, 2006; Titov and McDonald, 2008). Similarly, there is also a rich body of work devoted to segmentation of events/discourses/meetings via heuristics, machine learning, etc. (Hearst, 1993; Boykin and Merlino, 2000; Galley *et al.*, 2003; Dielmann and Renals, 2004).

Directly applying these current solutions to analyze the event and its associated tweets however has a major drawback: they treat event and tweets independently, thus ignoring the topical influences of the event on its associated tweets. In reality they are obviously inter-dependent. For example, in practice, when tweets are generated by the crowds to express their interests in the event, their content is essentially influenced by the topics covered in the event in some way. Based on such dependencies, i.e., topical influences, a person can respond to the event in a variety of ways. For example, she may choose to comment directly on a specific topic in the event which is of concern and/or interest to her. So, her tweets would be deeply influenced by that specific topic. In another situation, she could also comment broadly about the event. Therefore, the tweets would be less influenced by the specific topics but more by the general topics of the event.

In this chapter, we are interested in jointly characterizing the topics of the event and its associated tweets, as well as segmenting the event in one unified model. Our work is motivated by the observation that the topical influences from the event on its associated tweets are not only used for indicating the topics mentioned in the event but also indicating the content and topics in tweets and the tweeting behaviors of the crowd. Besides, by accounting for such influences on tweets, we can obtain a richer context about the evolution of topics and the topical boundaries in the event which is critical to the event segmentation, as mentioned in (Shamma *et al.*, 2009).

We build our joint Bayesian model ET-LDA (event tweet LDA). In our model, an event may consist of many paragraphs, each of which discusses a particular set of topics. These topics evolve over the timeline of the event. We assume that whether the topic mixture of a paragraph changes from the one in its preceding paragraph follows a binomial distribution parameterized by the similarity between their topic distributions. With some probability, the two paragraphs are merged to form a segment; otherwise, a new segment is created. Additionally, we assume the event (in fact the segments) can impose topical influences on the associated tweets. Intuitively, since tweets are generated by the crowd to express their interest in the event, they are essentially influenced by the topics covered in the event in some way. In order to characterize such influences, we first propose that rather than enforcing tweets to be correlated only with the topics of the event that occur within time-windows around the tweets' timestamps (a common approach in the literature, e.g., (Shamma *et al.*, 2009)), they should correlate to any topic in the event. Next, we claim that a person can compose her tweets in a variety of ways to respond to the event. To take an example, she may choose to comment directly on a specific topic in the event which is concerning and/or interesting to her. So, her tweets would be deeply influenced by that topic. In another situation, she could comment broadly about the event. In consequence, the tweets would be less influenced by the specific topics but more by the general themes of the event. Our approach models exactly these two distinct tweeting behaviors of the crowd, and the words in the tweets can belong to two distinct types of topics: general topics, which are high-level and constant across the entire event, and specific topics, which are detailed and relate to specific segments of the event.

We define a tweet in which most words belong to general topics as a “general tweet”, indicating a weak topical influence from the event, whereas a tweet with more words about the specific topics is defined as a “specific tweet”, indicating a strong topical influence from one segment of the event. Similar to the event segmentation, whether the event has strong or weak influence on tweets depends on a binomial distribution. To learn our model, we derive inference and estimate parameters using Gibbs sampling. In the update equations, we can observe how the tweets help regularize the topic modeling process via topical influences and *vice versa*.

3.0.1 Our Contributions.

Enabled by this model, people would gain much deeper insights about the event (e.g., which topic was most interesting to the crowd) and the tweets around it (e.g., what they were about). In addition, the model also sheds light on the *nature* of the crowd’s tweeting behaviors in the following ways: (1) Reveals the topical context of the tweets, and (2) Shows how the tweets evolve over the event’s timeline. Such work, to our knowledge, has not been investigated before and is not feasible with other alternative methods. For example, manual coding of the tweets is prohibitively expensive, and pure topic modeling (such as LDA (Blei *et al.*, 2003)) does not easily enable the segmentation of the event and distinguishing between two types of tweets.

3.0.2 Our Results.

We perform quantitative studies of the proposed model over two large sets of tweets in response to President Obama’s speech on the Middle East in May 2011 and a Republican Primary debate in Sept. 2011. Our results reveal several key insights into how people responded to the event: (1) We find the crowd’s tweeting behavior varies with the timeline of the event. More episodic tweets were witnessed during the event and less were found before or after the event (the percentages on average are 55%/35%/38%). (2) We also discover that people showed a greater level of engagement (the total number of tweets and the percentage of episodic tweets) in the Republican debate which centered around national issues as opposed to President Obama's Middle East speech. (3) We find that, as the event evolved, the crowd tended to comment on any topic in the event – that topic could have been discussed before, was being discussed currently, or was expected to be discussed later on.

We also address the issue of evaluating results in the absence of ground truth. This is accomplished with a user study with 31 active Twitter users in a university. We evaluate the goodness of sampled topics and episodic tweets by different methods based on the participants' perception of their quality. From the participant responses in the user study, we observe that our approach yields better quality, with improvements in the range of 18%–41% over the state-of-the-art.

The rest of the chapter is organized as follows. In Section 3.1 we discuss related work. Section 3.5.1 presents our observation of the crowd's tweeting patterns to an event. In Section 3.2 we present our approach. Section 3.4 and 3.5.4 present quantitative studies and subjective evaluations, followed by a discussion of their results. Section 7.6 concludes the chapter.

3.1 Related Work

In this section, we introduce relevant prior work with respect to understand, analyze and make sense of events and their Twitter responses. In specific, we provide a detailed overview of the realm of prior work corresponding to: (a) analyzing and sensemaking of tweets and events; (b) topic modeling of textual documents; and (c) the segmentation of documents and events.

Analyzing and Sensemaking of Tweets and Events: While the topic of analyzing and making sense of a crowd's responses to a media event is relatively new, there have been some recent attempts to characterize events based on the tweets contributed around them. Previous work includes: inferring structures and dynamics of events based on the usage patterns of Twitter (e.g., volumes of tweets over an event's timeline), the textual content of tweets (e.g., keywords) and Twitter users' social networks (e.g., followings/followees, relationships between hashtags) (Shamma *et al.*, 2009, 2010); detecting/reporting either planned events or breaking events from tweets (Weng and Lee, 2011; Petrović *et al.*, 2010; Becker *et al.*, 2011; Sakaki *et al.*, 2010); summarizing events using tweets (Chakrabarti and Punera, 2011); sentimental analysis of tweets to understand the events (Diakopoulos and Shamma, 2010a). A slightly different angle to characterize events is through exploring types of Twitter users that posted messages about them. For example, in (Starbird *et al.*, 2010; Vieweg *et al.*, 2010; De Choudhury *et al.*, 2012) the authors studied events by the classification of Twitter audience types and categories.

There is also a rich body of work that investigates tweets outside the context of events. This includes studies of why people tweet (Java *et al.*, 2007; Zhao and Rosson, 2009), representations of tweet content using a labeled topic model (Ramage *et al.*, 2010), characterizations of individuals' activity on Twitter through a content-based categorization of the type of their tweets (Naaman *et al.*, 2010) or through network analysis (Wu *et al.*, 2011; Kwak *et al.*, 2010), and also predictions of social influence on Twitter and other social media (Cui *et al.*, 2011; Bakshy *et al.*, 2011).

Topic Modeling: Our work is also informed by prior work on topic modeling methods such as Latent Dirichlet Allocation (Blei *et al.*, 2003). Such methods have achieved great success in discovering underlying topics from text documents. Recently, there has been increasing interest in developing better and sophisticated topic modeling schemes for various scenarios. One line of such research is to extend topic models on networked documents or short documents, e.g., research publications, blogs, social networks, etc. For example, PHITS (Hofmann, 2001) models the documents and their inter-connectivity based on topic-specific distributions; RTM (Chang and Blei, 2009) also studies document networks but is based on a hierarchical relational model; (Dietz *et al.*, 2007) models the influences in a citation networks built over publications; similarly, LinkPLSA-LDA characterizes topics and influences of blogs (Nallapati *et al.*, 2008); (Ramage *et al.*, 2010) proposes a labeled LDA model specifically for tweets. In addition to these work which assumes the learnt topic distribution is static in document collections, some other work considers the dynamics of topics where each topic distribution keeps evolving over a timeline. This includes the dynamic topic model (Blei and Lafferty, 2006) and the topic over time (TOT) model (Wang and McCallum, 2006). Also note that these topic models require to set the number of topics in advance. Therefore, various nonparametric topic models have been proposed to tackle this limitation such as (Blei *et al.*, 2010; Teh *et al.*, 2006).

Event Segmentation: In parallel, there is a rich body of related work on automatic topic segmentation of events, texts, and meetings. Many approaches have been developed based on ideas from information retrieval, natural language processing, and recently machine learning. For example, in (Hearst, 1993) the authors use a measure of lexical cohesion between adjoining paragraphs for segmenting texts. LCSeg (Galley *et al.*, 2003) uses a similar approach on both text and meeting transcripts and gains better performance than that achieved by applying text/monologue-based techniques. In addition to lexical approaches, machine learning methods have also been considered. (Beeferman *et al.*, 1999) combines a variety of features such as statistical language modeling, cue phrases, discourse information to segment broadcast news. Similarly, (Maskey and Hirschberg, 2003) use entirely non-lexical features. Recent advances have considered using generative models (Barzilay and Lee, 2004; Purver *et al.*, 2006). These methods enable the segmentation of topics through a generative process of building lexical models of the topics.

Limitation of Previous Work: The focus of most of the above work is either to model topics in documents (where documents are assumed to be the *homogenous*, e.g, tweets, research papers) or segment the events alone. However, they do not know provide insights into how to characterize one source of text (tweets) in response to another (event). A key distinct difference in our work is that, our method considers the event and the associated tweets to be *heterogenous*: the topics in a tweet may be sampled from different types of topic mixtures (general or specific). Additionally, the topic mixtures in an event evolve over its timeline. Besides, previous work only focuses on better understanding of events, or isolated analysis of tweets. Thus, they do not provide insights into how to extract sense from the tweets around the events. As a result, another distinct difference in our work is that we provide a comprehensive and in-depth analysis of the event and its associated tweets based on its topical influences (Hu *et al.*, 2011, 2012b,c).

3.2 Modeling Topical Influences

The observations mentioned above highlight the importance of developing models that can characterize the crowd's involvement with the event. Since such involvement (tweeting) is topically influenced by the event which itself is topically evolving, we model this complexity by a hierarchical Bayesian model based on Latent Dirichlet Allocation (LDA). With this model, the topic modeling of the event/tweets and event segmentation can be achieved concurrently. In next sections, we first introduce a conceptualized view of our model. We then describe the mathematical representations of the model in detail.

3.2.1 Conceptual Model

Our proposed model is called the joint Event and Tweets LDA (ET-LDA), which generalizes LDA (Blei *et al.*, 2003) by jointly modeling the topic segmentation of an event and two distinct types of topics within associated tweets. The conceptual model of ET-LDA is shown in Figure 3.1. ET-LDA assumes that: (1) An event is formed by discrete sequentially-ordered segments, each of which discusses a particular set of topics. A segment consists of one or many coherent paragraphs available from the transcript of the event [1] . Creating these segments follows a generative process in ET-LDA: First, we treat each paragraph in an event as being generated from a particular distribution of topics, where each topic is a probability distribution over a vocabulary. Next, we apply the Markov assumption on the distribution over topics covered in the event: with some probability, the topic distribution for paragraph s is the same as the previous paragraph $s-1$; otherwise, a new distribution is sampled over topics for s. This pattern of dependency is produced by associating a binary variable with each paragraph, indicating whether its topic is the same as that of the previous paragraph or different. If the topic remains the same, these paragraphs are merged to form one segment.

Furthermore, ET-LDA assumes that: (2) A tweet consists of words which can belong to two distinct types of topics: *general* topics, which are high-level and constant across the entire event, and *specific* topics, which are concrete and relate to specific segments of the event. A tweet in which most words belong to general topics is defined as a *steady* tweet, indicating a weak topical influence from the event, whereas a tweet with more words from specific topics is defined as an *episodic* tweet, indicating

[1] For many public televised events, transcripts are readily published by news services like the New York Times, etc. Paragraph outlines in the transcripts are usually determined through human interpretation and may not necessarily correspond to topic changes in the event.

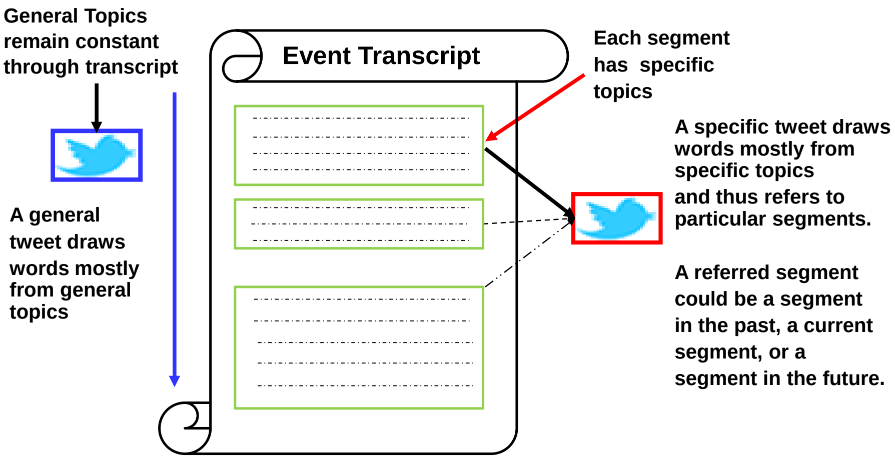

General Topics remain constant through transcript
A general tweet draws words mostly from general topics
Event Transcript
Each segment has specific topics
A specific tweet draws words mostly from specific topics and thus refers to particular segments.
A referred segment could be a segment in the past, a current segment, or a segment in the future.

Figure 3.1: Conceptual Model Of ET-LDA

a strong topical influence from a segment of the event. In other words, an episodic tweet *refers* to a segment of the event. Similar to the event segmentation, composing tweets also follows a generative process in ET-LDA. To begin with, we assume that the distribution of general topics is fixed for a tweet since it is a response tagged with the official hashtag of the event (hence it should be related to the event). On the contrary, the distribution of specific topics keeps varying with respect to the evolution of the event, because it is a more directed and intended response. So, when a person wants to compose a tweet to comment on the on-going event, she has two choices on picking the appropriate words: with some probability, a word w is sampled from the mixture of general topics about the event, otherwise, it is sampled from the mixture of specific topics which occurs "locally" in the parts of the event that w refers to. The hypothesis behind the second case is that, the audience may be influenced by a set of topics that are covered by a particular part (i.e., a segment) of the event. As a result, when she picks a word to respond to that part of the event, its topic is likely to be among the topics that specifically appeared in that segment. For example, consider a tweet which was posted at the beginning of President Obama's Middle

East speech: "*Sec Clinton introducing President Obama on #Mideast #StateDept #MESpeech*". It can be viewed as a mixture of general topics "*Middle East*" that was shared across the entire tweets corpus (words: *"#Mideast" and "#MESpeech"*), and specific topic "*Foreign policy*", sensitive to the part of the event when the Secretary of State, Hillary Clinton was introducing President Obama (words: *"Sec"*, *"Clinton"* and *"#StateDept"*). Note that this specific topic only occurred in the tweets that were posted at beginning of the event. Similar to the segmentation of the event, the preference of specific topics versus general topics is controlled by a binary variable associated with each word of a tweet.

3.2.2 Graphical Model

The graphical model representation of ET-LDA is shown in Figure 3.2. It has two major components with each capturing one perspective of our targets: the left part captures the event's topics and their evolution (event segmentation), whereas the right part captures the associated tweets' topics and the crowd's tweeting behaviors. Both parts have the LDA-like model and are connected by the link which defines the topical influences from the event on its Twitter feeds. Table 3.1 lists the notation used in this model. We have the generative process in ET-LDA in Algorithm 1.

Mathematically, in the event part, we assume an event's transcript S consists of many paragraphs. Each paragraph s ($s \in S$) is associated with a particular distribution of topics $\theta^{(s)}$ which assigns each word in s of a topic z_s^i. Note that $\theta^{(s)}$ is a multinomial distribution over K topics, determined by a binary variable $c^{(s)}$ under the governance of a binomial distribution $\delta^{(s)}$. This distribution is then parameterized by a symmetric beta prior α_δ. To model the topic evolutions in the event, we apply the Markov assumption on $\theta^{(s)}$: when $c^{(s)} = 0$, $\theta^{(s)}$ is the same as the distribution of topics of previous paragraph $s-1$, i.e., $\theta^{(s)} = \theta^{(s-1)}$, measured by the delta

Table 3.1: Mathematical Notation in ET-LDA

Notation	Description
S	a set of paragraphs in the event's transcript
N_s	the number of words in paragraph s
T	a set of tweets associated with the event
M_t	the number of words in tweet t
$\theta^{(s)}$	topic mixture of the specific topics from a paragraph s of the event
$\psi^{(t)}$	topic mixture of the general topics from tweets corpus
$\delta^{(s)}$	parameter for choosing to draw topics in paragraph s from $\theta^{(s)}$ or $\theta^{(s-1)}$
$c^{(s)}$	indicates whether the topic of a paragraph is drawn from current or previous segment's topics.
$\lambda^{(t)}$	parameter for choosing to draw topics in t from θ or ψ
$c^{(t)}$	indicates whether the topic of a tweet is drawn from specific or general topics
$s^{(t)}$	a referred segment, to which a specific topic in a tweet is associated
w_s, w_t	words in event's transcript, tweets, respectively
z_s, z_t	topic assignments of words in event, tweets, respectively.
α, β	Dirichlet/beta parameters of the Multinomial/Bernoulli distributions

function $\delta(\theta^{(s-1)}, \theta^{(s)})$. As a result, s and its preceding paragraph $s-1$ are merged into a segment. Otherwise, when $c^{(s)} = 1$, $\theta^{(s)}$ is drawn for s from a Dirichlet prior with parameter α_θ for creating a new segment. Therefore, $c^{(s)}$ can be viewed as an segmentation indicator of whether a paragraph should be merged into the previous paragraph or not.

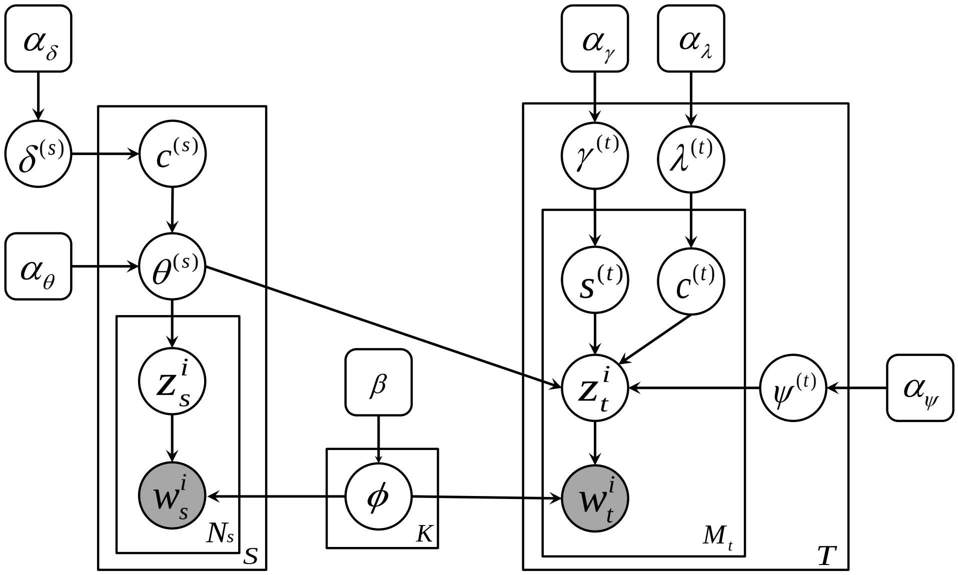


Figure 3.2: Graph Model of Et-LDA. $S(T)$ Is a Set of Paragraphs (Tweets). Z_s (Z_t) Is the Topic for Paragraph s (Tweet t), Which Can Be Drawn Either from Topic Mixture $\theta^{(s)}$ of the Event or Topic Mixture $\psi^{(t)}$ of the Tweets Corpus. Shaded Variables W_s^i and W_t^i Are the ith Word in s and t and Are Observed in the Dataset.

In the tweets part, the topic for a word in a tweet can be sampled from a mixture of specific topics $\theta^{(s)}$ or a mixture of general topics $\psi^{(t)}$ over K topics given a distribution $c^{(t)}$ defining the preference. $c^{(t)}$ is sampled from a binomial distribution $\lambda^{(t)}$. In the first case, $\theta^{(s)}$ is from a referring segment s of the event, where s is chosen according to a categorical distribution $s^{(t)}$. Although $c^{(t)}$ and $c^{(s)}$ share almost the same functionality, $c^{(t)}$ is controlled by an asymmetrical beta prior, which sets the preference parameter α_{λ_γ} (for specific topics) and α_{λ_ψ} (for general topics) accordingly. Besides, an important property of the categorical distribution $s^{(t)}$ is to allow choosing any segment. This reflects the fact that a person may compose a tweet on topics discussed in a segment that (1) was in the past (2) is currently occurring, or (3) will occur after the tweet is posted (usually when she expects certain topics to be discussed in the event). Last, ϕ is the word distribution over a vocabulary with corresponding parameter β.

3.3 Inference in the Model via Gibbs Sampling

The extract inference of the posterior distribution of the hidden variables $\mathbf{z}_s$, $\mathbf{z}_t$, $\mathbf{c}_s$, $\mathbf{c}_t$ and $\mathbf{s}_t$ in ET-LDA is intractable because of the coupling between the hyperparameters. Therefore, we utilize approximate methods like collapsed Gibbs sampling algorithm (Griffiths *et al.*, 2004) for parameter estimation. Note that Gibbs sampling allows the learning of a model by iteratively updating each latent variable given the remaining variables. More details of the Gibbs sampling rules and how we derive these rules can be found in Appendix A.

In specific, we want to estimate the posterior distribution of the following hidden variables: (**i**) $\mathbf{z}_s$, evaluated for each word in every paragraph s in the event transcript and then used to infer $\theta^{(s)}$; (**ii**) $\mathbf{z}_t$, evaluated for each word in each tweet t written by the Twitter users and then used the results to infer $\psi^{(t)}$; (**iii**) $\mathbf{c}_t$, evaluated for the topic types in tweet t; (**iv**) $\mathbf{s}_t$, evaluated for selecting segments from the event's transcript for t; and (**v**) $\mathbf{c}_s$, evaluated for each paragraph to indicate segmentation of the event.

Following the Gibbs sampling scheme, we begin with the joint distribution of all tweets and paragraphs of the event in ET-LDA as $P(\mathbf{z}_t, \mathbf{z}_s, \mathbf{w}_s, \mathbf{w}_t, \mathbf{c}_s, \mathbf{c}_t, \mathbf{s}_t | \mathbf{z}'_t, \mathbf{z}'_s, \mathbf{w}_s, \mathbf{w}_t, \mathbf{c}'_s, \mathbf{c}'_t, \mathbf{s}'_t)$, where $\mathbf{z}'_t, \mathbf{z}'_s, \mathbf{c}'_s, \mathbf{c}'_t, \mathbf{s}'_t$ are vectors of assignments of topics, segment indicators, topic

Algorithm 1: Generation Process in ET-LDA

```
foreach paragraph $s \in S$ do
    draw a segment choice indicator $c^{(s)} \sim Bernoulli(\delta^{(s)})$
    if $c^{(s)} = 1$ then
        draw a topic mixture $\theta^{(s)} \sim Dirichlet(\alpha_\theta)$
    else
        draw a topic mixture $\theta^{(s)} \sim \delta(\theta^{(s-1)}, \theta^{(s)})$
    foreach word $w_s^i \in s$ do
        draw a topic $z_s^i \sim Multinomial(\theta^{(s)})$
        draw a word $w_s^i \sim \phi_{z_s^i}$
foreach tweet $t \in T$ do
    foreach word $w_t^i \in t$ do
        draw a topic changing indicator $c^{(t)} \sim Bernoulli(\lambda^{(t)})$
        if $c^{(t)} = 1$ then
            draw a topic mixture $\psi^{(t)} \sim Dirichlet(\alpha_\psi)$
            draw a general topic $z_t^i \sim Multinomial(\psi^{(t)})$
        else
            draw a paragraph $s \sim Categorical(\gamma^{(t)})$
            draw a specific topic $z_t^i \sim Multinomial(\theta^{(s)})$
        draw a word from its associated topic $w_t^i \sim \phi_{z_t^i}$
```

switching indicators and segment choice indicators for all words in the collection except for the one at position i in a tweet or in a paragraph of the event's transcript. Then, by using the chain rule and integrating out the parameters $\phi, \gamma^{(t)}, \theta^{(s)}, \psi^{(t)}, \delta^{(s)}, \lambda^{(t)}$ (because the model only uses conjugate priors (Bishop *et al.*, 2006)), we can obtain the posterior probability of aforementioned hidden variables. Below, we give a brief overview of how these probabilities are inferred. Please refer to Appendix A for detailed derivation.

For parameter estimation in (**i**), we have the estimation of the topic distribution $\mathbf{z}_s$ for each word in paragraph s as:

$$P(\mathbf{z}_{s,i}|\mathbf{z}_{-(s,i)}, \mathbf{z}_t, \mathbf{w}_s, \mathbf{w}_t, \mathbf{c}_s, \mathbf{c}_t, \mathbf{s}_t) = \frac{n^k_{sw} + n^k_{tw} + \beta - 1}{n^k_{s(.)} + n^k_{t(.)} + W\beta - 1} \times \frac{n^{\mathcal{S}_i}_k + nt^{\mathcal{S}_i}_k + \alpha_\theta - 1}{n^{\mathcal{S}_i}_{(.)} + nt^{\mathcal{S}_i}_{(.)} + K\alpha_\theta - 1} \tag{3.1}$$

where $\mathcal{S}$ is a set of segments of the event's transcript. Each element ($\mathcal{S}_i$) in this set contains multiple paragraphs. $n^{\mathcal{S}_i}_k$ is the number of times topic k assigned to words in segment $\mathcal{S}_i$. $nt^{\mathcal{S}_i}_k$ is the number of times topic k appears in tweets, where these tweets are direct response to the sentences in segment $\mathcal{S}_i$ (i.e., $c_t = 0$). W is the size of the vocabulary and K is the number of topics.

For parameter estimation in (**ii**), we consider a two-step Gibbs sampling since the topic distribution $\mathbf{z}_t$ for each word in tweet t have two cases based on whether $\mathbf{z}_t$ is a specific topic ($\mathbf{c}_t = 0$) or a general topic ($\mathbf{c}_t = 1$):

$$P(\mathbf{z}_{t,i}|\mathbf{z}_{-(t,i)}, \mathbf{z}_s, \mathbf{w}_s, \mathbf{w}_t, \mathbf{c}_s, \mathbf{c}_t, \mathbf{s}_t) = \begin{cases} \dfrac{n^k_{sw} + n^k_{tw} + \beta - 1}{n^k_{s(.)} + n^k_{t(.)} + W\beta - 1} \times \dfrac{n^{\mathcal{S}_i}_k + nt^{\mathcal{S}_i}_k + \alpha_\theta - 1}{n^{\mathcal{S}_i}_{(.)} + nt^{\mathcal{S}_i}_{(.)} + K\alpha_\theta - 1} & \mathbf{c}_t = 0 \qquad (3.2) \\ \dfrac{n^k_{sw} + n^k_{tw} + \beta - 1}{n^k_{s(.)} + n^k_{t(.)} + W\beta - 1} \times \dfrac{n^i_k + \alpha_\psi - 1}{n^i_{(.)} + K\alpha_\psi - 1} & \mathbf{c}_t = 1 \qquad (3.3) \end{cases}$$

where n_k^i is the number of times topic k appears in tweet t, where t is about a general topics sampled from $\psi^{(t)}$. $n_{(.)}^i = \sum_{k=1}^K n_k^i$ is the total number of times all topics $1...k$ appear in t.

Similarly, for parameter estimation in (**iii**), we also have two cases for the posterior distribution of the topic switching indicator $\mathbf{c}_t$:

$$P(\mathbf{c}_{(t,i)}|\mathbf{c}_{-(t,i)}, \mathbf{c}_s, \mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t, \mathbf{s}_t) = \begin{cases} \dfrac{M_t^0 + \alpha_{\lambda_\lambda} - 1}{M_t + \alpha_{\lambda_\gamma} + \alpha_{\lambda_\psi} - 1} \times \dfrac{n_k^{\mathcal{S}_i} + nt_k^{\mathcal{S}_i} + \alpha_\theta - 1}{n_{(.)}^{\mathcal{S}_i} + nt_{(.)}^{\mathcal{S}_i} + K\alpha_\theta - 1} & \mathbf{c}_t = 0 \quad (3.4) \\ \dfrac{M_t^1 + \alpha_{\lambda_\psi} - 1}{M_t + \alpha_{\lambda_\gamma} + \alpha_{\lambda_\psi} - 1} \times \dfrac{n_k^i + \alpha_\psi - 1}{n_{(.)}^i + K\alpha_\psi - 1} & \mathbf{c}_t = 1 \quad (3.5) \end{cases}$$

where M_t^0 is the number of words in tweet t whose topics are specific topics. On the other hand, M_t^1 is the number of words in tweet t whose topics are general topics. $M_t = M_t^0 + M_t^1$ is the number of words in t.

Next, for parameter estimation of segment choice indicator $\mathbf{s}_t$ in (**iv**), we have:

$$P(\mathbf{s}_t|\mathbf{s}_{-(t,i)}, \mathbf{c}_t, \mathbf{c}_s, \mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t) = \frac{n_k^{\mathcal{S}_i} + nt_k^{\mathcal{S}_i} + \alpha_\theta - 1}{n_{(.)}^{\mathcal{S}_i} + nt_{(.)}^{\mathcal{S}_i} + K\alpha_\theta - 1} \times \frac{n_s^i + \alpha_\gamma - 1}{n_{(.)}^i + S\alpha_\gamma - 1} \tag{3.6}$$

where S is the set of segments in the event's transcript. $s \in S$ is the segment that tweet t refers to. n_s^i is the number of times segment s is refereed by words in t.

Last, we present Gibbs update rules for the estimation of segmentation indicator $\mathbf{c}_s$ in ($\mathbf{v}$). Since this variable is sampled from a Binomial distribution, it has two possible values that control whether a paragraph s should have the same topic distribution as its preceding paragraph $s-1$ (when $\mathbf{c}_s = 0$) or have a new topic distribution sampled from a Multinomial (when $\mathbf{c}_s = 1$). Thus, we have the following update rules:

$$P(\mathbf{c}_s|\mathbf{c}_{-(s,i)}, \mathbf{c}_t, \mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t, \mathbf{s}_t) \propto \begin{cases} \frac{S_t^0 + \alpha_\delta - 1}{S + 2\alpha_\delta - 1} \times \frac{\prod_{k=1}^K \Gamma(n_k^{\mathcal{S}_i} + nt_k^{\mathcal{S}_i} + \alpha_\theta)}{\Gamma(n_{(.)}^{\mathcal{S}_i} + nt_{(.)}^{\mathcal{S}_i} + K\alpha_\theta)}, \quad \mathbf{c}_s = 0 & (3.7) \\ \frac{S_t^1 + \alpha_\delta - 1}{S + 2\alpha_\delta - 1} \times \frac{\Gamma(K\alpha_\theta)}{\Gamma(\theta)^K} \times \frac{\prod_{k=1}^K \Gamma(n_k^{\mathcal{S}_{(s-1)}} + nt_k^{\mathcal{S}_{(s-1)}} + \alpha_\theta)}{\Gamma(n_{(.)}^{\mathcal{S}_{(s-1)}} + nt_{(.)}^{\mathcal{S}_{(s-1)}} + K\alpha_\theta)} \\ \times \frac{\prod_{k=1}^K \Gamma(n_k^{\mathcal{S}_{(s)}} + nt_k^{\mathcal{S}_{(s)}} + \alpha_\theta)}{\Gamma(n_{(.)}^{\mathcal{S}_{(s)}} + nt_{(.)}^{\mathcal{S}_{(s)}} + K\alpha_\theta)}, \quad \mathbf{c}_s = 1 & (3.8) \end{cases}$$

where S_s^1 is the number of times paragraph's topic changes (i.e., $\mathbf{c}_s = 1$). $S = S_s^0 + S_s^1$ is the number of paragraphs in an event's transcript. Note that excluding $\mathbf{c}_{(s,i)} = 1$ from the sequence $\mathbf{c}_s$ makes a segment $\mathcal{S}_i$ split into two small segments: $\mathcal{S}_{(i-1)}$, which contains paragraphs from the beginning of original segment $\mathcal{S}_i$ to the one (i.e., $s-1$) right before s; and $\mathcal{S}_i$ which contains sentences from s to the end of $\mathcal{S}_i$. In contrast, excluding $\mathbf{c}_{(s,i)} = 0$ from the sequence will only cause the segment $\mathcal{S}_{(i)}$ missing the count for paragraph s.

During the parameter estimation, our inference algorithm keeps track of a $W \times K$ (words by topics) count matrix for both event and tweets, a $S \times 2$ (paragraphs by coin) count matrix for indicating segments in the event, a $T \times 2$ (tweets by coin) count matrix for indicating topic switch in tweets, and a vector of T length for keeping the choice of segment. Given these matrices, we can estimate the probabilities of $\theta^{(s)}$, $\psi^{(s)}$, $\delta^{(s)}$, $\lambda^{(t)}$, and $\gamma^{(t)}$.

3.4 Experiments

In this section, we present the performance of ET-LDA model based on quantitative studies and subjective evaluations. Our study is mainly comprised of two parts. In the first part, we evaluate the effectiveness of ET-LDA against other baselines. Four main tasks are undertaken: (**1**) the topics extracted from the whole corpus (tweets and transcripts of events) are compared with those separately extracted from LDA model, (**2**) the alignment between events and their associated tweets is compared with those separately aligned by the LDA model, (**3**) the capability of predicting topical influences of the events on unseen tweets in the test set is compared with LDA, and (**4**) the quality of event segmentation is compared with LCSeg – a popular HMM-based segmenting tool in the literature (Galley *et al.*, 2003).

In the second part, we apply ET-LDA to gain insights into two public televised events and their associated tweets. We first present our conjectures about Twitter users' posting behavior in responding to the event. Then we show more insights about the tweeting behavior in terms of how people responded to the event over its timeline, and how the segments of the events were referred to by the episodic tweets. As a result, we present (**5**) the evolution of episodic tweets over the event's timeline, and (**6**) the distribution of segments of the events as they were referred to by the episodic tweets. Through these results, we confirm our conjectures in a discussion section.

3.4.1 Experimental Setup

3.4.1.1 Data collection

To perform the experiments, we crawled tweets for two events using the Twitter API. The first event is President Obama's speech on the Middle East, where we obtained the tweets tagged with "*#MESpeech*". The second is the Republican Primary debate, where the tweets were tagged with "*#ReaganDebate*". Note that we only consider tweets with these hashtags, officially posted by the White House and NBC News, respectively, before the events. We obtained the transcripts of both events from the New York Times [2] [3] . We preprocessed both datasets and the transcripts by removing non-English tweets, retweets, punctuation and stopwords and stemming all terms. Table 3.2 summarizes the properties of these datasets after preprocessing. We use the hashtags to refer to these events in the rest of this work.

Table 3.2: Properties of Datasets Used in Our Experiments

Events	**MESpeech**	**ReaganDebate**
Event Air Time	05/19/2011 12:14PM-1:10PM	09/07/2011 8:00PM-10:00PM
Time span of tweets	05/18 - 05/25	09/06 - 09/13
Total #Tweets	11,988	112,414
#Tweets before event	1,916	42,561
#Tweets during event	4,629	46,672
#Tweets after event	5,443	23,181

3.4.1.2 Expanding tweets

It is known that topic modeling methods behave badly when applied to short documents such as tweets. To remedy this, we need to expand the tweets in some way to augment their context. Current efforts include using Wikipedia to enrich tweets (Hu *et al.*, 2009), grouping tweets by same authors (Zhao *et al.*, 2011), etc. Inspired by (Sahami and Heilman, 2006), our approach here treats tweet t as a query and sends it to a search engine. After generating a set of top-n query snippets $d_1, ...d_n$, we compute the TF-IDF term vector v_i for each d_i. Finally, we pick the top-m terms

[2] http://www.nytimes.com/2011/05/20/world/middleeast/20prexy-text.html

[3] http://www.nytimes.com/2011/09/08/us/politics/08republican-debate-text.html

from v_i and concatenate them to t to form an expanded tweet. In the experiments, we used the Google search engine [4] for retrieving snippets and set $n = 5$ and $m = 10$.

3.4.1.3 Model settings

We used the Gibbs sampling algorithm for training ET-LDA on the tweets datasets with the transcript. The sampler was run for 2000 iterations for both datasets. Coarse parameter tuning for the prior distributions was performed. We varied the number of topics K in ET-LDA and chose the one which maximizes the log-likehood $P(\mathbf{w}_s, \mathbf{w}_t|K)$, a standard approach in Bayesian statistics (Griffiths and Steyvers, 2004). As a result, we set $K = 20$. In addition, we set model hyperparameter $\alpha_\delta = 0.1, \alpha_\theta = 0.1, \alpha_\gamma = 0.1, \alpha_{\lambda_\gamma} = \alpha_{\lambda_\psi} = 0.5, \alpha_\psi = 0.1$, and $\beta = 0.01$.

3.4.2 Effectiveness Evaluation of the ET-LDA

In following four sections, we show the evaluations of (1) extracted topics, (2) alignment, (3) model prediction, (4) event segmention by ET-LDA against several baseline models.

3.4.2.1 Topics from the ET-LDA Model

We first study the topics discovered from the two datasets by our proposed model ET-LDA. Table 3.3 and Table 3.5 present the highest probability words from two distinct types of topics – in the rest of this work, we refer to them as **top words**. The topics extracted by baseline LDA model (which was trained on the event transcripts and tweet datasets separately with $K = 20$) are presented in Table 3.4 and Table 3.6. For the specific topics (under the column **Specific**), we directly pick the top 2 from the distribution of topics for each segment of the event. The topics that are

[4] http://www.google.com/cse/

ubiquitously and consistently mentioned in the entire tweets dataset are considered as the general topics (under the column **General**) because their distributions are fixed for the event (recall Section 3.2). Note that all of the topics have been manually labeled for identification (e.g. "*Arab Spring*") to reflect our interpretation of their meaning from their top words. These settings are also applied to the results by LDA.

For MESpeech (see Table 3.3), all specific topics in 7 segments seem to correlate well with the event from a reading of the transcript. Furthermore, it is clear that these topics are sensitive to the event's context and keep evolving as the event progresses (in the sense that topics from most segments are different). The only exceptions are "*Human rights*" and "*Foreign policy*", which occur in two segments (S1 and S7). This can be explained by the fact that these two segments serve as the opening and ending of the event. Usually, the content of these two parts tends to be similar since they are either related to the outline or the summarization of the event. On the other hand, general topics and their top words capture the overall themes of the event well. But unlike specific topics, these general topics are repeatedly used across the entire event by the crowd, in particular when expressing their views on the themes of the event (e.g., "*Arab spring*") or even on some popular issues that are neither related nor discussed in the event (e.g., "*Obama*").

For ReaganDebate (see Table 3.5), we show a sample of 7 (out of 14) segments. All specific topics and their top words from these segments look well related to the event. However, compared to MESpeech where the specific topics were discussed in sequence (except for segments (S1 and S7) which we discussed above), we discover that here the specific topics are rather disordered and occur repeatedly. For example, "*Healthcare*" is mentioned in both segments S3 and S10, and "*Immigration*" is mentioned in segments S6 and S13, etc. This interesting observation is mainly due to the structure of the debate. Note that ReaganDebate is a multi-way conversation. Although it was

Table 3.3: Top Words from Topics of Mespeech. Top 2 Specific Topics per Segment (S1–s7). Top 3 General Topics from the Entire Tweet Corpus.

S	Specific	Top Words
S1	Human rights	Rights Transition People Power
	Foreign policy	Secure Mideast Arab Clinton State
S2	Terrorism	Bin Laden Mass Murderer Cry
	People	Dignity Power Street Square people
S3	Arab democracy	Democracy Yemen Syrian Bahrain
	Egypt revolution	Mubarak Resign Policy Reform
S4	Youth	Promote Untapped Arab talent youth
	Free speech	Open Internet Mind Access Paper
S5	Economics	Aid Trade Debt Billion Dollar
	Reform	Egypt Reform Support Resource
S6	Border problem	Israel Palestine Borders State Jordan
	Peace treaty	Palestine Peace Jewish Agreed treaty
S7	Human rights	Rights Transition People Power
	Foreign policy	Secure Mideast Arab Clinton State

General	Top Words
Arab spring	Arabia Bahrain Iran Mosques Syrian Leader Government Stepped
Israel & Palestine	Israel Palestine Borders Lines Hamas Negotiate Permanent Occupation
Obama	President Job Tough Critique Jews Policies Attacking Weakness

led by two anchors, sometimes a presidential candidate also expounded his claims and attacked the other candidates' records on some topics, resulting in rebuttals among

Table 3.4: Top Words from Topics of Mespeech by the Baseline Lda Model (3 out of a Total of 20 Topics Are Shown Here)

	Topics	**Top words**
LDA on event	“MiddleEast/Arab”	Young People Deny Country Democracy Women Violent Cairo
	“Security/Terrorism”	Many Home Transition State Security Conflict Blood Al Qaeda
	“Israel Palestine issues”	Palestinian Israel Know Between Leader resolve Issue Boarder
LDA on tweets	“Obama”	Wonderful Obama Job Approval GOP Talking Economics
	“Arab Spring”	Obama Town Assad Month Syria Libya Leave Must Jews
	“Security/Peace”	Iran Bin Laden Dead Oil War Murder Iraq Risk Nuclear

the candidates. Besides, the event partnered with an online medium (Politico.com) through which readers wrote down their questions to the candidates which were then selected by the anchors. Therefore, common concerns such as “*Healthcare*”, “*Economics*”, and “*Immigration*” were discussed back and forth heavily throughout the entire debate, producing many more segments than MESpeech (14 vs. 7) and the reoccurrence of the specific topics.

It is clear that all specific and general topics from the ET-LDA model are very reasonable from a reading of the transcripts. Furthermore, we observe that the specific topics are sensitive to the event’s context and keep evolving as the event progresses.

Table 3.5: Top Words from Topics of Reagandebate. Top 2 Specific Topics per Segment. Top 3 General Topics from the Entire Tweet Corpus.

Specific		**Top Words**
S1	Campaign	Tonight Republicans Campaign Leadership
	Candidates	Perry Romney Michele Huntsman Governor
S2	Job market	Job Payroll Market Crisis Monstrous
	Taxes	Income Tax Pledges Taxpayer Committed
S3	Healthcare	Obamacare Wrong Unconstitutional Deal
	Economics	Debt Fence Economics Commitment Cured
S6	Candidates	Perry Romney Michele Huntsman Governor
	Immigration	Legal Mexico Immigrants citizen Solution
S9	Debts	Government Financially Failure China
	Regulations	Fed Up Wrong Funding Expenditures
S10	Social Sec.	Social Security Benefits Ponzi Scheme
	Healthcare	Obamacare Wrong Unconstitutional Deal
S13	Immigration	Legal Mexico Immigrants citizens Solution
	Economics	Debt Fence Economics Commitment Cured

General	**Top Words**
Social security	Perry Social Security Ponzi scheme Check Constitutional Lowest Wage Vote Wrong
Economics	Private Sector Obama Conservative Budget Amendment Growth Employment Taxes Job
Health Care	Legislative Legal Solution Homelessness
	Obamacare Jeopardizes Medicare Doctor

Table 3.6: Top Words from Topics of Reagandebate by the Baseline LDA Model (3 out of a Total of 20 Topics Are Shown Here

	Topics	Top words
LDA on event	"Social Sec."	Constitution Law Government Wrong Federal Funding Monstrous Financially
	"Regulations"	Fed Funding Expenditures Devastating Economy Policies Hurt President
	"Health care"	Romney Cheaper Free Mandate Individual Obamacare Question Better
LDA on tweets	"Social Sec."	Perry Ridiculous Ponzi Exaggerated Provocative Tcot Investment Reformed
	"Obama"	Warfare Job Creation Obamacare Approval Poll Troop Withdraw Vote
	"Economics"	Monster Stupid Bloody Obama Jobs Lost Sided Down Blasting Hopeless

On the other hand, general topics and their top words capture the overall themes of the event well. But unlike specific topics, these are repeatedly used across the entire event by the crowd in their tweets, in particular when expressing their views on the themes of the event (e.g., "Arab spring", "Immigration") or even on some popular issues that are neither related nor explicitly discussed in the event (e.g., "Obama" in **MESpeech**, "Conservative" in **ReaganDebate**).

The results of LDA seem less reasonable by comparison. Although LDA may extract general topics like "Israel Palestine issues" just like ET-LDA, since these topics remain constant throughout the document, LDA cannot extract specific topics for the event. In fact, “Israel Palestine issues” shows the advantage of ET-LDA: it is the top general topic for entire tweet collection (which is very relevant to and influenced by the event) whereas LDA fails to identify that (its top topic is ‘Obama’ which is less relevant). The data showed that people tweeted about this issue a lot. Besides, some top words for LDA topics are not so related to the event. This lack of correspondence is more pronounced for LDA when it is applied to the tweet datasets, e.g., *GOP Job Approval* in topic “Obama” of the tweets corpus by LDA. This is mainly because ET-LDA successfully characterizes the topical influences from the event on its Twitter feeds such that the content/topics of tweets are regularized, whereas the LDA method ignores these influences and thus gives less reasonable results.

Since the evaluation of topics extracted by topic models often lacks of ground truth, we depend on a user study to further evaluate the quality of these topics. We recruited 31 participants who are graduate students from the engineering school of the first author’s university. As a part of our selection criteria, they were required to follow the news closely and tweet at least three times per week (same participants for other user study tasks in this work, which are described in next sections). Median age of participants was 26 years (range: 21-37 years). The procedure of our user study is the following: each participant was provided (**i**) 5 samples of segments per event (recall MESpeech has 7 and ReaganDebate has 14 segments), together with short summaries for both events. For the comparison of the quality of topics, participants were provided with top 3 specific topics per segment and top 5 general topics of the event extracted by ET-LDA and top 2 topics extracted from the segment (determined by ET-LDA in advance) by a traditional LDA model ($K = 20$). After each sample, the participant was asked to rate the quality of topics, on a Likert scale of 1 to 5 rating. The duration of the study was 10-20 minutes.

After that, we compare the performance of our proposed method against the

Table 3.7: Performance of Methods on the Quality of Topics (**T**) for Each Sampled Segment (S1-s5) Based on Likert Scale. The Higher Values Are Better.

MESpeech

		S1	S2	S3	S4	S5
T	ET-LDA	0.51	0.45	0.55	0.62	0.68
	LDA	0.43	0.41	0.47	0.44	0.51

ReaganDebate

		S1	S2	S3	S4	S5
T	ET-LDA	0.51	0.61	0.69	0.67	0.68
	LDA	0.48	0.51	0.52	0.54	0.57

baseline method (LDA) using the qualitative responses obtained in the user study. In Table 3.7, we show the measure of the Likert scale for the results of two methods, averaged over the value diversity. We observe that the best ratings are obtained by our proposed method ET-LDA (on an average 22% improvement over the baseline LDA method). We also study the statistical significance of ET-LDA with respect to LDA. We performed a paired t-test on the participant ratings obtained from the user study. We find that the comparisons of ET-LDA to LDA yield low p-values. For MESpeech, we get $p = 0.0163$ and for ReaganDebate we have $p = 0.0092$. This indicats that the improvement in the performance of ET-LDA is statistically significant (against significance level of 0.05).

3.4.2.2 Alignment of Event and Tweets

Next, we evaluate the performance of event-tweets alignment. Basically, we want to evaluate whether the episodic tweets (i.e., tweets that are strongly influenced by the events) are correctly identified for each segment. Determining whether a response is an episodic tweet depends on its associated preference parameter c(t). As defined in ET-LDA, a response is an episodic tweet only if the sampled probability $P(c^{(t)}) > 0.5$, meaning that the majority of its topics are specific topics, influenced by the content of the segment it refers to.

Due to the lack of ground truth, we again rely on the user study. We give each participant a different questionnaire, which contained 3 parts: (**i**) 5 samples of segments per event (recall MESpeech has 7 and ReaganDebate has 14 segments), together with short summaries for both events. (**ii**) 5 samples of episodic tweets of each segment. (**iii**) 5 samples of steady tweets to the event. All topics and segments were generated by ET-LDA during the training time and the ordering of the samples was randomized. For the comparison of the soundness of episodic tweets and the alignment, participants were provided with top 5 tweets per segment measured by the distance (Jensen-Shannon divergence) of their topics to the ones of the referred segment. The JS divergence was calculated as $D_{JS} = \frac{1}{2}D_{KL}(P||R) + \frac{1}{2}D_{KL}(Q||R)$, where $R = (\frac{1}{2}P + \frac{1}{2}Q)$, P is a mixture of topics for tweets and Q is a mixture of topics for the referred segment, both are found by the LDA. Note these tweets are neither episodic nor steady, they are only similar/relevant to the segment of the event. After each sample, the participant was asked to rate the soundness of the alignment as compared to the ones described in (**iii**), on a Likert scale of 1 to 5 rating.

In Table 3.8, we show the measure of the Likert scale for the results of two methods, averaged over the value diversity. Again, our proposed method ET-LDA improves the baseline LDA method on an average 18%-41%. Besides, the difference between the methods is more obvious in ReaganDebate rather than MESpeech, because the crowd was topically influenced by ReaganDebate more (describe later) and only our proposed model can characterize such a relationship (while LDA ignores such influences). Besides, we study the statistical significance of ET-LDA with respect to LDA. As expected, ET-LDA yield low p-values to LDA ($p = 0.0408$ for **MESpeech** and $p = 0.0291$ for **ReaganDebate**), indicating that the improvement in performance of ET-LDA is statistically significant (against significance level of 0.05), particularly for the quality of topics in ReaganDebate. This is in conformity with our observations that ET-LDA outperforms LDA more if there exists a strong influence from the event on the crowd's responses.

Table 3.8: Performance of Methods on the Soundness of Event-Tweet Alignment (**Al**) Based on Likert Scale. The Higher Values Are Better.

MESpeech

		S1	S2	S3	S4	S5
AL	ET-LDA	0.49	0.51	0.56	0.58	0.63
	LDA	0.48	0.49	0.54	0.51	0.57

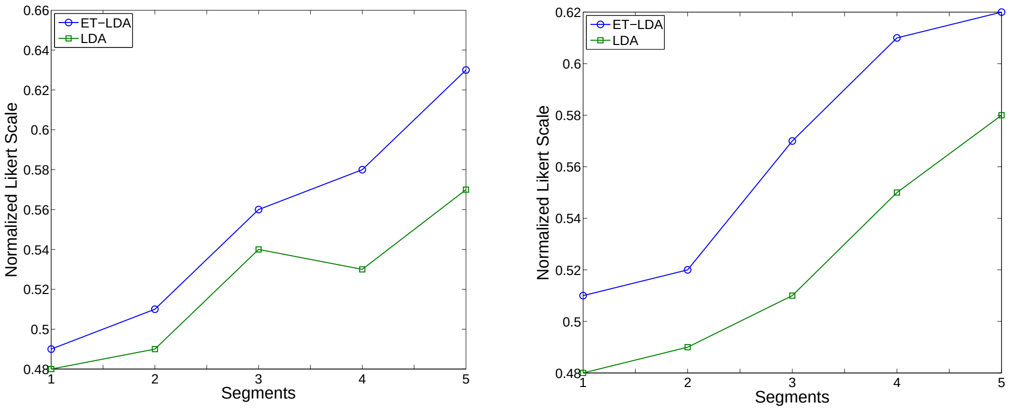
ReaganDebate

		S1	S2	S3	S4	S5
AL	ET-LDA	0.51	0.52	0.57	0.62	0.61
	LDA	0.48	0.49	0.51	0.51	0.58

3.4.2.3 Prediction Performance

Next, we study the prediction performance of ET-LDA. Specifically, we are interested in the prediction of topical influences from the event on the unseen tweets in our test set (20% of total tweets). Thus, we first run the Gibbs sampling algorithm, described in previous section, on the training set for each event/tweet dataset. Then we extend the sampler state with samples from the test set. For comparison, we used

LDA as our baseline approach. However, since LDA treats the event and tweets individually, we measure the topical influences by the distance of topic mixtures of the unseen tweets to the ones of the segments of the event (as determined by ET-LDA in advance). Similarly to evaluation of LDA in event-tweets alignment, we measure this distance by the Jensen-Shannon divergence.



(a) MESpeech

(b) ReaganDebate

Figure 3.3: Predictive Performance of ET-LDA Compared with LDA Model on 5 Randomly Sampled Segments.

To evaluate the "goodness" of prediction results by our proposed model, we again depend on users feedback. We asked our 31 participants to manually label the quality and strength of the predicted topical influences from events on the unseen tweet datasets on a Likert scale of 1 to 5 rating. We then averaged these ratings over the value diversity (i.e., normalization). In Fig. 3.3a and 3.3b, we present the results of the two methods on 5 randomly sampled segments.

In light of the observed differences in Fig. 3.3a and 3.3b, we study the statistical significance of ET-LDA with respect to LDA in terms of their predicting performance. We computed paired-t-tests for models with a significance level of $\alpha = 0.05$ and we obtained $p = 0.0161$ for **MESpeech** and $p = 0.0029$ for **ReaganDebate**. This reveals that the improvement in prediction performance of ET-LDA is statistically significant.

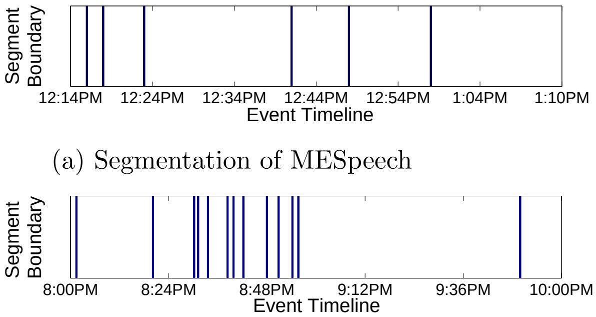
3.4.2.4 Event Segmentation

Finally, we study the quality and effectiveness of ET-LDA on the segmentation of the two events based on their transcripts. The results of the event segmentation (obtained using $K = 20$ in ET-LDA) are shown in Fig. 6.2b and 3.4b. To evaluate our model, we compare its results with the ones from a popular HMM-based tool *LCSeg* (trained on 15-state HMM) on the P_k measure (Beeferman *et al.*, 1999). Note that this measure is the probability that a randomly chosen pair of words from the event will be incorrectly separated by a hypothesized segment boundary. Therefore, the lower P_k indicates better agreement with the human-annotated segmentation results, i.e., better performance. In practice, we first ask four graduate students in our department to annotate the segments of the events based on their transcripts (two for each event) and later ask another graduate student to judge, for one event, which human annotation is better. We pick the better one of each event and treat it as the hypothesized segmentation. Then, we compute the P_k value. The results of two methods are shown in Table. 3.9.

Table 3.9: Comparisons of Segmentation Results on Two Events

	MESpeech		**ReaganDebate**	
	ET-LDA	LCSeg	ET-LDA	LCSeg
P_k	0.295	0.361	0.31	0.397

The results show that our model significantly outperforms the *LCSeg* – as the



(a) Segmentation of MESpeech

(b) Segmentation of ReaganDebate

Figure 3.4: Segmentation of the Event by ET-LDA

latter cannot merge topic mixtures in paragraphs according to their similarity, and thus places a lot of segmentation boundaries (i.e., over-segmented), resulting in poor performance.

3.5 Insights from the Application of ET-LDA: Sensemaking of Events and Tweets

Our previous evaluations prove the effectiveness of our proposed ET-LDA model in topics extraction, event-tweets alignment, prediction of topical influences, and event segmenting. Now, we apply ET-LDA to the two events and their associated tweets to make sense of them. In the next sections, we first propose several conjectures about the crowd's tweeting behavior in responding to an event based on our manual inspection on a small sample of data. Next, we present two key insights about the tweeting behavior gained from ET-LDA analysis. Finally, we confirm our conjunctures through these insights.

3.5.1 Understanding Tweeting Behavior

We first present a preliminary understanding of a crowd's response to an event they are interested in. As an example, Figure 3.5 shows how the crowd interacted over the timeline of the Republican Primary debate, namely, *before*, *during* and *after* the event. The total number of tweets we collected for this event was over 110,000.

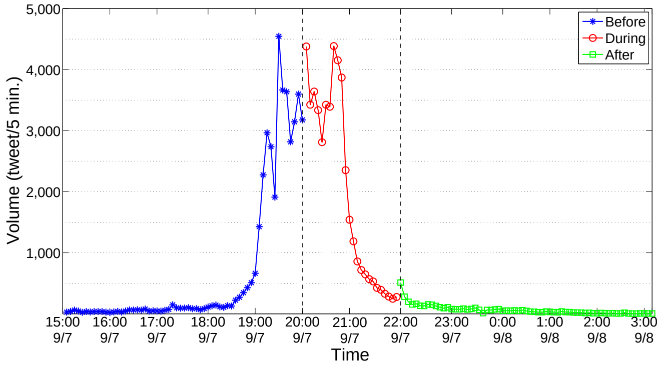


Figure 3.5: The Volume of Tweets (Number of Tweets Posted Within 5 Minutes Time Window) During 09/07/2011 15:00 – 09/08/2011 3:00. The Debate Was During 09/07/2011 20:00 – 22:00. All Tweets Were Tagged *#reagandebate*.

Based on the graph, we make three observations: (1) The swell of conversation occurred mostly within 1 hour right *before* the debate started, indicating that a large number of people began to pay attention to it then. Since the debate had not started yet, we conjecture their responses were mostly tangential (e.g., posted for presence) or commentaries about the general themes of the debate (which were known in advance). (2) The volume of tweets fluctuated *during* the debate, indicating different levels of involvement of the crowd with the evolving debate. We conjecture these changes were due to the fact that an event is made up of segments in sequence. Each segment covers a set of topics which may be uniquely interesting to the crowd and may influence their responses to be very specific to the content of the event. (3) A much smaller volume of tweets was witnessed right *after* the debate ended, indicating that most people quickly lost interest. We conjecture these tweets were of a different nature (e.g. slightly more specific to the content of the event) from the ones posted before the event, as the crowd had just listened to or experienced the event.

In addition to the above observations as reflected by the Twitter volume, we can further understand the crowd's responses from a different angle by analyzing their content. As mentioned earlier, this is nontrivial due to the vast amount of tweets. Hence, we first analyzed a small sample of tweets through manual inspection. We find that a tweet's content can be either weakly or strongly influenced by the debate's content. Tweets with weak correlations used words that were mostly about the general topics of the debate. So they seemed to be steady and less affected by the debate's progress. On the other hand, the words used in tweets with strong correlations were mostly related to specific topics, particularly influenced by the part of the debate that they responded to. Consequently, they seemed to be more episodic. Moreover, we find the pattern of steady versus episodic complies with the timeline of the debate. *Before* (and *after*) debate, most tweets were steady, while the episodic tweets were seen more frequently *during* the debate. According to these findings, our conjectures (which are mentioned earlier in this section) seem to be verified although the sample is limited.

3.5.2 Evolution of Episodic Tweets over the Event's Timeline

Next, to confirm these conjunctures, we study the crowd's responses that are strongly influenced by the event. Specifically, we are interested in how these responses evolve over the event's timeline. Determining whether a response is an episodic tweet depends on its associated preference parameter $c^{(t)}$. As defined in ET-LDA, a response is an episodic tweet only if the sampled probability $P(c^{(t)}) > 0.5$, meaning that the majority of its topics are specific topics, influenced by the content of the segment it refers to. Figure 3.6 and Figure 3.7 plot the percentage of those episodic tweets, split by 3 periods of the events. The tweets are presented in buckets, and the percentage of the episodic tweets refers to the proportion in a bucket. Note that the tweets in both figures were ordered by their time.

For MESpeech (see Figure 3.6), only 18% responses were episodic tweets initially, indicating that most responses at the time were either tangential or about the high-level themes of the event. This is because the responses (first 100 to 200 tweets) were contributed almost as early as 1 day before the event started. Then, a rapid increase of episodic tweets (from 18% to 39%) was witnessed just before the event, suggesting that people had gathered more information about it. We observe that interesting changes occur both when the event begins and as it is ending. In both cases, the percentage of episodic tweets rises up sharply (beginning: from 39% to 52%; ending: from 43% to 50%) and then drops down quickly. We believe this makes sense since people are often very excited when the event starts and ends. Under such circumstances, they tend to respond strongly to both parts. For example, a large number of the responses like "*Obama starts talking*", "*Here we go, Obama finally showed up*" were witnessed in response to the opening of MESpeech, and responses such as "*Obama's speech was finally over*" were seen mostly from the ending of the event. In fact, the beginning (the ending) part is usually determined by ET-LDA as the first (last) segment. More surprising to us was the fact that the percentage of episodic tweets remained mostly stable during the event. This might be because the most audience members had lower interest levels about specific topics about the Middle East, so their responses tended to be general even as the event was airing.

For ReaganDebate (see Figure 3.7), the graph for the percentage of episodic tweets shows a similar behavior to the one in MESpeech. However, we also discovered three key differences through the comparison. First, the responses are much more strongly influenced by the specific topics of the debate when compared to MESpeech, (33% vs. 18% in terms of the lowest percentage). We believe this is because ReaganDebate was about domestic issues that interested more people. Therefore, they tended to follow the debate closely and their responses were more episodic. Second and more

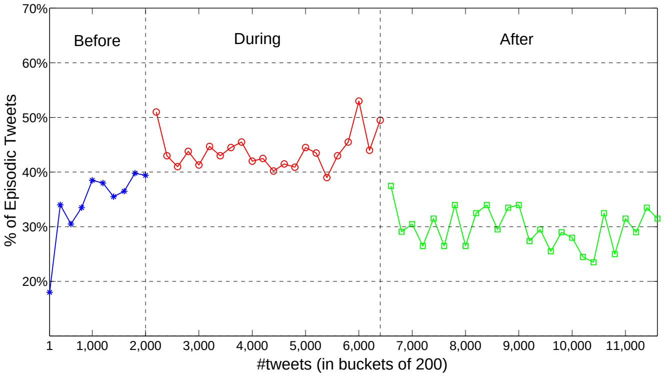


Figure 3.6: The Percentage of Episodic Tweets to Mespeech over Its Timeline. Tweets Were Ordered by Their Time.

interestingly, the crowd was less excited during the opening and ending of the debate. We attribute this to two reasons: (1) MESpeech was significantly delayed by 40 minutes. Therefore, responses were stronger when the event finally began, and (2) before ReaganDebate, there had been 4 Republican Primary debates already, so the crowd might have been less excited at the start. Lastly, we find the percentage of episodic tweets rises significantly during the debate (see the percentage rise around the 66,000th tweet). While looking through the content of the segments that these tweets referred to, we find topics like "*Healthcare*" and "*Economics*" were discussed. We conjecture that, since these topics are controversial and are a strong concern in the Primaries, the responses from the audience were pronounced.

3.5.3 Distribution of Segments Referred to by Episodic tweets

We now study how segments in the events were referred to by episodic tweets from the crowd. As defined in ET-LDA, an episodic tweet may refer to any segment of an event based on its associated categorical distribution governed by parameter $s^{(t)}$. We sample the highest probability segment from the distribution and deem it the *referred* segment. Figure 3.8 plots the results for both events, where each data point denotes

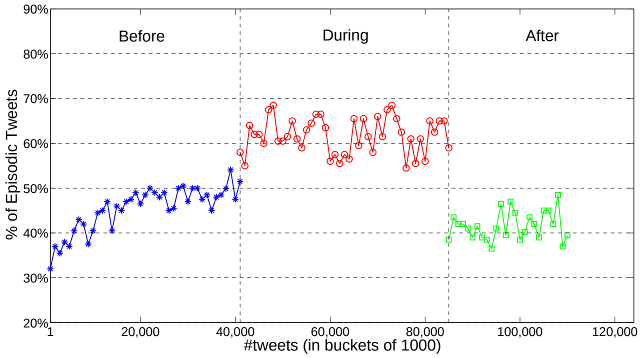
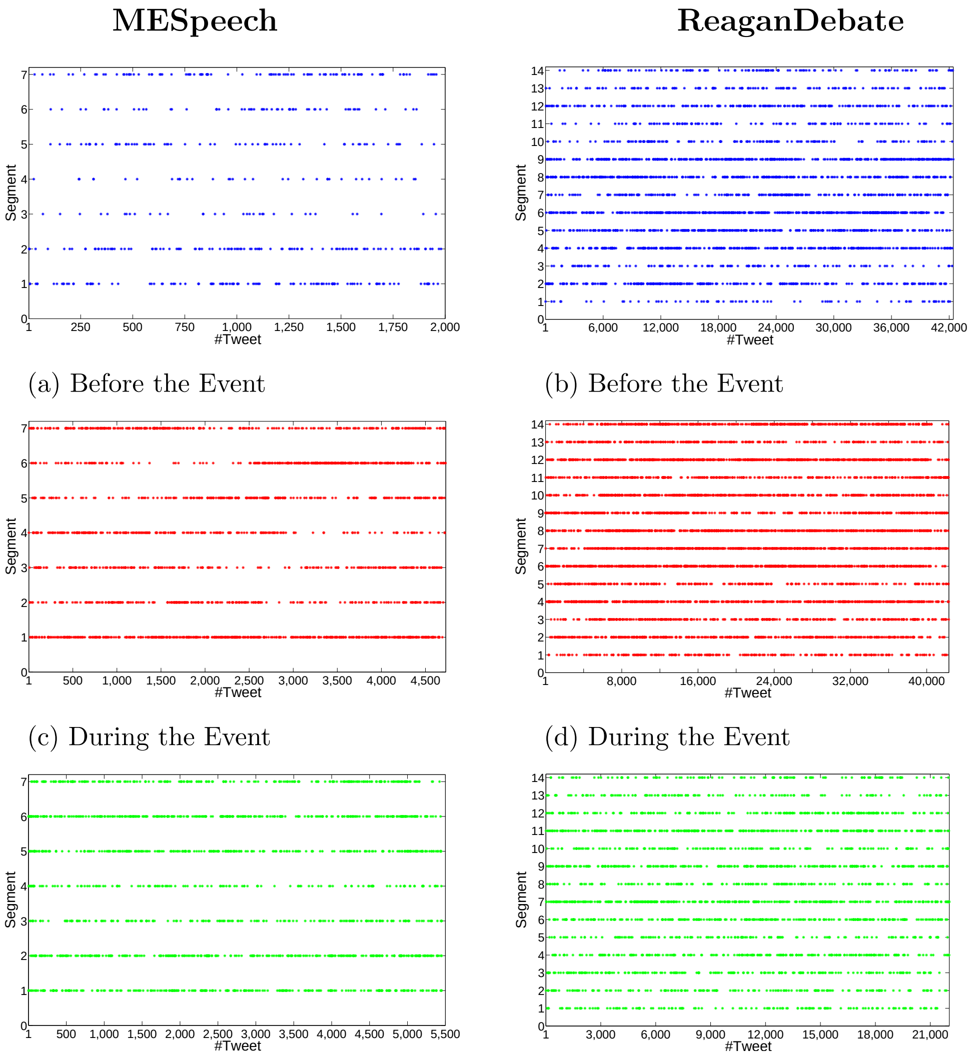


Figure 3.7: The Percentage of Episodic Tweets to Reagandebate over Its Timeline. Tweets Were Ordered by Their Time.

a tweet (which is an episodic tweet). Again, all tweets in both figures were ordered by their time.

For MESpeech, we first show how segments were referred to *during* the event. The results are shown in Figure 3.8c. As expected, we find the data points are quite dense for all segments, indicating that the crowd were following the event closely. Next, in Figure 3.8a we show how tweets talked about the event before it really started. Not surprisingly, the data points to all segments were pretty sparse. Among the segments, Segments 1 and 2 were referred to slightly more by the episodic tweets, since their focused topics (see Table 3.3) were mostly general (e.g., "*Human rights*") or popular (e.g., "*Terrorism*") so that people could respond specifically without knowing any content of the event. Based on the patterns of the data points in these figures, we make two *key* observations here: (1) Looking horizontally, we find that the crowd's attention tended to shift from one segment to the next as the event progressed. Our observation is based on the fact that the density of the data points of segments evolved over the event's timeline (see Segments 4-6 in Figure 3.8c). Initially, a segment starts out sparse since most people may still be focusing on other segments. Gradually, it becomes dense and stays dense (as more episodic tweets were contributed) during



(a) Before the Event

(b) Before the Event

(c) During the Event

(d) During the Event

(e) After the Event

(f) After the Event

Figure 3.8: The Distribution of Referred Segments by Episodic Tweets. Each Dot Presents a Tweet. All Tweets Were Ordered by Their Posted Time.

the time that the segment was occurring in the event. Afterwards, the density of the segment turns back to sparse because the audience may have lost interest in these topics. (2) More interestingly, when we look vertically in the graphs, we find the episodic tweets not only refer to the segments whose covered topics had been discussed before or were being discussed currently, but also refer to the segments whose topics are expected to be discussed later on in the event. We believe this is possible as long as the person has a high interest level, and expectation that these topics will be discussed. Lastly in Figure 3.8e, we see that the level of overall density of the segments lies between the ones in Figure 3.8a and Figure 3.8c. We believe this is because people had gained more information after the event (so they responded more specifically than before the event), but also they lost some interest in the event (so their responses were less specific than during the event).

For ReaganDebate, we observe two major differences from the results in MESpeech. First, there were significantly more episodic tweets regardless of the progress of the event (in Fig. 3.8b, 3.8d, 3.8f, the data points of every segment are much denser than the ones in Fig. 3.8a, 3.8c, 3.8e). Second, nearly all segments drew the crowd's attention (episodic tweets) consistently during and after the event as the segments are continuously dense, as opposed to the ones that have evolved over the timeline of MESpeech (graphically, every line has short periods of high density in Figure 3.8a, 3.8c, 3.8e). We attribute this to the fact that the crowd had a better background in domestic issues and was familiar with the topics covered in the event.

3.5.4 Summary of Insights

We now summarize the central findings of our case studies. The first finding is that the crowd's responses tended to be general and steady before the event and after the event, while during the event, they were more specific and episodic. Such findings confirm our conjectures in Section 3.5.1.

Secondly, the crowd showed different levels of engagement in different kinds of events. We attribute this to the reason that people may have greater interest levels about the general topics of certain events (e.g., topics in ReaganDebate). Our final finding is that the topical context of the tweets did not always correlate with the timeline of the event. We have seen that a segment in the event can be referred to by episodic tweets at any time irrespective of whether the segment has already occurred or is occurring currently or will occur later on. This finding is significant in light of the fact that current approaches such as (Shamma *et al.*, 2009) focus on correlating tweets to the event based on their timestamps, however our models enable a richer perspective.

3.6 Summary of Chapter

In this chapter, we have described a joint statistical model ET-LDA for aligning, analyzing and sensemaking of public events and their Twitter feeds. ET-LDA is developed based on the characterization of topical influences between an event and the tweets around it. Depending on such influences, tweets are labeled steady or episodic. Our model enables the topic modeling of the event/tweets and the segmentation of the event in one joint unified framework. We provided systematic evaluation of the effectiveness of ET-LDA against several baseline methods on two large sets of tweets in responses to two public events. Through both quantitative studies and subjective evaluations, our model showed significant improvements over the baseline methods. Furthermore, the application of ET-LDA on the two events also revealed interesting patterns on how users respond to events. Such patterns have not been investigated before as the needed analysis was not feasible with other alternative methods.

We believe ET-LDA presents the first step towards understanding complex interactions between events and social media feedback. In fact, beyond the transcripts of publicly televised events that we used in this chapter, ET-LDA can also handle other forms of text sources that describe an event. For example, by applying our model both to the news articles and the social media feedback about an event, we can not only explore how people respond to an event, but also how the social media responses differ from the journalists' responses in the mainstream media. We also believe that this chapter reveals a perspective that is useful for tools in event playback and the extraction of a variety of further dimensions such as sentiment and polarity. For example, one can examine how the crowd's mood is affected by the event based on the topical influences.

Chapter 4

EVENT CHARACTERIZATION VIA SENTIMENT ANALYSIS

In the previous chapter we introduced ET-LDA, a powerful Bayesian model to characterize events and their associated Twitter responses by analyzing the topics of the events and their evolutions and the crowd's event-related responding behaviors, based on event/tweet alignment and event segmentation. This chapter furthers these analyses by exploring other signals to make sense of events. In particular, we focus on the sentiment embedded in the crowd's event-related tweets.

It is known that given the ubiquity and immediacy of social media, individuals often express their opinions on Twitter and Facebook, in particular during live or breaking public events such as the U.S. Presidential debate and Apple products press conference. While viewers can see opinions one by one when watching, the collection of these posts provides an opportunity to understand the overall sentiment of people during the event. Gleaning insights from those posts is of increasing importance to many businesses. Recent studies have revealed that a massive number of people, news media, companies and political campaigns turn to social media to collect views about products and political candidates during and after the event. This guides their choices, decision-making, voting, and even stock market investments (Bollen *et al.*, 2011; Pak and Paroubek, 2010; Liu *et al.*, 2012; Jiang *et al.*, 2011).

In this chapter, we are interested in making sense of events by automatically characterizing segments and topics of that event in terms of the aggregate sentiments (positive [+] *v.s.* negative [-]) they elicited on Twitter (see Fig. 4.1). Classifying the sentiment behind textual content has received considerable attention during recent years. A standard approach would be to manually label comments (e.g., tweets) with

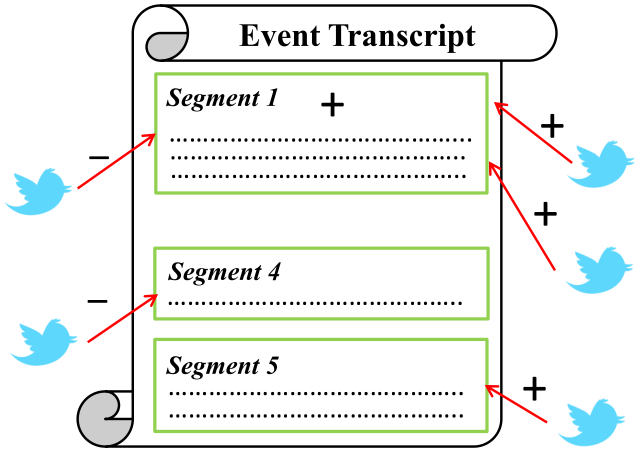


Figure 4.1: *Problem Setup*

their sentiment orientation and then apply off-the-shelf text classification techniques (Pang *et al.*, 2002). However, such a solution is inapplicable to our problem due to three reasons. First, manually annotating the sentiment of a vast amount of tweets is time consuming and error-prone, presenting a bottleneck in learning high quality models. Besides, sentiment is always conveyed with highly domain-specific contextual cues, and the idiosyncratic expressions in tweets may rapidly evolve over time, especially when tweets are posted live in response to the event. It can cause models to potentially lose performance and become stale. Last and most importantly, this approach is unable to relate aggregated Twitter sentiment to segments and topics of the event. One may consider enforcing tweets' correlation with the segment and topics from the event that occur within fixed time-windows around the tweets' timestamps (O'Connor *et al.*, 2010a) and classify the sentiment based on that. However, as pointed by the previous chapter and our recent work (Hu *et al.*, 2012a), this assumption is often not valid: a segment of the event can actually be referred to by tweets at any time irrespective of whether the segment has already occurred or is occurring currently or will occur later on.

The weaknesses discussed in the foregoing motivate the need for a fully automated framework to analyze events via aggregated twitter sentiment, with (1) little or no manual labeling of tweet sentiment, (2) ability to align tweets to the event, and (3) ability to handle the dynamics of tweets. While such a framework does not exist, the literature does provide partial solutions. For example, ET-LDA (Hu *et al.*, 2012c) provides an effective unsupervised framework for aligning tweets and events by jointly modeling both the tweets and events in a latent topic space. Similarly, while manual annotation of all tweets is infeasible, it is often possible to get sentiment labeling for small sets of tweets. Finally, there also exist domain-independent sentiment lexicons such as MPQA corpus (Wilson *et al.*, 2009).

We propose a flexible framework, named SocSent, for event analytics via Twitter sentiment that leverages these partial solutions. Specifically, our framework seeks low-rank representations of the Twitter sentiment and its correlations to the event by factorizing an input tweet-term matrix into four factors corresponding to tweets-segment, segment-topic, topic-sentiment and sentiment-words. The ET-LDA approach can be seen as providing the initial information ("prior knowledge") on the tweet-segment and segment-topic factors. Similarly, the availability of labeled tweets can be used to constrain the product of tweet-segment, segment-topic and topic-sentiment matrices. Finally, the sentiment lexicon is used to regulate the sentiment-words matrix. We pose this factorization as an optimization problem where, in addition to minimizing the reconstruction error, we also require that the factors respect the prior knowledge to the extent possible. We derive a set of multiplicative update rules that efficiently produce this factorization, and provide empirical comparisons with several competing methodologies on two real datasets, covering one recent U.S. presidential candidates debates in 2012 and one press conference. We examine the results both quantitatively and qualitatively to demonstrate that our method improves significantly over baseline approaches. In following sections, we describe SocSent in more details.

4.1 Related Work

Sentiment analysis has achieved great success in determining sentiment from underlaying text corpus like newspaper articles (Pang *et al.*, 2002) and product reviews (Hu and Liu, 2004). Various approaches, mostly learning-based, have been proposed, which include classification using sentiment lexicons (Wilson *et al.*, 2009), topic sentiment mixture model (Mei *et al.*, 2007), and nonnegative matrix factorization (Li *et al.*, 2009). Recently, there has been increasing interest in applying sentiment analysis to social media data like tweets such as (Bollen *et al.*, 2011; O'Connor *et al.*, 2010a). Some works also consider incorporating external social network information to improve the classification performance ((Tan *et al.*, 2011; Hu *et al.*, 2013a)).

Our work is also inspired by the research in characterizing events by the tweets around them. These works include inferring structures of events using Twitter usage patterns (Shamma *et al.*, 2009), exploring events by the classification of audience types on Twitter (Vieweg *et al.*, 2010), sentiment analysis of tweets to understand the events (Diakopoulos and Shamma, 2010b) and modeling the behavioral patterns between events and tweets (Hu *et al.*, 2012a).

The focus of the above work is mostly classifying sentiments of document sources or processing the tweets around the event. However, they do not provide insights into how to characterize the event's segments and topics through the aggregated Twitter sentiment, which is the main contribution of this work. Perhaps the closest work to us is (Diakopoulos and Shamma, 2010b). However, it depends on completely manual coding (via Amazon Mechanical Turk) to determine the sentiment. In contrast, we provide a fully automated and principled solution which can be used to handle the vast amount of tweets posted around an event.

4.2 SocSent Framework

In this section, we first present the basics of SocSent. We then describe how to obtain and leverage prior knowledge. Table 4.1 lists the notation. Note that although the primary sentiment we focus on is binary, our model can be easily extended to handle multiple types of sentiment.

Notation	Size	Description
$\mathbf{X}$	$n_t \times N$	Tweet-Term matrix
$\mathbf{G}$	$n_t \times n_s$	Tweet-Segment matrix
$\mathbf{T}$	$n_s \times K$	Segment-Topic matrix
$\mathbf{S}$	$K \times 2$	Topic-Sentiment matrix
$\mathbf{F}$	$N \times 2$	Term-Sentiment matrix
$\mathbf{G}_0$	$n_t \times n_s$	Prior knowledge on Tweet-Segment
$\mathbf{F}_0$	$N \times 2$	Prior knowledge on Term-Sentiment
$\mathbf{R}_0$	$n_t \times 2$	Prior knowledge on Tweet-Sentiment

Table 4.1: Notation in SocSent

4.2.1 Basic Framework

Let a public event be partitioned into n_s sequentially ordered segments, each of which discusses a particular set of topics. A segment consists of one or more coherent paragraphs available from the transcript of the event (I will discuss the segmentation in Section 4.2.2.3). There are also n_t tweets posted by the audience in response to the event, contributing to a vocabulary of N terms. As mentioned earlier, our goal is to identify segment and topics of the event that gained praise or criticism, according to how people reacted and appreciated them on Twitter. Accordingly, our basic framework takes those n_t tweets in terms of tweet-vocabulary matrix $\mathbf{X}$ as input and decomposes into four factors that specify soft membership of tweets and terms in three latent dimensions: segment, topic, and sentiment. In other words, our basic model tries to solve the following optimization problem:

$$\begin{aligned} \min_{\mathbf{G},\mathbf{S},\mathbf{F}} \quad & \left\|\mathbf{X} - \mathbf{GTSF}^\top\right\|_F^2 \\ s.t. \quad & \mathbf{G} \geqslant 0, \mathbf{T} \geqslant 0, \mathbf{S} \geqslant 0, \mathbf{F} \geqslant 0 \end{aligned} \tag{4.1}$$

where $\mathbf{G} \in \mathbb{R}^{n_t \times n_s}$ indicates the assignment of each tweet to the event segments based on the strength of their topic associations. That is, the I-th row of $\mathbf{G}$ corresponds to the posterior probability of tweet I referring to each of the n_s segments of the event. Similarly, $\mathbf{T} \in \mathbb{R}^{n_s \times K}$ indicates the posterior probability of a segment s belonging to the K topic clusters. Also, $\mathbf{S} \in \mathbb{R}^{K \times 2}$ encodes the sentiment distribution of each topic k. Finally, $\mathbf{F} \in \mathbb{R}^{N \times 2}$ represents the binary sentiment for each term in the vocabulary of tweets. Note that the non-negativity makes the factorized factors easy to interpret. As a result of this factorization, we can readily determine whether people appreciate the segments or topics of the event or dislike them. For example, from topic-sentiment matrix $\mathbf{S}$ I can directly obtain the crowd's opinion on each topic covered in the event. In addition, from segment-sentiment matrix $\mathbf{Q}$ (where $\mathbf{Q} = \mathbf{T} \times \mathbf{S}$), we can distill sentiment regarding each segment of the event. Finally, it is also feasible to characterize the sentiment for each tweet, through the new tweet-sentiment matrix $\mathbf{R}$ where $\mathbf{R} = \mathbf{G} \times \mathbf{T} \times \mathbf{S}$.

4.2.2 Constructing Prior Knowledge

So far, our basic matrix factorization framework provides potential solutions to infer the aggregated Twitter sentiment regarding the segment and topics of the event. However, it largely ignores a lot of prior knowledge on the learned factors. Previous literature (see (Pang *et al.*, 2002)) shows that leveraging such knowledge can help regulate the learning process and enhance the framework's performance (which is empirically verified in Section 4.3). Accordingly, we first show how to construct three types of prior knowledge: (a) sentiment lexicons of terms, (b) sentiment labels of tweets, and (c) alignment of tweets to the segment of the event. We then incorporate them into our framework in Section 4.2.3.

4.2.2.1 Sentiment Lexicon

Our first prior knowledge is from a sentiment lexicon, which is publicly available as a part of the MPQA corpus [1] . It contains 7,504 representative words that have been human-labeled as expressing positive or negative sentiment. In total, there are 2,721 positive (e.g., "*awesome*") and 4,783 negative (e.g., "*sad*") unique terms. It should be noted, that this list was constructed without any specific domain in mind; this is further motivation for using training examples and unlabeled data to learn domain specific connotations. To overcome the irregular English usage and out-of-vocabulary words in Twitter, we apply a lexicon normalization technique (Han and Baldwin, 2011) for the terms in our sentiment lexicon. This involves detecting ill-formed words and generates correction candidates based on morphophonemic similarity. As a result, "*happppppppppy*" is seen as a correct variant of "*happy*" thus sharing the same sentiment. We use those candidates to expand the original lexicon, making it adaptive

[1] http://mpqa.cs.pitt.edu/

to Twitter-related linguistic styles. Besides, we also add popular abbreviations and acronyms on Twitter such as "*smh*" (shake our head, negative) and "*lol*" (positive) to the lexicon. Eventually, we have 5,267 positive and 8,701 negative unique terms in the lexicon. We encode it in a term-sentiment matrix $\mathbf{F}_0$, where $\mathbf{F}_0(I, 1) = 1$ if the a word I has positive sentiment, and $\mathbf{F}_0(I, 2) = 1$ for negative sentiment.

4.2.2.2 Sentiment Label of Tweets

In addition to the lexicon, our second prior knowledge comes from human effort. We ask people to label the sentiment for a few tweets (e.g., less than 1000) for the purposes of capturing some domain-specific connotations, which later leads to a more domain-adapted model. The partial labels on documents can be described using a tweet-sentiment matrix $\mathbf{R}_0$ where $\mathbf{R}_0(I, 1) = 1$ if the tweet expresses positive sentiment, and $\mathbf{R}_0(I, 2) = 1$ for negative sentiment. One can use soft sentiment labeling for tweets, though our experiments are conducted with hard assignments.

4.2.2.3 Alignment of Tweets to the Event Segments

Our last prior knowledge focuses on the alignment between the event and the tweets which were posted in response to it. Like the sentiment label of tweets, this prior also tries making the model more domain-specific. To overcome the inherent drawbacks of the fixed time-window approach, we apply the ET-LDA model from our previous work (Hu *et al.*, 2012c). ET-LDA is a hierarchical Bayesian model based on Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003). It aims to model: (1) the event's topics and their evolution (event segmentation), as well as (2) the associated tweets' topics and the crowd's tweeting behaviors. The model has two major components with each capturing one perspective of the goals. Both parts have the LDA-like model, and are connected by the link which captures the topical influences from the event on its Twitter feeds. In practice, ET-LDA takes an event's transcript and all the event-related tweets and then concurrently partitions the speech into a number of homogeneous segments and aligns each tweet to event segments based on the strength of their topical associations. We encode the alignment results in a tweet-segment matrix $\mathbf{G}_0$ where its rows represent n_t tweets and its columns represent n_s segments of the event. As the content of $\mathbf{G}_0$ is the posterior probability of a tweet referring to the segments, we have $\sum_{1\leq j\leq n_s} \mathbf{G}_0(I,j) = 1$ for each tweet I.

4.2.3 Incorporating Prior Knowledge into the Framework

After defining and constructing the three types of prior knowledge, we can incorporate them into our basic factorization framework as supervision (see Eq. A.2). I later demonstrate in Section 4.3 that such supervision provides better regularization to the learned factors and significantly enhances the model's performance.

$$\begin{aligned}
\min_{\mathbf{F},\mathbf{T},\mathbf{G}} \quad & \mathcal{J} = \left\| \mathbf{X} - \mathbf{GTSF}^\top \right\|_F^2 \\
& + \alpha Tr\left((\mathbf{F}-\mathbf{F}_0)^\top \mathbf{\Lambda} (\mathbf{F}-\mathbf{F}_0) \right) \\
& + \beta Tr\left((\mathbf{GTS}-\mathbf{R}_0)^\top \mathbf{\Theta} (\mathbf{GTS}-\mathbf{R}_0) \right) \\
& + \gamma Tr\left((\mathbf{G}-\mathbf{G}_0)^\top \mathbf{\Gamma} (\mathbf{G}-\mathbf{G}_0) \right) \\
s.t. \quad & \mathbf{F} \geqslant 0, \mathbf{T} \geqslant 0, \mathbf{G} \geqslant 0, \mathbf{S} \geqslant 0 \qquad (4.2)
\end{aligned}$$

where $\alpha > 0$, $\beta > 0$, and $\gamma > 0$ are parameters which determine the extent to which I enforce $\mathbf{F} \approx \mathbf{F}_0$, $\mathbf{G} \approx \mathbf{G}_0$ and the multiplication $\mathbf{G} \times \mathbf{T} \times \mathbf{S} \approx \mathbf{R}_0$, respectively. $\mathbf{\Lambda} \in \mathbb{R}^{N\times N}$, $\mathbf{\Theta} \in \mathbb{R}^{n_t\times n_t}$ and $\mathbf{\Gamma} \in \mathbb{R}^{n_t\times n_t}$ are diagonal matrices, indicating the entries of $\mathbf{F}_0$, $\mathbf{G}_0$ and $\mathbf{R}_0$ that correspond to labeled entities. The squared loss terms ensure that the solution for $\mathbf{F}$, $\mathbf{G}$, $\mathbf{T}$, and $\mathbf{S}$, in the otherwise unsupervised learning problem, be close to the prior knowledge $\mathbf{F}_0$, $\mathbf{G}_0$ and $\mathbf{R}_0$.

It is worth noting the benefit of coupling $\mathbf{G}$, $\mathbf{T}$, and $\mathbf{S}$ in Eq. A.2. One may consider applying regularization to each of them. However, this will add additional computational cost during the model inference since the model gets more complex. In contrast, the supervision from $\mathbf{R_0}$ on the joint of $\mathbf{G}$, $\mathbf{T}$, and $\mathbf{S}$ can achieve the equivalent enforcement (while $\mathbf{G}$ is individually constrained). Besides, the above model is generic and it allows flexibility. For example, in some cases, our prior knowledge on $\mathbf{F}_0$ is not very accurate and I use smaller α so that the final results are not dependent on $\mathbf{F}$ very much. In addition, the introduction of $\mathbf{G}_0$ and $\mathbf{R}_0$ allows us to incorporate partial knowledge on tweet polarity and assignment information.

4.2.4 Model Inference

The coupling between $\mathbf{G}$, $\mathbf{T}$, $\mathbf{S}$, $\mathbf{F}$ makes it difficult to find optimal solutions for all factors simultaneously. In this work, we adopt an alternative optimization scheme (Ding *et al.*, 2006) for Eq. A.2, under which we update $\mathbf{G}$, $\mathbf{T}$, $\mathbf{S}$, $\mathbf{F}$ alternatingly. Blow, we show the inference process for obtaining the update rule for the tweets-segment matrix $\mathbf{G}$. Note that other factors can be inferred using the similar steps.

To infer the update rule for $\mathbf{G}$ in the framework, we first rewrite Eq. A.2 as:

$$\begin{aligned}\mathcal{J} =& Tr\left(\mathbf{X}^\top\mathbf{X} - 2\mathbf{X}^\top\mathbf{GTSF}^\top + \mathbf{FS}^\top\mathbf{T}^\top\mathbf{G}^\top\mathbf{GTSF}^\top\right)\\ &+ \alpha Tr\left(\mathbf{F}^\top\mathbf{\Lambda F} - 2\mathbf{F}^\top\mathbf{\Lambda F}_0 + \mathbf{F}_0^\top\mathbf{\Lambda F}_0\right)\\ &+ \beta Tr\left(\mathbf{S}^\top\mathbf{T}^\top\mathbf{G}^\top\mathbf{\Theta GTS} - 2\mathbf{S}^\top\mathbf{T}^\top\mathbf{G}^\top\mathbf{\Theta R}_0 + \mathbf{R}_0^\top\mathbf{\Theta R}_0\right)\\ &+ \gamma Tr\left(\mathbf{G}^\top\mathbf{\Gamma G} - 2\mathbf{G}^\top\mathbf{\Gamma G}_0 + \mathbf{G}_0^\top\mathbf{\Gamma G}_0\right) \qquad (4.3)\end{aligned}$$

Next, in order to infer the updating rule $\mathbf{G}$, we first construct the Lagrangian for $\mathbf{G}$ as:

$$
\begin{aligned}
\mathcal{L} &= \mathscr{J} - \Gamma\mathbf{G}^\top \qquad (4.4)\\
&= Tr\left(\mathbf{X}^\top\mathbf{X} - 2\mathbf{X}^\top\mathbf{GTSF}^\top + \mathbf{FS}^\top\mathbf{T}^\top\mathbf{G}^\top\mathbf{GTSF}^\top\right)\\
&\quad +\beta Tr\left(\mathbf{S}^\top\mathbf{T}^\top\mathbf{G}^\top\boldsymbol{\Theta}\mathbf{GTS} - 2\mathbf{S}^\top\mathbf{T}^\top\mathbf{G}^\top\boldsymbol{\Theta}\mathbf{R}_0 + \mathbf{R}_0^\top\boldsymbol{\Theta}\mathbf{R}_0\right)\\
&\quad +\gamma Tr\left(\mathbf{G}^\top\boldsymbol{\Gamma}\mathbf{G} - 2\mathbf{G}^\top\boldsymbol{\Gamma}\mathbf{G}_0 + \mathbf{G}_0^\top\boldsymbol{\Gamma}\mathbf{G}_0\right) - \Psi\mathbf{G}^\top + C \qquad (4.5)
\end{aligned}
$$

where the Lagrangian multipliers Ψ enforce the nonnegativity constraint on $\mathbf{G}_{i,j}$, and we use C to represent the terms that are irrelevant to $\mathbf{G}$. Then, we compute the first-order derivatives of $\mathcal{L}$ with respect to $\mathbf{G}$ as:

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathbf{G}} =& -2\left(\mathbf{XFS}^\top\mathbf{T}^\top + \beta\boldsymbol{\Theta}\mathbf{R}_0\mathbf{S}^\top\mathbf{T}^\top + \gamma\boldsymbol{\Gamma}\mathbf{G}_0\right)\\
&+2\left(\mathbf{GTSF}^\top\mathbf{FS}^\top\mathbf{T}^\top + \beta\boldsymbol{\Theta}\mathbf{GTSS}^\top\mathbf{T}^\top + \gamma\boldsymbol{\Gamma}\mathbf{G}\right) \qquad (4.6)
\end{aligned}
$$

From the complementary slackness condition Trefethen and Bau III (1997), we can obtain:

$$
\left(-\mathbf{XFS}^\top\mathbf{T}^\top - \beta\boldsymbol{\Theta}\mathbf{R}_0\mathbf{S}^\top\mathbf{T}^\top - \gamma\boldsymbol{\Gamma}\mathbf{G}_0 + \mathbf{GTSF}^\top\mathbf{FS}^\top\mathbf{T}^\top + \beta\boldsymbol{\Theta}\mathbf{GTSS}^\top\mathbf{T}^\top + \gamma\boldsymbol{\Gamma}\mathbf{G}\right)\mathbf{G} = 0 \qquad (4.7)
$$

This is the fixed point equation that the solution of $\mathbf{G}$ must satisfy at convergence. So we have the update rule for the tweets-segment matrix $\mathbf{G}$ as:

$$
G_{ij} \leftarrow G_{ij}\sqrt{\frac{\left[\mathbf{XFS}^\top\mathbf{T}^\top + \beta\boldsymbol{\Theta}\mathbf{R}_0\mathbf{S}^\top\mathbf{T}^\top + \gamma\boldsymbol{\Gamma}\mathbf{G}_0\right]_{ij}}{\left[\mathbf{GTSF}^\top\mathbf{FS}^\top\mathbf{T}^\top + \beta\boldsymbol{\Theta}\mathbf{GTSS}^\top\mathbf{T}^\top + \gamma\boldsymbol{\Gamma}\mathbf{G}\right]_{ij}}} \qquad (4.8)
$$

which is equivalent to

$$\left(-\mathbf{XFS}^{\top}\mathbf{T}^{\top} - \beta\boldsymbol{\Theta}\mathbf{R}_0\mathbf{S}^{\top}\mathbf{T}^{\top} - \gamma\boldsymbol{\Gamma}\mathbf{G}_0 + \mathbf{GTSF}^{\top}\mathbf{FS}^{\top}\mathbf{T}^{\top} + \beta\boldsymbol{\Theta}\mathbf{GTSS}^{\top}\mathbf{T}^{\top} + \gamma\boldsymbol{\Gamma}\mathbf{G}\right)\mathbf{G}^2 = 0 \tag{4.9}$$

Next, for the tweets-segment matrix $\mathbf{T}$, we have:

$$T_{ij} \leftarrow T_{ij}\sqrt{\frac{\left[\beta\mathbf{G}^{\top}\boldsymbol{\Theta}\mathbf{R}_0\mathbf{S}^{\top} + \mathbf{G}^{\top}\mathbf{XFS}^{\top}\right]_{ij}}{\left[\mathbf{G}^{\top}\mathbf{GTSF}^{\top}\mathbf{FS}^{\top} + \beta\mathbf{G}^{\top}\boldsymbol{\Theta}\mathbf{GTSS}^{\top}\right]_{ij}}} \tag{4.10}$$

In addition, for the tweets-segment matrix $\mathbf{S}$, we have:

$$S_{ij} \leftarrow S_{t_{ij}}\sqrt{\frac{[\mathbf{T}^{\top}\mathbf{G}^{\top}\mathbf{XF}]_{ij}}{[\mathbf{T}^{\top}\mathbf{G}^{\top}\mathbf{GTSF}^{\top}\mathbf{F} + \mathbf{T}^{\top}\mathbf{G}^{\top}\boldsymbol{\Theta}\mathbf{GTS}]_{ij}}} \tag{4.11}$$

Last, for the tweets-segment matrix $\mathbf{F}$, we have:

$$F_{ij} \leftarrow F_{ij}\sqrt{\frac{[\mathbf{X}^{\top}\mathbf{GTS} + \alpha\boldsymbol{\Lambda}\mathbf{F}_0]_{ij}}{[\mathbf{FS}^{\top}\mathbf{T}^{\top}\mathbf{G}^{\top}\mathbf{GTS} + \alpha\boldsymbol{\Lambda}\mathbf{F}]_{ij}}} \tag{4.12}$$

The learning algorithm consists of an iterative procedure using the above four rules until convergence. The outline of the specific steps is shown below.

Computational complexity The tweet-term matrix $\mathbf{X}$ is typically very sparse with $z \ll n_t \times N$ non-zero entries. Also, K and n_s are typically also much smaller than n_t and N. By using sparse matrix multiplications and avoiding dense intermediate matrices, the updates can be very efficiently and easily implemented. In particular, updating $\mathbf{G}$, $\mathbf{T}$, $\mathbf{S}$ and $\mathbf{F}$ each takes $O(C^2(n_t+n_s+N)+Cz)$ time per iteration which scales linearly with the dimensions and density of the data matrix. C is a constant. Empirically, the number of iterations before practical convergence is usually very small (less than 350). Thus, our approach can scale to large datasets.

Algorithm 2: Factorization with Prior Knowledge

```
input  : $\alpha$, $\beta$, $\gamma$
output: $\mathbf{G}$, $\mathbf{T}$, $\mathbf{S}$, $\mathbf{F}$
Initialize $\mathbf{G} \geqslant 0, \mathbf{T} \geqslant 0, \mathbf{S} \geqslant 0, \mathbf{F} \geqslant 0$
while Algorithm Not Converges do
    Update $\mathbf{G}$ with Eq.(A.4) while fixing $\mathbf{T}$,$\mathbf{S}$,$\mathbf{F}$
    Update $\mathbf{T}$ with Eq.(A.5) while fixing $\mathbf{G}$,$\mathbf{S}$,$\mathbf{F}$
    Update $\mathbf{S}$ with Eq.(A.7) while fixing $\mathbf{G}$,$\mathbf{T}$,$\mathbf{F}$
    Update $\mathbf{F}$ with Eq.(A.6) while fixing $\mathbf{G}$,$\mathbf{T}$,$\mathbf{S}$
```

4.3 Experiments

In this section, we examine the effectiveness of our proposed framework SocSent} against other baselines. Three sentiment classification tasks are undertaken on: 1) event segments, 2) event topics, and 3) tweets sentiment.

4.3.1 Datasets and Experimental Setup

I use two large scale tweet datasets associated with two events from different domains: (1) the first U.S. Presidential debate on Oct 3, 2012 and (2) President Obama's Middle East speech on May 19, 2011. The first tweet dataset consists of 181,568 tweets tagged with **DenverDebate** and the second dataset consists of 25,921 tweets tagged with **MEspeech**. Both datasets were crawled via the Twitter API using these two hashtags. In the rest of this chapter, we use the hashtags to refer to these events. We obtained the transcripts of both events from the New York Times, DenverDebate has 258 paragraphs and MESpeech has 73 paragraphs. Furthermore, we split both tweet datasets into a 80-20 training and test sets. For ET-LDA, we use the implementation from (Hu *et al.*, 2012c). Its parameters are set using the same procedure described in (Hu *et al.*, 2012c). Coarse parameter tuning for our framework SocSent was also performed. We varied α, β and γ and chose the combination which minimizes the reconstruction error in our training set. As a result, we set $\alpha = 2.8$, $\beta = 1.5$, $\gamma = 1.15$. All experimental results in this section are averaged over 20 independent runs.

Establishing Ground Truth: To quantitatively evaluate the performance of our framework, we need the ground truth of the sentiment for event segments, event topics and tweets. At first, we asked 14 graduate students in our school (but not affiliated with our project or group) to manually label the sentiment (I.e., positive or negative) of 1,500 randomly sampled tweets for each dataset. We then applied ET-LDA model to segment two events and establish the alignment between the labeled tweets and the event segments. So for each segment, we label its sentiment according to the majority aggregated Twitter sentiment that correlated to it. For example, if 60 out of 100 tweets that refer to segment S are positive, then S is considered to have received positive sentiment since people showed their appreciation for it. In addition, the top-5 topics of S (these top topics were also learned by ET-LDA) are also labeled as having positive sentiment. We aggregate this sentiment across all the event segments and assign the majority sentiment to each topic of the event. Finally, we obtained 35 segments of DenverDebate, where 20 segments were labeled as negative. Also, 62% labeled tweets and 12 out of 20 topics were negative. For MEspeech, we have 6 of 9 segments, 13 out of 20 topics, and 72% tweets marked as negative.

Baselines: To better understand the performance of SocSent, we implemented some competitive baseline approaches: 1) *LexRatio*: This method (Wilson *et al.*, 2009) counts the ratio of sentiment words from OpinionFinder subjectivity lexicon [2]

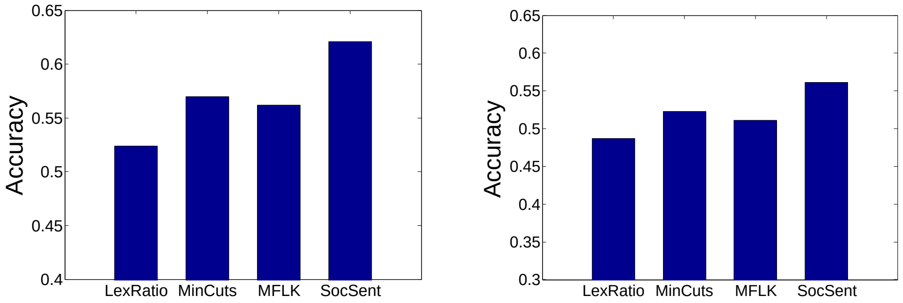
[2] http://mpqa.cs.pitt.edu/opinionfinder/

in a tweet to determine its sentiment orientation. Due to its unsupervised setting, we ignore the tweets which do not contain any sentiment words; 2) *MinCuts*: This method (Pang and Lee, 2004) utilizes contextual information via the minimum-cut framework to improve polarity-classification accuracy. 3) *MFLK*: This is a supervised matrix factorization method which decomposes an input term-document matrix into document-sentiment and sentiment-terms matrices. Supervision from a sentiment lexicon is enforced (Li *et al.*, 2009). I implemented it with our Twitter lexicon.

4.3.2 Classification Results

Classification of Sentiment of the Event Segment I first study the performance of SocSent on classifying the segments' sentiment for the two events via aggregated Twitter responses against the baseline methods. Note these baselines are inherently unable to relate their Twitter sentiment classification results to the event segments. To remedy this, we take a two-step approach. First, we split the whole event into several time windows (10-min. in our experiment). Then, we enforce the segments' sentiment to be correlated with the inferred tweet sentiment that occur within the time-windows around the tweets's timestamps. Figure 4.2 presents the classification results where accuracy is measured based on the manually labeled ground truth. It is clear that SocSentcan effectively utilize the partially available knowledge on tweet/event alignment from ET-LDA to improve the quality of sentiment classification in both events. In particular, it improves other approaches in the range of 7.3% to 18.8%.

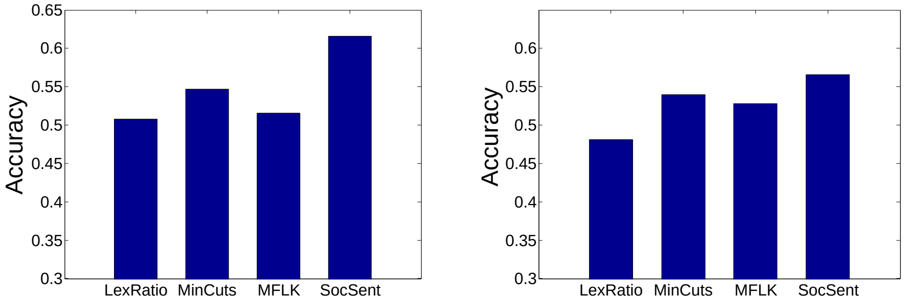
Classification of Sentiment of the Event Topics Next, we study sentiment classification of the topics covered in the event. The results are shown in Figure 4.3. Similar to the last task, we again use time window approach to correlate the event topics with the sentiment of tweets. Not surprisingly, SocSent improves the three



(a) DenverDebate (b) MEspeech

Figure 4.2: *Classification of Sentiment of the Event Segment.*

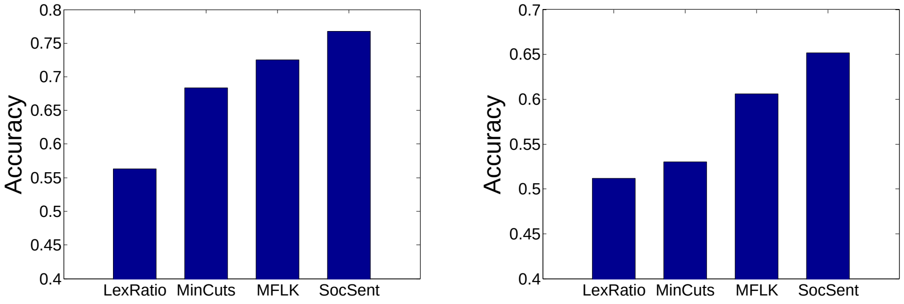
baselines with a range of 6.5% to 17.3% for both datasets.



(a) DenverDebate (b) MEspeech

Figure 4.3: *Classification of Sentiment of the Event Topics.*

Classifying the Sentiment of the Tweets In the third experiment, we evaluate the prediction accuracy of Twitter sentiment. Figure 4.4 illustrates the results for tweets posted in response to DenverDebate and MEspeech. As I can see, SocSent greatly outperforms other baselines on both datasets. In fact, it achieves the largest performance improvement margin (compared to results in Figure 4.2 and 4.3). I believe this is because SocSentadopts the direct supervision from the pre-labeled tweet sentiment. We also observe that all the methods have better performance on DenverDebate than on MEspeech (see Figure 4.2, 4.3 and 4.4). This is mainly because DenverDebate attracted a significant larger number of tweets than MEspeech. Therefore, the DenverDebate dataset is likely to be less sparse in the sense that more words from the sentiment lexicon can also be found in the training set. As a result, the effect of sentiment lexicon is fully utilized thus producing better results than on MEspeech.



(a) DenverDebate (b) MEspeech

Figure 4.4: *Classification of Sentiment of the Tweets.*

4.3.2.1 Varying Training Data Size

In Table 4.2, we show the performance of classifying segments' sentiment using various methods with respect to different size of training data. Note that LexRatio is an unsupervised approach so its performance is unchanged in this experiment. It is clear that the other three methods achieve better performance when more training data is supplied. Besides, on both `DenverDebate` and `MEspeech` datasets, we find that SocSent is more stable over other methods with various sizes of training data from 10% to 100%. In other words, SocSent does not show dramatic changes when the size of the training data changes. This demonstrates that our proposed method is robust to training data sizes.

4.3.2.2 Effectiveness of Prior Knowledge

Finally, given the available three types of prior knowledge – sentiment lexicon, tweet labels and tweet/event alignment by ET-LDA, it is interesting to explore their impact on the performance of SocSent. Table 4.3 presents the evaluation results on

Table 4.2: *Classification Accuracy on Segment's Sentiment Vs. Training Data Sizes. Notations:* ***LR*** *Is for LexRatio,* ***MC*** *Is for MinCuts,* ***MF*** *for MFLK, and* ***SS****, for Our Method* SocSent.

`DenverDebate`

	$T_{10\%}$(gain)	$T_{25\%}$(gain)	$T_{50\%}$(gain)	$T_{100\%}$(gain)
LR	0.524	0.524	0.524	0.524
MC	0.538 (+2.7%)	0.563 (+7.4%)	0.568 (+8.4%)	0.574 (+9.5%)
MF	0.532 (+1.5%)	0.536 (+2.3%)	0.558 (+6.5%)	0.562 (+7.3%)
SS	0.588 (+12.2%)	0.595 (+13.5%)	0.613 (+17.0%)	0.621 (+18.5%)

`MESpeech`

	$T_{10\%}$(gain)	$T_{25\%}$(gain)	$T_{50\%}$(gain)	$T_{100\%}$(gain)
LR	0.487	0.487	0.487	0.487
MC	0.502 (+3.1%)	0.520 (+6.8%)	0.521 (+6.9%)	0.523 (+7.4%)
MF	0.488 (0.2%)	0.504 (+3.5%)	0.509 (+4.5%)	0.511 (+4.9%)
SS	0.541 (+11.1%)	0.549 (+12.7%)	0.558 (+14.6%)	0.561 (+15.2%)

two datasets, where we judge SocSent on three aforementioned classification tasks with respect to different combinations of its prior knowledge. For each combination, we come up with separate update rules which have the similar form as Eq. A.4-Eq. A.7. Besides, we find optimal parameters using the same procedure described above in the setup of experiment. Several insights are gained here: First, using single type of prior knowledge is less effective than combining them. Especially, combining all three types of prior knowledge leads to the most significant improvement (an average of 29.8% gain over the baseline *N.A* on two datasets). Second, domain-specific knowledge (tweet labels, event/tweet alignment) is more effective than domain-independent knowledge (sentiment lexicon) in all three prediction tasks. Last, domain-specific knowledge is particulary helpful in its corresponding task. For example, having tweet/event alignment (denoted as $\mathbf{G}_0$ in Table 4.3) achieves more accurate results in classifying the sentiment of the event segments than without having it. For example, combinations with this prior knowledge such as $\mathbf{F}_0 + \mathbf{G}_0$ or $\mathbf{R}_0 + \mathbf{G}_0$ have better performance than $\mathbf{F}_0 + \mathbf{R}_0$ with 6.5% and 8.3% improvement, respectively. These insights demonstrate the advantage of SocSent's ability to seamlessly incorporate prior knowledge.

4.4 Summary of Chapter

In this chapter, we have described a flexible factorization framework, SocSent that characterizes the segment and topics of an event via aggregated Twitter sentiment. Our model leverages three types of prior knowledge: sentiment lexicon, manually labeled tweets and tweet/event alignment from ET-LDA, to regulate the learning process. We evaluated our framework quantitatively and qualitatively through various tasks. Based on the experimental results, our model shows significant improvements over the baseline methods. We believe that our work presents the first step towards understanding complex interactions between events and social media feedback and reveals a perspective that is useful for the extraction of a variety of further dimensions such as polarity and influence prediction.

Table 4.3: *Combinations of Prior Knowledge Vs. Accuracy. Notations:* $\mathbf{F}_0$ *for Sentiment Lexicon,* $\mathbf{R}_0$ *for Tweets Labels,* $\mathbf{G}_0$ *for Prior Tweet/event Alignment Knowledge from Et-LDA.* ***N.A*** *Refers to the Basic Framework Without Any Constraints.*

`DenverDebate`

	Segment (gain)	Topics (gain)	Tweets (gain)
N.A	0.486	0.502	0.498
$\mathbf{F}_0$	0.523 (+7.5%)	0.542 (+7.9%)	0.545 (+9.4%)
$\mathbf{R}_0$	0.532 (+9.5%)	0.548 (+9.2%)	0.578(+16.1 %)
$\mathbf{G}_0$	0.484 (-0.01%)	0.504 (+0.02%)	0.491 (-1.4 %)
$\mathbf{F}_0$+$\mathbf{R}_0$	0.572 (+17.7%)	0.564 (+12.4%)	0.735 (+47.8 %)
$\mathbf{F}_0$+$\mathbf{G}_0$	0.604 (+23.6%)	0.605 (+20.5%)	0.68(+36.5%)
$\mathbf{R}_0$+$\mathbf{G}_0$	0.612 (+25.7%)	0.612 (+21.9%)	0.687(+37.9%)
$\mathbf{F}_0$+$\mathbf{R}_0$+$\mathbf{G}_0$	0.618 (+27.2%)	0.628 (+25.1%)	0.768 (+54.2 %)

`MEspeech`

	Segment (gain)	Topics (gain)	Tweets (gain)
N.A	0.472	0.498	0.512
$\mathbf{F}_0$	0.493 (+4.4%)	0.503 (+1.1%)	0.557 (+8.7%)
$\mathbf{R}_0$	0.502 (+6.3%)	0.512 (+2.8%)	0.566 (+10.5%)
$\mathbf{G}_0$	0.467 (-1.1%)	0.494 (-0.8%)	0.515 (+0.5%)
$\mathbf{F}_0$+$\mathbf{R}_0$	0.542 (+14.8%)	0.552 (+10.2%)	0.606 (+18.3%)
$\mathbf{F}_0$+$\mathbf{G}_0$	0.568 (+20.3%)	0.578 (+16.0%)	0.632 (+23.4%)
$\mathbf{R}_0$+$\mathbf{G}_0$	0.578 (+22.4%)	0.588 (+18.1%)	0.642 (+25.3%)
$\mathbf{F}_0$+$\mathbf{R}_0$+$\mathbf{G}_0$	0.588 (+24.5%)	0.598 (+20.1%)	0.652 (+27.3%)

Chapter 5

TRENDING EVENT DETECTION IN SOCIAL MEDIA

In previous two chapters we introduced ET-LDA and SocSent, two powerful tools for handling the Event Characterization task. This chapter mainly focuses on the second task: Event Recognition. As we know that tweets can be seen as a dynamic source of information enabling individuals, companies and government organizations to stay informed about "what is happening right now". Besides, various examples have shown that events on social media are often reported earlier than by traditional news outlets (Lotan, 2012). Therefore, it is important to detect events on social media when they start trending, thereby facilitating better awareness for both regular social media users and news outlets.

However, it is know that social media is chaotic, noisy and overwhelming. Therefore, in order to develop our event detection algorithm we need to figure out how to separate the noisy information from interesting real-world events. Moreover, the event detection algorithm also requires to be highly scalable and efficient approaches in order to handle and process large amounts of tweets (especially for real-time event detection). In next sections, we provide our techniques to address these challenges.

5.1 Background of DeMa

Event detection has studied in the topic detection and tracking (TDT) community for a long period and recently in social community. In general, existing event detection approaches on Twitter (the ones that are adapted from TDT literature) can be broadly classified into two categories: document-pivot methods and feature-pivot methods, depending on whether they rely on document or temporal features. The former detects events by clustering documents based on the semantics distance between documents (Yang *et al.*, 1998), while the latter studies the features of words(Kleinberg, 2003) to make the event prediction. Although the document-pivot approach achieved great success in several TDT tasks, it has several drawbacks when applying to social media: First, measuring the distance between short documents like Twitter messages is a very challenge problem, and the state-of-the-art approaches still lack performance (Becker *et al.*, 2010). Second and more importantly, the document-pivot clustering approach is not effective in handling the cases where the events occur as bursts several times in a long time period. In other words, such approach may result in many small clusters, and can make it difficult to find the major events.

These weakness of document-pivot approaches give us the motivation for considering the feature-pivot approach (i.e., considering word distributions (assume each word is a feature) rather than document distributions during clustering). There has been a few work recently on feature-pivot methods for detecting trending events on Twitter (Becker *et al.*, 2010; Weng and Lee, 2011; Sakaki *et al.*, 2010; Petrović *et al.*, 2010). These approaches typically detect events by exploiting the temporal patterns or signal of Twitter streams. This is because, new and trending events often exhibit a burst of features in Twitter streams yielding, for instance, a sudden increased use of specific keywords. Bursty features that occur frequently together in tweets can then be grouped into trends. However, these event detection approaches suffer from two problems: (1) unable to locate the time periods when bursts happen, and (2) unable to differentiate whether the detected new event is trivial/endogenous or not (as mentioned in (Naaman *et al.*, 2011), in addition to trending events, endogenous or nonevent trends are also abundant on Twitter). Here, we present an effective unsupervised event detector, DeMa, to remedy these problems when detecting unplanned trending and novel events on social media. The DEMA framework is outlined in Figure 5.1. There are three major steps: (1) trending features identification, (2) trending features ranking, and (3) trending features grouping. Details are given in the following sections.

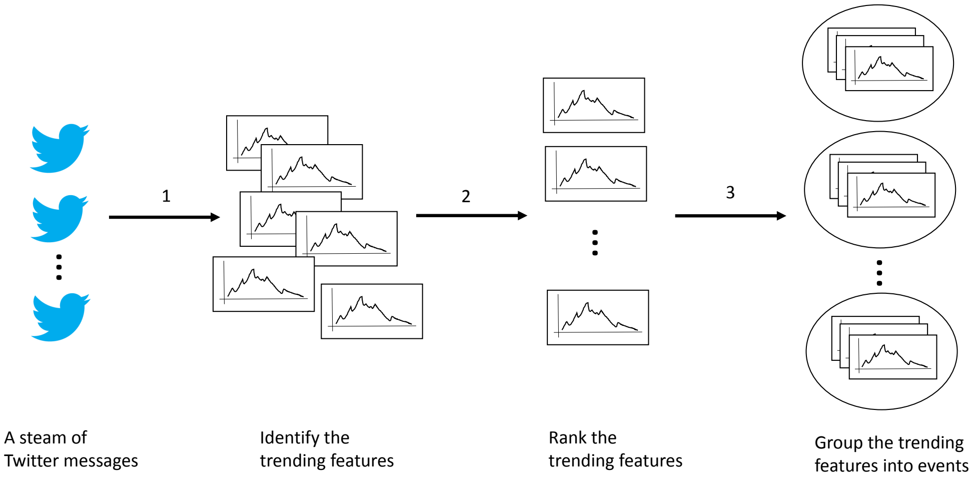

1
2
3
A steam of Twitter messages
Identify the trending features
Rank the trending features
Group the trending features into events

Figure 5.1: The Overview of the DEMA Approach

5.2 The DEMA framework

5.2.1 Problem Definition

Before moving forward. We first need to define the problem of trending event detection. Formally, given a stream of noisy Twitter messages (tweets), where each tweet consists of a set of terms (we call them *features* in this chapter), F_1, F_2, … F_m (e.g., gas, leak, danger, etc.), we define the problem of *trending event detection* as to find a set of trending events, where each one of them consists of a set of topically-related trending features, at a given time period. Furthermore, inspired by the model of theoretical "bursts" in streams of topics (He and Parker, 2010; Kleinberg, 2003), we define *trending* as a time interval over which the rate of change of momentum (product of mass and velocity – in classical mechanics) is positive. We further define that mass is the current importance of the feature and the velocity is the feature's average frequency in Twitter posts, during a time period. Next, we explain how the DeMa approach identifies these trending features from a substantial volume of Twitter posts.

5.2.2 Identify Trending Features

Since it is hard to directly measure the momentum based on the definition of mass and velocity in Twitter, we choose to use the trend analysis tools EMA (Exponential Moving Average), MACD (Moving Average Convergence Divergence), and MACD histogram from the quantitative finance literature (Murphy, 1999) to yield established measures of momentum. Next, we explain how these tools work to identify trending features from Twitter posts.

Given a feature F and its time series $S = S(F) = \{f_1, f_2, \ldots, f_m\}$, f_i denotes the frequency that F is mentioned by the Twitter posts posted within the i-th period. For example, the word "morning" can have a time series $S = \{248, 305, 154, 52, 24, 9\}$ from 8 a.m. to 2 p.m. of the day, in which it was mentioned 248 times by the Twitter posts from 8 a.m. to 9 a.m., 305 times from 9 a.m. to 10 a.m., and so on. Moving averages are commonly used with time series data to smooth out short-term fluctuations and highlight longer-term trend. Here, we compute the n-hour EMA for $S(F)$ as:

$$EMA(n)[S]_i = \alpha \times f_i + (1-\alpha) \times EMA(n-1)[S]_{(i-1)} \tag{5.1}$$

$$= \sum_{k=0}^{n} \alpha(1-\alpha)^k S_{i-k} \tag{5.2}$$

where $\alpha = 2/(n+1)$ is a smoothing factor, n is a time lag, and $1 \leq i \leq m$ is the index of time period. Essentially, the EMA smoothens out noises of F by averaging its time series over a specific number of periods. Next, to spot changes in the momentum of

F, we compute the MACD statistics, which is defined as the difference between the n_1- and n_2- hour EMA for $S(F)$:

$$MACD(n_1, n_2) = EMA_{(n_1)} - EMA_{(n_2)} \tag{5.3}$$

where n_1 and n_2 are time lags. Finally, to identify whether and when F is trending, we need to quantify the rate of change of its momentum. Therefore, we calculate the MACD histogram, defined as the difference between F's MACD and its signal line (the n-day EMA of MACD):

$$\text{signal}(n_1, n_2, n_3) = EMA(n_3)[MACD(n_1, n_2)] \tag{5.4}$$

$$\text{histogram}(n_1, n_2, n_3) = MACD(n_1, n_2) - \text{signal}(n_1, n_2, n_3) \tag{5.5}$$

As this difference measures the rate of change, the result at a given time period can be either positive (indicating F is trending up) or negative (indicating F is trending down). Therefore, based on the calculation results of MACD histograms, DeMa can easily locate the time periods when a trending event happens (i.e., MACD histogram becomes positive at time T).

5.2.3 Rank Trending Features

In some cases, the trending features may occur repeatedly. For example, "morning" can be trending from 8 a.m. to 11 a.m. every day. Such trending feature may be trivial, less interesting and more predictable compared to the ones which are single occurrences. To resolve this problem, we assign a "novelty" score to the identified trending feature according to their deviation from their expected trend. More specifically, for a trending feature F, we denote $R(h, d, w, F)$ as its MACD histogram result during hour h, day d, and week w. With this notation, we can compare F's trend in a specific day or hour in a given week to the same day or hour in other weeks (e.g., 9 a.m. on Monday, Aug 6, 2012, vs. the trend on other Mondays at 9 a.m.). Let $Mean(h, d, F)$ and $SD(h, d, F)$ denote the average trend and the standard deviation of F on hour h and day d over week w_1 to w_n, respectively. Then, the novelty score of feature F on hour h, day d, and week w is defined as:

$$\text{Score}(h, d, F) = \frac{R(h, d, w, F) - \text{Mean}(h, d, F)}{\text{SD}(h, d, F)} \tag{5.6}$$

Based on this score, we rank each feature to find the novel trending features.

In practice, to detect the *daily* trending events from social media, we first built a dictionary of features from all the Twitter posts of one day. Then, we created a time series for each feature by counting their frequencies in Twitter posts in every two hours. As a result, we have a 12-hour-long time series for every feature. Then, we applied the EMA, MACD, and MACD histogram over the time series data to identify whether and when a feature is trending. Finally, for every two hours, we picked the trending feature which (1) is least mentioned 20 times in the Twitter posts from that time period, and (2) has a novelty score among the top 25 scores for all trending features from that time period. Since these steps are computable in an online fashion (He and Parker, 2010), our DeMa approach is highly efficient and is able to work on identifying trending event features in real-time.

5.2.4 Trending Feature Clustering

Multiple events (and their associated features) can be trending within the same time period. Therefore, we need a way to separate these features and group them into topically-related event-clusters, respectively (i.e., each cluster represents one event). Since Twitter messages are constantly evolving and new events get added to the stream in real-time, we have no priori knowledge of the number of clusters that will be trending in a time period. To this end, we use the *shared nearest neighborhood* (SNN) clustering algorithm Jarvis and Patrick (1973). We chose this algorithm because it is scalable and does not require fixing the number of clusters in advance (as oppose to most clustering schemes like k-means).

The SNN algorithm is executed as follows: each trending feature is a node of the graph and each node is linked to another by an edge if it belongs to the k neighbor list of the second object. Here, we define feature F_1 is the neighbor of feature F_2 only if F_1 and F_2 are *topically*-related (e.g., "gas" can be a neighbor to "leak" but may not be to "party"). To learn a feature's topic, we use topic modeling (Blei *et al.*, 2003), a popular machine learning tool for getting topic distributions from text. In order to measure the topical relationship between two features (i.e., neighbors), we use the Jensen-Shannon divergence on their topic distributions. As a result, if the distance is above a threshold, the two features are neighbors.

5.3 Experimental Results

We evaluate the DeMa approaches on a large dataset of Twitter data. Below, we first describe the dataset and report the experimental settings, and then turn to the results of our experiments.

5.3.1 Experimental Settings

Dataset: Our dataset consists of over 2,600 Twitter messages randomly sampled from 2,585,000 Twitter messages posted during August 2012. Here, the sampling process is needed because we need human annotators to establish the ground truth in the evaluation. Besides, we collected these messages from the users who identified their location as Seattle, WA. Our intention here is to make the event annotations on collected Twitter messages more local-specific and hence it is more accurate.

Annotations: We used three human annotators from Seattle who are local neighborhood experts to label the tweets. In particular, we asked the annotators to label whether a tweet is event-related or not. We also calculated the Cohen's kappa to measure the inter-agreement ($\kappa = 0.77$ indicates the high quality of the annotation). As a result of this step, our annotators indicated that 878 of the total Twitter messages were event-related. Furthermore, for those tweets who are labeled as "event-related", we request the annotators to assign the importance of an event (0 = not important, 1 = important, 2 = very important).

Baselines: To evaluate the performance of DeMa, we implement some competitive baseline approaches: 1) **Fastest**: it is an event thread selection approach proposed by Petrovic et al. Petrović *et al.* (2010), which selects the fastest-growing threads in a stream of Twitter messages, and 2) **Random**: it is a very naive baseline which selects Twitter messages randomly as events.

5.3.2 Experimental Results

We begin by examining the performance of our DeMa against two baselines on the Twitter dataset. Table 5.1 shows the accuracy of three approaches, where the accuracy is measured based on the manually labeled ground truth. As we can see, DeMa outperformed Fastest over 50%, showing that it is overall more effective in identifying the trending events from social media. A deeper examination of our results revealed a strong, significant correspondence ($beta = 0.53$, $p < 0.001$) of the DeMa's results by a logistic regression. The event detector also produced a score for the importance to prioritize events, and this score was much higher for Twitter messages the participants identified as events ($t = 16.92$, $p < 0.001$). The participants' ratings of the importance of an event was significantly correlated with the event detectors ($r = 0.31$, $p < 0.001$).

Tools	Accuracy	Events identified	False positives
DeMa	0.78	684	17%
Fastest	0.52	456	35%
Random	0.24	211	85%

Table 5.1: Event Detection Accuracy of Various Approaches Against the Manually Labeled Ground Truth

In addition, we also provide a sample of events identified by DeMa, and their most frequent terms, are presented in Table 5.2.

Description of Events	Terms
Westminster Dog Show	westminster, dog, show, club
Gas leaked	Gas, Leak, Pike, 10th, St, Pine, Blocked, Siren
Openning of Gluten-free kitchen	Glueten, Free, Dedicated, Kitchen, Bar, Cap, Hill
Sea Toy Fair	toyfairsea, starwars, hasbro, lego

Table 5.2: A Sample of Events Extracted by DeMa

5.4 Application of DeMa

In this section, we are interested in applying DeMa to facilitate information seeking and improve civic engagement for hyperlocal community. More specifically, people rely on multiple sources of information to learn about the communities they live in (Rosenstiel *et al.*, 2011), either for the purpose of community awareness or participation (Newport and Jawahar, 2003). Hyperlocal information is comprised of the news, people, and events that are set within a particular locality, and is of particular interest primarily to the residents of that locality (Glaser, 2004). One of the most important sources of hyperlocal content is social media, such as blogs, microblogs, and social networking sites. Social media has many advantages over traditional media in assisting people's quest for hyperlocal content such ubiquity and immediacy as we discussed earlier in Chapter 1. Moreover, social media has emerged as a dominant platform for communication and connection. As hyperlocal content is mostly generated by and for a community, seamless communication and networking (through one's social networks) can increase exposure to timely peer generated content, raise people's community awareness, and potentially foster their sense of community (Ellison *et al.*, 2007).

In spite of these benefits, social media tends to be noisy, chaotic, and overwhelming, posing challenges to users in seeking and distilling high quality content from the noise. It should be no surprise that, regardless of the popularity of social media as a source of hyperlocal information, people are still using television and newspapers (among other traditional sources) as their main channels for local information [28]. People need help leveraging social media as a source of information about their hyperlocal communities. At one extreme are the fast-paced, uncurated social media streams: chaotic and overwhelming. At the other extreme are the traditional, authoritative, news sources: slow and less participatory than social media. In this paper, we present Whoo.ly, a novel web service balanced between these two extremes.

Whoo.ly automatically discovers, extracts, and summarizes relevant hyperlocal information contributed on Twitter to facilitate people's neighborhood information-seeking activities. Inspired by the core journalism questions (what, who, where, and when), Whoo.ly provides four types of hyperlocal content in a simple webbased interface (See Figure 1): (i) active events (events that are trending in the locality); (ii) top topics (most frequently mentioned terms and phrases from recent Twitter posts); (iii) popular places (most frequently checkedin/mentioned); and (iv) active people (Twitter users mentioned the most). It is important to note that it is not our goal with Whoo.ly to replace traditional news media. Instead, we want to provide hyperlocal information that is complementary to what both traditional news media and social media have to offer.

The unique features of Whoo.ly are the novel event detection and summarization based on DeMa we developed. Top neighborhood topics are inferred using a simple yet effective weighting scheme that finds the most important words and phrases from posts. To identify the most popular places in a neighborhood, we used both template-based information extractors and learning-based information extractors. Finally, to distill a ranked list of the active people in a community, we developed a ranking scheme on the social graph of Twitter users based on their mentioning and posting activities.

To evaluate Whoo.ly's utility as a tool for finding neighborhood information, including its user interface and our algorithms, we performed a user study with thirteen residents from three Seattle neighborhoods. Most of our participants believed Whoo.ly provided them with useful neighborhood information, and rated it easier to use than Twitter's native tools.

5.4.1 Related Work of Whoo.ly

Using new technologies to promote community awareness and participation has long been a research topic for the HCI community and (Wellman, 2005; Lewis and Lewis, 2012). Webmediated communities such as Netville and the Blacksburg Electronic Village have demonstrated how the Internet can enhance spatial immediacy, facilitate discussion, and quickly mobilize people around local issues (Hampton and Wellman, 2003).

The prevalence of “Web 2.0” has provided new opportunities for technologies to facilitate better information seeking and communication about local communities. In particular, social media tools have been used to report various activities including breaking news (Kwak *et al.*, 2010), public debates (Hu *et al.*, 2012b), crises like floods (Vieweg *et al.*, 2010), earthquakes (Sakaki *et al.*, 2010), or even during wartime (Monroy-Hernández *et al.*, 2012). Recently, leveraging social media resources for local communities has drawn considerable attention in both research and industry. Such efforts include Livehoods (Cranshaw *et al.*, 2012). Among them, CiVicinity (Hoffman *et al.*, 2012) provides a hyperlocal community portal that integrates information from Facebook, blogs, calendars, and other sources to promote civic awareness and participation. Virtual Town Square (VTS) (Kavanaugh *et al.*, 2014) also aggregates local information from a predefined set of information sources (government, schools, and news organizations) to improve community engagement. Our work uniquely builds on this line of research by exploring automatic solutions to the detection, extraction and summarization of neighborhood information from noisy Twitter posts.

The hyperlocal content in Whoo.ly is automatically mined from Twitter, which presents unique challenges not directly addressed by related work. (1) Our method of finding top topics was inspired by the TFIDF statistics that assign scores for terms based on their mentioned frequency within and across documents. Even though there are other efforts to find top topics from Twitter posts (O'Connor *et al.*, 2010b), such approaches often take a long time to run to discover meaningful topics, and we seek to provide reasonably realtime results. (2) Information extraction has been a longstanding research topic (Chang *et al.*, 2006). In Whoo.ly, we use a hybrid approach of both template-based and learning-based extractors to find popular places in Twitter posts.

5.4.2 Whoo.ly Overview and Design Process

In this section, we first provide an overview of Whoo.ly and its features. Then, we highlight the motivations underlying the choices we made in the design process.

Whoo.ly is a web service built on top of Twitter. Its goal is to provide people with relevant and reliable hyperlocal news content. By browsing the website, people immediately find what is happening in a specific neighborhood. Whoo.ly provides four hyperlocal content types: active events, top topics, active people, and popular places (See Figure 5.2). All of them are automatically extracted and summarized from Twitter using various approaches we developed, such as statistical event detector, graph-based ranking algorithm, and information extractors (see the System Design section for more detail).

Early in the design process for Whoo.ly, we examined local newspapers, community blogs, existing hyperlocal sites, and Twitter. The exploratory study revealed several interesting results that we used to motivate the design of Whoo.ly: (1) The majority of the people only consume information they do not produce it but only read it; (2) People become more active in reporting and disseminating local breaking events (e.g., shooting, water leak) on Twitter by reposting related tweets; (3) People tend to follow neighborhood curators or bloggers who are dedicated to posting hyperlocal content; and (4) Local media and local news services effectively cover important local topics. However, people further seek hyperlocal content generated by people in their communities.

We performed an additional preliminary analysis of Twitter data to help inform our design decisions, answering the following questions: (1) can we find a base of local Twitter posts based on neighborhoods; (2) were there enough messages to seed a neighborhood website; and (3) what do people care to talk about on Twitter regarding their neighborhoods? We first queried for all Twitter messages from people who claimed Seattle as their home town for the month of October of 2011. We then performed a simple extraction of Twitter messages that mentioned one of 83 Seattle neighborhoods. We found 50,609 unique Seattle users and produced 1.2 million messages (about 8% of total Seattle population), out of which 5% explicitly mentioned Seattle, and another 2% mentioned a Seattle neighborhood. On average 132 people posted per neighborhood over the month, averaging 1.8 messages each, which translates into about 8 messages per day per neighborhood. There was great variability across neighborhoods, but we considered the above averages to be a promising start and used them as the volume of neighborhood Twitter messages to expect.

To examine message content, we sampled 24% of the messages (424) from three neighborhoods preselected for being diverse from each other. We first coded the messages for whether they were erroneously assigned to the neighborhood. Surprising, only 21 messages (5%) were erroneously assigned, largely because of overlapping neighborhood names and other place names (e.g., the area ”Mount Baker” and the mountain “Mount Baker” it was named after). We then looked at how many were personal in nature, of little interest to anyone aside from the author’s friends. We found that 13% of messages were of this nature. Places checkins comprised another 55 messages (10%), which we expect might be interesting when aggregated but not at the individual level. Six items were impossible to interpret and were left unclassified. The remaining messages were 71% on topic, meaningfully pertaining to the neighborhood. We further inspected and coded by message type and whether or not they were about a current event. We defined a current event as a real world occurrence with an associated time period such that if it is not observed, experienced, or attended in that time period a person will not be able to do so later. Thus a crime, a fire, a festival, or a Friday happy hour are current events. In contrast, a photo shared online, a news story link, a recommendation to try a restaurant, or a shout out of thanks are not. We found that 55 % of the remaining Twitter messages were about an event.

All message types in our data sample were classified as depicted by Table 1. Topics such as crime reports, Yelp like recommendations, and local news were not surprising. The neighborhood affirmations and salutations were surprising, where people in the community post messages talking about how much they love their neighborhood, or community affirming, humorous messages reinforcing the neighborhood's stereotypical traits.

Based on these findings, we decided to focus first on detecting events and then to promote community enabling features such as a list of top users so that people can know and follow each other. To prevent information overload, we also provided top topics so that people can quickly learn the common neighborhood topics in the Twitter posts (tweets).

5.4.3 Data Collection of Whoo.ly

Whoo.ly is built on Twitter. We utilized the Twitter Firehose that is made available to us via our company's contract with Twitter. Since we are interested in discovering hyperlocal content for local communities in various geographic regions, we needed to obtain a set of Twitter posts from each region. Twitter offers two possible ways to infer a tweet's location: GPS coordinates associated with a tweet or the user's location in their profile. In this work we used the location information derived from the user profile since the number of Twitter posts found by GPS coordinates is very limited (about 0.6%). From our preliminary analysis using this method, we found a reasonable quantity of on topic neighborhood messages.

We observed that most Twitter users prefer to mention only their city instead of local community for the profile location, probably due to privacy concerns (Hecht *et al.*, 2011). As a result, we first obtained Twitter posts from the Firehose, where each associated user profile location matches one of the dictionary strings for a city, e.g., "Seattle" or "Sea". Next, we mapped these Twitter posts into different neighborhood regions by matching their textual content against a list of neighborhoods. Note that the neighborhood list for each city is created by domain experts who have comprehensive experience with the neighborhood development and boundaries in that city.

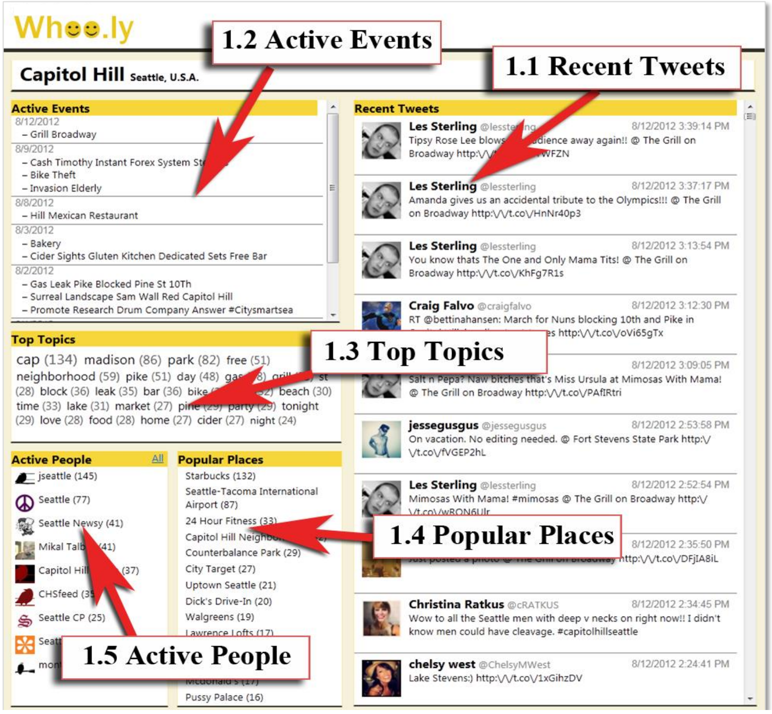
We used a dataset that included about 2.2 million Twitter posts in English from about 120,000 unique users whose profile location indicated they are from Seattle, over a three month period from June 1, 2012 to Aug 15, 2012. While we mainly used this static dataset for developing our prototype, our methods may easily be extended to handle realtime tweet streams

5.4.4 System Design of Whoo.ly

In this section we describe the system design of Whoo.ly (Figure 5.2), including the interface design of its components and the technical design behind them. Whoo.ly's interface is implemented in HTML, CSS, and Ajax controls toolkits, served by ASP.net on the cloud service Windows Azure. Whoo.ly first shows a start page, where a user selects his or her country, city, and neighborhood through dropdown lists. After selecting their location, users are taken to the results page (Figure 5.2), which displays recent Twitter posts, top topics, popular places, and active people.

5.4.4.1 Recent Twitter Posts

Whoo.ly presents recent Twitter posts in a scrolling list on the right side of the results page (Figure 5.2). Each row in this list contains a detailed Twitter profile for a user on the top, and his or her recent posts at the bottom. The profile includes



Whoo.ly

Capitol Hill Seattle, U.S.A.

Active Events
8/12/2012
– Grill Broadway
8/9/2012
– Cash Timothy Instant Forex System St
– Bike Theft
– Invasion Elderly
8/8/2012
– Hill Mexican Restaurant
8/3/2012
– Bakery
– Cider Sights Gluten Kitchen Dedicated Sets Free Bar
8/2/2012
– Gas Leak Pike Blocked Pine St 10Th
– Surreal Landscape Sam Wall Red Capitol Hill
– Promote Research Drum Company Answer #Citysmartsea

Top Topics
cap (134) madison (86) park (82) free (51) neighborhood (59) pike (51) day (48) gas 8) grill st (28) block (36) leak (35) bar (36) bike 2) beach (30) time (33) lake (31) market (27) pine party tonight (29) love (28) food (28) home (27) cider (27) night (24)

Active People All
jseattle (145)
Seattle (77)
Seattle Newsy (41)
Mikal Talb (41)
Capitol Hill (37)
CHSfeed (35
Seattle CP (25)
Seatt
mon

Popular Places
Starbucks (132)
Seattle-Tacoma International Airport (87)
24 Hour Fitness (33)
Capitol Hill Neighbo
Counterbalance Park (29)
City Target (27)
Uptown Seattle (21)
Dick's Drive-In (20)
Walgreens (19)
Lawrence Lofts (17)
Pussy Palace (16)

Recent Tweets

Les Sterling @lessterling 8/12/2012 3:39:14 PM
Tipsy Rose Lee blows dience away again!! @ The Grill on Broadway http:\/\/ WFZN

Les Sterling @lessterling 8/12/2012 3:37:17 PM
Amanda gives us an accidental tribute to the Olympics!!! @ The Grill on Broadway http:\/\/t.co\/HnNr40p3

Les Sterling @lessterling 8/12/2012 3:13:54 PM
You know thats The One and Only Mama Tits! @ The Grill on Broadway http:\/\/t.co\/KhFg7R1s

Craig Falvo @craigfalvo 8/12/2012 3:12:30 PM
RT @bettinahansen: March for Nuns blocking 10th and Pike in es http:\/\/t.co\/oVi65gTx

8/12/2012 3:09:05 PM
Salt n Pepa? Naw bitches that's Miss Ursula at Mimosas With Mama! @ The Grill on Broadway http:\/\/t.co\/PAfIRtri

jessegusgus @jessegusgus 8/12/2012 2:53:58 PM
On vacation. No editing needed. @ Fort Stevens State Park http:\/\/t.co\/fVGEP2hL

Les Sterling @lessterling 8/12/2012 2:52:54 PM
Mimosas With Mama! #mimosas @ The Grill on Broadway http:\/\/t.co\/wRON6Ulr

8/12/2012 2:35:50 PM
http:\/\/t.co\/DFjIA8iL

Christina Ratkus @cRATKUS 8/12/2012 2:34:45 PM
Wow to all the Seattle men with deep v necks on right now!! I didn't know men could have cleavage. #capitolhillseattle

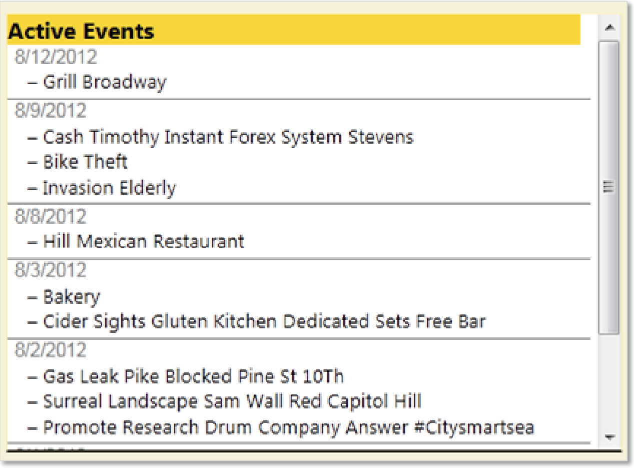
chelsy west @ChelsyMWest 8/12/2012 2:24:41 PM
Lake Stevens:) http:\/\/t.co\/1xGihzDV

Figure 5.2: The Main Whoo.Ly Interface, with the Recent Twitter Posts and Summaries of Events, Topics, Places, and People

standard elements retrieved from Twitter such as the user name, screen name, user's profile image, user's profile location, and the posting time of the messages. Whoo.ly only provides the most recent Twitter posts from a time window of 14 days mainly because people are usually only interested in most recent Twitter posts. Nevertheless, the length of the time span can be easily adjusted through a dropdown list at the bottom of the results page.

5.4.4.2 Active Events

Whoo.ly presents an active events list calendar (Figure 1.2) on the upper left side of the results page. Each entry shows the events organized by date. Every event is summarized by a list of terms and, by clicking on its name, the user is taken to a page (Figure 5.5) containing all the posts that are about that event, ranked by their relevance score using vector similarity (Manning *et al.*, 2008). The events are generated by the DeMa algorithm.



Active Events

8/12/2012
- Grill Broadway

8/9/2012
- Cash Timothy Instant Forex System Stevens
- Bike Theft
- Invasion Elderly

8/8/2012
- Hill Mexican Restaurant

8/3/2012
- Bakery
- Cider Sights Gluten Kitchen Dedicated Sets Free Bar

8/2/2012
- Gas Leak Pike Blocked Pine St 10Th
- Surreal Landscape Sam Wall Red Capitol Hill
- Promote Research Drum Company Answer #Citysmartsea

Figure 5.3: A Close-up View of the Active Events Pane. Events Are Organized by Date and Are Represented by a List of Terms Most Associated with Each Event

5.4.4.3 Top Topics

Below the trending events section, Whoo.ly shows a list of top topics (with their frequencies) that are being discussed in the recent Twitter posts (Figure 1.3). Clicking a topic leads to a page showing all the Twitter posts about it. This component helps people quickly understand and familiarize themselves with the most important topics about the neighborhood appearing in Twitter posts. We design a fast approach by applying normalized TFIDF statistics for each uni-, bi-, and tri-gram from the recent Twitter posts. We then rank these grams to render this component.

5.4.4.4 Popular Places

Beyond the event and topics, Whoo.ly shows a list of 15 most popular places (Figure 1.4) that people keep checking into and mentioning in Twitter posts. Similar to other components, clicking a place leads to a page showing all Twitter posts about this place. This component helps people discover interesting places in their neighborhood and learn what is happening there. Extracting these places from Twitter posts requires automated information extraction, which has been a longstanding research topic in NLP and machine learning Mahmud *et al.* (2014). In the next sections, we describe two types of extractors we use to build this component, namely a template-based extractor and learningbased extractor.

Template-based Information Extractor Through our manual inspection of the Twitter posts content (see the Overview section), we found there is a small percentage of Twitter posts (approximately 7%) that were posted by Foursquare check-ins. Such Twitter posts have a specific template in their content: begin with the phase "I'm at", followed by a place name (e.g., Space Needle), and followed by its address (e.g., 400 Broad Street, Seattle, WA 98102). Given this structure, we designed a templatebased extractor using regular expressions to distill the place information.

Learning-based Information Extractor For Twitter posts without explicit format for location inference, we used a statistical information extractor. It is built on top of an ngram language Markov model and previously trained on Wikipedia pages, Tweets, and Yelp data (Wang *et al.*, 2013). We apply it to analyze the Twitter posts to extract entities for places, e.g., restaurants, parks, streets, stadiums, etc.

5.4.4.5 Active People

Last, Whoo.ly displays a list of top 10 most active people (i.e., Twitter users) for the corresponding neighborhood (Figure 1.5). Each record in the list combines a user's profile and the frequency this user posts or was recently mentioned by other people. In addition, Whoo.ly also presents the profiles, latest Twitter posts, and activities of all the users who have recently posted Twitter messages (by clicking "All" on the up right corner of this division). With this component, one can easily identify who are the active and influential people in the neighborhood and can decide to follow their activity. To build this component, we developed a PageRank-like algorithm to rank the Twitter users based on their mentioning and posting activities. Specifically, a directed graph $D(V, E)$ is formed with the users and the "follower-followee" relationships among them. V is the vertex set, containing all the users. E is the edge set. There is an edge between two users if there is "following" relationship between them, and the edge is directed from follower to followee. Our algorithm performs an activity-specific random walk on graph D to calculate the rank. It visits each user with certain transition probability by following the appropriate edge in D. The probability is proportional to a linear combination of the interactions between two users (e.g., RT, mentioning, reply) and how many Twitter posts a user has posted recently. The idea is that the more activities a user has, the higher this user's rank is.

5.4.5 User Study

We evaluated Whoo.ly as a tool for users to learn about what is happening in their neighborhood using a within subjects comparison of Whoo.ly and Twitter, where users completed a series of information seeking tasks for each platform and then provided feedback. For our user study, we focused only on three Seattle neighborhoods for which we were able to recruit participants.

5.4.5.1 Participants

We introduced 13 Seattle residents into a private, pre-release version of Whoo.ly through five focus group sessions, with two or three people per session. Participants were recruited from a preexisting database of people who for the most part had expressed interest in user studies. Potential participants in the database were first filtered for address zip codes in our target neighborhoods. After receiving phone calls to screen for whether they continued to live in the neighborhood and had a Twitter account, they were scheduled to participate in one of five sessions. In exchange for their participation they received their choice of a software gratuity or gift card. Participants were on average 30 years of age (ranging from 23 to 48). 54% of them were female and 46% were male. Ten participants were white, one Asian, one Native American, and one had other ethnic identity. The majority of participants were from the Capital Hill neighborhood (69%), with 23% from Wallingford and 8% from Rainier Valley. These neighborhoods differed in density, SES, and level of existing community infrastructure.

5.4.5.2 Procedure

During two hour user sessions participants first completed a preliminary questionnaire. They then briefly discussed their current communication practices for finding and sharing neighborhood information in a semi-structured focused group. Participants then individually completed a series of tasks with both Whoo.ly and Twitter using laptops with an Internet connection we provided. After a brief discussion of participants' experiences, we ended the session by having them rate a series of Twitter messages for neighborhood content.

Preliminary Questionnaire

Participants first completed a brief preliminary questionnaire to assess demographic information, use of Internet, and social media, and measure of their current neighborhood including psychological sense of community, neighborhood communication efficacy, and civic engagement. We measured psychological sense of local community (Sarason, 1974), or the feeling of connection, belonging, and loyalty to a local community, with items such as "I feel loyal to the people in my neighborhood," "I really care about the fate of neighborhood," and "I feel like I belong in my neighborhood." Civic engagement was measured using items from the Civic Engagement Questionnaire (Keeter *et al.*, 2002), a standard measure asking how often respondents had engaged in various civic activities such as "Spending time participating in any neighborhood community service or volunteer activity" and "playing a leadership role in my neighborhood (such as local government or leadership in a club)." Neighborhood communication self-efficacy, including communication self-efficacy, was measured with items adapted from the California Civic Index (Kahne *et al.*, 2005) that addressed communication, including "I know how to collect information and be informed about neighborhood issues," and "I know how to get in touch with members of my neighborhood when I need to communicate with them." For each measure, items were rated on a Likert scale of 1 to 7, where 1 = not at all and 7 = extremely so, and then items were averaged for analysis.

Focus Group

To further elucidate existing information seeking and communication practices, we then had participants discuss their neighborhoods using a semi-structured group interview. Participants first described the character of their neighborhoods, how long they have been living there, and whether they had a sense of connection or community to their neighborhood. We then asked participants to discuss what kinds of information they cared to learn about in their neighborhoods. Participants then described the tools they currently use to seek out information or communicate with others around neighborhood issues and where they would like to see changes or improvements in the tools available.

5.4.6 Neighborhood Information Seeking Task

Following the focus group, participants individually completed a series of four information seeking tasks, once in Whoo.ly, and once in Twitter. Each participant completed the tasks separately on a laptop with an Internet connection following instructions in a paper packet. The order of the tasks (Whoo.ly vs. Twitter) was counterbalanced across sessions, ending with seven participants completing the Twitter tasks first and six participants completing the Whoo.ly tasks first. Participants were instructed, "for this part of the study we will have you explore what's happening in your neighborhood using [Twitter or Whoo.ly]." The four tasks were: 1) find neighborhood events: *"try to find three interesting or significant events that happened in your neighborhood the past couple of weeks"*; 2) find neighborhood reporters: *"imagine you wanted to try to follow three people to help you keep up to date with what's happening in your neighborhood try to find those three people you would follow"*; 3) find neighborhood topics: *"imagine you wanted to find out what kinds of topics your neighborhood tends to care about try to find three of these topics"*; and 4) find neighborhood friends: *"imagine you wanted to get to know some people in your neighborhood better find three people you might want to know more".*

Participants were instructed to spend only a few minutes on each task, to get a sense for the experience in the system they were evaluating. After completing each task participants rated the ease of the task to complete, how confident they felt about their answers, and how engaged they were by the task (that is, to what extent they found it fun or interesting).

Following the completion of these tasks, participants rated the overall usefulness and ease of each system (Twitter and Whoo.ly), the extent to which it provided a good overview of what is happening in their neighborhood, the extent to which it provided a sense of connection, and which system they would prefer to use for finding out what is happening in their neighborhoods. Finally, participants were asked to rank their preference for individual aspects of the Whoo.ly interface and provide opened ended feedback to questions about what they liked, disliked, and possible improvements.

Tweet Rating Task In order evaluate the event detection algorithms, participants were asked to rate a randomly selected series of Twitter posts from the period spanning two weeks prior to that of the current Whoo.ly system. For each tweet, participants rated if it was a about a neighborhood event and if so, how significant was the event to their neighborhood, where 1 = not at all, few people involved, and 7 = extremely so, entire neighborhood involved.

5.4.7 Results

In analyzing our results, we first examined our participants' existing neighborhood information seeking and communication practices to better shed light on their experience of Whoo.ly and potential considerations for a real world deployment of this system. We then assessed how well participants completed information seeking tasks in Whoo.ly, providing a comparison to Twitter as a baseline tool for searching and browsing Twitter messages. Finally, we further examined themes that emerged from participant ratings and discussions that would meaningfully influence the design of Whoo.ly and similar systems.

5.4.8 Existing Practices

In our preliminary questionnaire participants rated themselves as having high levels of overall Internet experience, with 39% categorizing themselves as intermediate, 45% as advanced, and 16% as expert. Seventy six percent of participants reported spending four or more hours a day using the Internet. For communicating and sharing with others, participants reported text messaging (M = 6.5, SD = 0.66) and email (M = 6.6, SD = 0.65) to be extremely important, then social networking sites such as Facebook (M = 5.9, SD = 1.00), blogs (M = 4.1, SD = 1.32), Twitter (M = 3.6, SD = 1.90), and mailing lists less so (where 1 = not at all, and 7 = extremely so).

Most of the participants in our study cared very much about their neighborhoods, reporting fairly high levels of psychological sense of community (M = 5.0, SD = 0.83). The few exceptions made apparent from our interviews were individuals new to the neighborhood, or one participant who felt his neighborhood was too transitional by nature to become attached to it. However, the participants had lower levels of civic engagement (3.0, 1.27) and communication self-efficacy (M = 3.8, SD = 1.8). When asked to what extent they could collect information and be informed about neighborhood issues, participants' responses were on average moderate (M = 3.9, SD = 1.8). An examination of the distribution of this variable suggests it is bimodal, for example people either were low (45% at 2 or 3) or high (39% at 5 or higher) in their ability to find information or communicate with their neighborhood. When participants were asked how exactly they learned about what was happening in their neighborhoods, resources were quite diverse, including local newspapers, local blogs, following business on Twitter, local meetings, Facebook groups, coffee shops, and services such as Reddit, Google, and Yelp. However, local blogs clearly played a prominent role and word of mouth was frequently mentioned as a source of information. Several people mentioned Facebook or Facebook groups, but these were groups of people they knew who were in their neighborhoods, rather than public Facebook groups for the entire neighborhood. Further, it was clear that some neighborhoods had many more resources available than others.

We further asked what kinds of neighborhood information participants wanted to know about. Emerging themes were events such as local festivals and block parties, crime, new restaurants and bars, building developments, people, and local business promotions such as happy hours and coupons. Events and crime were most frequently mentioned, particularly as they impacted the local community. One participant's response was, Community stuff like I heard about neighborhood night out but I didn't know about it, my street closed and people were out drinking and barbecuing and I didn't know about ityou know about the big things, but little community stuff, that stuff you should know.

On average, participants were not confident they knew how to get in touch with members of their neighborhood when they needed to communicate with them (M = 3.5, SD = 1.9). When participants were asked, if they needed to communicate with members of their neighborhood community about neighborhood issues, how would they do so, face-to-face was rated the most highly (M = 5.2, SD = 1.8), followed by Facebook groups. During the interviews across sessions participants similarly exhibited low confidence in how they would go about communicating with their neighbors, and expected they would resort to walking down the street. One participant replied, "old fashioned way, knock on door. Too many people in the neighborhood to have phone numbers and emails." More tech-savvy participants said they would contact the local blog or access their neighbors' email addresses.

We asked participants to discuss their Twitter usage in particular, given the focus of Twitter as a source of public information in Whoo.ly. All participants had an account, but the majority used it primarily to consume information, either the news or their friends' posts. Only a few used Twitter to follow their neighborhood bloggers or neighborhood businesses.

To summarize, we found that our participants were fairly tech-savvy and felt fairly attached to their neighborhoods. While only a few were more civically engaged, most reported they would want to be more so. However, the participants did not have a strong sense for how to find out about what was happening in their neighborhoods or how to get involved. Particularly, they were not sure how they would go about communicating with others in their neighborhood about issues they cared about. Participants were especially interested in learning about local community events and crimes and relied heavily on one or two hyperlocal bloggers to do so.

5.4.9 Whoo.ly Evaluation

Participants completed four tasks exploring their neighborhood – find recent events, find local neighborhood reporters, find neighborhood topics, and find potential neighborhood friends using both Whoo.ly and Twitter. We performed an omnibus repeated measures ANOVA (technology X task X type of rating) to test for the impact of type technology across measures of task ease, confidence in completing task, and task engagement. Overall, we found a significant effect of technology (F(1,11) = 3.02, p = 0.05 1tailed), with participants showing preference for Whoo.ly. As can be seen from Figure 3, people overall found Whoo.ly easy to use and found the tasks easier to complete in Whoo.ly than in Twitter. We found neighborhood communication self-efficacy to be a meaningful co-variate interacting with this effect (F(1,11) = 3.3, p = 0.04, for interaction of technology X task X self-efficacy), meaning participants with lower levels of self-efficacy were likely to favor Whoo.ly over Twitter, especially for the find friends task.

These results suggest that Whoo.ly is particularly easy for users to learn more about their neighborhood if they do not already have effective tools to find information and access people in their neighborhood.

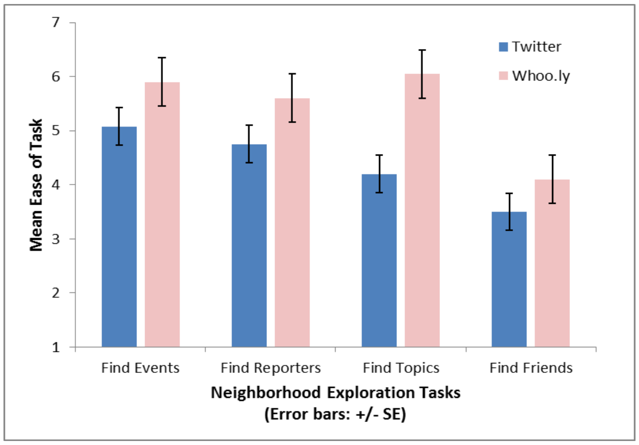


Figure 5.4: Participants Generally Found It Easier to Complete Neighborhood Exploration Tasks Using Whoo.Ly (Where 1 = Not at All, and 7 = Extremely so.)

Participants also completed overall ratings of Whoo.ly and Twitter, and again using an omnibus repeated measures ANOVA (technology X type of rating) we found an effect of technology (F(1,11) = 3.09, p = 0.06), such that participants reported it as more useful (F(1,11) = 2.24, p = 0.08) and easier to use (F(1,11) = 2.72, p = 9.07), that it provided a better overview (F(1,11) = 2.74, p = 0.07, and that it increased the sense of connection to their neighborhood community (F(1, 11) = 3.5, p = 0.04), as shown in Figure 4. Again, neighborhood communication self-efficacy had a marginally significant interaction such that people with lower levels self-efficacy were more impacted by Whoo.ly in their ratings of sense of connection (F(1, 11) = 2.81, p = 0.09).

To assess our event detector, we compared user ratings of 503 Twitter posts in our participants' neighborhoods to the event detectors. Users indicated that 170 of the total Twitter messages were event-related. Among these, the detector also identified 78% of messages as event-related, relative to 17% false positives. A logistic regression shows a strong, significant correspondence (beta = 0.53, p ¡ 0.001). The event detector also produced a score for the importance to prioritize events in the user interface, and this score was much higher for Twitter messages the participants identified as events (t = 16.92, p ¡ 0.001). The participants' ratings of the importance of an event was significantly correlated with the event detectors (r = 0.31, p ¡ 0.001).

In order to compare the relative value of the types of summarization provided by Whoo.ly, we asked participants to rank the five main sections by order of preference, where 1 = most preferred and 5 = least preferred. We found that participants rated recent events most highly (M = 1.6), followed by the Tweet stream (M = 2.8), the top topics (M = 3.2), active people (M = 3.5), and popular places (M = 3.5). After participants completed both sets of tasks, we asked them to choose which application they would prefer to use to find out what is happening in their neighborhood. Eight participants out of 13 preferred Whoo.ly. However, when asked to compare it to their favorite neighborhood blog, eight out of 13 said they would prefer their neighborhood blog. On average, participants indicated they were somewhat likely to actually use Whoo.ly if it were made publicly available (M = 4.4, SD = 1.62 where 1 = not at all, and 7 = extremely so). In order to shed light on some of our more quantitative findings, each participant was asked to provide feedback in writing about what they liked and disliked about Whoo.ly and how they would suggest improving it. Then, participants were asked to briefly discuss their experiences. When asked what they most liked about Whoo.ly, participants indicated the summarization and community features. Some of participants' answers were, *”Really liked it overall, definitely a lot* *easier to find stuff"*, *"Whoo.ly was set up specifically with the community in mind. It makes community news/events/issues/people etc. easily accessible".* When asked what they disliked, a few participants noted that a lot of the content felt like spam, such as the Craigslist postings, Foursquare checkins, or overly personal posts, which interfered with participants' ability to access meaningful content. One participant said, "Results. Mostly the furniture on craigslist. Need to filter out those, and be able to differentiate between the spammy 'top users' and the real top users."

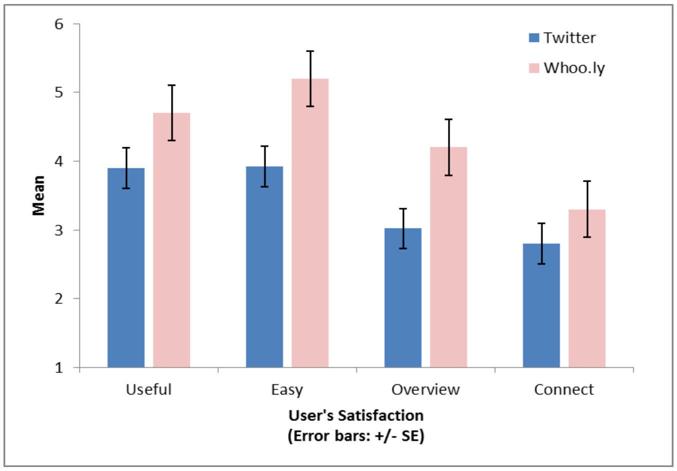


Figure 5.5: Whoo.Ly Was Found to Be More Useful, Easy to Use, With a Better Overview of the Users Neighborhoods, and A Sense of Connection to Their Neighborhood Communities.)

During the discussions, there were also several requests for further, personalized filters, to focus on the kind of content they cared about. When asked why they preferred Whoo.ly over Twitter, again participants noted the filtering, summarization, and community features. One participant's answer was, *"Twitter isn't set up for a community. Whool.y functions amazingly for this."* Consistent with our more quantitative findings, we found that participants who preferred Twitter over Whoo.ly did so because they were already well-connected to their neighborhoods and already using Twitter to follow neighborhood reporters. For example, a participant's reply was, *"If I didn't know my neighborhood as well I would use both and compare data. Since I am very embedded in my community Whoo.ly is just another aggregator."*

When asked why they would prefer their local blog over Whoo.ly, participants noted blogs had more extensive features such as calendars and they benefitted from social curation. When asked why they would prefer Whoo.ly, participants mentioned its ease consumption and community feel. Some of the participants' reasons were, *"like that it's short messageseasier than if browsing full blog with full messages; easier to figure what's going on." "Whoo.ly offers not only news/events, but also connects you with people. Like combining Twitter and a newspaper, I like it".*

5.4.10 Discussion

As shown above, the overall reaction to the information provided on Whoo.ly was quite positive. The participants to our study found Whoo.ly easier to use than Twitter and the majority said they would prefer it as a tool for exploring their neighborhoods.

As a prototype system, Whoo.ly has advanced the state of the art for information seeking in hyperlocal communities, but many opportunities for improvement remain. As people cross the line from consuming hyperlocal information to engaging with their local community, they seek to know as much about the people as about the news. Thus, from a hyperlocal community perspective, it is also important to recommend potential similar friends such as "people like me in my neighborhood" as a feature to improve neighborhood connections. Besides, exploring the sentiments behind people's response/reactions to neighborhood issues can be useful (). Furthermore, it is interesting to note the unique characteristic of consuming social media when embedded in a geographical location people could easily walk out their front doors and down the street to experience, for example, the local event they had just read about online.

It is worth noting that we deliberately placed the questionnaire and the focus group prior to the information seeking user tasks to frame the tasks specifically on neighborhood seeking behaviors. Our intention was to give users the opportunity to have access to each other's neighborhood seeking experiences in evaluating the technology's effectiveness. We recognized a discussion could have systematically and artificially affected preferences towards or against Whoo.ly across all participants. However, there is no indication that this is the case. To further assess potential discussion confound, we tested for group size (2 vs. 3) on preference for Whoo.ly vs. Twitter, and found no effect. Moreover, we also found there were no session and level of Twitter usage effects.

5.5 Summary of Chapter

In this chapter we introduced DeMa, a novel event detection algorithm to discover trending events from Twitter posts for any given time period. Moreover, we developed Whoo.ly, an application of DeMa. Whoo.ly is a web service that facilitates information seeking in hyperlocal communities by finding and summarizing neighborhood Twitter messages. We presented several computational approaches used in Whoo.ly to discover hyperlocal content from noisy and overwhelming Twitter posts. In particular, activity based ranking algorithms and information extractors provided additional insights into the most active people and popular places in a local community. We performed a user study to evaluate Whoo.ly, and we found that (1) our event detector accurately identified events and (2) the local residents who participated in our study found Whoo.ly to be an easier tool for finding hyperlocal information than Twitter. Social media such as Twitter has altered society's information and communication fabric and will continue to be increasingly integrated in our daily lives. We believe this paper presents a promising approach to leveraging Twitter messages to better support hyperlocal community awareness and engagement.

Chapter 6

SPATIAL CROWDSOUCING FOR ENRICHMENT OF EVENT CONTEXT

In this chapter, we focus on how to enrich an event context by gather first-hand information (e.g., photos, videos) about that event from the field using a principled crowdsourced approach. Research on crowdsourcing thus far has primarily concentrated on tasks that can be accomplished entirely *online*, such as image labeling (Von Ahn, 2006), language translation (Shahaf and Horvitz, 2010) and visual recognition (Bigham *et al.*, 2010). Studies have also shown that workers can solve these online tasks more effectively and accurately than a single expert or computer algorithms (Snow *et al.*, 2008). Here, we argue that the crowdsourcing paradigm is also useful be applied to solve *spatial* tasks, i.e., tasks associated with locations such as gather information about an event, in our scenario. More specifically, to collect data about a trending event in a city, the requester (i.e., a journalist) can crowdsource this task (with monetary incentives) to a group of workers who are near these spots and have interests, time, and skills. Once they travel to their assigned spots and complete their tasks there, the collected data (i.e., images, videos) are sent back to the requester.

Spatial crowdsourcing has attracted increasing interests recently (Alt *et al.*, 2010; Kazemi and Shahabi, 2012; Benouaret *et al.*, 2013; Sadilek *et al.*, 2013; Kim *et al.*, 2014). However, most existing solutions are inapplicable to our scenarios due to three reasons: First, they can work only with self-incentivised workers (i.e., volunteers in (Burke *et al.*, 2006; Krause *et al.*, 2008)). The situations in which each worker requires to be rewarded are largely ignored. Second, they assume the crowdsourced task has a uniform utility regardless of its location or the dispatched worker. However, in reality, some locations can be more informative than the others and workers can have different levels of skills. This will becomes a problem especially when we have a limited budget, since only a subset of tasks can be selected and accomplished. Last and most importantly, these studies assume that workers will always accept and complete the task requests. In fact, a lot features can affect a work's decision such as time availability, the task's location and the traveling cost. Therefore, such certainty never exists.

This chapter proposes a generic spatial crowdsourcing framework, *CrowdX* (Hu *et al.*, 2014), for enrich an event context by collecting photos and videos about the event from the field. It harnesses programmatic access to workers like traditional crowdsourcing. More importantly, it avoids the weaknesses discussed in the foregoing by constructing 1) a model to assess the non-uniform *utility* of each task assignment, 2) a model to estimate the expected *cost* of each worker, and 3) a probabilistic model to quantify the worker's *uncertainty*. Before going into details of these models, let's explore a very specific case study of CrowdX based on the motivating scenario mentioned in Chapter 1: A requester (e.g., a journalist) wants to collect photos of a protest parade from various spots in West Phoenix, under a limited budget. Instead of traveling to each spot herself, she posts the task on CrowdX. Consequently, CrowdX generates the requests separately for each potential worker. The request contains a task description (e.g., take 5 photos), directions to the worker's assigned spot S, and an anticipated reward. Once the worker accepts and completes the task request at S, the photos are sent back to the requester who will reward the worker subsequently. In case a spot is miss-covered or new spots pop up, the request can post a new task on CrowdX if the budget is still available, and so forth. There are several challenges in realizing CrowdX. First, due to the limited budget, dispatching all workers to every event spot is infeasible. So on what criteria should potential workers and spots be selected? Next, after selecting workers and spots, how to determine which worker is to be assigned to which spot? Last, it is uncertain whether a worker will accept and complete the request. Since such uncertainty largely affects the worker-spots assignment and the cost estimation, how to quantify it?

We address the first two challenges by a utility-theoretic approach which considers workers and event spots jointly and selects worker-spot assignments that simultaneously maximize the overall utility, and achieve low cost within the given budget. The utility of each worker-spot assignment is assessed based on two properties: *representativeness* (an ideal spot should cover the most representative and diverse aspects of the event at the same time) and *quality* (an ideal worker should complete her task with high quality). In practice, we rely on both location information of the event spots and personal information of the workers, and develop statistical models to measure these two properties and compute the utility. Next, to address the third challenge, we develop a Bayesian model to quantify each worker's uncertainty by predicting the likelihood of her accepting a task request and finishing the task, based on several features such as the distance between her present location and the assigned spot's location. Based on this, we also develop a cost function that takes into account different cost metrics and estimates the expected cost for each work-spot assignment. Finally, balancing utility of user-spot selection and assignment with the need to achieve a probabilistic budget-aware cost efficiently can be formalized as a discrete optimization problem. Exploiting the concept of submodularity, we develop a greedy algorithm which is guaranteed to provide near-optimal solution for this hard problem. We also provide extensive experimental validation of the proposed CrowdX system to show its efficiency and effectiveness.

6.1 Problem Formulation

In this section, we formally define the research problem of this paper using CrowdX's case study as a running example. We denote $\mathcal{U} = \{u_1, ..., u_m\}$ as a ground set of workers and $\mathcal{S} = \{s_1, ..., s_n\}$ as a ground set of event spots. $\mathcal{W}$ is a Cartesian product of these two sets $\mathcal{W} = \{w_{u_1,s_1}, ..., w_{u_m,s_n}\}$, where w_{u_i,s_j} indicates that ith worker u_i is assigned to jth spot s_j. We also use $\mathcal{W}_{\mathcal{U}}$ ($\mathcal{W}_{\mathcal{S}}$) to represent a union of the worker (spots) of $\mathcal{W}$. Now, let $\mathcal{W}' \subseteq \mathcal{W}$ be a finite subset of all worker-spot assignments, and so let $\mathcal{W}_{\mathcal{U}'} \subseteq \mathcal{W}_{\mathcal{U}}$ and $\mathcal{W}_{\mathcal{S}'} \subseteq \mathcal{W}_{\mathcal{S}}$. Any possible subset is associated with a utility $F(\mathcal{W}') \geq 0$, and a cost $C(\mathcal{W}') \geq 0$, where functions $F(\cdot)$ and $C(\cdot)$ will be defined next. Recall in the case study, our goal is to select the best subset of work-spot assignments so as to take photos of an emerging event, in strict accordance with budget constraints on the cost. Accordingly, we model this as an optimization problem

$$\max_{\mathcal{W}' \subseteq \mathcal{W}} F(\mathcal{W}') \quad \text{subject to} \quad C(\mathcal{W}') \leq B \tag{6.1}$$

from some budget $B > 0$. This optimization problem aims at finding the subset which achieves the highest utility and subjects to a budget on the rewards.

6.1.1 Modeling the Utility Function

We assert that the utility of $\mathcal{W}'$ should have two properties: *representativeness* and *quality*. First, inspired by the classic MMR principle (Carbonell and Goldstein, 1998), we assess the representativeness of $\mathcal{W}'$ from two perspectives: First, it is a function of the similarity of the subset spots $\mathcal{W}_{\mathcal{S}'}$ of $\mathcal{W}'$ to the ground set of spots $\mathcal{W}_{\mathcal{S}}$ of $\mathcal{W}$, or as a function representing some form of "coverage" of $\mathcal{W}_{\mathcal{S}}$ by $\mathcal{W}_{\mathcal{S}'}$. For example, given a ground set $\mathcal{W}_{\mathcal{S}} = \{s_1, s_2, s_3, s_4\}$, we call $\mathcal{W}_{\mathcal{S}'_1} = \{s_1, s_2\}$ has better coverage than $\mathcal{W}_{\mathcal{S}'_2} = \{s_3, s_4\}$ if $Sim(\mathcal{W}_{\mathcal{S}'_1}, \mathcal{W}_{\mathcal{S}}) > Sim(\mathcal{W}_{\mathcal{S}'_2}, \mathcal{W}_{\mathcal{S}})$. As a result, we only need to dispatch workers to spots s_1 and s_2 since what happens at these two places are informative enough to cover the whole event. On the other hand, the representativeness of $\mathcal{W}'$ is also related to the "diversity". This is because an event often has multiple aspects. For example, two nearby spots s_1 and s_3 represent one aspect of the event and what happens at s_2 alone represents another. Therefore, we call $\mathcal{W}_{\mathcal{S}'_1} = \{s_1, s_2\}$ has better diversity than $\mathcal{W}_{\mathcal{S}'_2} = \{s_1, s_3\}$ since its event spots capture more aspects of the event.

The second property of the utility of $\mathcal{W}'$ is its quality, or more specifically, the quality of the workers' accomplishments. Here, we make a reasonable assumption that every worker will perform equally at all of their assigned spots. Then, the quality of $\mathcal{W}'$ will only depend on two factors: 1) the quality of the worker herself and 2) whether she will accept the request and complete the task. Since our case study is about collecting photos of an event, we measure the first factor in terms of how skilled the worker is at taking photos. So, for the same spot s_1, the worker-spot assignment $\mathcal{W}'_1 = \{w_{u_1,s_1}\}$ can achieve higher utility than $\mathcal{W}'_2 = \{w_{u_2,s_1}\}$ if u_1 is more skilled than u_2. Next, for the second factor, we will present a probabilistic model in Sec. 3.2 to quantify the uncertainty of the worker's actions.

Finally, merging the notations of representativeness and quality into one measure, we define the utility of $\mathcal{W}'$ as:

$$F(\mathcal{W}') = L(\mathcal{W}') + \alpha R(\mathcal{W}') + \beta Q(\mathcal{W}') \tag{6.2}$$

where $L(\mathcal{W}')$ measures the coverage, $R(\mathcal{W}')$ rewards the diversity, and $Q(\mathcal{W}')$ quantifies the quality of the workers, $\alpha \geq 0$ and $\beta \geq 0$ are trade-off coefficients. Below, we present the definitions for each term.

6.1.1.1 Coverage function

As we discussed above, $L(\mathcal{W}')$ measures the similarity between the spots of $\mathcal{W}'$ and the spots of $\mathcal{W}$. So formally we have:

$$L(\mathcal{W}') = \sum_{i \in \mathcal{W}_{\mathcal{S}'}, j \in \mathcal{W}_{\mathcal{S}}} Sim(i, j) \tag{6.3}$$

where $Sim(,)$ is the similarity function for spot i from the subset and j from the ground set. There are many ways to measure this similarity. Here, we measure it based on the distance between the spots (e.g., Manhattan distance). We say i is more similar to j than another spot k if i and j are geographically closer. As a result, i can cover j better than covering k. Of course, one can consider other similarity measurements such as textual similarity for i and j based on their text descriptions (e.g, tweets that mention i and j).

6.1.1.2 Diversity function

Next, we define the diversity function $R(\mathcal{W}')$ as:

$$R(\mathcal{W}') = \sum_{i \in \mathcal{W}_{\mathcal{S}'}} \sum_{j=1}^{K} Sim(i, P_j) \tag{6.4}$$

where $P_j, j = 1, ...K$ is a partition of the ground set $\mathcal{W}_{\mathcal{S}}$ into separate disjoint clusters. We can apply any clustering algorithm (e.g., K-means) to generate those clusters on a 2D map. Then, the similarity function $Sim(i, P_j)$ is defined as the mean distance of a spot i to every element of cluster P_j. So i and P_j are more similar if their distance is shorter. Besides, $R(\mathcal{W}')$ rewards diversity in that there is usually more benefit to selecting a spot from a cluster not yet having one of its elements already chosen. Also, $R(\mathcal{W}')$ is distinct from $L(\mathcal{W}')$ in that $R(\mathcal{W}')$ might wish to include certain outlier material that the coverage function $L(\mathcal{W}')$ could ignore.

6.1.1.3 Quality function

Recall that our assumption that the quality of each worker-spot assignment in $\mathcal{W}'$ only depends on the quality of the worker herself and whether she will accept and complete the task. So essentially, we have to compute the *expected quality of the assignment* in $Q(\mathcal{W}')$ as:

$$Q(\mathcal{W}') \equiv \mathbb{E}[Q(\mathcal{W}')] = \sum_{w_{u,s} \in \mathcal{W}'} P_{u \to s} \cdot q(u) \tag{6.5}$$

where $P_{u \to s}$ is the probability that worker u will accept the request, travel to spot the assigned spot s from her present location, and complete the task there (we will define this probability in next section). Besides, $q(u)$ represents the skill level u has at taking photos. In our context, this is specified by the worker herself in CrowdX.

6.1.2 Modeling the Cost Function

Central to CrowdX is the coupling of maximizing the overall utility of assigning worker to a set of event spots, with constraints defined by limited budgets on rewarding these workers after they complete their tasks. We achieve this coupling by introducing a cost function C which associates each worker-spot assignment $w_{u,s}$ with a non-negative cost $c(u, s)$. Note that the cost $c(u, s)$ is incurred only if the worker u accepts and completes the task request at s (since now the requester has to reward u). As we mentioned earlier, however, these actions are uncertain. Therefore, we have to compute the *expected cost* as:

$$C(\mathcal{W}') \equiv \mathbb{E}[C(\mathcal{W}')] = \sum_{w_{u,s} \in \mathcal{W}'} P_{u \to s} \cdot c(u, s) \tag{6.6}$$

Now, let's define the nonnegative assignment cost $c(u, s)$ under different situations. Obviously, the simplest case is to consider the unit cost, i.e., $c(u, s) = 1$, then cost

$C(\mathcal{W}') = \sum_{w_{u,s} \in \mathcal{W}'} P_{u \to s}$ is equal to the total probability of the workers accepting the requests and completing the tasks. Of course, by defining more complex cost functions, we can make our problem conform to more general, and expressive policies. For example, we might limit a worker's travel distances. So we can define $c(u, s) \propto c_u \cdot \exp(Dist(u, s))$, i.e., the cost is exponentially proportional to the distance between worker u's present location and spot s. And c_u can be seen as the base rate for u such as "50 cents per mile". As a result of this cost function, CrowdX aims to minimize the incurred cost by assigning u to nearby spots if possible.

6.1.3 Quantifying the uncertainty

Next, the more interesting part is to define and estimate $P_{u \to s}$, the probability underlying user u accepting the request and completing the task subsequently at spot s. Such uncertainty plays an important role in both the quality function (Eq. 5) and the cost function (Eq. 6). Clearly, there are many features that can affect this probability. For example, whether u will accept and complete the task request can depend on the distance between her present location and s (i.e., how long she needs to travel), time requirement of the task (e.g., u needs to visit s in 10 minutes) and of course, the amount of rewards she will get (if she accepts and completes the task). Taking these features into account, we model $P_{u \to s}$ as a conditional probability:

$$P_{u \to s} = P(AC = 1 | \mathcal{X}) \tag{6.7}$$

where $AC = 1$ indicates u will accept the request and complete the task subsequently at the assigned spot s ($AC = 0$ indicates otherwise), $\mathcal{X}$ is a set of features that may affect the outcomes of AC. We can rewrite Eq. 7 as:

$$P(AC = 1 | \mathcal{X}) = \frac{P(\mathcal{X} | AC = 1) P(AC = 1)}{\sum\limits_{i=0,1} P(\mathcal{X} | AC = i) P(AC = i)} \tag{6.8}$$

Next, we present our approaches to estimate the likelihood $P(\mathcal{X}|AC = 1)$ and the prior $P(AC = 1)$ of Eq. 8. Note that the counterparts $P(\mathcal{X}|AC = 0)$ and $P(AC = 0)$ can be estimated in the same way and thus are omitted due to the space limit. We start with the likelihood $P(\mathcal{X}|AC = 1)$. As we discussed earlier, many features can affect this likelihood. Here, we consider three typical ones in $\mathcal{X}$, namely, $Dist(u, s)$ – the distance between u's location and s; Δt – the task's time requirement and $c(u, s)$ – the anticipated reward. One can imagine that the closer the worker is to spot s, the more time the assigned task permits, or the higher reward will lead to a higher likelihood of u accepting the request. Based on an independence assumption between travel distance, time, and reward, we further decompose the likelihood into two parts:

$$\begin{aligned} P(\mathcal{X}|AC = 1) &= P(Dist(u, s), \Delta t|AC = 1) \\ &\quad \cdot P(c(u, s)|AC = 1) \end{aligned} \tag{6.9}$$

We begin with the estimation of the first term of Eq. 9, the likelihood of travel distance under required time given the worker will accept and complete the task. We say a worker u is likely to accept and complete the task as long as he can reach spot s more or less in time. In other words, we wish $Dist(u, s)/v_u \leq \Delta t + \varepsilon$, where $Dist(u, s)/v_u$ is the travel time for u to reach s, v_u is the worker u's current travel speed, and ε is a time buffer. So essentially, we can compute this likelihood as:

$$P(Dist(u, s), \Delta t|AC = 1) = P\left(v_u \geq \frac{Dist(u, s)}{\Delta t + \varepsilon}\right) \tag{6.10}$$

To estimate this probability, we follow the basic probability principles and compute:

$$P\left(v_u \geq \frac{Dist(u, s)}{\Delta t + \varepsilon}\right) = \int_{v^*}^{\infty} f(\tilde{v}) d\tilde{v} \tag{6.11}$$

where $v^* = \frac{Dist(u,s)}{\Delta t + \varepsilon}$, $f(v)$ is the probability density function (PDF) for the distribution of travel speed v. Extensive research in the transportation research community (Rakha *et al.*, 2010; Emam and Ai-Deek, 2006; Westgate *et al.*, 2013; Wang *et al.*, 2012) has found that the traffic/vehicle speed typically follows a log-normal distribution. Given the cumulative distribution function (CDF) of a log-normal distribution $\mathcal{LN}(x;\mu,\sigma)$ is $\int \mathcal{LN}(x;\mu,\sigma) = \frac{1}{2} + \frac{1}{2}\text{erf}(\frac{\ln x - \mu}{\sqrt{2}\sigma})$, where erf() is the Gauss error function (Abramowitz and Stegun, 1972), we can solve Eq 11. as:

$$\int_{v^*}^{\infty} f(\tilde{v})d\tilde{v} = \frac{1}{2} \cdot \left(1 - \text{erf}\left(\frac{\ln a - \mu}{\sqrt{2}\sigma}\right)\right) \tag{6.12}$$

where $a = \Delta t + \varepsilon$. Next, to estimate the second term of Eq. 9 we simply define:

$$P(c(u,s)|AC = 1) = \frac{\#T(u, AC = 1 \wedge \geq c(u,s))}{\#T(u, AC = 1)} \tag{6.13}$$

Here $\#T(u, AC = 1)$ denotes the number of task requests that were accepted and completed by u before, where $\#T(u, AC = 1 \wedge \geq \$)$ denotes the number of times u accepted a request with its task reward greater or equal to the cost $c(u,s)$.

Finally, we estimate the prior knowledge $P(AC = 1)$ of user u accepting and completing her task request without knowing the features:

$$P(AC = 1) = \frac{\#T(u, AC = 1)}{\#T(u,)} \tag{6.14}$$

where $\#T(u,)$ is the total number of tasks u received. In our context, all counts $\#T(u)$ are learnt directly from the worker history in CrowdX.

6.2 Algorithm and Workflow in CrowdX

We first present an algorithm for solving Eq. 1. Note that it is NP-hard and the exact solution is intractable [1] . Fortunately, we can prove that Eq. 1 contains a unique property which allows us to obtain provably near-optimal solutions. Recall our goal behind Eq. 1 is to select a subset of worker-spot assignments to optimize value of information subject to budget constraints. Intuitively, this selection problem satisfies the following diminishing returns property: The higher the utility worker-spot assignment already selected, the less the addition of a new assignment helping us. This property can be formalized by the concept of *submodularity.* A set function f defined on subsets of V is called submodular, if $f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$ for $A \subseteq B \subseteq V$ and $s \in V \setminus B$. We can prove that the utility function F of Eq. 1 is submodular: its coverage function L and diversity function R clearly are submodular because they both have diminishing effects on additional worker-spot assignments. Besides, we can prove the quality function Q is modular (i.e., $Q(A \cup \{s\}) - Q(A) = Q(B \cup \{s\}) - Q(B)$). Since a sum of submodular/modular functions is also submodular (Lovász, 1983), then F is submodular.

Exploiting the concept of submodularity, and inspired by (Khuller *et al.*, 1999) we propose Algorithm 1 to solve Eq. 1. This greedy algorithm (line 2-7) sequentially finds a particular worker-spot assignment w^* ($w^*, w \in \mathcal{W}$) with the largest ratio of utility function gain to the cost, i.e., $F(G \cup \{w\}) - F(G)/c(w)$. If adding w increases the utility while not violating the budget constraint, it is then selected. The most important aspect of a greedy algorithm is the design of its greedy heuristic. As discussed in (Khuller *et al.*, 1999), the greedy heuristic in line 3 has an unbounded approximation factor. For example, let $V = \{a, b\}$, $F(\{a\}) = 1$, $F(\{b\}) = p$, $w_a = 1$,

[1] Eq.1 can be reduced to Budgeted Maximum Coverage problem which is known to be NP-hard (Khuller *et al.*, 1999)

Algorithm 3: Greedy Algorithm

```
1 $G \leftarrow \emptyset$, $U \leftarrow \mathcal{W}$
2 while $U \neq \emptyset$ do
3     $w^* \leftarrow \arg\max \frac{F(G \cup w) - F(G)}{c(w)}$
4     if $\sum_{i \in G} c(w_i) + c(w^*) \leq B$ then
5         $G \leftarrow G \cup \{w^*\}$
6     $U \leftarrow U \setminus \{w^*\}$
7 end while
8 $v^* \leftarrow \arg\max_{v \in \mathcal{W}, c(v) \leq B} f(\{v\})$
9 return $G_f = \arg\max_{S \in \{v^*, G\}} F(S)$
```

$w_b = p + 1$, and $B = p + 1$. The solution obtained by the greedy heuristic is $\{a\}$ with objective function value 1, while the true optimal objective function value is p. The approximation factor for this example is then p and thus unbounded. We address this issue by adding line 8 and 9. So after the sequential selection, set G is compared to the within-budget singleton with the largest utility value, and the larger of the two becomes the final output, i.e., the cost-effective assignments that achieve the highest utility. This step ensures that we can obtain a constant approximation factor (see Theorem 1).

Theorem 1. *Algorithm 1 achieves a (1-1/e)-approximation.*

Proof. Omitted due to space limit. Please refer to (Khuller *et al.*, 1999) for a similar proof. □

So Theorem 1 basically states the solution G_f found by Algorithm 1 can be at least as good as $(1 - 1/e)$ of the optimal solution even in the worst case. It is also easy to prove the computational cost for Algorithm 1 is $O(|\mathcal{W}| \log |\mathcal{W}|)$, which is quite feasible for even for large datasets.

6.2.1 Workflow of CrowdX

Next, we describe the workflow in CrowdX for a requester collecting photos of a breaking event. Here, the inputs are a set of event spots $\mathcal{S}$, a set of workers $\mathcal{U}$, and a budget B. The output is a collection of event photo taken by the workers at their assigned event spots.

Step 1 Given $\mathcal{U}$, $\mathcal{S}$ and B, CrowdX finds the best worker-spot assignment $\mathcal{W}'$ using Algorithm 1.

Step 2 For each assignment $w_{u,s} \in \mathcal{W}'$, CrowdX sends u a task request. u is required to respond (accept or reject) within a limited time t. After t, any unresponsive request will be discarded. We denote $\mathcal{U}' \subseteq \mathcal{U}$ as a set of workers who accept and complete task requests.

Step 3 The requester will reward each $u \in \mathcal{U}'$ with reward $c(u)$ once u's photos are received, and update the budget such that $B = B - c(u)$. After that, in case the requester wants to collect photos from a missing spot or new event spots (which just popped up), CrowdX will first check if $B > 0$ (adding new budget is also allowed). If so, CrowdX will then update $\mathcal{U}$ with new spots and S with new worker information. Finally, with updated $\mathcal{U}$, $\mathcal{S}$ and B, CrowdX will go back to Step 1 and so forth.

6.3 Evaluation

In this section, we examine the performance of the greedy algorithm (Algorithm 1) of CrowdX, against other baselines. Two main tasks are undertaken: 1) scalability, and 2) robustness w.r.t different task budgets, task's requirements, and workers' costs.

6.3.0.1 Datasets and Experimental Setup

We use both real world data and synthetic data for our experiments. The real world data is obtained from Gowalla (Gowalla, 2010), a popular location-based social network, where users are able to check in to different spots. Each check-in record includes a user's check-in time and the location (GPS coordinates) of the spot. Here, we assume Gowalla users are the workers of our CrowdX system, and we use the location of the user's last check-in of the day as her present location. In our experiment, we randomly select workers (i.e., Gowalla users with their locations) of a random day from a bounding box of 50 miles by 50 miles in the San Francisco area. Moreover, the event spots (and their locations) are also randomly generated for the same bounding box. Recall that the travel distance of a worker u to an event spot s plays a vital role in estimating the worker's uncertainly ($P_{u \to s}$, see Eq. 7). To estimate this distance for the real dataset, we call Google Directions API to retrieve the actual distance given the origin (i.e., the worker's GPS coordinates) and the destination (i.e., the spot's GPS coordinates). Finally, our real dataset has 200 workers, 200 spots, and 40,000 worker-spot assignments and their actual distances. [2] . On the other hand, we also generate a cheaper and much larger scale synthetic dataset which has 20,000 workers and 20,000 spots. The distance between a worker's location to a spot's location is then measured by manhattan distance. For both datasets, we randomly generate a worker's task history in CrowdX, which includes the total number of task requests the worker received, the number of tasks requests the worker accepted and completed, and the rewards of these tasks. We then use this information to estimate the priors for each worker (see Eq. 13 and 14). Last, we also simulate workers' levels of skills at taking high quality photos from 1 to 10 where 10 is most skilled.

[2] note that Google's API quota limited us by making this dataset bigger, we remedy the size problem in the synthetic dataset

There are several parameters we need to setup. First, given that the transportation research community has researched extensively on applying lognormal distribution on the real world data to model traffic patterns and vehicle speeds (Jenelius and Koutsopoulos, 2014; Westgate *et al.*, 2013; Wang *et al.*, 2012), we directly adopt their results and set $\mu = 2$ and $\sigma = 1.5$. Second, For α and β (trade-off coefficients, see Eq. 2), coarse parameter tuning was performed. We vary α and β to normalize the coverage, the diversity and the quality since they often have different scales. As a result, we set up $\alpha = 1.87$ and $\beta = 6.15$.

6.3.0.2 Baselines

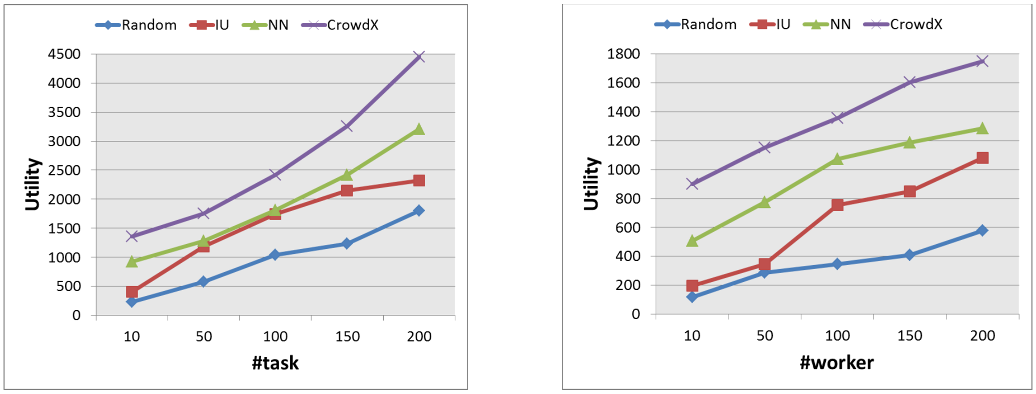
To better understand the performance of our approach, we implemented some competitive baselines:

- *Random*: Let $\mathcal{U}$ be the set of workers and $\mathcal{S}$ be the set of even spots in our datasets, we first build a ground set of $\mathcal{W}$ as a Cartesian product $\mathcal{U}$ and $\mathcal{S}$. Then we *randomly* select a subset of $\mathcal{W}' \subseteq \mathcal{W}$ such that the total cost of $\mathcal{W}'$ is below the given budget. We calculate the cost using the cost function defined in Eq. 6.

- *Ignore uncertainty* (IU): We use the same joint Cartesian product $\mathcal{W}$ as the input here. The basic idea of this approach is that it assumes every worker will accept and complete the task request (a common assumption in present spatial crowdsourcing literature). Therefore, when computing the utility $F(\mathcal{W})$, we ignore the uncertainty and set $P_{u \to s} = 1$ in Eq. 5 and Eq. 6.

- *Nearest neighbor* (NN): Instead of modeling workers $\mathcal{U}$ and spots $\mathcal{S}$ jointly, this approach solves a similar optimization problem as Eq. 1 with redefined utility and quality function for spots only: $F(\mathcal{S}') = L(\mathcal{S}') + \alpha R(\mathcal{S}') + \beta Q(\mathcal{S}')$ where the quality of each spot depends on its k-nearest workers $\mathcal{U}_s^k$, i.e., $Q(\mathcal{S}') = \sum_{s\in\mathcal{S}'}\sum_{u\in\mathcal{U}_s^k} P_{u\to s}\cdot q_u$, where q_u represents each worker's quality/skills. Similarly, we redefine the cost function as $C(\mathcal{S}') = \sum_{s\in\mathcal{S}'}\sum_{u\in\mathcal{U}_s^k} P_{u\to s}\cdot c_u$, where c_u represents each worker's cost.

6.3.1 Scalability



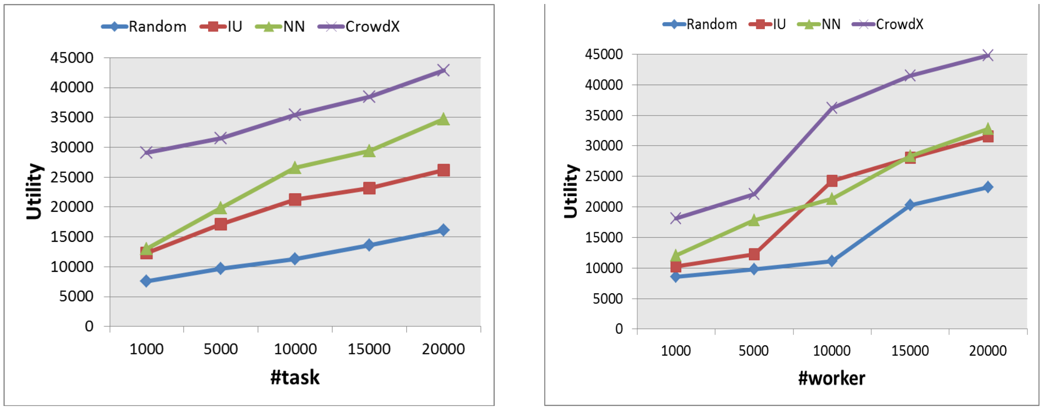
(a) Varying #tasks, when #user = 200

(b) Varying #workers, when #task=50

Figure 6.1: Performance on the Real Dataset

We first evaluate the scalability of CrowdX by varying 1) the number of spatial tasks (i.e., event spots), and 2) the number of workers, against three baselines using both datasets. Note that, the total budget is proportional to number of tasks (we will explore the effect of varying the budget later). The results are depicted in Fig 1–2. Several observations can be made. First, for all baselines, the utility increases as the number of tasks and the the number of workers grow. Second and more importantly, it is also clear that our method performs better than baselines consistently, by gaining 20% to 45% for both datasets. It is worth noting that, among all the baselines, NN performs better than IU and Random in most cases. This makes sense because both CrowdX and NN take into account the worker's uncertainty, whereas IU and

Random largely ignore it. Besides, it is also interesting to highlight the reason behind CrowdX's superiority over NN. Note that NN only considers k-nearest neighbors U_s^k for each spot $s \in \mathcal{S}'$. In other words, NN considers worker-spot assignment with size $k \times |\mathcal{S}|$. However, since CrowdX consider jointly modeling worker-spot inherently, it essentially models $|\mathcal{U}| \times |\mathcal{S}|$ combinations (where $|\mathcal{U}| \gg k$) which results a much larger search space than NN. Therefore, CrowdX can often outperform NN especially when the neighbors of a spot are low quality.



(a) Varying #tasks, when #user = 1000

(b) Varying #workers, when #task = 500

Figure 6.2: Performance on the Synthetic Dataset

6.3.2 Robustness

Next, we evaluate the robustness and effectiveness of our proposed Algorithm 1 by exploring the impact of three important parameters: budget B, time requirement Δt and base cost rate c_u on affecting the total information utility. Note that the following experiments are conducted on the real data with $\#spot = 50$ and $\#worker = 200$. Besides, we also set the time buffer $\varepsilon = 10$ minutes.

6.3.2.1 Varying budget B

We first consider varying the budget B from \$10 to \$100 while fixing $c_u = 0.5$ and $\Delta t = 30$ minutes. Fig. 3 shows the results w.r.t the total utilty. Clearly, when the requester has more budget, she can always get better results, i.e, the utility of the results obtained by all approaches grows with the budget, with CrowdX again performs of the best. Interestingly, we also observe that the utility of CrowdX increases near exponentially after $B >$ \$50. One explanation could be: recall the cost function for a given user u and a spot s is exponentially proportion to the distance u will travel to complete her assigned task at s. Therefore, when $B > 50$, the worker can visit the event spots which are very far away from her present location whose utilities maybe significantly higher than the nearby spots that she has already visited (when $B \leq 50$).

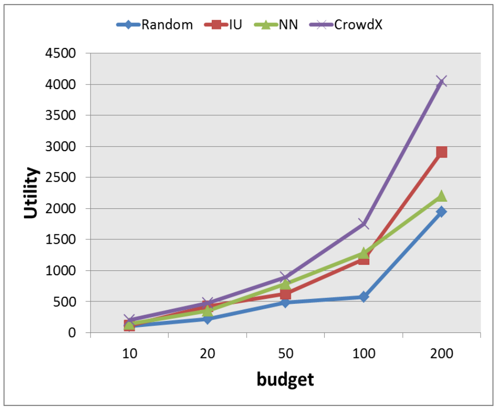


Figure 6.3: Varying Budget B $w.r.t$ Total Utility

6.3.2.2 Varying time requirement Δt

Next, we vary the time requirement Δt from 5 minutes to 90 minutes while fixing $c_u = 0.5$ and $B =$ \$100. The results are shown in Fig. 4. Again, we observe that all approaches can achieve much higher utility when Δt increases. This is because the travel likelihood $P((Dist(u,s),\Delta t|AC=1)$ becomes much higher when worker u has more time allowances (i.e., the time she can spend on her trip to spot s) given a fixed travel distance $Dist(u,s)$ (recall Eq. 10). As a result, the probability that u will accept and complete the task will also become higher, implying more workers are likely to accept and complete more tasks, which in turn will result a much higher total utility.

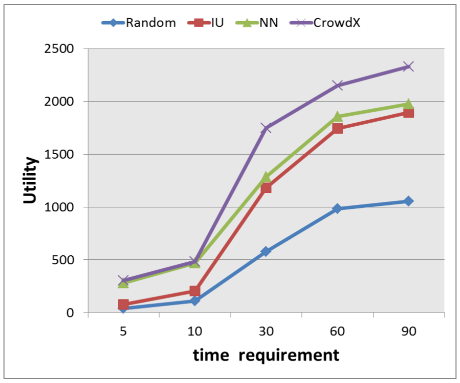


Figure 6.4: Varying Time Requirement Δt *w.r.t* Total Utility

6.3.2.3 Varying base cost rate c_u

Finally, we explore the performance of all approaches w.r.t to increasing the base cost rate (i.e., cost per mile, recall Eq. 6) while fixing $\Delta t = 30$ minutes and $B = \$100$. The results are shown in Fig. 5. Interestingly, the total utility significantly decreases when we raise $c(u,s)$ (especially when $c_u \geq 1$). This can be seen as the "reverse" result of varying budget in Fig. 3: due to the cost which is exponentially proportional to the travel distance, a fixed budget can quickly "burn out". As a result, one can image only a few spots will be visited, resulting a low total utility.

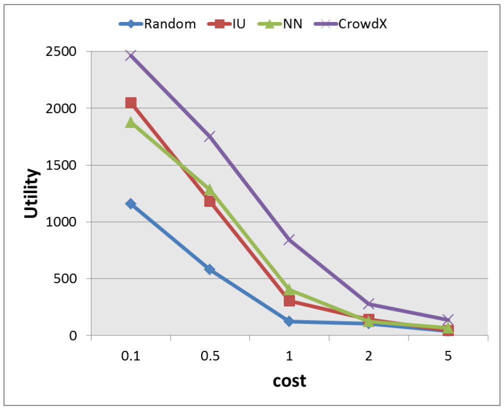


Figure 6.5: Varying Base Cost Rate c_u $w.r.t$ Total Utility

6.4 Summary of Chapter

In this chapter, we have described a spatial crowdsourcing system CrowdX for handling the Event Enrichment task. To achieve the best utility of assigning a worker to an event spot, CrowdX takes into account three features: 1) the coverage and diversity of the spot, and the quality of the worker, 2) the expected cost of each assignment, and 3) the uncertainty of the worker. We designed an efficient utility-theoretic approach that finds a set of user-spot assignments that simultaneously maximizes overall utility, and achieves low cost in strict accordance with budget constraints. Based on the experimental results, our model shows significant improvements over the baseline methods. We believe that our work presents the first step towards building a generic spatial crowdsourcing framework for solving spatial tasks. In the future, we want to extend our framework by supporting more complex tasks and more quality control mechanisms.

Chapter 7

MAKING SENSE OF PEOPLE'S ENGAGEMENT IN REAL-WORLD EVENTS ON SOCIAL MEDIA

Previous chapters have aimed to develop tools for automatic event characterization based on event topics and the crowd's tweeting behaviors (Chapter 3) and based on the sentiment analysis of the crowd's tweet content (Chapter 4), real-time event recognition of events from social media (Chapter 5), and sensing and gathering information to enrich events using the wisdom of crowds (Chapter 6). These tools, as encapsulated in the EventRadar toolbox, provide various effective ways to analyze an event using its social media responses. Enabled by the EventRadar toolbox, we are able to conduct more in-depth event analysis and reveal deep insights into people's behavior when engaging with events and understand the motivation behind their behavior – the main focus of this chapter. Our discoveries contribute to the scientific contributions of this dissertation.

As we have discussed earlier, recent years have witnessed a growing interest in research and industry that aims to develop various computational tools for event analysis. Unfortunately, little is understood thus far about the factors that affect people's engagement with real-world events on social media (e.g. posting or exchanging event-related tweets). *Does a person post tweets about an event because they are interested in the topics pertaining to that event?* Or *are they instead engaged because their friends are also posting tweets about it?* Or *is their engagement a reflection of the fact that this is a local event happening in their neighborhood?* Furthermore, *how and to what extent do the different topics of events affect the degree of a person's engagement*? Answering these questions holds the key to developing a wide range of applications in marketing, political campaigns, and citizen journalism (Dahlgren, 2009; Gil de Zúñiga *et al.*, 2012). Consider this: a personalized event recommendation engine can automatically recommend a list of new events (as they are happening and trending on Twitter) to a person, based on a prediction of that a person's Twitter engagement – this can help people learn about and engage with more such events.

In this chapter, we aim to answer the questions put forth previously by exploring multiple predictive variables, and quantifying their potential influence on predicting a person's *presence* and *degree* of Twitter engagement with various real-world events. Specifically, we operationalize a person's Twitter engagement with a real-world event as the posting of tweets about that event, including retweets and replies related to the event. The presence of the person's Twitter engagement in response to an event can be defined as the existence of at least one tweet (or RT or mention) that mentions that event. The degree of the person's Twitter engagement is measured by the number of tweets posted regarding that event; more such tweets indicate that the person is more engaged. Next, inspired by prior theoretical constructs that bridge social science, linguistics, and computer mediated communication, we collect factors that could potentially affect the person's Twitter engagement in a real-world event from five broad categories. These are: (i) *Twitter activities* (i.e., her activities on Twitter prior to her current engagement with the event), (ii) *Tweets' topics* (i.e., the topical interests extracted from her prior tweets), (iii) *Twitter user types* (i.e., the Twitter user category she falls into), (iv) *Geolocation* (her geographical proximity to the event), and (v) *Social network structure* (her Twitter followers, following, and common neighbors).

We map these dimensions into 17 numeric predictive variables manifested on Twitter, spanning the volume of tweets produced, burstiness of tweets, frequency of retweets, usage of hashtags, communication mode (direct versus broadcast), topical interest extracted from the person's tweets and those of their following lists, her approximate location and geographic proximity, and social network structure. We construct two statistical models to assess the relative contributions of these variables towards predicting the presence of a person's Twitter engagement and the degree of that Twitter engagement in 406 real-wold events.

Using our models, several insights about the factors and their influence on predicting the presence and degree of people's Twitter engagement in real-world events are revealed. For example, in terms of the presence of engagement, we find that, among all the predictive factors, *a person's prior Twitter activity* and *the person's social network* most significantly impacts the presence of the person's engagement with events. For example, having posted more tweets in the past, or following Twitter accounts that are dedicated to news leads to people's engagement. Similarly, we also find that measures of topical interest have strong and statistically significant levels of impact on a person's degree of engagement, especially during political events.

This chapter is organized as follows: we begin with a comprehensive survey of related works that inspired this work. Next, we present the data collection policy, followed by a detailed description of our statistical models. We present the results of our models with a discussion at the end of this chapter.

7.1 Background

7.1.1 Twitter and Real-World Events

As social media has become prominent in daily life, the evolving ways in which information is generated, viewed, shared, and exchanged have inevitably transformed people's activities when engaging with real-world events (Lenhart *et al.*, 2010). Recent years have witnessed a growing research interest in developing tools for real-time identification of real-world events and the tweets associated with those events (Sakaki *et al.*, 2010; Becker *et al.*, 2011) (we also developed DeMa event detection algorithm (Hu *et al.*, 2013b) and presented it in Chapter 5. On the one hand, the vast number of tweets posted around an event enrich the user experience of that live event. However, that very same scale poses tremendous challenges for methods that attempt to extract sense from those tweets. Various research efforts have thus focused on how to make sense of tweets that are posted in response to real-world events, including inferring the events' structure of events from tweets volume (Shamma *et al.*, 2009), visualizing events using tweets (Diakopoulos *et al.*, 2010), and sentiment analysis of tweets to understand events (Diakopoulos and Shamma, 2010a).

7.1.2 Making Sense of Twitter Engagement in Real-World Events

To date, we are aware of little research that directly addresses the issue studied in this chapter; however, there has been an extensive amount of literature on exploiting predictive factors on social media for various other issues. For example, Golder et al. (Golder and Yardi, 2010) and Kivran-Swaine et al. (Kivran-Swaine *et al.*, 2011) focused on factors that influenced tie formation/tie break-up; Gilbert and Karahalios (Gilbert and Karahalios, 2009) focused on factors that affect tie strength. Moreover, (Suh *et al.*, 2010) examined factors that affect retweets; and Hong et al. (Hong *et al.*, 2011) explored factors that influence the popularity of tweets.

Our effort differs from the past work in that we are exploring factors that may affect people's Twitter *engagement* in response to *real-world events*. Below, we discuss related work showing how a person's prior Twitter activities (e.g., communicating with others), her tweets' content (e.g., topical interests, linguistic styles), her geographical location, and her social networks relate to her Twitter engagement with real-world events. Inspired by these works, we collect corresponding predictive factors and build statistical models upon those factors to assess their contributions to predicting the presence and degree of a person's event engagement on Twitter.

7.1.3 Social Capital and Event Engagement

Social capital has been identified as a collection of resources that either an individual or an organization gains through a set of communal norms, networks, and sanctions (Bourdieu, 1986; Putnam, 1995; Wellman and Wortley, 1990). The relationship between social capital and event engagement (as a special case of civic engagement) has long been a research topic (Putnam, 1995; Gil de Zúñiga *et al.*, 2012; Shah, 1998). In particular, Putnan found that engaging in events (e.g., participating in events, discussing events with others) can be seen as a precursor to social capital, and that it offers a forum for facilitating the growth of social capital (Putnam, 1995). Many other researchers have shared similar views by claiming that social capital is created when civic engagement is "excited" by events and directed toward a particular end or purpose (Hyman, 2002). In addition, prior research has also identified several kinds of social activities and behaviors that can affect social capital on social media. These include directed communications with targeted individuals (e.g., Facebook private messages; Twitter replies, mentions, and favorites), broadcast communications which are not targeted at anyone in particular (e.g., Facebook wall updates or tweets with no "@" in them), and passive consumption of content (Burke *et al.*, 2011). Moreover, the volume of social media posts (e.g., total number of tweets in a period) and the posting rate have also been shown to influence social capital (Hutto *et al.*, 2013).

However, following Putnan's views on event engagement, we posit that instead of affecting social capital directly, these activities may first affect people's *engagement* in events on social media. Such engagement will *in turn* affect social capital. Therefore, such social activities and Twitter event engagement are closely related. Here, we empirically test whether a person's prior Twitter activities help in predicting their engagement with an event.

7.1.4 Tweet Topics and Event Engagement

The "endurability" theory (Read *et al.*, 2002) shows that people are likely to remember a good experience and are willing to repeat it. Application of this theory to Twitter event engagement indicates that a person may be more likely to get engaged in an event if the topics related to that event are the same as – or at least similar to – the topic that the person is interested in on Twitter (which can be inferred from their previous tweets). For example, assume that Sam is a sports fan and had previously posted a large number of tweets in response to various sports-related events (while they were live on the air). Subsequently, given a football game event and a movie awards ceremony event, both of which are happening currently and trending on Twitter, Sam is more likely to be engaged with the former rather than the latter. Besides, a person's topical interests can be inferred not just from their posted tweets, but from other resources as well. The principle of *homophily* asserts that similarity between individuals leads to a greater potential for interpersonal connections; when establishing connections, people tend to build relationships with others who are like them (McPherson *et al.*, 2001). Sharing interests with another person is one form of similarity (Feld, 1981) that can be used to build relationships; this can lead to the follow relationship being established.

Based on these, we posit that a person's topical interests can be inferred from the people they are following. Returning to our running example – although Sam may only have posted a few tweets about sports and/or football, the fact that Sam also follows a number of Twitter accounts that are dedicated to football (e.g. NFL official accounts, football analysts, etc.) makes an engagement with a football event more likely than one with a movie awards ceremony. Here, we empirically study how the topical interests of a person (inferred both from their tweets as well as the users they follow) affect their engagement with events on Twitter.

7.1.5 Twitter User Types and Event Engagement

Naaman et al. (Naaman *et al.*, 2010) found that there are two basic categorizations of Twitter users: *informers*, who share informational content; and *meformers*, who share tweets about themselves. One effective way to distinguish informers and meformers is based on the linguistic styles of their tweets. Those who share information or describe things tend to use more third person pronouns (*She, He, It, They, etc*), while meformers, who post mostly about themselves tend to use first person pronouns (*I, We, Us, etc*) more often.

Here, we posit that informers are more likely to engage in events through the posting or sharing of information than meformers. We explore linguistic styles of tweets and examine whether different types can predict a person's event engagement on Twitter.

7.1.6 Geolocation and Event Engagement

It is known that a person's geographical location (geolocation) significantly affects her social connections and activities in the offline world. Recent research has also found evidence to show that offline geography has a significant impact on user interactions, tie formation, and information diffusion on online social media like Twitter (Kulshrestha *et al.*, 2012). In particular, researchers have discovered that users preferentially connect and exchange information with other users from their own country, and lesser information is exchanged across national boundaries. However, even such transnational links and interactions occur between users in geographically and linguistically proximal countries within their network. Similarly, researchers also identified that geographical proximity plays a key role in trend/innovation adoption (Toole *et al.*, 2012). Based on these results, we *conjecture* that a person's geolocation may affect their engagement with real-world events on Twitter if that person's location is geographically proximate to the event's location (e.g., a user may only care the events that happen in her neighborhood).

7.1.7 Social Networks and Event Engagement

The correlation between social network influence (e.g., network size and social ties) and user engagement has been studied extensively. For example, Zuniga et al. showed that the relationship between online and offline network size and people's engagement with civic events is positive (de Zúñiga and Valenzuela, 2011). They further found that network structure and social ties (especially weak ties) are determined to be strong predictors of the engagement. There are many different ways to form ties on Twitter (Golder and Yardi, 2010), and ties can be formed either directly or indirectly. For example, following a person on Twitter can be seen as a direct tie. In such cases, dyadic properties such as reciprocity play key roles in the process of tie formation. On the other hand, ties can be formed indirectly such as through common network neighbors (known as transitive ties). For example, consider the case where three people form an undirected network: A and C are both friends of B, but A and C are not friends. However, as the number of common neighbors (occurrences of B) between A and C increases, the likelihood of an A-C tie being formed and the corresponding tie strength also increase (Cartwright and Harary, 1956). In this paper we *explore* the extent to which these network sizes and tie formations impact a person's engagement in real-world events as compared to the person's Twitter activities, topical interests, user types and geolocation information.

7.2 Data Collection

In this section, we describe our data collection strategies. Note that in order to show how people's Twitter activities, their tweets' topics, their Twitter user types, their geolocations, and their social networks affect their Twitter engagement with real-world events, we needed to collect: 1) a list of real-world events and their associated tweets, and 2) profiles of Twitter users (who post event-related tweets). Moreover, since we want to evaluate the influence of people's geolocations on their event engagement, we needed to infer the geolocations of both Twitter users and events.

7.2.1 Obtaining Real-world Events and Events' Geolocations

To identify real-world events from tweets, one possible solution is to first obtain an event list directly from newspapers (since reporters often tend to mention the location of the event in their news articles about that event) and then fetch the corresponding tweets. However, such an approach is not applicable for several reasons. First, not every event reported by newspapers is popular/trending on Twitter. As Hong et al. (Hong *et al.*, 2011) pointed out, the popularity of tweets is affected by multiple reasons aside from newsworthiness. Second and more importantly, such an approach will be significantly biased towards larger, more broadly newsworthy events due to the nature of newspapers, which could potentially misguide our analysis. To avoid this, we follow a different approach: 1), by automatically detecting real-world events from Twitter streams, and 2), manually inferring their geographical location, i.e., geolocations later.

To automatically detect real-world events from Twitter, we directly use the DeMa event detection algorithm (more details on DeMa can be found in Chapter 5). After obtaining the real-word event clusters (each cluster refers to one event), we ask annotators to individually read a sample tweet from each real-world event's cluster to gain an understanding of what the event is really about. Then the annotators are asked to find the geolocation of the event cluster via search engines with their own search keywords such as event-related hashtags, timestamps, based on their event understanding. Our assumption here is that many real-world events will be covered by news, blogs, and other media, and their geolocations will often be mentioned. Surprisingly, this simple approach yields a very satisfying result.

7.2.2 Obtaining Twitter Users' Geolocations

Inferring Twitter users' geolocation based on their tweets has been an emerging research topic in recent years (Cheng *et al.*, 2010). In this work, we follow the methods mentioned in (Mahmud *et al.*, 2014) to infer the geolocations of Twitter users [1] . Specifically, the location inference algorithm uses tweet contents, tweeting behavior and other auxiliary information such as time zone to predict the home location of Twitter users. We then verify the extracted location information with the diurnal patterns of the user's tweets (Naaman *et al.*, 2012). For example, most people in New York City will tweet about having dinner and the nightlife between 5:00PM EST to 1:00AM EST. So if a person regularly posts tweets about lunch around 12:00AM EST, they probably are not from the New York City area. Based on our preliminary testing, we found this algorithm together with the diurnal pattern verification yield a stable performance (78.4% for cities).

[1] One may consider to infer a Twitter user's geolocation based on the information from her Twitter profile, i.e., she may mention her location in her profile. However, this method may result unbiased samples since those Twitter users who have relatively more complete profiles, i.e., location, might also be more active in general and tend to have more friends.

7.2.3 Constructing the dataset

In practice, we first obtained about 20 million tweets in English from the Twitter firehose during June to September of 2014. We then used the automated event detection on these tweets to find real-world events. As we mentioned earlier, as the first step in our event detection framework, we ran a hierarchical clustering algorithm to cluster tweets into roughly 19,000 clusters. After that, we used a classifier to classify these clusters into two classes: an event cluster and a non-event cluster. We used 30 human annotators to label both the training set (300 clusters sampled from the entire clusters) and the testing set (478 clusters sampled from the entire clusters) in our study (note that if there is any conflict between annotators, a third party judge will step in to resolve conflicts, i.e., the Fleiss' kappa $\kappa = 1$). After applying a sampling to balance the class distribution (which was skewed towards the non-event class), we trained and validated the classifier using 10-fold cross validation. We tested different classifiers such as logistic regression, SVM, nearest-neighbor, and random forest; and chose SVM because of its performance. As a result of event detection, we obtained 1014 event clusters.

Next, we needed to infer the geolocations of these 1014 event clusters. We asked 20 annotators to individually read a sample of 100 tweets from each real-world event's cluster and find the geolocation of the event cluster via search engines based on their understanding of that the event. This simple approach yielded a very satisfying result: our annotators were able to find the geolocations (on city level) for 40% of the event clusters (N=406). Among these 406 event clusters with geolocations identified, the majority (N=353) events happened in the U.S (e.g., New York City, NY, Beverly Hills, CA, Ferguson, MO), and the rest were in Europe, Middle East and Asia.

Finally, based on 406 events used in this work, we obtained a total number of

10,895 Twitter users who posted at least three tweet in response to one of these events. We applied the location inference algorithm (see above) to predict the location (on City level) of each user. Besides, in order to calculate the measures for the predictor variables, we collected all the tweets posted by each user in the most recent six months preceding their first ever event engagement in any of 406 events.

7.3 Methods

With 406 events obtained, in this section, we provide more details on the statistical models that are used to predict the presence and magnitude of a person's engagement on Twitter with a given event. We first present the dependent variables used in our predictive models, followed by a description of the predictor variables.

7.3.1 Dependent Variables

- *Presence of a person's Twitter engagement in a real-world event:* This is a binary measure that indicates whether or not a person posts, replies to, or retweets tweets in relation to a particular event on Twitter (1: engaged; 0: not engaged).

- *Degree of a person's Twitter engagement in a real-world event:* This is a continuous measure that indicates the number of tweets that the person generates (via post, reply to, or retweet) relating to the event.

7.3.2 Predictor Variables

The literature reviewed in the previous section pointed to five major kinds of predictor variables: Twitter activities, tweets' topics, Twitter user types, geolocation, and social network structure. Using these categories as a guide, we collected 17 variables that are manifested on Twitter as potential predictors of a person's Twitter engagement with a real-world event.

7.3.2.1 Variables related to Twitter activities

- *Total number of tweets.* The total number of tweets a person has posted. These tweets include new posts, retweets, and replies.

- *Maximum tweets per hour.* The maximum rate of tweets per hour, which captures the "burstiness" of a person's activities.

- *Average tweets per hour.* The average rate of tweets per hour, which gives a general idea of a person's level of activity.

- *Directed communications.* The number of tweets with "@" (including both @mentions and @replies) plus the number of favorite tweets divided by the total number of tweets. This measure indicates interpersonal activities between the person and other Twitter users.

- *Broadcast communications.* The ratio of tweets with no "@" at all in the tweet to total number of tweets in a period.

- *Ratio of retweets.* The total number of times a person reposts other Twitter users' tweets, relative to the total number of tweets produced by the person in a period. This measure complements the direct and broadcast communication measures by indicating how often the person interacts with other Twitter users and broadcast their tweets to their own social circle (i.e., their followers).

- *Hashtag usage.* This is defined as the ratio of tweets that contain at least one hashtag to the total number of tweets from a person in a period.

7.3.2.2 Variables related to tweets' content

- *Topical interests from tweets' content.* This measure is calculated as the cosine similarity between two normalized TFIDF vectors: the first is computed based on a person's tweets in a period, while the second is computed based on all the event-related tweets (from other users) posted prior to the person's engagement with that event. In practice, assume a person u has posted T_u tweets in the past three months. Now, assume an event starts at 8:00PM and u gets engaged in this event on Twitter (i.e., user u posts their first event-related tweet) at 8:30PM. Additionally, between 8:00PM and 8:30PM, there are T_Q event-related tweets posted by a set of other users Q. We then compute the topical interests measure as $sim(TFIDF(T_u), TFIDF(T_Q))$. Note that computing the dot product between two TFIDF vectors is a standard technique for measuring topic relevance between text corpora in information retrieval (Manning *et al.*, 2008). Intuitively, higher similarly indicates that the person's prior exhibited topical interests (reflected from their prior tweets' content) are closer to the event's topics (which are inferred from other people's event-related tweets).

- *Topical interests from the person's following list.* This measure is calculated based on the topical similarity between the topics used by the users that a person follows (following list), and the event's topics. More specifically, the following list's topics are computed using methods mentioned in (Burgess *et al.*, 2013): first, given the following list of a person, we obtain the 200 most recent tweets from each user on that list. Next, we distill topic distributions from these tweets using topic models (Blei *et al.*, 2003). On the other hand, as mentioned above, we are also able to get T_Q event-related tweets posted by other users prior to the target person's engagement in the event (i.e., their first event-related tweet). For the tweets T_Q, we run topic models to obtain the same number of topics as the vectors in the previous analysis, and the corresponding topical distribution for each topic. We then compute the similarities between the two topical distributions – one learned from the following list, and the other from the event – using JensenShannon divergence. As with the cosine similarity measure, higher similarity here indicates that the person's topical interests (reflected from the list of people that they follow) are closer to the event's topics.

7.3.2.3 Variables related to Twitter user types

- *Meformer.* This is computed as the ratio of meformer tweets to the total number of tweets by a person in a period. As mentioned earlier, we detected meformer tweets based on linguistic styles. More specifically, if a tweet contains any of the 24 self-referencing pronouns (e.g., words like "I", "me", "we", "us") identified in LIWC, then it is classified as a meformer tweet.

- *Informer.* This is computed as the ratio of informer tweets to the total number of tweets by a person in a period. We identified informer tweets as those containing any of the 20 third-person pronouns (e.g., words like "He", "She", "it", "them") defined in LIWC. In addition, if a tweet contains either a URL, "RT", "MT", or "via", we deem it an informer tweet as well.

7.3.2.4 Variables related to geolocation information

- *Geographical proximity.* The first measure considers the geographical proximity between a person's location and the event's location. As indicated in the previous section, the dataset used in this study only includes Twitter users and events whose geolocations could be identified. Note that it is impractical to model the proximity of two geolocations continuously in terms of their physical distance in miles, because the effect of geographical proximity may not be linearly proportional to physical distance. In fact, such an effect is more likely to be exponential according to recent research (Kulshrestha *et al.*, 2012). Therefore, in practice, we only consider measuring the geographical proximity in terms of two discrete bins: local (distance within $\leq$ 50 miles) and non-local (distance $>$ 50 miles).

7.3.2.5 Variables related to network structure

- *Number of followings.* The number of Twitter users that a person was following (prior to the person's Twitter engagement in an event).

- *Number of followers.* The number of Twitter users who were following the person (prior to the person's Twitter engagement in an event).

- *Followings posted prior.* The number of a person's followings who had already posted event-related tweets before the person posted to that event. As discussed earlier, since following (e.g., A follows B) forms a directed tie, it is possible that the person will be influenced to post tweets when a lot of their followings post about an event prior to their own engagement.

- *Average common neighbor prior.* This measure examines the overlaps between the followings of a person a and the followings of user b, where a has already engaged in the event on Twitter while b has not. In the context of Twitter, a person's followings often represents their interests. Therefore, the common neighbor factor essentially measures the shared interests between two people. According to triadic closure, such a measure also indicates the tie strength between A and B (Wasserman, 1994). In practice, for the user a, we compute this feature as $\sum_{b \in B} CommonNeighborhood(a, b)/|B|$, where B is a set of users who have already engaged in the event ($a \notin B$).

- *Number of followings about news.* This measure is defined as the total number of a person's followings who are deeply involved in news. To identify those news related accounts, we first obtain Twitter profiles for all of the person's friends. We then look at each profile to check which ones contain news related keywords such as "news", "reporter", "journalist", "TV" and so on. We deem those users news related accounts. One motivation for this measure is that news agencies are often authorities and first-hand resources for reporting events. It is possible that if a person followers a lot of news agency accounts, then they will likely be interested in knowing about and engaging with real-world events.

7.4 Results

In the following section, we first provide descriptive statistics for the variables used in our statistical models. Following this, we present the contribution of these variables in predicting the presence and degree of people's Twitter engagement with real-world events.

7.4.1 Descriptive Statistics

Table 7.1 shows descriptive statistics (mean, standard deviation) for the number of events that one person engages with, and the event-related tweets that person posts – along with 17 predictor variables – based on the event data we collected in June to September 2014 (see the 'Data Collection' section). For comparison, we also generate statistics based on an event participant's regular tweets six months prior the the events (i.e., December 2013 to May 2014). We calculate the significance of the difference between these two situations. Note that some of the predictor variables are compared pair-wise, such as topical interests, geographic proximity and so on. Therefore, we only report the pair-wise statistics for the event data.

Also note that we excluded users that were extreme outliers (z-score > 4.0) with respect to our metrics for activity levels and follower/following counts. As a result, we had a total of 10,638 people (we removed 257 "outlier" users from a total of 10,895 users in our dataset, see Data Collection section) participate, by posting Twitter messages over the course of 406 events. Within these messages, 28% of the messages had hashtags, 48% retweets, 27% direct replies, 33% links, and 68% mentions, indicating that the event participants were highly interactive.

7.4.1.1 Twitter activities

On average, a user engaged in 12.1 events over a month, and they posted 2.33 tweets per event. In terms of burstiness, users posted no more than 6.57 tweets within an hour (average). This seems to indicate that over the course of an event, people tend to post using a stable pace (as avg. tweet per hr is very different between tweets from the event a person engaged in and norma tweets from the person's prior tweets history. The Broadcast Communication shows the average number of tweets that are not directed to any specific person. During events, this rate is significantly higher. Such changes are also reflected in directed communication. The ratios of retweets and hashtag usage to the total number of tweets in a period are moderate for the majority of users retweets comprised about 15% of users' messages, and hashtags were used in about 20% of tweets. Compared to these, we witness significant changes during events – where the ratio of hashtags and retweets increases to 44% and 42% respectively. Combining these discoveries, we conclude that people tend to communicate more with others during an event that they are engaged in, thus showing a deeper involvement and engagement with the topics related to that event.

7.4.1.2 Tweet content

In general, users show a fairly diverse range of topics that they post in relation to, which is reflected in and manifested as the relatively low topical similarity to actual event topics. In particular, the topic similarity inferred from a user's tweet content is 0.5, while topic similarity inferred from their followings is 0.3.

7.4.1.3 Tweet user types

Besides, nearly half of users' regular tweets are identified as "meformer" (41%), and the "informer" category accounts for 24% of tweets. However, in the context of event engagement, the percentage value of "informer" tweets witnessed a sharp increase to 43%, and "meformer" tweets decreased to 29%. This indicates that people tend to share more information (e.g., through retweets, third person comments about the event) during the course of an event. However, people do also continue posting information about their thoughts and their presence during the event.

7.4.1.4 Geolocation

In terms of the geographical proximity between the event participants' location and the event's location, we found that most events were non-local to the event participants – this is reflected in that measure's relatively high value (i.e., 418 miles between the inferred event participants' locations and the events' locations), accompanied with high standard deviation (189.8 miles).

7.4.1.5 Social network

The majority of users have an average of 387 followers, and 117 friends. About 4.33 event participants who joined in the event prior to the target user's engagement are the followings of that user. Moreover, for the people who posted prior to the target user but are not part of the following set, it is seen that there are around 10 common friends between those users and the target user. This indicates that one-hop weak ties do exist between event participants. Later we will demonstrate the strength of these predictors.

7.4.2 Prediction of presence

We now turn to the core question examined in this study: *to what extent do the 17 variables used predict the presence, and degree, of a person's Twitter engagement with a real world event?*

In order to examine the relative impact of these variables, we first standardized the measures, and then examined whether they predicted a user's participation / engagement using a repeated measures (406 trials, or events) logistic regression. The question of whether or not the user participated was modeled as a binary dependent variable. Table 7.2 shows the results of this regression. An immediate insight that can be gleaned is that the total tweets posted by a user prior to her event engagement is a significant predictor of whether the user will take up or engage with an event. Specifically, as far as communication oriented tweets are concerned, both directed and broadcast communication are good indicators, albeit in opposite senses. The coefficients for those variables seem to indicate respectively that lower directed communication or higher broadcast communication correlate directly with higher engagement. This is fairly intuitive, since directed communication tends to be among a user's friends and about non-event topics, and in most cases can only be seen by the mentioned users; while broadcast communication is intended for a wider audience consisting of all of the user's followers. Finally, both the ratio of hashtags used and the ratio of retweets are positive indicators of event engagement; this is easy to see since RTs and hashtags respectively are two key ways in which a user can signal their active interest and affiliation with an event.

As far as the tweet content variables and Twitter user types variables are concerned, we did not find evidence of the topical interests being good predictors of engagement with events that display those same topics. However, we will show later in our analysis that when the tweets are broken down by topic and not considered as a single monolithic set, these topic-specific correlations become stronger predictors of engagement. As regards meformer versus informer tweets, the meformer tweets are not very good predictors of engagement, which is obvious since such tweets mostly involve the user talking about things that are highly personalized and hyper-local to their own lives. Informer tweets, on the other hand, display a positive correlation to engagement; since such tweets are usually in the third person, this result combined with the broadcast communication considered previously indicate that a user who posts such tweets will usually engage with something that multiple other users are also interested in (hence an event as against a personalized happening).

As concerns geolocation, we did not find any significant evidence – in contrast with prior research (Kulshrestha *et al.*, 2012) – that the geographical proximity has any effect on a user's engagement with an event. This would seem to indicate that users will choose to engage with an event whether or not it is "local" (in their surrounding vicinity) or non-local.

Finally, where the social network variables are concerned, we find that all of the variables are predictors with at least some degree of significance (and some more so than others). Interestingly, the only positive correlation is with the number of new friends. Note here that the measure of news friends' posts is strongly correlated to the measure of how many friends posted tweets related to the event before the user (i.e., *Followings posted prior*, $r = 0.62, p < .001$). This suggests that most of the news friends' posts are occurring before the user starts contributing messages and engaging with the event. This indicates that users are inspired and motivated to engage with events when they see tweets relating to those events on their timelines. However, this only goes so far – as the negatively correlated variables show, a large number of friends/followers and neighbors may bring down awareness, engagement, and subsequent participation (i.e., their coefficients are negative). We argue that this can be possibly attributed to a variety of factors. Some of these may include cognitive overload on the part of the target user, higher noise, posts being perceived as less personal, and most importantly, a perception that the topic is already sufficiently covered, e.g., posted by friends (thus reducing an "informer" user's motivation in engaging with it).

7.4.3 Prediction of Degree

To further explore the relative impact of these variables in predicting the *degree* of prediction in new events, we performed a linear regression, using participation levels in past events to predict the level of participation in a final, target event. The results are shown in Table 7.3.

We find that the most significant predictors of the degree of a user's engagement happen to be the social network variables, followed by the twitter activity variables. Specifically, the only social network variable that shows a significant positive correlation is the number of posts from the user's friends prior to the user's engagement with the event, which can be explained in terms of the activity that a user sees on their timeline with regard to that event. However, as in the previous case, increases in the user's network size seem to dampen the degree of engagement somewhat (which can be attributed to many of the same reasons described previously). The participant's own past tweet content seemed to have no significant effect on the predicted degree of engagement, save for the total and broadcast tweets, which offer a historical window into how active the user was in general.

7.4.4 Prediction of Degree w.r.t Different Topics

Finally, we are interested in understanding how, and to what extent, the decomposition of events into their constituent topics affects the performance of our predictors (for predicting the degree of people's engagement). To this end, the first task is to infer topics from our event clusters. We obtained six event categories from a large news agency: politics, technology & science, entertainment, sport, local, and odd news. We then asked 30 coders to code the event clusters manually, and resolved conflicts later Note that we only allow one label for a given event. Subsequent to this, we ran the linear regression again – these results are displayed in Table 7.4.

Interestingly, we observe that some predictors do indeed change with respect to different event topics. For example, we witness that for events related to politics, the effect of social activities such as the max number of tweets per hour exhibits a higher β value when contrasted with the findings from the general events in Table 4. We also found that the effect of a person's topic interest is stronger for politics and sports events, but relatively lesser for entertainment events. These results suggest that people who are devoted to politics and sports tend to be more recognizable and explicit (e.g. political junkies, analysts, and followers; sports fans). However, entertainment and science & tech events may consist of event topics – and subsequently user engagement – that varies thick and fast. As regards news users and tweets, following these becomes imperative for politics, tech & science, sports, and entertainment; while friends usually tend to post before users engage with local events and odd events. More generally, these results demonstrate the different pathways of information within a social network structure such as Twitter's. For news events, people first learn about them (and thus engage with them) via information posted from news accounts; if they find the event and its topics interesting, they tend to intensify the level of engagement. However, for local events and odd news, people tend to get engaged more via their friends' tweets, and thus the effect of information from friends is shown as more important. Finally, one of the most interesting contrasts occurs with respect to geolocation and geographic proximity – in Tables 3 and 4, geolocation information rarely affects the presence and degree of event engagement. However, when we look at the effect of geolocation with respect to the various event topics, it is shown to be more important for sports and local events. This makes complete intuitive sense: an overwhelming number of users tend to care, to a very large extent, about their own local sports teams and about local events that they may directly affect them.

7.5 Discussion and Implications

At the beginning of this paper, we posed five important questions relating to the engagement of users on social media with real-world events; and whether such engagement (and its level) could be effectively and practically predicted based on information available from that social media. In this section, we consider possible answers to those questions that are suggested by the data and revisit the related theories to examine our answers.

Does a person post tweets about an event because they are interested in the topics pertaining to that event?

Our analysis confirms that this is indeed the case. To highlight this, we point the reader to the analysis concerning prediction of presence, and the contrast with the same prediction analysis given a breakdown of the events into different topics. In the former case, there is no significant indicator of correlation from the content of a user's tweets to their engagement with an event. However, in the latter, there is a marked increase in the significance of the correlation between the content of tweets related to events in specific topics, and the user's engagement with those events (particularly for politics and entertainment). This is exactly what the "endurability" theory (Read *et al.*, 2002) proves: people are likely to remember a good experience and are willing to repeat it. In other words, people like to repeatedly talk about the topics that they are most familiar with/interested in. Therefore, they will show deeper engagement in those specific topics, in contrast to boarder and more general topics.

Are they instead engaged because their friends are also posting tweets about it?

The answer to this is positive as well, conditioned on the type of event that the user is engaging with. We have shown in the previous section that certain kinds of events – local events, as well as odd news – users tend to engage more due to their friends (following list) posting content relating to those events prior to the user's own engagement. This verifies the discoveries by Zuniga et al. (de Zúñiga and Valenzuela, 2011) network structure and social ties (especially weak ties) are determined to be strong predictors of the civic engagement. We also extend their theory by discovering the social network and time affects on the engagement with real-world events (indeed, some events are about civic issues).

Perhaps they are just a very active user of Twitter?

The degree to which a user was active on Twitter (the number of tweets posted by them) does indeed show a strong correlation across all cases to their predicted engagement with an event. This correlation seems to be agnostic of the type of event (as against the previous two questions, above), and hence it seems likely that more active users are more likely to be interested and engaged in a new event, across the board. This finding validates our earlier conjecture that these activities will first directly affect people's engagement in events on social media; such engagement will later indirectly affect social capital. Our finding extends existing literatures on the relationship between social media activities and social capital (Putnam, 1995; Burke *et al.*, 2011; Hyman, 2002) by exploring the role of user engagement.

Is their engagement a reflection of the fact that this is a local event?

The answer to this question reverts to the pattern of dependence on the kind of event observed in the answers to the first two questions. There are certain *kinds* of events that can be classified as engaging to a user primarily due to their local nature – as described in the previous section, these tend to be sports and local events. The connection to local events is obvious and trivial; a user in New York City is unlikely by and large to care about events that are happening in (say) far-off Tulsa, Oklahoma. For sports, it is likely that users within a given geographical area are more likely to care about teams that call that particular area home (although of course there will always be outliers; however, our analysis is focused on the typical user).

How and to what extent do the different topics of events affect the degree of a user's engagement?

The answer to this question can be found in the aggregation of the answers to all of the previous questions – it does certainly seem like the different genres of events (even among the typical genres that we considered) affect the degree, and nature, of a user's engagement with an event. While engagement with political, entertainment, and science & technology events seems to depend more on the content of past tweets (both of the user as well as the people they follow), engagement with sports and local events tends to correspond more closely with the user's geolocation.

7.5.0.1 Limitations

Although the data that we use and the results produced from that data seem to imply some rather strong conclusions, certain limitations of the study must also be considered when going forward. The first of these is the categorization of events: although the categories we use in this study are quite general, and capture a large portion of the posts on Twitter, arguments can certainly be made in support of finer-grained categories that will support more nuanced analysis with respect to users' potential engagement with events. Additionally, the event detection and classification process that is currently used by us can be further improved – both to classify events better, and to allot events across different categories (as against just a single category, as is the case currently). Finally, in this study, we did not consider the fact that there may exist different *kinds* of target users when engagement with events is under consideration. While we did partition a target user's following list coarsely (in terms of friends, news accounts, etc.), the target users themselves may also be distributed across various categories that exhibit some correlation (and hence predictive power) with respect to event engagement.

7.6 Summary of Chapter

In this chapter, we developed statistical models to explore and understand people's Twitter engagement with real-world events. Categories of engagement predictors were conceptually developed, operationalized, and assessed for their relative impact on users' engagement presence, and the degree of that engagement. We explored the relative impact of multiple measures collected from five different user perspectives. In particular, we found several key factors that predict the users' presence in engagement with real- world events, including total number of tweets, communication modes, friends' engagement in events, etc. We also examined the effects of these predictors in predicting the degree of engagement. We also examined the effects of these factors with respect to the different types of events predicated on their topics. We concluded that users' prior activities, as well as their social network structure, can be very good predictors for both the presence and the degree of their engagement with real-world events. Given a finer granularity of events (according to their topics), the content of tweets and the geographic proximity provide additional predictive power with respect to various different event categories.

	Not engaged with events	**Engaged with events**	**Difference**		
Measure	***Mean***	***SD***	***Mean***	***SD***	***p-value***
Event engagement					
Event count	–	–	12.1	7.78	–
Tweet count	–	–	2.33	.322	–
Twitter activity variables					
Total tweets	278.2	167.2	28.2	5.82	$***$
Max tweets per hr	6.39	5.78	6.57	5.44	ns
Avg. tweets per hr	1.14	.004	1.74	1.23	ns
Directed communications	2.81	6.4	1.83	5.33	$***$
Broadcast communications	.83	.22	1.48	1.11	$***$
Hashtag ratio	.2	0.24	.42	.006	$***$
RT ratio	.15	.41	.44	.054	$***$
Twitter content variables					
Topical interests from tweets content	–	–	.05	.01	–
Topical interests from following	–	–	.03	.00	–
Twitter user type variables					
Meformer tweets	.41	.14	.29	.219	$***$
Informer tweets	.24	.23	.43	.31	$***$
Geolocation variable					
Geographical proximity	–	–	318mi	189.8mi	–
Social network variables					
Num. followers	387	150.1	387	150.1	ns
Num. friends	117	109.78	117	109.78	ns
Followings posted prior	–	–	4.33	5.00	–
Average common neighbor prior	–	–	10.33	10.42	–
News friends	5.73	8.22	5.73	8.22	ns

Table 7.1: Mean and Sd Values for Twitter Users' Event Engagement, Compared to Averaged Values of These Twitter Users' Non-event Tweeting Behavior, and Paired Sample t-tests for the Difference. $***\ p < 0.001$, $**\ p < 0.01$, $*\ p < 0.05$.

	β	***SE***	***p-value***
Twitter activity variables			
Total tweets	.37	.045	$< .001^{***}$
Max tweets per hr	.01	.039	0.21
Avg. tweets per hr	-.08	.033	0.11
Directed communications	-.17	.071	$< .001^{***}$
Broadcast communications	.04	.059	$< .001^{***}$
Hashtag ratio	.09	.045	$< .001^{***}$
RT ratio	.069	.039	$< .001^{***}$
Tweet content variables			
Topical interests from tweets content	.12	.039	.22
Topical interests from followings	.07	.017	.17
Twitter user types variables			
Meformer tweets	.06	.013	.46
Informer tweets	.02	.016	$< .001^{***}$
Geolocation variables			
Geographical proximity	.01	.032	0.59
Social network variables			
Num. followers	-.04	.023	$< .001^{***}$
Num. friends	-.07	.016	$< .001^{***}$
Followings posted prior	.22	.024	$< .05^{*}$
Average common neighbor prior	-.22	.015	$< .01^{**}$
News friends	.30	.022	$< .001^{***}$

Table 7.2: Prediction of Presence: Logistic Regression Coefficients for Standardized Variables in Simultaneous Repeated Measures Logistic Regression Predicting Participation in Events over 406 “Trials”. Adjusted $R^2 = 0.67$, $*** \, p < 0.001$, $** \, p < 0.01$, $* \, p < 0.05$

	β	***SE***	***p-value***
Twitter activity variables			
Total tweets	.087	.055	$< .001^{***}$
Max tweets per hr	.02	.043	.32
Avg. tweets per hr	.11	.033	$< .01^{**}$
Directed communications	-.17	.028	.24
Broadcast communications	-.01	.076	$< .001^{***}$
Hashtag ratio	.11	.045	$< .01^{**}$
RT ratio	.49	.012	$< .01^{**}$
Tweet content variables			
Topical interests from tweets content	.11	.029	0.12
Topical interests from followings	.06	.02	0.21
Twitter user type variables			
Meformer tweets	-.11	.014	.28
Informer tweets	.21	.026	$< .001^{***}$
Geolocation variables			
Geographical proximity	.02	.032	.52
Social network variables			
Num. followers	-.04	.033	$< .001^{**}$
Num. friends	-.11	.046	$< .001^{**}$
Followings posted prior	.02	.024	$< .001^{**}$
Average common neighbor prior	-.02	.015	$< .001^{**}$
News friends	.13	.022	$< .14$

Table 7.3: Prediction of Degree: Linear Regression Coefficients for Standardized Variables in Simultaneous Repeated Measures Logistic Regression Predicting Participation in Events over 506 “Trials”. Adjusted $R^2 = 0.56$, $* * * \, p < 0.001$, $** \, p < 0.01$, $* \, p < 0.05$

	politics		tech and science		entertain		sports		local		odd events	
	β	***p-value***	β	***p-value***	β	***p-value***	β	***p-value***	β	***p-value***	β	***p-value***
Twitter activity variables												
Total tweets	.051	***	.074	**	.071	**	.066	**	.078	**	.11	**
Max tweets per hr	.12	**	.02	ns	.01	ns	.12	ns	.04	ns	.02	ns
Avg. tweets per hr	.03	ns	.04	ns	.11	ns	.07	ns	.08	ns	.12	ns
Directed communications	-.12	ns	-.12	ns	.11	**	-.17	**	.01	ns	-.101	ns
Broadcast communications	-.02	***	-.04	***	.02	***	-.11	**	-.02	***	-.08	***
Hashtag ratio	.08	*	.09	**	.11	***	.21	***	.07	**	.106	***
RT ratio	.09	**	.06	**	.071	**	.087	***	.08	*	.19	***
Tweet content variables												
Topical interests from tweets content	.22	**	.02	*ns*	.02	*ns*	.62	***	.12	ns	.12	ns
Topical interests from followings	.08	**	.07	ns	.07	*ns*	.54	**	.081	ns	.067	ns
Tweet content variables												
Meformer tweets	-.06	ns	-.09	ns	-.05	ns	-.01	ns	.01	ns	-.02	ns
Informer tweets	.21	***	.102	**	.12	***	.08	***	.09	***	.11	***
Geolocation variables												
Geographical proximity	.01	ns	.02	ns	.01	ns	.20	***	.42	**	.00	ns
Social network variables												
Num. followers	-.18	***	-.12	***	-.21	***	-.104	**	-.04	**	-.07	***
Num. friends	-.11	***	-.21	**	.08	***	.107	***	-.11	***	-.02	***
Followings posted prior	0.2	**	-.02	*ns*	-.12	*ns*	-.11	*ns*	.71	***	.14	***
Average common neighbor prior	-.12	***	-.02	*	-.15	***	-.02	**	-.011	***	-.033	***
News friends	.51	***	.39	**	.22	***	.21	***	-.01	ns	.04	ns

Table 7.4: Prediction of Degree of Twitter Engagement given Different Topics: Linear Regression Coefficients for Standardized Variables in Simultaneous Repeated Measures Logistic Regression Predicting Participation in Events over 406 "Trials", $* * * p < 0.001$, $** p < 0.01$, $* p < 0.05$

Chapter 8

CONCLUSION AND FUTURE WORKS

This chapter concludes the dissertation by summarizing the contributions of the work, highlight the future directions and consider some of the broader implications and impact resulting from of this work.

8.1 Summary of Contributions

Social media platforms such as Twitter, Facebook, and blogs have emerged as valuable – in fact, the de facto – virtual town halls for people to discover, report, share and communicate with others about various types of events. These events range from widely-known events (e.g., the U.S Presidential debate) to smaller scale, local events (e.g., a local Halloween block party). During these events, we often witness a large amount of commentary contributed by crowds on social media. This burst of social media information greatly enriches the user experience when interacting with the event and also enriches people's awareness of the event. This dissertation explored ways to leverage social media as a source of information and analyze events based on their social media responses collectively. The technical contributions of this work is the development of EventRadar, an event analysis toolbox that is able to automatically identify, enrich, and characterize events using the massive amounts of social media responses. EventRadar contains four tools to handle three core event analysis tasks. Specifically, for the Event Characterization task, we first developed ET-LDA, a joint Bayesian statistical model that characterizes topical influences between an event and its associated Twitter feeds (tweets). Our model enables the topic modeling of the event/tweets and the segmentation of the event in one unified framework. Based on the alignment established by ET-LDA, we developed SocSent, a flexible factorization framework that characterizes the segment and topics of an event via aggregated Twitter sentiment. SocSent leverages three types of prior knowledge: sentiment lexicon, manually labeled tweets and tweet/event alignment from ET-LDA, to regulate the learning process. Next, for the Event Recognition task, we developed DeMa, an unsupervised event detection algorithm which detect trending events from a stream of noisy social media posts based on time-series analysis and rank the trend events based on their novelty score. Based on DeMa, we also developed Whoo.ly, a web service that facilitates information seek-ing in hyperlocal communities by finding and summarizing neighborhood Twitter messages. Whoo.ly uses several computational approaches to discover hyperlocal content from noisy and overwhelming Twitter posts, such as DeMa for detecting local trending events. In addition, we also developed activity-based ranking algorithms and information extractors in Whoo.ly to provide additional insights into the most active people and popular places in a local community. Last, for the Event Enrichment task, we developed a spatial crowdsourcing system CrowdX for helping journalists gather more first-hand event information from the field. To achieve the best utility of assigning a worker to an event spot, CrowdX takes into account three features: 1) the coverage and diversity of the spot, and the quality of the worker, 2) the expected cost of each assignment, and 3) the uncertainty of the worker. We designed an efficient utility-theoretic approach that finds a set of user-spot assignments that simultaneously maximizes over-all utility, and achieves low cost in strict accordance with budget constraints.

Enabled by EventRadar, this dissertation also made scientific contributions by uncovering insights that have not been explored previously and re-validating existing social theories with new evidence. We found that the crowd's tweeting behavior varies significantly with the timeline of an event. Besides, we also discovered that people show different levels of engagement in different kinds of events. Furthermore, we found the topical context of the tweets does not always correlate with the timeline of the event. This dissertation also explored factors that affect people's engagement in the event. Based on the predictive analysis, we found that people engage in an event because they are interested in the topics pertaining to that event. Also, people tend to engage more in an event due to their friends (following list) are also posting tweets about it. Last, features like event locations do not affect people's engagement.

8.2 Future Work

This work can be extended in several promising directions. First, in ET-LDA and SocSent, we primarily focused on modeling and characterization of the hidden effect and topical influences of an event on its audience via their social media responses. We want to further investigate 1) the robustness, and 2) the predictive power of these models. Therefore, the main future direction here is to extend ET-LDA and SocSent to handle the online analysis and predictive analysis rather than the "after-the-fact" analysis of the entire event and the associated tweets, retrospectively. Besides, note that one limitation in ET-LDA and SocSent is that these models need to have event transcripts. In reality, however, not every event has a transcript or a document (to record its time). Therefore, our second line of future work involves developing automated tools and extending EventRadar to handle the events that are not transcribable. Last, we are interested in extending DeMa event detection to detect events that are both trending and interesting. Based on that, we would like to build a personalized event recommendation engine. The main motivation of this is that, given a large volume of event trending on social media everyday, people need a filtering tool that can help them decide which events on social media they should pay attention to (given the information overload problems in social media).

8.3 Broader Implications

In addition to our contributions that are related to a specific application (event analysis on social media), techniques presented in this dissertation also have some broader impacts to various other areas, especially in journalism and civic engagement.

As we mentioned earlier, technology is rapidly shifting the ways in which information about news and events gathered, processed, and disseminated. Computational Journalism is the application of computing to the activities of journalism including information gathering, organization and sensemaking, communication and presentation, and dissemination and public response to news information, all while upholding the core values of journalism such as accuracy and verifiability. Our EventRadar toolbox can readily contribute to computational journalism realm with tools to handle many challenges in computational journalism such as information gather (via CrowdX), organization and sensemaking (via ET-LDA, SocSent, and DeMa).

On the other hand, social media such as Twitter has altered society's information and communication fabric and will continue to be increasingly integrated in our daily lives. We also contribute to this direction by developing DeMa and its application, Whoo.ly, which automatically extracts and summarizes hyperlocal information about events, topics, people, and places from the Twitter posts. We believe this paper presents a promising approach to leveraging Twitter messages to better support hyperlocal community aware-ness and engagement.

Additionally, work that has been presented as part of this dissertation has resulted in several publications at top conferences (see References). The DeMa work also won a best paper nomination at ACM CHI 2013. Besides, the work has been featured in various press including ABC news, PBS, The Seattle Times, FastCompany, Computer Magazine, Neowin, ASU news

REFERENCES

Abramowitz, M. and I. A. Stegun, *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, no. 55 (Courier Dover Publications, 1972).

Allan, J., *Topic detection and tracking: event-based information organization*, vol. 12 (Springer, 2002).

Alt, F., A. S. Shirazi, A. Schmidt, U. Kramer and Z. Nawaz, "Location-based crowdsourcing: extending crowdsourcing to the real world", in "Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries", pp. 13–22 (ACM, 2010).

Anderson, C. W., "Towards a sociology of computational and algorithmic journalism", new media & society **15**, 7, 1005–1021 (2013).

Bakshy, E., J. Hofman, W. Mason and D. Watts, "Everyone's an influencer: quantifying influence on twitter", Proceedings of the fourth ACM international conference on Web search and data mining pp. 65–74 (2011).

Barzilay, R. and L. Lee, "Catching the drift: Probabilistic content models, with applications to generation and summarization", Proceedings of HLT-NAACL **2004** (2004).

Becker, H., M. Naaman and L. Gravano, "Learning similarity metrics for event identification in social media", Proceedings of the third ACM international conference on Web search and data mining pp. 291–300 (2010).

Becker, H., M. Naaman and L. Gravano, "Beyond trending topics: Real-world event identification on twitter", Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM11) (2011).

Beeferman, D., A. Berger and J. Lafferty, "Statistical models for text segmentation", Machine learning **34**, 1-3, 177–210 (1999).

Benouaret, K., R. Valliyur-Ramalingam and F. Charoy, "Crowdsc: Building smart cities with large scale citizen participation", (2013).

Bigham, J. P., C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White *et al.*, "Vizwiz: nearly real-time answers to visual questions", in "Proceedings of the 23nd annual ACM symposium on User interface software and technology", pp. 333–342 (ACM, 2010).

Bishop, C. *et al.*, *Pattern recognition and machine learning*, vol. 4 (springer New York, 2006).

Blei, D., T. Griffiths and M. Jordan, "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies", Journal of the ACM (JACM) **57**, 2, 7 (2010).

Blei, D. and J. Lafferty, “Dynamic topic models”, Proceedings of the 23rd international conference on Machine learning pp. 113–120 (2006).

Blei, D., A. Ng and M. Jordan, “Latent dirichlet allocation”, the Journal of machine Learning research **3**, 993–1022 (2003).

Blume, L., D. Easley and M. O’hara, “Market statistics and technical analysis: The role of volume”, The Journal of Finance **49**, 1, 153–181 (1994).

Bollen, J., H. Mao and X. Zeng, “Twitter mood predicts the stock market”, Journal of Computational Science **2**, 1, 1–8 (2011).

Bourdieu, P., “The forms of capital”, Handbook of theory and research for the sociology of education (1986).

Boykin, S. and A. Merlino, “Machine learning of event segmentation for news on demand”, Communications of the ACM (2000).

Burgess, M., A. Mazzia, E. Adar and M. Cafarella, “Leveraging noisy lists for social feed ranking”, Seventh International AAAI Conference on Weblogs and Social Media (2013).

Burke, J. A., D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy and M. B. Srivastava, “Participatory sensing”, Center for Embedded Network Sensing (2006).

Burke, M., R. Kraut and C. Marlow, “Social capital on facebook: Differentiating uses and users”, CHI (2011).

Carbonell, J. and J. Goldstein, “The use of mmr, diversity-based reranking for reordering documents and producing summaries”, in “Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval”, pp. 335–336 (ACM, 1998).

Cartwright, D. and F. Harary, “Structural balance: a generalization of heider’s theory.”, Psychological review **63**, 5, 277 (1956).

Chakrabarti, D. and K. Punera, “Event summarization using tweets”, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media pp. 66–73 (2011).

Chang, C. H., M. Kayed, M. R. Girgis and K. F. Shaalan, “A survey of web information extraction systems”, Knowledge and Data Engineering, IEEE Transactions on **18**, 10, 1411–1428 (2006).

Chang, J. and D. Blei, “Relational topic models for document networks”, Artificial Intelligence and Statistics pp. 81–88 (2009).

Cheng, Z., J. Caverlee and K. Lee, “You are where you tweet: a content-based approach to geo-locating twitter users”, in “Proceedings of the 19th ACM international conference on Information and knowledge management”, pp. 759–768 (ACM, 2010).

Cohen, S., J. T. Hamilton and F. Turner, “Computational journalism”, Communications of the ACM **54**, 10, 66–71 (2011a).

Cohen, S., C. Li, J. Yang and C. Yu, “Computational journalism: A call to arms to database researchers.”, in “CIDR”, vol. 2011, pp. 148–151 (2011b).

Cranshaw, J., R. Schwartz, J. I. Hong and N. M. Sadeh, “The livehoods project: Utilizing social media to understand the dynamics of a city.”, ICWSM (2012).

Cui, P., F. Wang, S. Liu, M. Ou, S. Yang and L. Sun, “Who should share what?: item-level social influence prediction for users and posts ranking”, Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval pp. 185–194 (2011).

Dahlgren, P., *Media and political engagement* (Cambridge University Press Cambridge, 2009).

Das, T., P. Mohan, V. N. Padmanabhan, R. Ramjee and A. Sharma, “Prism: platform for remote sensing using smartphones”, in “Proceedings of the 8th international conference on Mobile systems, applications, and services”, pp. 63–76 (ACM, 2010).

De Choudhury, M., N. Diakopoulos and M. Naaman, “Unfolding the event landscape on twitter: classification and exploration of user categories”, Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work pp. 241–244 (2012).

de Zúñiga, H. G. and S. Valenzuela, “The mediating path to a stronger citizenship: Online and offline networks, weak ties, and civic engagement”, Communication Research (2011).

Diakopoulos, N., M. De Choudhury and M. Naaman, “Finding and assessing social media information sources in the context of journalism”, in “Proceedings of the SIGCHI Conference on Human Factors in Computing Systems”, pp. 2451–2460 (ACM, 2012).

Diakopoulos, N., M. Naaman and F. Kivran-Swaine, “Diamonds in the rough: Social media visual analytics for journalistic inquiry”, VAST’10 (2010).

Diakopoulos, N. and D. Shamma, “Characterizing debate performance via aggregated twitter sentiment”, Proceedings of the 28th international conference on Human factors in computing systems pp. 1195–1198 (2010a).

Diakopoulos, N. A. and D. A. Shamma, “Characterizing debate performance via aggregated twitter sentiment”, Proceedings of the 28th international conference on Human factors in computing systems CHI 10 p. 1195 (2010b).

Dielmann, A. and S. Renals, “Dynamic bayesian networks for meeting structuring”, ICASSP (2004).

Dietz, L., S. Bickel and T. Scheffer, "Unsupervised prediction of citation influences", Proceedings of the 24th international conference on Machine learning pp. 233–240 (2007).

Ding, C. H. Q., T. Li, W. Peng and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering", Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining pp. 126–135 (2006).

Ellison, N. B. *et al.*, "Social network sites: Definition, history, and scholarship", JCMC (2007).

Emam, E. B. and H. Ai-Deek, "Using real-life dual-loop detector data to develop new methodology for estimating freeway travel time reliability", Transportation Research Record: Journal of the Transportation Research Board **1959**, 1, 140–150 (2006).

eMarketer, "What Do TV-Social Media Multitaskers Talk About?", `http://bit.ly/1yaCj8s` (2011).

Feld, S. L., "The focused organization of social ties", American journal of sociology pp. 1015–1035 (1981).

Fillion, R., *How Social Media Covered the Empire State Shooting* (http://blogs.wsj.com/digits/2012/08/24/how-social-media-covered-the-empire-state-shooting/, 2012).

Flaounas, I., O. Ali, T. Lansdall-Welfare, T. De Bie, N. Mosdell, J. Lewis and N. Cristianini, "Research methods in the age of digital journalism: Massive-scale automated analysis of news-contenttopics, style and gender", Digital Journalism **1**, 1, 102–116 (2013).

Galley, M., K. McKeown, E. Fosler-Lussier and H. Jing, "Discourse segmentation of multi-party conversation", Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1 pp. 562–569 (2003).

Garrison, B., *Computer-assisted reporting* (Lawrence Erlbaum Associates, 1998).

Gerner, D. J., P. A. Schrodt, R. A. Francisco and J. L. Weddle, "Machine coding of event data using regional and international sources", International Studies Quarterly pp. 91–119 (1994).

Gil de Zúñiga, H., N. Jung and S. Valenzuela, "Social media use for news and individuals' social capital, civic engagement and political participation", Journal of Computer-Mediated Communication **17**, 3, 319–336 (2012).

Gilbert, E. and K. Karahalios, "Predicting tie strength with social media", in "Proceedings of the SIGCHI Conference on Human Factors in Computing Systems", pp. 211–220 (ACM, 2009).

Gilbert, P., L. P. Cox, J. Jung and D. Wetherall, "Toward trustworthy mobile sensing", in "Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications", pp. 31–36 (ACM, 2010).

Glaser, M., "The new voices: Hyperlocal citizen media sites want you (to write)", Online Journalism Review **17** (2004).

Golder, S. A. and S. Yardi, "Structural predictors of tie formation in twitter: Transitivity and mutuality", in "Social Computing (SocialCom), 2010 IEEE Second International Conference on", pp. 88–95 (IEEE, 2010).

Google, "The New Multi-Screen World Study", `http://bit.ly/1cWdiT6` (2012).

Gowalla, in "http://en.wikipedia.org/wiki/Gowalla", (2010).

Griffiths, T. and M. Steyvers, "Finding scientific topics", Proceedings of the National Academy of Sciences (2004).

Griffiths, T. L., M. Steyvers, D. M. Blei and J. B. Tenenbaum, "Integrating topics and syntax", NIPS (2004).

Hampton, K. and B. Wellman, "Neighboring in netville: How the internet supports community and social capital in a wired suburb", City & Community **2**, 4, 277–311 (2003).

Han, B. and T. Baldwin, "Lexical normalisation of short text messages: Makn sens a# twitter", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies **1**, 368–378 (2011).

He, D. and D. S. Parker, "Topic dynamics: an alternative model of bursts in streams of topics", Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining pp. 443–452 (2010).

Hearst, M., "Texttiling: A quantitative approach to discourse segmentation", Sequoia (1993).

Hecht, B., L. Hong, B. Suh and E. H. Chi, "Tweets from justin bieber's heart: the dynamics of the location field in user profiles", in "Proceedings of the SIGCHI Conference on Human Factors in Computing Systems", pp. 237–246 (ACM, 2011).

Hoffman, B., H. Robinson, K. Han and J. Carroll, "Civicinity events: pairing geolocation tools with a community calendar", in "Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications", p. 14 (ACM, 2012).

Hofmann, D., "The missing link-a probabilistic model of document content and hypertext connectivity", NIPS (2001).

Hong, L., O. Dan and B. D. Davison, "Predicting popular messages in twitter", Proceedings of the 20th international conference companion on World wide web pp. 57–58 (2011).

Hu, M. and B. Liu, “Mining and summarizing customer reviews”, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining pp. 168–177 (2004).

Hu, X., N. Sun, C. Zhang and T. Chua, “Exploiting internal and external semantics for the clustering of short texts using world knowledge”, CIKM’09 (2009).

Hu, X., L. Tang, J. Tang and H. Liu, “Exploiting social relations for sentiment analysis in microblogging”, Proceedings of WSDM (2013a).

Hu, Y. and S. Farnham, “Predicting peoples engagement in real-world events on social media”, in “Under review”, (2014).

Hu, Y., S. D. Farnham and A. Monroy-Hernández, “Whoo. ly: Facilitating information seeking for hyperlocal communities using social media”, Proceedings of the 2013 ACM annual conference on Human factors in computing systems pp. 3481–3490 (2013b).

Hu, Y., E. Horvitz, J. Krumm and S. Kambhampati, “Planning in support of crowdsourcing tasks in the physical world”, in “Under review”, (2014).

Hu, Y., A. John and D. Seligmann, “Event analytics via social media”, SIGMM workshop on Social and Behavioural Networked Media Access (SBNMA’11) (2011).

Hu, Y., A. John, D. Seligmann and F. Wang, “What were the tweets about? topical associations between public events and twitter feeds”, Proceedings from ICWSM, Dublin, Ireland (2012a).

Hu, Y., A. John, D. Seligmann and F. Wang, “What were the tweets about? topical associations between public events and twitter feeds.”, Proceedings of ICWSM (2012b).

Hu, Y., A. John, F. Wang and S. Kambhampati, “Et-lda: Joint topic modeling for aligning events and their twitter feedback”, Twenty-Sixth AAAI Conference on Artificial Intelligence (2012c).

Hu, Y., F. Wang and S. Kambhampati, “Listening to the crowd: automated analysis of events via aggregated twitter sentiment”, in “Proceedings of the Twenty-Third international joint conference on Artificial Intelligence”, pp. 2640–2646 (AAAI Press, 2013c).

Hutto, C., S. Yardi and E. Gilbert, “A longitudinal study of follow predictors on twitter”, in “Proceedings of the SIGCHI Conference on Human Factors in Computing Systems”, pp. 821–830 (ACM, 2013).

Hyman, J. B., “Exploring social capital and civic engagement to create a framework for community building”, Applied Developmental Science **6**, 4, 196–202 (2002).

Jarvis, R. A. and E. A. Patrick, “Clustering using a similarity measure based on shared near neighbors”, Computers, IEEE Transactions on **100**, 11, 1025–1034 (1973).

Java, A., X. Song, T. Finin and B. Tseng, “Why we twitter: understanding microblogging usage and communities”, Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis pp. 56–65 (2007).

Jenelius, E. and H. N. Koutsopoulos, “Probe vehicle data sampled by time or space: Consistent travel time allocation and estimation”, (2014).

Jiang, L., M. Yu, M. Zhou, X. Liu and T. Zhao, “Target-dependent twitter sentiment classification”, in “Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1”, pp. 151–160 (Association for Computational Linguistics, 2011).

Kahne, J., E. Middaugh and K. Schutjer-Mance, “California civic index [monograph]”, New York: Carnegie Corporation and Annenberg Foundation (2005).

Kavanaugh, A., A. Ahuja, S. Gad, S. Neidig, M. A. Pérez-Quiñones, N. Ramakrishnan and J. Tedesco, “(hyper) local news aggregation: designing for social affordances”, Government Information Quarterly (2014).

Kazemi, L. and C. Shahabi, “Geocrowd: enabling query answering with spatial crowdsourcing”, in “Proceedings of the 20th International Conference on Advances in Geographic Information Systems”, pp. 189–198 (ACM, 2012).

Kazemi, L., C. Shahabi and L. Chen, “Geotrucrowd: trustworthy query answering with spatial crowdsourcing”, in “Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems”, pp. 304–313 (ACM, 2013).

Keeter, S., C. Zukin, M. Andolina and K. Jenkins, “Improving the measurement of political participation”, in “Annual meeting of the Midwest Political Science Association”, (2002).

Khuller, S., A. Moss and J. S. Naor, “The budgeted maximum coverage problem”, Information Processing Letters **70**, 1, 39–45 (1999).

Kim, S. H., Y. Lu, G. Constantinou, C. Shahabi, G. Wang and R. Zimmermann, “Mediaq: mobile multimedia management system”, in “Proceedings of the 5th ACM Multimedia Systems Conference”, pp. 224–235 (ACM, 2014).

Kivran-Swaine, F., P. Govindan and M. Naaman, “The impact of network structure on breaking ties in online social networks: unfollowing on twitter”, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2011).

Kleinberg, J., “Bursty and hierarchical structure in streams”, Data Mining and Knowledge Discovery **7**, 4, 373–397 (2003).

Krause, A., E. Horvitz, A. Kansal and F. Zhao, “Toward community sensing”, in “Proceedings of the 7th international conference on Information processing in sensor networks”, pp. 481–492 (IEEE Computer Society, 2008).

Krishnamurthy, B., P. Gill and M. Arlitt, “A few chirps about twitter”, in “Proceedings of the first workshop on Online social networks”, pp. 19–24 (ACM, 2008).

Kulshrestha, J., F. Kooti, A. Nikravesh and P. K. Gummadi, “Geographic dissection of the twitter network.”, ICWSM (2012).

Kwak, H., C. Lee, H. Park and S. Moon, “What is twitter, a social network or a news media?”, in “Proceedings of the 19th international conference on World wide web”, pp. 591–600 (ACM, 2010).

Lenhart, A., K. Purcell, A. Smith and K. Zickuhr, “Social media & mobile internet use among teens and young adults. millennials.”, Pew Internet & American Life Project (2010).

Lewis, S. and D. A. Lewis, “Examining technology that supports community policing”, in “Proceedings of the SIGCHI Conference on Human Factors in Computing Systems”, pp. 1371–1380 (ACM, 2012).

Li, T., Y. Zhang and V. Sindhwani, “A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge”, Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1 pp. 244–252 (2009).

Liu, K.-L., W.-J. Li and M. Guo, “Emoticon smoothed language models for twitter sentiment analysis.”, in “AAAI”, (2012).

Livne, A., M. P. Simmons, E. Adar and L. A. Adamic, “The party is over here: Structure and content in the 2010 election.”, ICWSM (2011).

Lotan, G., *Big Data for Breaking News: Lessons from #Aurora, Colorado* (http://giladlotan.com/2012/08/big-data-breaking-news-aurora-colorado/, 2012).

Lovász, L., “Submodular functions and convexity”, in “Mathematical Programming The State of the Art”, pp. 235–257 (Springer, 1983).

Mahmud, J., J. Nichols and C. Drews, “Home location identification of twitter users”, arXiv preprint arXiv:1403.2345 (2014).

Makkonen, J., “Investigations on event evolution in tdt”, Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Proceedings of the HLT-NAACL 2003 student research workshop-Volume 3 pp. 43–48 (2003).

Manning, C. D., P. Raghavan and H. Schütze, *Introduction to information retrieval*, vol. 1 (Cambridge university press Cambridge, 2008).

Maskey, S. and J. Hirschberg, “Automatic summarization of broadcast news using structural features.”, INTERSPEECH (2003).

McPherson, M., L. Smith-Lovin and J. M. Cook, “Birds of a feather: Homophily in social networks”, Annual review of sociology (2001).

Mei, Q., X. Ling, M. Wondra, H. Su and C. Zhai, “Topic sentiment mixture: modeling facets and opinions in weblogs”, Proceedings of the 16th international conference on World Wide Web pp. 171–180 (2007).

Monroy-Hernández, A., E. Kiciman, D. Boyd and S. Counts, “Narcotweets: Social media in wartime”, in “Sixth International AAAI Conference on Weblogs and Social Media”, (2012).

Murphy, J., *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications* (Penguin. com, 1999).

Naaman, M., H. Becker and L. Gravano, “Hip and trendy: Characterizing emerging trends on twitter”, Journal of the American Society for Information Science and Technology **62**, 5, 902–918 (2011).

Naaman, M., J. Boase and C. Lai, “Is it really about me?: message content in social awareness streams”, Proceedings of the 2010 ACM conference on Computer supported cooperative work pp. 189–192 (2010).

Naaman, M., A. X. Zhang, S. Brody and G. Lotan, “On the study of diurnal urban routines on twitter.”, ICWSM (2012).

Nallapati, R., A. Ahmed, E. Xing and W. Cohen, “Joint latent topic models for text and citations”, Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining pp. 542–550 (2008).

Newport, J. K. and G. G. Jawahar, “Community participation and public awareness in disaster mitigation”, Disaster prevention and management **12**, 1, 33–36 (2003).

Nichols, J., J. Mahmud and C. Drews, “Summarizing sporting events using twitter”, in “Proceedings of the 2012 ACM international conference on Intelligent User Interfaces”, pp. 189–198 (ACM, 2012).

O’Connor, B., R. Balasubramanyan, B. Routledge and N. Smith, “From tweets to polls: Linking text sentiment to public opinion time series”, Proceedings of the International AAAI Conference on Weblogs and Social Media pp. 122–129 (2010a).

O’Connor, B., M. Krieger and D. Ahn, “Tweetmotif: Exploratory search and topic summarization for twitter.”, in “ICWSM”, (2010b).

Pak, A. and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining.”, in “LREC”, (2010).

Pang, B. and L. Lee, “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts”, Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics p. 271 (2004).

Pang, B., L. Lee and S. Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques”, Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 pp. 79–86 (2002).

Petrović, S., M. Osborne and V. Lavrenko, “Streaming first story detection with application to twitter”, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics pp. 181–189 (2010).

Purver, M., T. Griffiths, K. Körding and J. Tenenbaum, “Unsupervised topic modelling for multi-party spoken discourse”, ACL (2006).

Putnam, R. D., “Bowling alone: America’s declining social capital”, Journal of democracy (1995).

Rakha, H., I. El-Shawarby and M. Arafeh, “Trip travel-time reliability: issues and proposed solutions”, Journal of Intelligent Transportation Systems **14**, 4, 232–250 (2010).

Ramage, D., S. Dumais and D. Liebling, “Characterizing microblogs with topic models”, ICWSM’10 (2010).

Read, J., S. MacFarlane and C. Casey, “Endurability, engagement and expectations: Measuring children’s fun”, in “Interaction design and children”, vol. 2, pp. 1–23 (Shaker Publishing Eindhoven, 2002).

Reavy, M., *Introduction to computer-assisted reporting: A journalist’s guide* (McGraw-Hill Higher Education, 2001).

Ritter, A., O. Etzioni, S. Clark *et al.*, “Open domain event extraction from twitter”, Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining pp. 1104–1112 (2012).

Rosenstiel, T., A. Mitchell, K. Purcell and L. Rainie, “How people learn about their local community”, Pew Research Center’s Project for Excellence in Journalism and the Pew Internet & American Life Project (2011).

Sadilek, A., J. Krumm and E. Horvitz, “Crowdphysics: Planned and opportunistic crowdsourcing for physical tasks”, SEA **21**, 10,424, 125–620 (2013).

Sahami, M. and T. Heilman, “A web based kernel function for measuring the similarity of short text snippets”, WWW’06 (2006).

Sakaki, T., M. Okazaki and Y. Matsuo, “Earthquake shakes twitter users: real-time event detection by social sensors”, Proceedings of the 19th international conference on World wide web pp. 851–860 (2010).

Sarason, S. B., *The psychological sense of community: Prospects for a community psychology.* (Jossey-Bass, 1974).

Shah, D. V., “Civic engagement, interpersonal trust, and television use: An individual-level assessment of social capital”, Political Psychology **19**, 3, 469–496 (1998).

Shahaf, D. and E. Horvitz, “Generalized task markets for human and machine computation.”, in “AAAI”, (2010).

Shamma, D., L. Kennedy and E. Churchill, “Tweet the debates: understanding community annotation of uncollected sources”, SIGMM workshop on Social media (2009).

Shamma, D., L. Kennedy and E. Churchill, “Conversational shadows: Describing live media events using short messages”, ICWSM’10 (2010).

Snow, R., B. O’Connor, D. Jurafsky and A. Y. Ng, “Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks”, in “Proceedings of the conference on empirical methods in natural language processing”, pp. 254–263 (Association for Computational Linguistics, 2008).

Starbird, K., L. Palen, A. Hughes and S. Vieweg, “Chatter on the red: what hazards threat reveals about the social life of microblogged information”, CSCW’10 (2010).

Suh, B., L. Hong, P. Pirolli and E. H. Chi, “Want to be retweeted? large scale analytics on factors impacting retweet in twitter network”, Social Computing (SocialCom), 2010 IEEE Second International Conference on pp. 177–184 (2010).

Surowiecki, J., *The wisdom of crowds* (Random House LLC, 2005).

Tan, C., L. Lee, J. Tang, L. Jiang, M. Zhou and P. Li, “User-level sentiment analysis incorporating social networks”, arXiv preprint arXiv:1109.6018 (2011).

Tanev, H., J. Piskorski and M. Atkinson, “Real-time news event extraction for global crisis monitoring”, pp. 207–218 (Springer, 2008).

Taylor, M. P. and H. Allen, “The use of technical analysis in the foreign exchange market”, Journal of international Money and Finance **11**, 3, 304–314 (1992).

Teh, Y., M. Jordan, M. Beal and D. Blei, “Hierarchical dirichlet processes”, Journal of the American Statistical Association **101**, 476, 1566–1581 (2006).

Times, N. Y., “Not Waiting for Pundits Take, Web Audience Scores the Candidates in an Instant”, `http://nyti.ms/1rrOU2l` (2012).

Titov, I. and R. McDonald, “Modeling online reviews with multi-grain topic models”, WWW (2008).

Toole, J. L., M. Cha and M. C. González, “Modeling the adoption of innovations in the presence of geographic and media influences”, PloS one **7**, 1, e29528 (2012).

Trefethen, L. N. and D. Bau III, *Numerical linear algebra*, vol. 50 (Siam, 1997).

Troncy, R., B. Malocha and A. T. Fialho, "Linking events with media", in "Proceedings of the 6th International Conference on Semantic Systems", p. 42 (ACM, 2010).

Väätäjä, H., T. Vainio, E. Sirkkunen and K. Salo, "Crowdsourced news reporting: supporting news content creation with mobile phones", in "Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services", pp. 435–444 (ACM, 2011).

Vieweg, S., A. Hughes, K. Starbird and L. Palen, "Microblogging during two natural hazards events: what twitter may contribute to situational awareness", CHI'10 (2010).

Von Ahn, L., "Games with a purpose", Computer **39**, 6, 92–94 (2006).

Wang, C.-K., B. Hsu, M.-W. Chang and E. Kiciman, "Simple and knowledge-intensive generative model for named entity recognition", Microsoft Research (2013).

Wang, X. and A. McCallum, "Topics over time: a non-markov continuous-time model of topical trends", Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining pp. 424–433 (2006).

Wang, Y., W. Dong, L. Zhang, D. Chin, M. Papageorgiou, G. Rose and W. Young, "Speed modeling and travel time estimation based on truncated normal and lognormal distributions", Transportation Research Record: Journal of the Transportation Research Board **2315**, 1, 66–72 (2012).

Wasserman, S., *Social network analysis: Methods and applications*, vol. 8 (Cambridge university press, 1994).

Wellman, B., "Community: from neighborhood to network", Communications of the ACM **48**, 10, 53–55 (2005).

Wellman, B. and S. Wortley, "Different strokes from different folks: Community ties and social support", American journal of Sociology pp. 558–588 (1990).

Weng, J. and B.-S. Lee, "Event detection in twitter.", ICWSM **11**, 401–408 (2011).

Westermann, U. and R. Jain, "Toward a common event model for multimedia applications", Multimedia, IEEE **14**, 1, 19–29 (2007).

Westgate, B. S., D. B. Woodard, D. S. Matteson, S. G. Henderson *et al.*, "Travel time estimation for ambulances using bayesian data augmentation", The Annals of Applied Statistics **7**, 2, 1139–1161 (2013).

Wilson, T., J. Wiebe and P. Hoffmann, "Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis", Computational linguistics **35**, 3, 399–433 (2009).

Wu, S., J. Hofman, W. Mason and D. Watts, "Who says what to whom on twitter", Proceedings of the 20th international conference on World wide web pp. 705–714 (2011).

Xie, L., H. Sundaram and M. Campbell, “Event mining in multimedia streams”, Proceedings of the IEEE **96**, 4, 623–647 (2008).

Yang, Y., T. Pierce and J. Carbonell, “A study of retrospective and on-line event detection”, Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval pp. 28–36 (1998).

Yates, D. and S. Paquette, “Emergency knowledge management and social media technologies: A case study of the 2010 haitian earthquake”, International Journal of Information Management **31**, 1, 6–13 (2011).

Zacks, J. M. and B. Tversky, “Event structure in perception and conception.”, Psychological bulletin **127**, 1, 3 (2001).

Zalta, E. N. and S. Abramsky, “Stanford encyclopedia of philosophy”, (2003).

Zhao, D. and M. Rosson, “How and why people twitter: the role that micro-blogging plays in informal communication at work”, Proceedings of the ACM 2009 international conference on Supporting group work pp. 243–252 (2009).

Zhao, W., J. Jiang, J. Weng, J. He, E. Lim, H. Yan and X. Li, “Comparing twitter and traditional media using topic models”, ECIR’11 (2011).

APPENDIX A

APPROXIMATE INFERENCE IN ET-LDA VIA GIBBS SAMPLING

Since the exact inference of the hidden variables in the LDA family is intractable, we utilize approximate methods like the collapsed Gibbs sampling algorithm Griffiths *et al.* (2004) for parameter estimation. In specific, we want to evaluate the posterior distribution of following hidden variables: (**i**) $\mathbf{z}_s$, evaluated for each word in every paragraph s in the event transcript and then used to infer $\theta^{(s)}$; (**ii**) $\mathbf{z}_t$, evaluated for each word in each tweet t written by the Twitter users and then used the results to infer $\psi^{(t)}$; (**iii**) $\mathbf{c}_s$, evaluated for each paragraph to indicate segmentation of the event; (**iv**) $\mathbf{c}_t$, evaluated for the topic types in tweet t; and (**v**) $\mathbf{s}_t$, evaluated for selecting segments from the event's transcript for t.

To begin with, we first write the joint distribution for ET-LDA based the generative process and the model's graphical structure. Note that we could integrate out the parameters $\phi, \gamma^{(t)}, \theta^{(s)}, \psi^{(t)}, \delta^{(s)}, \lambda^{(t)}$ because the model only uses conjugate priors Bishop *et al.* (2006). As a result, we have:

$$\begin{aligned}
&P(\mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t, \mathbf{c}_s, \mathbf{c}_t, \mathbf{s}_t | \alpha_\delta, \alpha_\theta, \alpha_\gamma, \alpha_\lambda, \alpha_\psi, \beta) \\
= &\int P(\mathbf{w}_s|\mathbf{z}_s, \phi) P(\mathbf{w}_t|\mathbf{z}_t, \phi) P(\phi|\beta) d\phi \int P(\mathbf{s}_t|\gamma^{(t)}) P(\gamma^{(t)}|\alpha_\gamma) d\gamma^{(t)} \\
&\int P(\mathbf{z}_s|\theta^{(s)}) P(\mathbf{z}_t|\theta^{(s)}, \mathbf{s}_t, \mathbf{c}_t = 0) P(\theta^{(s)}|\alpha_\theta, \mathbf{c}_s) d\theta^{(s)} \int P(\mathbf{z}_t|\psi^{(t)}, \mathbf{c}_t = 1) P(\psi^{(t)}|\alpha_\psi) d\psi^{(t)} \\
&\int P(\mathbf{c}_s|\delta^{(s)}) P(\delta^{(s)}|\alpha_\delta) d\delta^{(s)} \int P(\mathbf{c}_t|\lambda^{(t)}) P(\lambda^{(t)}|\alpha_{\lambda_\gamma}, \alpha_{\lambda_\psi}) d\lambda^{(t)} \qquad \text{(A.1)}
\end{aligned}$$

Inference of $\mathbf{Z}_S$

Gibbs sampling allows the learning of a model by iteratively updating each latent variable given the remaining variables. So, instead of estimating the posterior distribution $P(\mathbf{z}_s|\mathbf{z}_t, \mathbf{w}_s, \mathbf{w}_t, \mathbf{c}_s, \mathbf{c}_t, \mathbf{s}_t)$, we estimate the probably $P(\mathbf{z}_{s,i}|\mathbf{z}_{-(s,i)}, \mathbf{z}_t, \mathbf{w}_s, \mathbf{w}_t, \mathbf{c}_s, \mathbf{c}_t, \mathbf{s}_t)$ using the Gibbs sampler. Note that $\mathbf{z}_{-(s,i)}$ is a vector of assignments of topics, for all words in the paragraph s in an event's transcript except for the one at position i. According to the Bayes rule, we can compute this conditional probability by dividing the joint distribution (see Eq. A.1) by the joint with all variables but $z_{(s,i)}$ (denoted by $\mathbf{z}_{-(s,i)}$). We further cancel factors in the faction that do not depend on $z_{(s,i)}$, as follows:

$$\begin{aligned}
P(\mathbf{z}_{s,i}|\mathbf{z}_{-(s,i)}, \mathbf{z}_t, \mathbf{w}_s, \mathbf{w}_t, \mathbf{c}_s, \mathbf{c}_t, \mathbf{s}_t) &= \frac{P(\mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t, \mathbf{c}_s, \mathbf{c}_t, \mathbf{s}_t)}{P(\mathbf{z}_{-(s,i)}, \mathbf{z}_t, \mathbf{w}_s, \mathbf{w}_t, \mathbf{c}_s, \mathbf{c}_t, \mathbf{s}_t)} \\
&= \frac{\int P(\mathbf{w}_s|\mathbf{z}_s, \phi) P(\mathbf{w}_t|\mathbf{z}_t, \phi) P(\phi|\beta) d\phi}{\int P(\mathbf{w}_s|\mathbf{z}_{-(s,i)}, \phi) P(\mathbf{w}_t|\mathbf{z}_t, \phi) P(\phi|\beta) d\phi} \\
&\cdot \frac{\int P(\mathbf{z}_s|\theta^{(s)}) P(\mathbf{z}_t|\theta^{(s)}, \mathbf{s}_t, \mathbf{c}_t = 0) P(\theta^{(s)}|\alpha_\theta, \mathbf{c}_s) d\theta^{(s)}}{\int P(\mathbf{z}_{-(s,i)}|\theta^{(s)}) P(\mathbf{z}_t|\theta^{(s)}, \mathbf{s}_t, \mathbf{c}_t = 0) P(\theta^{(s)}|\alpha_\theta, \mathbf{c}_s) d\theta^{(s)}} \qquad \text{(A.2)}
\end{aligned}$$

We now derive the first fraction of Eq. A.2. As defined in the generative process, a topic $z_{(s,i)}$ is generated from a Multinomial distribution whose prior ϕ is a Dirichlet.

Because Dirichlet is the conjugate prior of Multinomial, we could solve the Dirichlet-Multinomial integral in a straightforward way. Specifically, we have:

$$\begin{aligned}\int P(\mathbf{w}_s|\mathbf{z}_s,\phi)P(\mathbf{w}_t|\mathbf{z}_t,\phi)P(\phi|\beta)d\phi &= \int \prod_{k=1}^{K} \frac{1}{\Delta(\beta)} \prod_{w=1}^{W} \phi^{n_{sw}^k + n_{tw}^k + \beta - 1} d\phi \\ &= \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^K \prod_{k=1}^{K} \frac{\prod_{w=1}^{W} \Gamma(n_{sw}^k + n_{tw}^k + \beta)}{\Gamma(n_{s(.)}^k + n_{t(.)}^k + W\beta)}\end{aligned} \tag{A.3}$$

where $\Delta(\beta) = \frac{\Gamma(\beta)^W}{\Gamma(W\beta)}$, W is the size of the vocabulary, $\Gamma(\cdot)$ is the gamma function, n_{sw}^k and n_{tw}^k are the numbers of times topic k assigned to word w in the event and the tweets. $n_{s(.)}^k$ and $n_{t(.)}^k$ are the total number of words in the event and tweets assigned to topic k.

To yield the first fraction of Eq. A.2, we apply the above equation twice, and given the fact that $\Gamma(x+1) = x\Gamma(x)$, we obtain the following equation:

$$\begin{aligned}\frac{\int P(\mathbf{w}_s|\mathbf{z}_s,\phi)P(\mathbf{w}_t|\mathbf{z}_t,\phi)P(\phi|\beta)d\phi}{\int P(\mathbf{w}_s|\mathbf{z}_{-(s,i)},\phi)P(\mathbf{w}_t|\mathbf{z}_t,\phi)P(\phi|\beta)d\phi} =& \frac{\left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^K \prod_{k=1}^{K} \frac{\prod_{w=1}^{W} \Gamma(n_{sw}^k + n_{tw}^k + \beta)}{\Gamma(n_{s(.)}^k + n_{t(.)}^k + W\beta)}}{\left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^K \prod_{k=1}^{K} \frac{\prod_{w=1}^{W} \Gamma(n_{sw}^{k-(s,i)} + n_{tw}^k + \beta)}{\Gamma(n_{s(.)}^{k-(s,i)} + n_{t(.)}^k + W\beta)}} \\ =& \frac{n_{sw}^k + n_{tw}^k + \beta - 1}{n_{s(.)}^k + n_{t(.)}^k + W\beta - 1}\end{aligned} \tag{A.4}$$

where n_{sw}^k is the number of times that topic k is assigned to word w. So $n_{sw}^{k-(s,i)}$ is the the number of times that topic k is assigned to word w, but with the i-th word in paragraph s (which happens to be w) and its topic assignment (which happens to be k) excluded. Therefore, $n_{sw}^k = n_{sw}^{k-(s,i)}+1$. Similarly, $n_{s(.)}^k = \sum_{w=1}^{W} n_{sk}^w = \sum_{w=1}^{W} n_{sw}^{k-(d,i)}+1$.

Next, we derive the second fraction of Eq. 2. As defined in the generative process, $\mathbf{z}_s$ and $\mathbf{z}_t$ are both generated from Multinomial and share the same Dirichlet prior if and only if $\mathbf{c}_t = 0$, meaning that the topic of a word in a tweet t is influenced by the topic of the segment that t refers to. In such case, $\theta^{(t)} = \theta^{(s)}$. Besides, according to the Markov assumption on the event generation, the topic distribution $\theta^(s)$ associated with paragraph s is depended on the value of the binary segmenting indicator $\mathbf{c}_s$: it can be drawn from a new Dirichlet (when $\mathbf{c}_s = 1$) or can be just as same as the one in the preceding paragraph (when $\mathbf{c}_s = 0$). The distribution is thus:

$$P(\theta^{(s)}|\mathbf{c}_s, \theta^{(s-1)}) = \begin{cases} \delta(\theta_s, \theta_{s-1}), & \text{if } \mathbf{c}_s = 0\text{, merge with preceding paragraph } s-1 \\ \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} (\theta^{(s)^{n^k-1}}), & \text{if } \mathbf{c}_s = 1\text{, a new segment starts} \end{cases}$$

where $\delta(\cdot,\cdot)$ is the Dirac delta function.

Based on these, we now expand and solve the integral (the numerator of the second fraction) as:

$$
\begin{aligned}
&\int P(\mathbf{z}_s|\theta^{(s)})P(\mathbf{z}_t|\theta^{(t)},\mathbf{s}_t,\mathbf{c}_t=0)P(\theta^{(s)}|\mathbf{c}_s)d\theta^{(s)} \\
=&\int P(z_{s_1},...z_{s_n}|\theta_1^{(s)},...\theta_n^{(s)})P(z_{t_1},...z_{t_m}|\theta_1^{(t)},...\theta_m^{(t)},\mathbf{s}_t,\mathbf{c}_t=0)P(\theta_1^{(s)},...\theta_n^{(s)}|c_{s_1},...c_{s_n})d\theta^{(s)} \\
=&\int \underbrace{\prod_{i=1}^{n_1}P(z_i|\theta_i^{(s)})\prod_{j=1}^{m_1}P(z_j|\theta_j^{(t)},\mathbf{s}_t,\mathbf{c}_t=0)\prod_{i=1}^{n_1}P(\theta_i^{(s)}|c_{s_i})}_{\text{first segment}}\times \\
&\qquad\qquad\vdots \\
&\underbrace{\prod_{i=1}^{n_s}P(z_i|\theta_i^{(s)})\prod_{j=1}^{m_s}P(z_j|\theta_j^{(t)},\mathbf{s}_t,\mathbf{c}_t=0)\prod_{i=1}^{n_s}P(\theta_i^{(s)}|c_{n_{s_1}})\,d\theta^{(s)}}_{\text{last segment}} \\
=&\int \prod_{s=1}^{\mathcal{S}}\prod_{i=1}^{n_s}P(z_i|\theta_i^{(s)})\prod_{j=1}^{m_s}P(z_j|\theta_j^{(t)},\mathbf{s}_t,\mathbf{c}_t=0)\prod_{i=1}^{n_s}P(\theta_i^{(s)}|c_{n_{s_1}})d\theta^{(s)} \\
=&\int \prod_{i=1}^{\mathcal{S}}\frac{1}{\Delta(\alpha_\theta)}\prod_{k=1}^{K}\theta^{n_k^{\mathcal{S}_i}+nt_k^{\mathcal{S}_i}+\alpha_\theta-1}d\theta^{(s)} \\
=&\left(\frac{\Gamma(K\alpha_\theta)}{\Gamma(\alpha_\theta)^K}\right)^{\mathcal{S}}\prod_{i=1}^{\mathcal{S}}\frac{\prod_{k=1}^{K}\Gamma(n_k^{\mathcal{S}_i}+nt_k^{\mathcal{S}_i}+\alpha_\theta)}{\Gamma(n_{(.)}^{\mathcal{S}_i}+nt_{(.)}^{\mathcal{S}_i}+K\alpha_\theta)} \qquad \text{(A.5)}
\end{aligned}
$$

where $\mathcal{S}$ is a set of segments of the event's transcript. Each element ($\mathcal{S}_i$) in this set contains multiple paragraphs. $n_k^{\mathcal{S}_i}$ is the number of times topic k assigned to words in segment $\mathcal{S}_i$. $nt_k^{\mathcal{S}_i}$ is the number of times topic k appears in tweets, where these tweets are direct response to the sentences in segment $\mathcal{S}_i$ (i.e., $c_t = 0$).

To yield the second fraction of Eq. A.2, we again apply the above equation twice and we cancel most factors but only leave some certain factors with $z_{d,i} = k$ and $v_{d,i} = w$. Thus we have:

$$
\begin{aligned}
\frac{\int P(\mathbf{z}_s|\theta^{(s)})P(\mathbf{z}_t|\theta^{(t)},\mathbf{s}_t,\mathbf{c}_t=0)P(\theta^{(s)}|\mathbf{c}_s)d\theta^{(s)}}{\int P(\mathbf{z}_{-(s,i)}|\theta^{(s)})P(\mathbf{z}_t|\theta^{(t)},\mathbf{s}_t,\mathbf{c}_t=0)P(\theta^{(s)}|\mathbf{c}_s)d\theta^{(s)}}
&=\frac{\left(\frac{\Gamma(K\alpha_\theta)}{\Gamma(\alpha_\theta)^K}\right)^{\mathcal{S}}\prod_{i=1}^{\mathcal{S}}\frac{\prod_{k=1}^{K}\Gamma(n_k^{\mathcal{S}_i}+nt_k^{\mathcal{S}_i}+\alpha_\theta)}{\Gamma(n_{(.)}^{\mathcal{S}_i}+nt_{(.)}^{\mathcal{S}_i}+K\alpha_\theta)}}{\left(\frac{\Gamma(K\alpha_\theta)}{\Gamma(\alpha_\theta)^K}\right)^{\mathcal{S}}\prod_{i=1}^{\mathcal{S}}\frac{\prod_{k=1}^{K}\Gamma(n_k^{\mathcal{S}_i-(s,i)}+nt_k^{\mathcal{S}_i}+\alpha_\theta)}{\Gamma(n_{(.)}^{\mathcal{S}_i-(s,i)}+nt_{(.)}^{\mathcal{S}_i}+K\alpha_\theta)}} \\
&=\frac{n_k^{\mathcal{S}_i}+nt_k^{\mathcal{S}_i}+\alpha_\theta-1}{n_{(.)}^{\mathcal{S}_i}+nt_{(.)}^{\mathcal{S}_i}+K\alpha_\theta-1} \qquad \text{(A.6)}
\end{aligned}
$$

where $k = \mathbf{z}_{(s,i)}$, $\mathcal{S}_i$ is the segment that sentence s belongs to.

So we have for:

$$P(\mathbf{z}_{s,i}|\mathbf{z}_{-(s,i)},\mathbf{z}_t,\mathbf{w}_s,\mathbf{w}_t,\mathbf{c}_s,\mathbf{c}_t,\mathbf{s}_t) = \frac{n_{sw}^k + n_{tw}^k + \beta - 1}{n_{s(.)}^k + n_{t(.)}^k + W\beta - 1} \times \frac{n_k^{\mathcal{S}_i} + nt_k^{\mathcal{S}_i} + \alpha_\theta - 1}{n_{(.)}^{\mathcal{S}_i} + nt_{(.)}^{\mathcal{S}_i} + K\alpha_\theta - 1} \tag{A.7}$$

Inference of $\mathbf{Z}_t$
Next, we examine $P(\mathbf{z}_t|\mathbf{z}_s,\mathbf{w}_s,\mathbf{w}_t,\mathbf{c}_s,\mathbf{c}_t,\mathbf{s}_t)$. Similar to the estimation of $\mathbf{z}_s$, we also use the Gibbs sampling to compute the probability $P(\mathbf{z}_{(t,i)}|\mathbf{z}_{-(t,i)},\mathbf{z}_s,\mathbf{w}_s,\mathbf{w}_t,\mathbf{c}_s,\mathbf{c}_t,\mathbf{s}_t)$. According to the generative process of ET-LDA, the topic distribution $\mathbf{z}_t$ of a tweet t can have two conditions: it can be either drawn from a Multinomial governed by the Dirichlet prior θ_s for specific topics (when $\mathbf{c}_t = 0$) or drawn from a Multinomial governed by the Dirichlet prior ψ_s for general topics (when $\mathbf{c}_t = 1$). Therefore, the conditional probability $P(\mathbf{z}_{(t,i)}|\mathbf{z}_{-(t,i)},\mathbf{z}_s,\mathbf{w}_s,\mathbf{w}_t,\mathbf{c}_s,\mathbf{c}_t,\mathbf{s}_t)$ has two cases with each capturing one condition. Using the Bayes rule and only consider factors that depend on $\mathbf{z}_t$, we have for $\mathbf{c}_t = 0$:

$$\begin{aligned} P(\mathbf{z}_{t,i}|\mathbf{z}_{-(t,i)},\mathbf{z}_t,\mathbf{w}_s,\mathbf{w}_t,\mathbf{c}_s,\mathbf{c}_t=0,\mathbf{s}_t) =& \frac{P(\mathbf{w}_s,\mathbf{w}_t,\mathbf{z}_s,\mathbf{z}_t,\mathbf{c}_s,\mathbf{c}_t=0,\mathbf{s}_t)}{P(\mathbf{z}_{-(t,i)},\mathbf{z}_s,\mathbf{w}_s,\mathbf{w}_t,\mathbf{c}_s,\mathbf{c}_t=0,\mathbf{s}_t)} \\ =& \frac{\int P(\mathbf{w}_s|\mathbf{z}_s,\phi)P(\mathbf{w}_t|\mathbf{z}_t,\phi)P(\phi|\beta)d\phi}{\int P(\mathbf{w}_s|\mathbf{z},\phi)P(\mathbf{w}_t|\mathbf{z}_{-(t,i)},\phi)P(\phi|\beta)d\phi} \\ &\cdot \frac{\int P(\mathbf{z}_s|\theta^{(s)})P(\mathbf{z}_t|\theta^{(s)},\mathbf{s}_t,\mathbf{c}_t=0)P(\theta^{(s)}|\alpha_\theta,\mathbf{c}_s)d\theta^{(s)}}{\int P(\mathbf{z}_s|\theta^{(s)})P(\mathbf{z}_{-(t,i)}|\theta^{(s)},\mathbf{s}_t,\mathbf{c}_t=0)P(\theta^{(s)}|\alpha_\theta,\mathbf{c}_s)d\theta^{(s)}} \end{aligned} \tag{A.8}$$

And for $\mathbf{c}_t = 1$, we have:

$$\begin{aligned} P(\mathbf{z}_{t,i}|\mathbf{z}_{-(t,i)},\mathbf{z}_s,\mathbf{w}_s,\mathbf{w}_t,\mathbf{c}_s,\mathbf{c}_t=1,\mathbf{s}_t) =& \frac{P(\mathbf{w}_s,\mathbf{w}_t,\mathbf{z}_s,\mathbf{z}_t,\mathbf{c}_s,\mathbf{c}_t=0,\mathbf{s}_t)}{P(\mathbf{z}_{-(t,i)},\mathbf{z}_s,\mathbf{w}_s,\mathbf{w}_t,\mathbf{c}_s,\mathbf{c}_t=0,\mathbf{s}_t)} \\ =& \frac{\int P(\mathbf{w}_s|\mathbf{z}_s,\phi)P(\mathbf{w}_t|\mathbf{z}_t,\phi)P(\phi|\beta)d\phi}{\int P(\mathbf{w}_s|\mathbf{z},\phi)P(\mathbf{w}_t|\mathbf{z}_{-(t,i)},\phi)P(\phi|\beta)d\phi} \\ &\cdot \frac{\int P(\mathbf{z}_t|\psi^{(t)},\mathbf{c}_t=1)P(\psi^{(t)}|\alpha_\psi)d\psi^{(t)}}{\int P(\mathbf{z}_{-(t,i)}|\psi^{(t)},\mathbf{c}_t=1)P(\psi^{(t)}|\alpha_\psi)d\psi^{(t)}} \end{aligned} \tag{A.9}$$

Next, we describe how to derive Eq. A.8 and Eq. A.9. First, note that in both Eq. A.8 and Eq. A.9 the first fraction is basically as same as the one in Eq. A.4. Besides, the second fraction in Eq. A.8 is very similar to Eq. A.6 so that we can derive it analogously. Therefore, we focus on how to compute the integrals in the second fraction of Eq. A.9 in the rest part.

In specific, as $P(\mathbf{z}_t|\psi^{(t)},\mathbf{c}_t = 1)$ and $P(\psi^{(t)}|\alpha_\psi)$ are conjugate pair of Multinomial-Dirichlet, we could solve this integral as:

$$\begin{aligned} \int P(\mathbf{z}_t|\psi^{(t)},\mathbf{c}_t=1)P(\psi^{(t)}|\alpha_\psi)d\psi^{(t)} &= \int \prod_{i=1}^{T} \frac{1}{\Delta(\alpha_\psi)} \prod_{k=1}^{K} {\psi^{(t)}}^{n_k^i+\alpha_\psi-1} d\psi^{(t)} \\ &= \left(\frac{\Gamma(K\alpha_\psi)}{\Gamma(\alpha_\psi)^K}\right)^T \prod_{i=1}^{T} \frac{\prod_{k=1}^{K}\Gamma(n_k^i+\alpha_\psi)}{\Gamma(n_{(.)}^i+K\alpha_\psi)} \end{aligned} \tag{A.10}$$

where n_k^i is the number of times topic k appears in tweet t, where t is about a general topics sampled from $\psi^{(t)}$. $n_{(.)}^i = \sum_{k=1}^{K} n_k^i$ is the total number of times all topics $1...k$ appear in t.

So, to yield the second fraction of Eq. A.9, we apply the above equation twice. We obtain the following equation:

$$\begin{aligned}\frac{\int P(\mathbf{z}_t|\psi^{(t)}, \mathbf{c}_t = 1)P(\psi^{(t)}|\alpha_\psi)d\psi^{(t)}}{\int P(\mathbf{z}_{(-t,i)}|\psi^{(t)}, \mathbf{c}_t = 1)P(\psi^{(t)}|\alpha_\psi)d\psi^{(t)}} =& \frac{\left(\frac{\Gamma(K\alpha_\psi)}{\Gamma(\alpha_\psi)^K}\right)^T \prod_{i=1}^{T} \frac{\prod_{k=1}^{K}\Gamma(n_k^i+\alpha_\psi)}{\Gamma(n_{(.)}^i+K\alpha_\psi)}}{\left(\frac{\Gamma(K\alpha_\psi)}{\Gamma(\alpha_\psi)^K}\right)^T \prod_{i=1}^{T} \frac{\prod_{k=1}^{K}\Gamma(n_k^{i^{-(t,i)}}+\alpha_\psi)}{\Gamma(n_{(.)}^{i^{-(t,i)}}+K\alpha_\psi)}} \\ =& \frac{n_k^i + \alpha_\psi - 1}{n_{(.)}^i + K\alpha_\psi - 1} \qquad \text{(A.11)}\end{aligned}$$

where $\mathbf{c}_{(t,i)} = 1$, i is the tweet t, $\mathbf{z}_{(t,i)} = k$

Finally, by combining the derivations in Eq.A.4, Eq.A.6 and Eq.A.11, we obtain the Gibbs sampling update rules for $P(\mathbf{z}_t|\mathbf{z}_s, \mathbf{w}_s, \mathbf{w}_t, \mathbf{c}_s, \mathbf{c}_t, \mathbf{s}_t)$, which have two cases based on whether $\mathbf{z}_t$ is a specific topic ($\mathbf{c}_t = 0$) or a general topic ($\mathbf{c}_t = 1$):

$$P(\mathbf{z}_{t,i}|\mathbf{z}_{-(t,i)}, \mathbf{z}_s, \mathbf{w}_s, \mathbf{w}_t, \mathbf{c}_s, \mathbf{c}_t, \mathbf{s}_t) = \begin{cases} \frac{n_{sw}^k + n_{tw}^k + \beta - 1}{n_{s(.)}^k + n_{t(.)}^k + W\beta - 1} \times \frac{n_k^{\mathcal{S}_i} + nt_k^{\mathcal{S}_i} + \alpha_\theta - 1}{n_{(.)}^{\mathcal{S}_i} + nt_{(.)}^{\mathcal{S}_i} + K\alpha_\theta - 1} & \mathbf{c}_t = 0 \qquad \text{(A.12)} \\ \frac{n_{sw}^k + n_{tw}^k + \beta - 1}{n_{s(.)}^k + n_{t(.)}^k + W\beta - 1} \times \frac{n_k^i + \alpha_\psi - 1}{n_{(.)}^i + K\alpha_\psi - 1} & \mathbf{c}_t = 1 \qquad \text{(A.13)} \end{cases}$$

Inference of $\mathbf{C}_t$

Next, we examine $P(\mathbf{c}_t|\mathbf{c}_s, \mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t, \mathbf{s}_t)$. Based on the generative process of ET-LDA, $\mathbf{c}_t$ is a Binomial distribution which controls whether a tweet t should have general topic sampled from tweets corpus or specific topics sampled from a specific segment of the event. It has a beta prior with asymmetric parameters α_{λ_γ} and α_{λ_ψ}. As a result, the estimation of $\mathbf{c}_t$ can be two cases: (1) when $\mathbf{c}_t = 0$ (for specific topics), and (2) when $\mathbf{c}_t = 1$ (for general topics). Similar to $\mathbf{z}_s$ and $\mathbf{z}_t$, we have the conditional probability $P(\mathbf{c}_{t,i}|\mathbf{c}_{-(t,i)}, \mathbf{z}_s, \mathbf{z}_t, \mathbf{w}_s, \mathbf{w}_t, \mathbf{c}_s, \mathbf{s}_t)$. So when $\mathbf{c}_t = 0$ we have:

$$\begin{aligned}P(\mathbf{c}_{t,i}|\mathbf{c}_{-(t,i)}, \mathbf{c}_s, \mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t, \mathbf{s}_t) =& \frac{P(\mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t, \mathbf{c}_s, \mathbf{c}_t, \mathbf{s}_t)}{P(\mathbf{z}_s, \mathbf{z}_t, \mathbf{w}_s, \mathbf{w}_t, \mathbf{c}_s, \mathbf{c}_{-(t,i)}, \mathbf{s}_t)} \\ =& \frac{\int P(\mathbf{z}_s|\theta^{(s)})P(\mathbf{z}_t|\theta^{(s)}, \mathbf{s}_t, \mathbf{c}_t = 0)P(\theta^{(s)}|\alpha_\theta, \mathbf{c}_s)d\theta^{(s)}}{\int P(\mathbf{z}_s|\theta^{(s)})P(\mathbf{z}_t|\theta^{(s)}, \mathbf{s}_t, \mathbf{c}_{-(t,i)} = 0)P(\theta^{(s)}|\alpha_\theta, \mathbf{c}_s)d\theta^{(s)}} \\ &\cdot \frac{\int P(\mathbf{c}_t|\lambda^{(t)})P(\lambda^{(t)}|\alpha_{\lambda_\gamma}, \alpha_{\lambda_\psi})d\lambda^{(t)}}{\int P(\mathbf{c}_{-(t,i)}|\lambda^{(t)})P(\lambda^{(t)}|\alpha_{\lambda_\gamma}, \alpha_{\lambda_\psi})d\lambda^{(t)}} \qquad \text{(A.14)}\end{aligned}$$

And when $\mathbf{c}_t = 1$ we have:

$$\begin{aligned}P(\mathbf{c}_{(t,i)}|\mathbf{c}_{-(t,i)},\mathbf{c}_s,\mathbf{w}_s,\mathbf{w}_t,\mathbf{z}_s,\mathbf{z}_t,\mathbf{s}_t) =&\frac{P(\mathbf{w}_s,\mathbf{w}_t,\mathbf{z}_s,\mathbf{z}_t,\mathbf{c}_s,\mathbf{c}_t,\mathbf{s}_t)}{P(\mathbf{z}_s,\mathbf{z}_t,\mathbf{w}_s,\mathbf{w}_t,\mathbf{c}_s,\mathbf{c}_{-(t,i)},\mathbf{s}_t)}\\ =&\frac{\int P(\mathbf{z}_t|\psi^{(t)},\mathbf{c}_t=1)P(\psi^{(t)}|\alpha_\psi)d\psi^{(t)}}{\int P(\mathbf{z}_t|\psi^{(t)},\mathbf{c}_{-(t,i)}=1)P(\psi^{(t)}|\alpha_\psi)d\psi^{(t)}}\\ &\cdot\frac{\int P(\mathbf{c}_t|\lambda^{(t)})P(\lambda^{(t)}|\alpha_{\lambda_\gamma},\alpha_{\lambda_\psi})d\lambda^{(t)}}{\int P(\mathbf{c}_{-(t,i)}|\lambda^{(t)})P(\lambda^{(t)}|\alpha_{\lambda_\gamma},\alpha_{\lambda_\psi})d\lambda^{(t)}}\end{aligned}\tag{A.15}$$

Note that in Eq. A.14 the derivation of the first fraction is essentially as same as Eq. A.6. Similarly, the first fraction in Eq. A.15 is as same as the one in Eq. A.11. Therefore, next we focus on how to derive the second part of both Eq. A.14 and Eq. A.15.

Given the fact $P(\mathbf{c}_t|\lambda^{(t)})$ is a Binomial and $P(\lambda^{(t)}|\alpha_{\lambda_\gamma},\alpha_{\lambda_\psi})$ is a Beta prior, we can easily solve the integral as:

$$\begin{aligned}\int P(\mathbf{c}_t|\lambda^{(t)})P(\lambda^{(t)}|\alpha_{\lambda_\gamma},\alpha_{\lambda_\psi})d\lambda^{(t)} =& \int\prod_{t=1}^{T}P(\mathbf{c}_t|\lambda^{(t)})P(\lambda^{(t)}|\alpha_{\lambda_\gamma},\alpha_{\lambda_\psi})d\lambda^{(t)}\\ =&\prod_{t\in T}\left[\frac{\Gamma(\alpha_{\lambda_\gamma}+\alpha_{\lambda_\psi})}{\Gamma(\alpha_{\lambda_\gamma})\Gamma(\alpha_{\lambda_\psi})}\frac{\Gamma(M_t^0+\alpha_{\lambda_\gamma})\Gamma(M_t^1+\alpha_{\lambda_\psi})}{\Gamma(M_t+\alpha_{\lambda_\gamma}+\alpha_{\lambda_\psi})}\right]\end{aligned}\tag{A.16}$$

where M_t^0 is the number of words in tweet t whose topics are specific topics. On the other hand, M_t^1 is the number of words in tweet t whose topics are general topics. $M_t=M_t^0+M_t^1$ is the number of words in t.

As mentioned earlier, in order to yield the second fraction of Eq. A.14 using Gibbs sampling, we need to consider the situation about the excluded variable. Here, since the excluded variable is $c_{t,i}=0$, we can have $M_t^0=M^0_{t_{-(t,i)}}+1$ and $M_t=M_{t_{-(t,i)}}+1$. Therefore, we can obtain the following equation for $c_t=0$:

$$\begin{aligned}\frac{\int P(\mathbf{c}_t=0|\lambda^{(t)})P(\lambda^{(t)}|\alpha_{\lambda_\gamma},\alpha_{\lambda_\psi})d\lambda^{(t)}}{\int P(\mathbf{c}_{-(t,i)}=0|\lambda^{(t)})P(\lambda^{(t)}|\alpha_{\lambda_\gamma},\alpha_{\lambda_\psi})d\lambda^{(t)}} =&\frac{\prod_{t\in T}\left[\frac{\Gamma(\alpha_{\lambda_\gamma}+\alpha_{\lambda_\psi})}{\Gamma(\alpha_{\lambda_\gamma})\Gamma(\alpha_{\lambda_\psi})}\frac{\Gamma(M_t^0+\alpha_{\lambda_\gamma})\Gamma(M_t^1+\alpha_{\lambda_\psi})}{\Gamma(M_t+\alpha_{\lambda_\gamma}+\alpha_{\lambda_\psi})}\right]}{\prod_{t\in T}\left[\frac{\Gamma(\alpha_{\lambda_\gamma}+\alpha_{\lambda_\psi})}{\Gamma(\alpha_{\lambda_\gamma})\Gamma(\alpha_{\lambda_\psi})}\frac{\Gamma(M^0_{t_{-(t,i)}}+\alpha_{\lambda_\gamma})\Gamma(M_t^1+\alpha_{\lambda_\psi})}{\Gamma(M_{t_{-(t,i)}}+\alpha_{\lambda_\gamma}+\alpha_{\lambda_\psi})}\right]}\\ =&\frac{M_t^0+\alpha_{\lambda_\gamma}-1}{M_t+\alpha_{\lambda_\gamma}+\alpha_{\lambda_\psi}-1}\end{aligned}\tag{A.17}$$

Similarly, when $c_t=1$ we have $M_t^1=M^1_{t_{-(t,i)}}+1$ and also $M_t=M_{t_{-(t,i)}}+1$. Therefore, we can obtain:

$$
\begin{aligned}
\frac{\int P(\mathbf{c}_t = 1|\lambda^{(t)})P(\lambda^{(t)}|\alpha_{\lambda_\gamma}, \alpha_{\lambda_\psi})d\lambda^{(t)}}{\int P(\mathbf{c}_{-(t,i)} = 1|\lambda^{(t)})P(\lambda^{(t)}|\alpha_{\lambda_\gamma}, \alpha_{\lambda_\psi})d\lambda^{(t)}} =& \frac{\prod_{t\in T}\left[\frac{\Gamma(\alpha_{\lambda_\gamma}+\alpha_{\lambda_\psi})}{\Gamma(\alpha_{\lambda_\gamma})\Gamma(\alpha_{\lambda_\psi})}\frac{\Gamma(M_t^0+\alpha_{\lambda_\gamma})\Gamma(M_t^1+\alpha_{\lambda_\psi})}{\Gamma(M_t+\alpha_{\lambda_\gamma}+\alpha_{\lambda_\psi})}\right]}{\prod_{t\in T}\left[\frac{\Gamma(\alpha_{\lambda_\gamma}+\alpha_{\lambda_\psi})}{\Gamma(\alpha_{\lambda_\gamma})\Gamma(\alpha_{\lambda_\psi})}\frac{\Gamma(M_t^0+\alpha_{\lambda_\gamma})\Gamma(M^1_{t_{-(t,i)}}+\alpha_{\lambda_\psi})}{\Gamma(M_{t_{-(t,i)}}+\alpha_{\lambda_\gamma}+\alpha_{\lambda_\psi})}\right]} \\
=& \frac{M_t^1 + \alpha_{\lambda_\gamma} - 1}{M_t + \alpha_{\lambda_\gamma} + \alpha_{\lambda_\psi} - 1} \qquad \text{(A.18)}
\end{aligned}
$$

Finally, by combining the derivations in Eq.A.6, Eq.A.11, Eq.A.17 and Eq.A.18, we obtain the Gibbs sampling update rules for $P(\mathbf{c}_{(t,i)}|\mathbf{c}_{-(t,i)}, \mathbf{c}_s, \mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t, \mathbf{s}_t)$, which have two cases based on whether $\mathbf{c}_t$ draws a specific topic ($\mathbf{c}_t = 1$) or a general topic ($\mathbf{c}_s = 0$):

$$
P(\mathbf{c}_{(t,i)}|\mathbf{c}_{-(t,i)}, \mathbf{c}_s, \mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t, \mathbf{s}_t) = \begin{cases} \dfrac{M_t^0 + \alpha_{\lambda_\lambda} - 1}{M_t + \alpha_{\lambda_\gamma} + \alpha_{\lambda_\psi} - 1} \times \dfrac{n_k^{\mathcal{S}_i} + nt_k^{\mathcal{S}_i} + \alpha_\theta - 1}{n_{(.)}^{\mathcal{S}_i} + nt_{(.)}^{\mathcal{S}_i} + K\alpha_\theta - 1} & \mathbf{c}_t = 0 \qquad \text{(A.19)} \\[2ex] \dfrac{M_t^1 + \alpha_{\lambda_\psi} - 1}{M_t + \alpha_{\lambda_\gamma} + \alpha_{\lambda_\psi} - 1} \times \dfrac{n_k^i + \alpha_\psi - 1}{n_{(.)}^i + K\alpha_\psi - 1} & \mathbf{c}_t = 1 \qquad \text{(A.20)} \end{cases}
$$

Inference of $\mathbf{S}_t$

Now, let's look at $P(\mathbf{s}_t|\mathbf{c}_t, \mathbf{c}_s, \mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t)$. Similar to other variables, we cancel the factors that do not depend on $\mathbf{s}_t$ and write the conditional probability below. Note that $\mathbf{s}_t$ is a categorical distribution deciding which segment s of the event should be chosen for dawning the specific topics of tweet t. Therefore, the usage of $\mathbf{s}_t$ implies $\mathbf{c}_t = 0$.

$$
\begin{aligned}
P(\mathbf{s}_{(t,i)}|\mathbf{s}_{-(t,i)}, \mathbf{c}_t, \mathbf{c}_s, \mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t) =& \frac{P(\mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t, \mathbf{c}_s, \mathbf{c}_t, \mathbf{s}_t)}{P(\mathbf{z}_s, \mathbf{z}_t, \mathbf{w}_s, \mathbf{w}_t, \mathbf{c}_s, \mathbf{c}, \mathbf{s}_{t_{-(t,i)}})} \\
=& \frac{\int P(\mathbf{z}_s|\theta^{(s)})P(\mathbf{z}_t|\theta^{(s)}, \mathbf{s}_t, \mathbf{c}_t = 0)P(\theta^{(s)}|\alpha_\theta, \mathbf{c}_s)d\theta^{(s)}}{\int P(\mathbf{z}|\theta^{(s)})P(\mathbf{z}_t|\theta^{(s)}, \mathbf{s}_{t_{-(t,i)}}, \mathbf{c}_t = 0)P(\theta^{(s)}|\alpha_\theta, \mathbf{c}_s)d\theta^{(s)}} \\
&\cdot \frac{\int P(\mathbf{s}_t|\gamma^{(t)})P(\gamma^{(t)}|\alpha_\gamma)d\gamma^{(t)}}{\int P(\mathbf{s}_{t_{-(t,i)}}|\gamma^{(t)})P(\gamma^{(t)}|\alpha_\gamma)d\gamma^{(t)}} \qquad \text{(A.21)}
\end{aligned}
$$

For the first fraction in Eq. A.21, we solve it by following the same derivation procedure as Eq. A.6. Then for the second fraction, since it comprises of a Multinomial-Dirichlet pair, we solve it by the same procedure described in Eq. A.10. Thus, we have:

$$
\int P(\mathbf{s}_t|\gamma^{(t)})P(\gamma^{(t)}|\alpha_\gamma)d\gamma^{(t)} = \left(\frac{\Gamma(K\alpha_\gamma)}{\Gamma(\alpha_\gamma)^K}\right)^T \prod_{i=1}^{T} \frac{\prod_{s=1}^{S}\Gamma(n_s^i + \alpha_\gamma)}{\Gamma(n_{(.)}^i + S\alpha_\gamma)} \qquad \text{(A.22)}
$$

where S is the set of segments in the event's transcript. $s \in S$ is the segment that tweet t refers to. n_s^i is the number of times segment s is refereed by words in t.

By applying the Eq. A.22, we obtain the following equation for the second fraction of Eq. A.21:

$$\begin{aligned}\frac{\int P(\mathbf{s}_t|\gamma^{(t)})P(\gamma^{(t)}|\alpha_\gamma)d\gamma^{(t)}}{\int P(\mathbf{s}_{t_{-(t,i)}}|\gamma^{(t)})P(\gamma^{(t)}|\alpha_\gamma)d\gamma^{(t)}} =& \frac{\left(\frac{\Gamma(K\alpha_\gamma)}{\Gamma(\alpha_\gamma)^K}\right)^T \prod_{i=1}^T \frac{\prod_{s=1}^S \Gamma(n_s^i+\alpha_\gamma)}{\Gamma(n_{(.)}^i+S\alpha_\gamma)}}{\left(\frac{\Gamma(K\alpha_\gamma)}{\Gamma(\alpha_\gamma)^K}\right)^T \prod_{i=1}^T \frac{\prod_{s=1}^S \Gamma(n_{s_{(-t,i)}}^i+\alpha_\gamma)}{\Gamma(n_{{(.)}_{(-t,i)}^i}+S\alpha_\gamma)}} \\ =& \frac{\frac{\Gamma(n_s^i+\alpha_\gamma)}{\Gamma(n_{(.)^i}+S\alpha_\gamma)}}{\frac{\Gamma(n_{s_{(-t,i)}}^i+\alpha_\gamma)}{\Gamma(n_{{(.)}_{(-t,i)}^i}+S\alpha_\gamma)}} \\ =& \frac{n_s^i+\alpha_\gamma-1}{n_{(.)}^i+S\alpha_\gamma-1} \qquad \text{(A.23)}\end{aligned}$$

Finally, by plugging Eq. A.6 and Eq. A.23 into Eq. A.21 we have the update rule for the distribution of $\mathbf{s}_t$ as follows:

$$P(\mathbf{s}_{(t,i)}|\mathbf{s}_{-(t,i)}, \mathbf{c}_t, \mathbf{c}_s, \mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t) = \frac{n_k^{\mathcal{S}_i} + nt_k^{\mathcal{S}_i} + \alpha_\theta - 1}{n_{(.)}^{\mathcal{S}_i} + nt_{(.)}^{\mathcal{S}_i} + K\alpha_\theta - 1} \times \frac{n_s^i + \alpha_\gamma - 1}{n_{(.)}^i + S\alpha_\gamma - 1} \qquad \text{(A.24)}$$

Inference of $\mathbf{C}_S$

Last, we examine $P(\mathbf{c}_s|\mathbf{c}_t, \mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t, \mathbf{s}_t)$, which is the posterior distribution of the binary event-segmentation indicator $\mathbf{c}_s$. Similar to other variables we derived above, here we only consider the factors that depend on $\mathbf{c}_s$ in the joint distribution Eq. A.1. So we have:

$$\begin{aligned}P(\mathbf{c}_{(s,i)}|\mathbf{c}_{-(s,i)}, \mathbf{c}_t, \mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t, \mathbf{s}_t) =& \frac{P(\mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t, \mathbf{c}_s, \mathbf{c}_t, \mathbf{s}_t)}{P(\mathbf{z}_s, \mathbf{z}_t, \mathbf{w}_s, \mathbf{w}_t, \mathbf{c}_t, \mathbf{c}_{-(s,i)}, \mathbf{s}_t)} \\ =& \frac{\int P(\mathbf{c}_s|\delta^{(s)})P(\delta^{(t)}|\alpha_\delta)d\delta^{(s)}}{\int P(\mathbf{c}_{(-s,i)}|\delta^{(s)})P(\delta^{(t)}|\alpha_\delta)d\delta^{(s)}} \\ &\cdot \frac{\int P(\mathbf{z}_s|\theta^{(s)})P(\mathbf{z}_t|\theta^{(s)}, \mathbf{s}_t, \mathbf{c}_t = 0)P(\theta^{(s)}|\alpha_\theta, \mathbf{c}_s)d\theta^{(s)}}{\int P(\mathbf{z}_s|\theta^{(s)})P(\mathbf{z}_t|\theta^{(s)}, \mathbf{s}_t, \mathbf{c}_t = 0)P(\theta^{(s)}|\alpha_\theta, \mathbf{c}_{(-s,i)})d\theta^{(s)}} \\ & \text{(A.25)}\end{aligned}$$

For the first fraction in Eq. 21, we solve the integrals by following the same derivation procedure as Eq. A.16. The only difference is that in Eq. A.16 the beta prior is parameterized by asymmetric hyperparameters but here the beta prior $P(\delta^{(t)}|\alpha_{\lambda_\delta})$ is symmetric. Thus we have:

$$\int P(\mathbf{c}_s|\delta^{(s)})P(\delta^{(t)}|\alpha_\delta)d\delta^{(s)} = \frac{\Gamma(2\alpha_\delta)}{\Gamma(\alpha_\delta)^2}\frac{\Gamma(S_s^0+\alpha_\delta)\Gamma(S_s^1+\alpha_\delta)}{\Gamma(S+2\alpha_\delta)} \qquad \text{(A.26)}$$

where S_s^1 is the number of times paragraph's topic changes (i.e., $\mathbf{c}_s = 1$). $S = S_s^0 + S_s^1$ is the number of paragraphs in an event's transcript.

We begin with deriving the first fraction in Eq. A.26. As defined in the generative process, $\mathbf{c}_s$ is a Binomial distribution which controls whether a paragraph s should have the same topic distribution as its preceding paragraph $s-1$ ($\mathbf{c}_s = 0$) or have a new topic distribution sampled from a Multinomial ($\mathbf{c}_s = 1$). Therefore, the estimation of $\mathbf{c}_s$ can have two cases: (1) $\mathbf{c}_s = 0$ (i.e., paragraph merges into old segment), and (2) $\mathbf{c}_s = 1$ (starts a new segment).

Let us first consider the condition when $c_s = 1$. Given $\mathbf{c}_{(-s,i)}$ is the sequence excluded c_s, we can get $S_s^1 = S_{-(s,i)}^1 - 1$ and $S = S_{-(s,i)} - 1$. So we have:

$$\begin{aligned}\frac{\int P(\mathbf{c}_s|\delta^{(s)})P(\delta^{(t)}|\alpha_\delta)d\delta^{(s)}}{\int P(\mathbf{c}_{(-s,i)}|\delta^{(s)})P(\delta^{(t)}|\alpha_\delta)d\delta^{(s)}} =& \frac{\frac{\Gamma(S_s^0+\alpha_\delta)\Gamma(S_s^1+\alpha_\delta)}{\Gamma(S+2\alpha_\delta)}}{\frac{\Gamma(S_s^0+\alpha_\delta)\Gamma(S_{(-s,i)}^1+\alpha_\delta)}{\Gamma(S_{-(s,i)}+2\alpha_\delta)}} \\ =& \frac{S_t^1 + \alpha_\delta - 1}{S + 2\alpha_\delta - 1}\end{aligned} \tag{A.27}$$

Similarly, when $c_s = 0$. Given $\mathbf{c}_{(-s,i)}$ is the sequence excluded c_s, we can get $S_s^0 = S_{-(s,i)}^0 - 1$ and $S = S_{-(s,i)} - 1$. So we have:

$$\begin{aligned}\frac{\int P(\mathbf{c}_s|\delta^{(s)})P(\delta^{(t)}|\alpha_\delta)d\delta^{(s)}}{\int P(\mathbf{c}_{(-s,i)}|\delta^{(s)})P(\delta^{(t)}|\alpha_\delta)d\delta^{(s)}} =& \frac{\frac{\Gamma(S_s^0+\alpha_\delta)\Gamma(S_s^1+\alpha_\delta)}{\Gamma(S+2\alpha_\delta)}}{\frac{\Gamma(S_{(-s,i)}^0+\alpha_\delta)\Gamma(S_s^1+\alpha_\delta)}{\Gamma(S_{-(s,i)}+2\alpha_\delta)}} \\ =& \frac{S_t^0 + \alpha_\delta - 1}{S + 2\alpha_\delta - 1}\end{aligned} \tag{A.28}$$

Next, we calculate the second fraction in Eq. A.25. We first consider the situation when $\mathbf{c}_{(s,i)} = 0$ and is excluded from the sequence $\mathbf{c}_s$. As a result, for a segment $\mathcal{S}_i$ which does not contain paragraph s, we have $n_k^{\mathcal{S}_{i(-s,i)}} = n_k^{\mathcal{S}_i} - n_k^{(s,i)}$, where $n_k^{(s,i)}$ is the number of times top k appears in paragraph s. Thus, for the second fraction, based on the same procedures illustrated in Eq. A.5 and Eq. A.6 we have:

$$\frac{\int P(\mathbf{z}_s|\theta^{(s)})P(\mathbf{z}_t|\theta^{(s)}, \mathbf{s}_t, \mathbf{c}_t = 0)P(\theta^{(s)}|\alpha_\theta, \mathbf{c}_s)d\theta^{(s)}}{\int P(\mathbf{z}_s|\theta^{(s)})P(\mathbf{z}_t|\theta^{(s)}, \mathbf{s}_t, \mathbf{c}_t = 0)P(\theta^{(s)}|\alpha_\theta, \mathbf{c}_{(-s,i)})d\theta^{(s)}} \propto \frac{\prod_{k=1}^K \Gamma(n_k^{\mathcal{S}_i} + nt_k^{\mathcal{S}_i} + \alpha_\theta)}{\Gamma(n_{(.)}^{\mathcal{S}_i} + nt_{(.)}^{\mathcal{S}_i} + K\alpha_\theta)} \tag{A.29}$$

where $\mathcal{S}_i$ is the segment that paragraph s belongs to.

Similarly, since excluding $\mathbf{c}_{(s,i)} = 1$ from the sequence $\mathbf{c}_s$ makes a segment $\mathcal{S}_i$ split into two small segments: $\mathcal{S}_{(i-1)}$, which contains paragraphs from the beginning of original segment $\mathcal{S}_i$ to the one (i.e., $s-1$) right before s; and $\mathcal{S}_i$ which contains sentences from s to the end of $\mathcal{S}_i$. Given on this situation, based on the derivation procedures in Eq. A.5 and Eq. A.6, we have:

$$\frac{\int P(\mathbf{z}_s|\theta^{(s)})P(\mathbf{z}_t|\theta^{(s)}, \mathbf{s}_t, \mathbf{c}_t = 0)P(\theta^{(s)}|\alpha_\theta, \mathbf{c}_s)d\theta^{(s)}}{\int P(\mathbf{z}_s|\theta^{(s)})P(\mathbf{z}_t|\theta^{(s)}, \mathbf{s}_t, \mathbf{c}_t = 0)P(\theta^{(s)}|\alpha_\theta, \mathbf{c}_{(-s,i)})d\theta^{(s)}}$$

$$\propto \frac{\Gamma(K\alpha_\theta)}{\Gamma(\theta)^K} \frac{\prod_{k=1}^{K} \Gamma(n_k^{\mathcal{S}_{(s-1)}} + nt_k^{\mathcal{S}_{(s-1)}} + \alpha_\theta)}{\Gamma(n_{(.)}^{\mathcal{S}_{(i-1)}} + nt_{(.)}^{\mathcal{S}_{(i-1)}} + K\alpha_\theta)} \frac{\prod_{k=1}^{K} \Gamma(n_k^{\mathcal{S}_{(i)}} + nt_k^{\mathcal{S}_{(i)}} + \alpha_\theta)}{\Gamma(n_{(.)}^{\mathcal{S}_{(i)}} + nt_{(.)}^{\mathcal{S}_{(i)}} + K\alpha_\theta)} \tag{A.30}$$

Finally, we plug Eq. A.27, Eq. A.28, Eq. A.29 and Eq. A.30 into Eq. A.25 we have the update rule for the distribution of $\mathbf{c}_s$:

$$P(\mathbf{c}_s|\mathbf{c}_{-(s,i)}, \mathbf{c}_t, \mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t, \mathbf{s}_t) \propto \begin{cases} \frac{S_t^0 + \alpha_\delta - 1}{S + 2\alpha_\delta - 1} \times \frac{\prod_{k=1}^{K} \Gamma(n_k^{\mathcal{S}_i} + nt_k^{\mathcal{S}_i} + \alpha_\theta)}{\Gamma(n_{(.)}^{\mathcal{S}_i} + nt_{(.)}^{\mathcal{S}_i} + K\alpha_\theta)}, \quad \mathbf{c}_s = 0 & \text{(A.31)} \\ \frac{S_t^1 + \alpha_\delta - 1}{S + 2\alpha_\delta - 1} \times \frac{\Gamma(K\alpha_\theta)}{\Gamma(\theta)^K} \times \frac{\prod_{k=1}^{K} \Gamma(n_k^{\mathcal{S}_{(s-1)}} + nt_k^{\mathcal{S}_{(s-1)}} + \alpha_\theta)}{\Gamma(n_{(.)}^{\mathcal{S}_{(s-1)}} + nt_{(.)}^{\mathcal{S}_{(s-1)}} + K\alpha_\theta)} \\ \times \frac{\prod_{k=1}^{K} \Gamma(n_k^{\mathcal{S}_{(s)}} + nt_k^{\mathcal{S}_{(s)}} + \alpha_\theta)}{\Gamma(n_{(.)}^{\mathcal{S}_{(s)}} + nt_{(.)}^{\mathcal{S}_{(s)}} + K\alpha_\theta)}, \quad \mathbf{c}_s = 1 & \text{(A.32)} \end{cases}$$