Title
Joint Image-Text Topic Detection and Tracking for Analyzing Social and Political News Events

Author
Li, Weixin

Publication Date
2017

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

Joint Image-Text Topic Detection and Tracking for Analyzing Social and Political News Events

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Weixin Li

2017

Abstract of the Dissertation

Joint Image-Text Topic Detection and Tracking for
Analyzing Social and Political News Events

by

Weixin Li
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2017
Professor Song-Chun Zhu, Chair

News plays a vital role in informing citizens, affecting public opinion, and influencing policy making. The analyses of information flow in the news information ecosystem are important issues in social and political science research. However, the sheer amount of news data overwhelms manual analysis. In this dissertation, we present an automatic topic detection and tracking method which can be used to analyze the real world events and their relationships from multimodal TV news data. We propose a Multimodal Topic And-Or Graph (MT-AOG) to jointly represent textual and visual elements of news stories and their latent topic structures. An MT-AOG leverages a context sensitive grammar that can describe the hierarchical composition of news topics by semantic elements about people involved, related places and what happened, and model contextual relationships between elements in the hierarchy. We detect news topics through a cluster sampling process which groups stories about closely related events together. Swendsen-Wang Cuts, an effective cluster sampling algorithm, is adopted for traversing the solution space and obtaining optimal clustering solutions by maximizing a Bayesian posterior probability. The detected topics are then continuously tracked and updated with incoming news streams. We generate topic trajectories to show how topics emerge, evolve and disappear over time. We conduct both qualitative and quantitative evaluations to show the effectiveness and efficiency of the proposed approach over existing methods. We further expand our work to the analysis of campaign communication in recent presidential elections. Specifically, we apply fully automated coding on a massive collection

of news and other campaign information to track which candidates are discussed on Twitter and in traditional television news coverage; what topics are being discussed in relation to the candidates and by which news outlets; and which candidates were treated most favorably across news outlets and media. Our methods, which rely on machine learning and digital visual processing, offer promising new methods for social and political science scholars hoping to study large-scale information datasets.

The dissertation of Weixin Li is approved.

Wei Wang

Junghoo Cho

Francis F. Steen

Song-Chun Zhu, Committee Chair

University of California, Los Angeles

2017

To my family...

for their encouragement and support.

Table of Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Research Questions 2

1.3 Our Method and Contribution 3

2 Literature Review **6**

2.1 Topic Detection and Tracking 6

2.1.1 Topic Modeling 6

2.1.2 Document Clustering or Topic Detection 7

2.1.3 Topic Tracking 8

2.1.4 News Gathering and Delivering System 9

2.1.5 Markov Chain Monte Carlo Methods 10

2.1.6 Datasets 11

2.2 Presidential Election Visualization 11

2.2.1 Media Bias Research 11

2.2.2 Automated Visual Analysis of Mass Media 14

3 Joint Image-Text News Topic Detection and Tracking by Multimodal Topic And-Or Graph **15**

3.1 Introduction 15

3.1.1 Motivation and Objective 15

3.1.2 Overview of Our Method 17

3.1.3 Summary of Contributions 22

3.2 Topic Representation 22

3.2.1 Overall Representation . . . 22
3.2.2 Text Representation . . . 24
3.2.3 Image Representation . . . 25
3.2.4 Joint Image-Text Representation . . . 28
3.2.5 Empirical Evaluations of Assumptions in MT-AOG . . . 28
3.3 Topic Detection . . . 29
3.3.1 Problem Formulation . . . 29
3.3.2 Inference by Swendsen-Wang Cuts . . . 31
3.4 Topic Tracking . . . 34
3.5 Experiment . . . 35
3.5.1 Datasets . . . 35
3.5.2 Experiment I: Topic Detection . . . 37
3.5.3 Experiment II: Topic Tracking . . . 47
3.5.4 Experiment III: Large-Scale Topic Detection and Tracking . . . 52

4 Visualizing the US Presidential Elections . . . 56
4.1 Introduction . . . 56
4.2 Research Questions . . . 57
4.3 Data . . . 58
4.3.1 Television News Content . . . 58
4.3.2 Twitter Data . . . 60
4.3.3 Candidates . . . 60
4.4 Methods . . . 60
4.5 Results . . . 63
4.5.1 Which Candidates Received the Most Coverage? . . . 63

4.5.2 What Topics Are being Discussed? 67

4.5.3 How Favorably Are the Candidates Being Portrayed? 73

5 Conclusions **78**

References **80**

List of Figures

3.1 Overview of the proposed topic detection and tracking method. 18

3.2 Illustration of our Multimodal Topic And-Or Graph. 19

3.3 A common example pair of a face and a object cluster discovered by our algorithm. 27

3.4 The histogram of the number of stories in each topic and the fitting result. . 29

3.5 Adjacency graph. 30

3.6 Swendsen-Wang Cuts flips one selected component. 33

3.7 One topic tracking trajectory example. 35

3.8 An example showing the preprocessing results. 42

3.9 Top five detected topics. 43

3.10 Precision-recall curves of topic detection methods on UCLA Broadcast News Dataset. 46

3.11 Precision-recall curves of our topic detection methods with/without different contextual relations. 47

3.12 Topic tracking result of the event Santa Monica Shooting. 48

3.13 The text part of topics in one topic tracking trajectory. 49

3.14 The image part of topics in one topic tracking trajectory. 50

3.15 Emotion analysis for the Santa Monica Shooting event. 51

3.16 Precision-recall curves of topic tracking methods on UCLA Broadcast News Dataset. 52

3.17 Topic trajectories for 2012 CNN news. 54

3.18 Emotion analysis for more gun shooting events. 55

4.1 Candidate tracking, Viz2016.com site. 61

4.2 Volume of tweets about final four 2016 presidential candidates. 64

4.3 Monthly total seconds of candidate coverage by outlet. 66

4.4 Use of candidate still images per hour, by outlet and election year. 69

4.5 Viz2016.com topic tracking. 71

4.6 Viz2016.com topic trajectory, detailed daily view. 72

4.7 Volume and tone of tweets about Trump, Clinton, Sanders and Cruz. 75

4.8 Percent of still images containing smiles, by candidate and outlet. 77

List of Tables

3.1 Clustering Performance of different methods on Reuters-21578. 39

4.1 UCLA Communication Studies News Archive descriptive statistics. 59

4.2 Percent of time devoted to discussing candidates by outlet. 65

4.3 Uses of candidate images, by year, outlet, and candidate. 68

Acknowledgments

First, I would like to express my deepest gratitude and highest respect to my advisor, Prof. Song-chun Zhu, for his guidance and supervision during my study at UCLA. He is a true scholar who devotes all of his career to the marvelous field of computer vision and machine learning. He truly cares for his student and is very passionate about his research. From him, I learned how to do research with a high standard including defining problems, developing elegant algorithms, and presenting our work.

I would also like to thank Prof. Francis Steen, Tim Groeling, and Jungseock Joo for their guidance and insightful suggestions which helped me a lot with my research. My gratitude also goes to my other doctorate committee members, Prof. Junghoo Cho and Prof. Wei Wang, for their valuable advice and discussions which improved this dissertation.

I feel very fortunate to have worked at the VCLA lab with my friendly group members. They have helped me a lot through these difficult years. It is a great pleasure to work with them, and I deeply appreciate their friendship and insights. Particularly, I would like to thank Hang Qi, Tao Yuan, Leili Tavabi who I worked with in the news analysis project. I also thank Quanshi Zhang, and Tianfu Wu for their assistance. My thanks also go to Yang Lu, Xiaohan Nie, Siyuan Qi, Tianmin Shu, Nishant Shukla, Ping Wei, Dan Xie, Yuanlu Xu, Chengcheng Yu, Yibiao Zhao, Hanlin Zhu, Yixin Zhu.

I would like to thank my parents for their unwavering support, and my beloved husband for sharing every moment with me.

Vita

2010	B.S. in Computer Science and Technology, Beihang University, Beijing, China
2014	M.S. in Computer Science, UCLA, Los Angeles, CA

Publications

Weixin Li, Jungseock Joo, Hang Qi, and Song-Chun Zhu, "Joint Image-Text News Topic Detection and Tracking by Multimodal Topic And-Or Graph", in *IEEE Transactions on Multimedia (TMM)*, vol. 19, no. 2, pp. 367-381, 2017.

Tim Groeling, Jungseock Joo, Weixin Li, Francis Steen, "Visualizing the 2016 Presidential Election", *American Political Science Association (APSA) Annual Meeting*, 2016.

Jungseock Joo, Weixin Li, Francis F. Steen, and Song-Chun Zhu, "Visual Persuasion: Inferring Communicative Intents of Images", *Computer Vision and Pattern Recognition (CVPR)*, 2014.

CHAPTER 1

Introduction

1.1 Motivation

News plays an important role in providing us information about real-world events. Having access to up-to-date information about current events is critical to people to live in the constantly changing world. The information that news provides affects our lives in many ways. It affects how we think about the world—most of our views about the society, about most countries and regions are formed based on news. It also affects how we make decisions—from the most basic thing such as what to wear (according to the weather reports), to vital things such as what jobs to choose (based on news reports of different industries), which stock to buy (according to the company's financial news), and which presidential candidate to vote for (based on the news reports about their personalities, past experience, attitudes towards social and economic issues, etc.).

With the rapid change of information and communication technologies, and the rise of citizen journalism nowadays, we can get massive information from various platforms, such as newspapers, broadcast news (TV news), news websites, apps, blogs, social networks, etc. People can easily access what they want to know, and get real-time updates about important events through these platforms. These platforms together form a complex information ecosystem with rich heterogeneous data.

The analysis of information flow in the ecosystem is an important topic in social and political science research. Firstly, the mass media has tremendous power in selecting what to disclose and how it is reported. For instance, it is well known that FOX News Channel has conservative bias in their news coverage, and MSNBC has a bias towards left-leaning politics.

They can choose what to report and what not to report about certain presidential candidates in the Democratic or Republican Party. They can also show their attitudes towards these candidates in various ways such as manipulating their images and video footage. Moreover, what and how the mass media present the information can have great impacts on the public. For instance, people will consider certain issues as more important if they are reported frequently and prominently. The news stories covered by the media can also cause certain reactions in the stock market. These are all important research issues for social and political science researchers.

The sheer amount of textual and visual data in the information ecosystem bring serious challenges to the social and political science research, in which most previous work mainly uses manual human coding. Each news outlet, e.g. CNN, can generate 24 hours of video per day, and on Twitter, thousands of tweets will be tweeted per second on average. Such a great amount of data make manual large-scale and quantitative analysis extremely difficult and expensive. The objective of this dissertation is to develop automatic methods that can provide promising news parsing solutions to serve as the basis for further information flow analyses in the ecosystem.

1.2 Research Questions

News provides information about current events. News stories are created and recorded everyday by various news platforms. Accordingly, for mass media analysis, it is desirable to have an automatic system which can summarize and organize the large and continuously updated news collection into events. The main purpose of this dissertation is to develop an automatic news topic detection and tracking method to facilitate further media analysis.

This dissertation mainly focuses on the broadcast TV news, which is the dominant source where people get their news for decades. Compared to other news sources, TV news has two distinct properties—multimodal and event-centric. TV news disseminates current events and provides updates as the events develop. Both textual and visual cues are used in TV news to present a detailed description of the reported events. Most existing methods in topic

detection and tracking are unimodal and typically only deal with the textual data. Different from these method, our approach aims at jointly modeling both textual and visual data. Moreover, to understand news events, it is important to know the information related to the four Ws: who, what, when, and where. While most previous work usually model all these four aspects as a whole, we aim at clearly representing the topic's semantic structures and modeling their changes as the events evolve.

Moreover, in TV news, data streams are continuously updated, involving the reports about the development of previously existing events and the emergence of new events. Periodically rebuilding the topic models on the entire updated collection can take a long time period and lead to heavy re-computation. Thus solving the topic detection and tracking task efficiently becomes another main objective of our method.

Based on the news events obtained from the automatic topic detection and tracking, we further visualize the US presidential elections as a case study. Campaign communication aims at shaping candidates' images and managing the media to help candidates reach voters. It is an important research topic in social and political science. We aim at conducting large-scale quantitative analysis of campaign communication, which social and political researchers seriously long for, especially when facing the sheer amount of data. We study several important research questions in the campaign communication, including: (1) what topics are reported and discussed about different candidates in different news outlets, (2) which candidates are discussed on Twitter and in the news outlets, and (3) how different news outlets favored different candidates reflected by the images that they select to represent these candidates. By studying these questions, we can get an idea of the media bias in the campaign communication.

1.3 Our Method and Contribution

To detect news events and track their evolution over time, we propose a joint image-text topic detection and tracking method based on Multimodal Topic And-Or Graph.

To jointly model both textual and visual data in the news collection and the event's

semantic structures, we propose a novel multimodal topic representation, i.e. Multimodal Topic And-Or Graph (MT-AOG), based on the And-Or Graph [ZM06] that is generally used for computer vision problems. The MT-AOG is a hierarchical and compositional model with a context-sensitive grammar, which jointly models hierarchical topic compositions of texts and visuals. It is suitable to represent the fine-grained news event structures and the event relationships.

To deal with the massive and continuously updated news data, we detect topics within short time periods and further discover long-time topic trajectories, so that we can show both detail descriptions about each topic in different time periods, and how these topics evolve over time.

For topic detection, we group stories that elaborate the same topics using cluster sampling methods by maximizing a Bayesian posterior probability. A graph-partitioning based cluster sampling algorithm, i.e. Swendsen-Wang Cuts, is adopted for its efficiency in sampling the solution space.

For topic tracking, after detecting topics within short time periods, we further link them to generate long-time topic trajectories by considering topic similarities in terms of both textual and visual channels, and the topics' temporal relations.

Our joint image-text topic detection and tracking method is described in Chapter 3.

For visualizing the US presidential election, our study uses a coding scheme similar to previous manual studies in the social and political research but replaces all manual hand-coding with the help of automatic computer vision and machine learning methods. And due to the use of fully automatic analysis methods, we can easily process tremendously more data compared to previous manual studies.

Our main data source is the UCLA Communication Studies News Archive, which contains daily US television news contents from October 2006. So our analysis focuses on the 2008, 2012, and 2016 US presidential elections, which have corresponding news reports recorded in the archive. We also collect Twitter data about the 2016 election candidates daily starting from August 2015.

To answer the first question we mentioned in the previous section, i.e. what topics are discussed about the candidates in TV news, we use our proposed joint image-text topic detection and tracking method to track what topics appear in news everyday, and how these topics evolve over time. We chart the top daily topics and their relationship with prior and subsequent topics. We also track the top weekly topics by different news outlets to generate a weekly summary.

For the question about which candidates are discussed on Twitter and in TV news, we count different candidates' mentions on both platforms. We use the number of daily tweets that contain one specific candidate's name to represent his/her mention counts on Twitter. For TV news, we calculated the mention time each candidate of interests is covered in the election related news. Coreference resolution is used for including sentences where pronoun appears.

To solve the third question of how news outlets favored different candidates especially reflected by images which are carefully selected, we use smiling as a straight measure of the image favorability. We detect faces in the news video frames, and further do face recognition and smile classification by Convolutional Neural Networks. We then examine the visual sentiment of our news data for different candidates by calculating the proportion of smiling images.

Our work of visualizing the US presidential elections is introduced in Chapter 4.

CHAPTER 2

Literature Review

2.1 Topic Detection and Tracking

Our topic detection and tracking work is mainly related to the following five research streams: topic modeling, document clustering or topic detection, topic tracking, news gathering and delivering systems, and Markov Chain Monte Carlo methods. We will also briefly introduce some multimedia datasets.

2.1.1 Topic Modeling

Probabilistic topic models [Hof99, Ble12] have been widely used for detecting and analyzing latent topics, such as the latent Dirichlet allocation (LDA) model [BNJ03, GS04] and its extensions [BL06, TJB06, WB11]. Even though these methods are effective in general topic modeling, they typically rely on the bag-of-words (BoW) representation. The BoW representation is computationally efficient, but it ignores the semantic and compositional structures of news events. Some methods have also been proposed to relax BoW assumption. In Boyd-graber and Blei's work [BB09], the linguistic structures of sentences are considered in the topic model, constraining word both thematically and syntactically. In Wallach's work [Wal06], word order is incorporated in the proposed model. However, news stories are generally driven by events, so information from aspects like "who", "where" and "what" is crucial for summarizing these stories and generating meaningful news topics. Newman et al. [NCS06] considered these aspects but included them as a whole. Li et al. [LWL05] used this information but assume that these components are independent. Moreover, all the aforementioned methods are unimodal methods which only use texts.

Multimodal probabilistic topic models have also been proposed in the literature [BJ03, PAN10, ZL12]. To detect Twitter events, Cai et al. [CYL15] proposed a Spatial-Temporal Multimodal TwitterLDA model which uses five Twitter cues including text, image, location, timestamp, and hashtag, and modeled topics as location-specific distributions. Qian et al. [QZX16] proposed a multimodal event topic model for social event analysis. But in their model, no compositional structures are considered for the textual or visual modality. Chen et al. [CSH15] proposed a Visual-Emotional LDA (VELDA) model which relates tweet images and texts both visually and emotionally for image retrieval. Jia et al. [JSD11] proposed a Multimodal Document Random Field (MDRF) model for image retrieval, which is built using a Markov random field over LDA. For both VELDA and MDRF, there is only one image for one document. Our method is designed to detect and track news topics using broadcast news videos.

We pose the topic detection problem as a graph partitioning problem, and organize news stories in a graph. Some probabilistic topic models also build document networks. The Rational Topic Model (RTM) proposed by Chang et al. [CB09], and Semi-Supervised Relational Topic Model (ss-RTM) proposed by Niu et al. [NHG14] are both extensions of LDA which account for links between documents when modeling topics. RTM models networks of text data, e.g. citation networks of documents. Ss-RTM is designed for recognizing images with text tags in social media. It jointly models image contents and their links (two images are linked if they share one or more common text tags). Both RTM and ssRTM use data from one modality to build links, and use data from another modality in nodes, while our method jointly models both texts and visuals in nodes and links. Our model for graph partitioning also considers the total partition number and partition size distributions.

2.1.2 Document Clustering or Topic Detection

Clustering based methods are also widely used for the task of news topic detection. A large number of methods for topic detection in the Topic Detection and Tracking (TDT) research [All02] (e.g. [ACD98, YCG99]) use clustering methods for detecting news topics,

where stories on the same topic are gathered. Traditional document clustering methods [AZ12, SKK00] can also be used for topic detection. However, most of these methods work on unimodal data and mainly focus on the text domain.

Multimodal topic clustering methods have been proposed by taking both texts and visuals into consideration. In most of these methods, texts are represented using the BoW representation [WNL06, WNH08, CZL14, ZS05]. For visual representation, some methods use color histograms of the keyframes [WNL06]. Other methods detect the near-duplicate keyframes (NDK) first and then use them to build visual relations between news stories [WNH08, CZL14]. Even though these methods can compute the visual similarities between stories, they are not capable of modeling the decomposition of visual parts in news topics. In terms of the clustering methods, [WNL06] and [WNH08] used co-clustering algorithm and one of its extensions with constraints added respectively. [ZS05] groups news stories based on the linear combination of textual and visual similarities. [CZL14] detects topics within one multimodal graph, which is obtained by merging one text graph and another visual graph constructed based on LDA and NDK respectively.

Some work also combined topic modeling and document clustering together, such as the multi-grain clustering topic model (MGCTM) proposed by Xie et al. [XX13]. They showed that these two tasks are closely related and can help each other as both performances are improved. This work still remains in the pure text domain and uses the BoW representation.

2.1.3 Topic Tracking

The traditional topic tracking problem in TDT ([All02, ACD98]) is defined as the process of finding related additional stories for some pre-learned topics. Many methods have been proposed for solving this problem such as those in [All02, MAS04, HC06]. However, deciding the topic of each incoming story based on the previous learned topics can take a long time in a large data collection.

In the probabilistic modeling community, some models incorporate time information, such as the Dynamic Topic Model (DTM) [BL06] which models topic evolution over time, and

the temporal Dirichlet process mixture model (TDPM) [AX08] for evolutionary clustering. In DTM, it is assumed that topics exist throughout the whole time period, which is usually not the case in the news domain. TDPM generates clusters that fit the data during each time period as much as possible while preserves the smoothness of clustering results over time. Both DTM and TDPM are unimodal.

Instead of using the previous two methods, we choose to do topic tracking by linking topics detected in different time periods. Some linking methods, such as those by Mei et al. [MZ05] and Kim et al. [KO11], are closely related to our topic tracking task. However, the method in [MZ05] is designed for news about some specific topics such as "tsunami." The similarity matrices used in [KO11] are based on the topics obtained by the original LDA model with BoW assumption. Moreover, both of the two methods are based on textual information only.

2.1.4 News Gathering and Delivering System

Several news gathering and delivering systems have been presented recently, such as News Rover [JLE13, LJE13] and EigenNews [YVC13, DVC13]. News Rover relies on external sources (e.g. Google News, which presumably uses user-click data, etc.) to get corresponding topics for TV news stories. TV news stories and collected topics are linked using the combination of NDK based visual similarity and BoW based textual similarity. EigenNews focuses on individual stories without the notion of topic. It discovers links among news stories and online articles by matching keyframes based on local visual features or matching texts based on BoW histograms and named entities. Different from these two systems, we learn topics solely from TV news data. Another difference of our method is that we use a joint probabilistic model of images and texts, and perform learning and inference on this unified representation.

Besides the previous four research streams, our work is also related to event coreference resolution [BH14, ZLJ15]. Zhang et al. [ZLJ15] proposed to detect coreferential news event pairs by incorporating textual and visual similarities. However, coreferential events are

defined to be the specific event occurrence mentioned in different sentences/documents with exactly the same characteristics (location, time, involved people, etc.), so event coreference resolution is not designed to deal with event evolutions, which is the goal of this work.

Since we use entities in the topic representation (i.e. "who", and "where"), our work is also related to another problem in the literature: Knowledge Base Population (KBP), which is the task of discovering facts about entities to augment a knowledge base (KB) [JG11, JGD10]. There are two tasks in KBP: Entity Linking - linking names in context to entities in the KB, and Slot Filling - adding information about an entity to the KB. In our work, we consider entities ("who" and "where") in the topic representation and model related contextual relations to get more meaningful topics. Different from the KBP problem, we focus on detecting and tracking topics using these entities as features of news stories, instead of gathering information about these entities from a corpus and expanding a knowledge base.

2.1.5 Markov Chain Monte Carlo Methods

Markov Chain Monte Carlo (MCMC) method is intensively used to traverse the space and sample the optimal solution in this work. MCMC method is especially effective when sampling a high-dimensional space. As an early MCMC algorithm in graph partition, Gibbs Sampler [GG84] requires exponential waiting time for the Markov Chain to converge to the target distribution.

Swendsen-Wang's algorithm [SW87] is a clustering sampling method to facilitate the convergence process in classic Ising and Potts model. In [BZ05], Swendsen-Wang Cuts (SWC) is proposed to generalize the original Swendsen-Wang's algorithm to arbitrary probability. The correctness can be proved from the perspective of Metroplis-Hastings algorithm. Swendsen-Wang Cuts algorithm has been successfully applied to several image analysis problems, such as image segmentation and stereo [BZ05]. Our work will add to the list of successful applications of SWC.

2.1.6 Datasets

In our work, we collect a new dataset named UCLA Broadcast News Dataset since there is a lack of publicly available multimedia dataset for news topic detection and tracking. Even though some multimedia news datasets have been used in previous work, such as the Topic Detection and Tracking (TDT) datasets [All02], and the TRECVID corpus [SOK06], they are not publicly available, and some of them do not have ground-truth annotations. News video datasets for other tasks have also been presented in the literature, e.g. the REPERE corpus for multimodal person recognition [GCM12] and Stanford I2V dataset for image-to-video visual search [ACC15].

2.2 Presidential Election Visualization

Our research for visualizing the presidential election touches on two main threads in the scholarly literature:

- empirical studies of media bias, and
- methodological explorations of the utility of machine learning for social science topics.

2.2.1 Media Bias Research

Scholars attempting to understand the characteristics of campaign news face limitations in their ability to observe information that hasn't been selected to be news ([Ham06]; see also [Gro13], on the difficulty of analyzing selection bias in news). However, there is a long tradition of studying coverage across news organizations in order to understand the idiosyncrasies of any single news organization's coverage. For example, Hofstetter [Hof76] attempted to use similarities in coverage across the three broadcast networks to identify the impact of structural (non-ideological) factors. As Schiffer [Sch06] observed, however, such an analysis "cannot distinguish whether a consistent slant... results from uniformity in non-ideological journalistic norms and constraints or from uniformity in bias".

Other scholars have attempted to use varying media coverage to generate a more comprehensive population of stories. For example, Stovall [Sto88] conducted an extensive initial search across news organizations to aggregate a comprehensive list of presidential campaign events, then used that list to search for mentions of those events in other outlets. Barrett and Peake [BP07] also chose to use a known set of presidential events (in this case, out of town travel by George W. Bush in 2001) to examine patterns in the coverage of different news organizations. Their design used the *Washington Post*'s coverage of these trips to provide a baseline for examining the volume and tone of local newspaper coverage of those same trips.

Covert and Wasburn [CW07] also use some media organizations as exemplars to help interpret the choices of others. In this case, however, they use overtly partisan news magazines (The *Progressive* and the *National Review*) to provide examples of what liberal and conservative coverage would look like, and then test for similarity in the coverage of *Time* and *Newsweek* over 25 years.

In one of the most influential recent studies of political news, Gentzkow and Shapiro [GS10] use a big data approach to understand ideological slant in political news. In their study, they assembled a corpus "of all phrases used by members of Congress in the 2005 *Congressional Record*, and identify those that are used much more frequently by one party than by another". They then searched the news coverage of hundreds of newspapers from the same time period to determine "whether the newspaper's language is more similar to that of a congressional Republican or a congressional Democrat." Their technique has the advantage of massive scalability: they report that on average, each of their 433 newspapers used their sample phrases more than 13,000 times in a given year, which would have been challenging to code through conventional means.

Another good example of this new approach is Ho and Quinn [HQ08], who infer the ideology of news organizations by comparing their editorial stances on major non-unanimous Supreme Court cases. They conceive of such editorials as "votes on the same issue facing the governmental decisionmakers in question. Combining this insight with well-developed statistical methods for ideal point estimation allows us to jointly analyze governmental actors and newspaper editorial boards, placing newspapers on a long-validated, substantively

meaningful, and transparent scale of political preferences".

Despite the importance of local, network, and cable television news, a surprisingly large proportion of media bias studies continue to focus on newspapers as their main or exclusive medium of study. Undoubtedly, part of this stems from the relatively broad availability of full text content from newspapers (versus the erratic availability of such transcripts for other media). However, even when television transcripts are available, they present researchers with an incomplete picture of reality by excluding the visual dimension of what are fundamentally visual media. There have been several projects that have attempted to tackle the relatively difficult goal of systematically analyzing the visuals used in the news. For example, Waldman and Devitt [WD98] analyzed photographs of presidential nominees in the 1996 presidential race used in five major newspapers, finding that the favorability of the images seemed to track the candidates standing in the polls more than the editorial slant of individual newspapers. Barrett and Barrington [BB05] tracked the presentation of candidate visuals in closely-matched races, finding a high degree of correspondence between a paper's current and historic editorial slant and the relative favorability of images used for each candidate.

In perhaps the most ambitious study in this area, Grabe and Bucy [GB09] attempt to systematically test for visual bias in television news. Their detailed hand-coded content analysis scheme codes the visual framing of coverage, visual weight, and various metrics of beneficial (getting the "last say," low angles, close-ups, zoom-ins, and eyewitness camera perspectives) or detrimental packaging ("lip flap" clips, in which the candidate is shown speaking but not heard, high angles, extreme close-ups, long shots, and zoom-outs). Similarly, Banning and Coleman [BC09] performed a content analysis of the nonverbal messages in 1159 shots drawn from TV coverage of the 2000 election. In so doing, they coded the facial expressions, appearance, and nonverbal behavior of the candidates, and the structural features of television edited into news stories by journalists (camera angle, distance, and movement). Finally, Hehman [HGH12] examined over 400 presidential images from five online media sites and coded them for warmth and competence, finding more favorable portrayals on sites sharing the president's ideology.

2.2.2 Automated Visual Analysis of Mass Media

An important component of our analysis is centered around the faces of politicians in TV news. The human face is a well-established research target in many disciplines across social sciences as well as computer vision and machine learning. Some researchers have argued that there exists a salient link among facial appearance of people, the perception of their personalities, and the outcomes of real world events such as elections [TMG05] or criminal sentencing [BJC04]. In computer science and statistics, it has been also reported that the perceived personality traits can be automatically predicted by machine learning approaches [ZFM03, RMT11, VSY14]) and the election results can be also predicted from the automatically rated facial traits [JSZ15]. These studies demonstrate the utility of computer vision and machine learning approaches as a data analysis tool to social science research.

Such an automated, data-driven, and computational analysis is especially powerful when the data corpus size is large, as it easily enables one to investigate the whole dataset—unlike a sampling-based inspection. For instance, Joo, Li, Steen, and Zhu [JLS14] systematically analyzed a huge amount of online newspaper articles and associated photographs of politicians using a hierarchical model and compared the visual favorability of the U.S. President in mass media with the public opinion toward him. Zhu, Luo, You, and Smith [ZLY13] also analyzed election related images in social media and linked particular visual features to viewer engagement.

In this dissertation, we use a fully automated system to find and infer the emotional state of major party presidential candidates in images aired in television news coverage during the 2008, 2012, and 2016 presidential election years. For each candidate, we then test for differences in the positivity of their images across news outlets and types of image. We believe this approach is more scalable and objective than prior attempts to code similar content using human coders alone, and should open new research questions to systematic analysis.

CHAPTER 3

Joint Image-Text News Topic Detection and Tracking by Multimodal Topic And-Or Graph

News stories provide information about real-world events and play a vital role in people's everyday life. The analyses of information flow in mass media, such as selection and presentation biases, agenda-setting patterns, persuasion techniques, or causal analysis are important issues in social and political science research. Our primary objective is to develop an automatic *topic detection and tracking* method which can be used to analyze the real world events and their relationships.

In this chapter, we introduce the proposed Multimodal Topic And-Or Graph based joint image-text topic detection and tracking method. The chapter is primarily based on the material published in [LJQ17].

3.1 Introduction

3.1.1 Motivation and Objective

News deals with an event and is presented in real-time as the event progresses. It updates and revises what have been reported. It also predicts the potential changes that may or may not follow in the future. Therefore, its narratives mostly focus on the temporal and causal relationships between events and how each event is dynamically transformed, based on observations made in particular points in time. Consequently, the most important thing in studying news is to understand how news stories are connected to each other over time, and this is our primal concern in this chapter – to identify news stories about the same event

and to monitor how they evolve.

Accordingly, we consider two related tasks in this work: topic detection and tracking [All02]. First, topic detection is aimed at clustering relevant news stories together on the fly where a topic is defined as each cluster and the corresponding multimodal model learned from it. Then we track these topics with continuously updated news data. Our objective is to generate topic trajectories to show how topics emerge, evolve, and disappear, and how their components change over time.

Our method specifically targets the domain of TV news, having two distinct properties from other types of corpora – multimodal and event-centric.

First of all, TV is a multimodal medium and TV news uses both verbal and non-verbal modalities via audio and video channels (our speech data is encoded as text via closed-captioning). Both textual and visual cues are important to understand the events described in the news. The visual dimension of mass media can be especially critical in relation to public response and engagement [JLS14], [JSZ15]. Our model jointly captures both dimensions unlike most existing approaches in topic detection/tracking which only use text inputs.

Secondly, TV news presents stories on real-world events. For those events, the key things to understand are "who did what, when, and where." Barack Obama's winning 2008 election is a completely different event than his re-election in 2012; but they are closely related. These events dynamically introduce new people or new places involved and are eventually connected to other events. Therefore, the model to deal with TV news should be able to clearly represent the semantic structure of an event as well as its local and global changes and relations with other events.

Moreover, despite the decades of study, there is a lack of publicly available multimedia dataset for evaluating news topic detection and tracking methods. Even though some multimedia news datasets have been used in previous work, such as the TDT datasets [All02], and the TRECVID corpus [SOK06], they are not publicly available, and some of them do not have ground-truth annotations.

To address these issues, we propose a novel multimodal topic representation, i.e. Mul-

timodal Topic And-Or Graph (MT-AOG), based on And-Or Graph (AOG), which is commonly used for various visual models [ZM06]. The core idea of AOG is hierarchical and compositional model, which is suitable to represent the news event structures and the event relationships. To discover topics and learn the model, we also adopt a graph-partitioning based cluster sampling method, Swendsen-Wang Cuts (SWC) [BZ05], which was originally developed for image parsing.

For evaluation, we use data from the UCLA Library Broadcast NewsScape[1], which contains a large number of broadcast news programs from the U.S. and the world since 2005. To collect the ground-truth data, we annotate a subset from the large collection.

3.1.2 Overview of Our Method

Fig. 3.1 shows an overview of our topic detection and tracking method. Both news videos and closed captions are the inputs to our method. After pre-processing steps such as story segmentation, we detect topics using a cluster sampling method, Swendsen-Wang Cuts (SWC), based on the proposed Multimodal Topic And-Or Graph (MT-AOG) which jointly models texts and images and organizes news topic components in a hierarchical structure. We further link topics detected in different time periods to generate topic trajectories which show how topics evolve over time. We describe our core representation and the main tasks in the following subsections.

3.1.2.1 Multimodal Topic And-Or Graph (MT-AOG)

We briefly introduce the proposed MT-AOG here. AOG has been used for modeling humans, objects and scenes in computer vision [JWZ12, YNL14]. MT-AOG embeds a context-sensitive grammar that jointly models hierarchical topic compositions of texts and images. There are three types of nodes in MT-AOG: AND-nodes representing compositions of sub-components (e.g. a topic is composed of the text part and the image part), OR-nodes for alternative structures (e.g. different configurations of a component in the topic structure),

[1] http://newsscape.library.ucla.edu

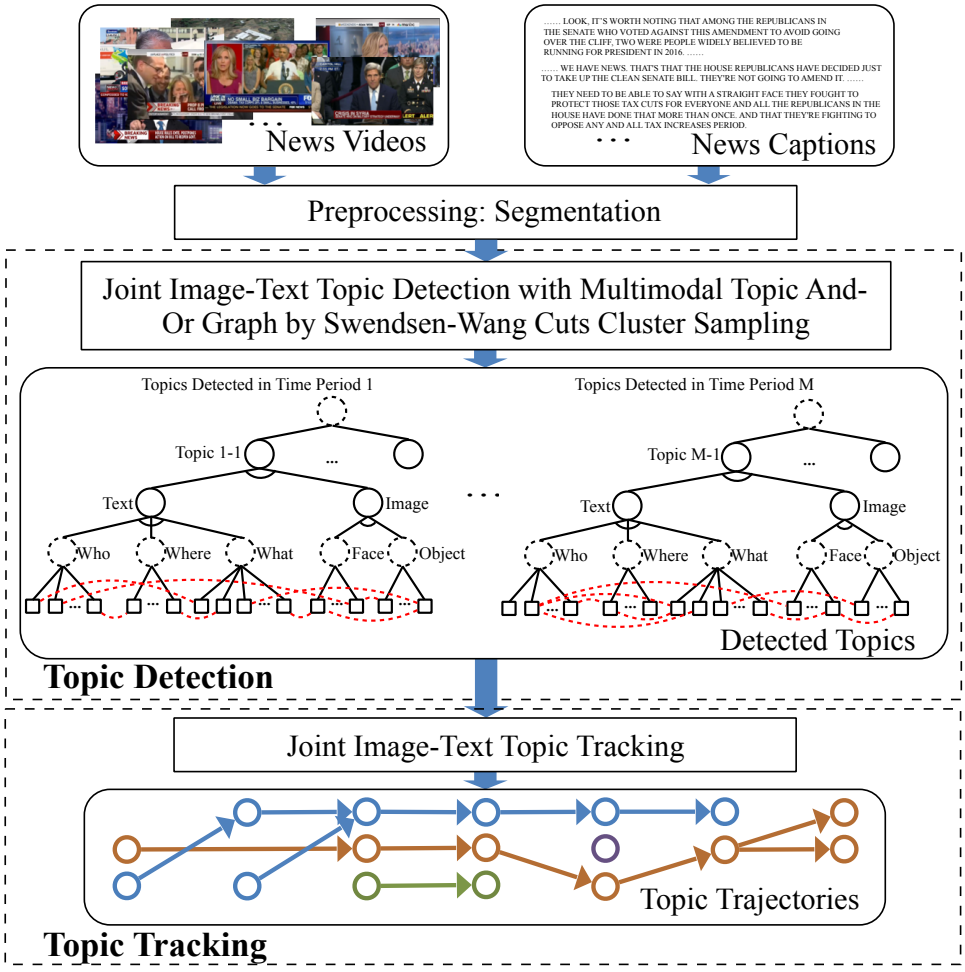


Figure 3.1: Overview of the proposed topic detection and tracking method. The inputs include both news videos and closed captions (texts). We detect topics through a joint image-text cluster sampling method within each time window. Then detected topics are tracked over time to form topic trajectories.

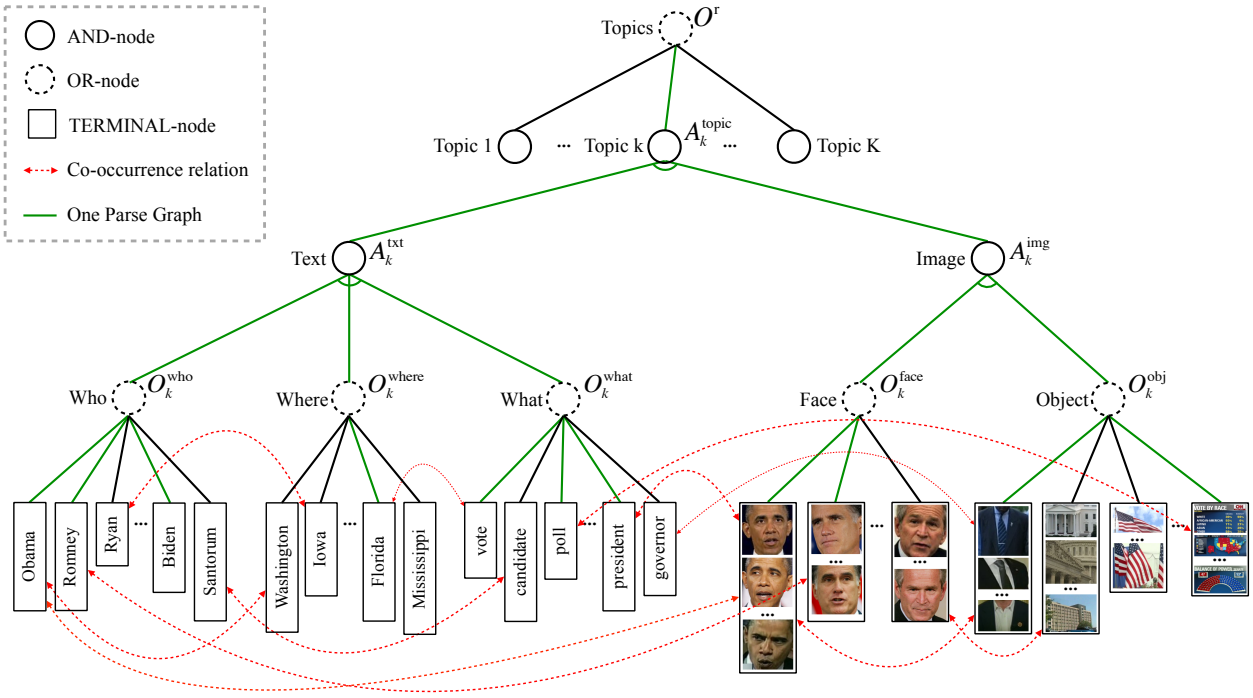


Figure 3.2: Illustration of our Multimodal Topic And-Or Graph. Three types of nodes are included: 1) AND-nodes representing topics' compositions, 2) OR-nodes for alternative structures, and 3) TERMINAL-nodes representing the most elementary, atomic components. The dashed red lines represent different components' co-occurring pairs. The green lines show an example of the parse graph.

and TERMINAL-nodes representing the most elementary components.

Fig. 3.2 illustrates the MT-AOG:

- *The root OR-node* in the top layer represents a number of distinct topic configurations. Each topic configuration specifies the actual contents of the topic.
- Each topic configuration is then represented by *a single topic AND-node* A_k^{topic} ($k = 1, ..., K$ where K is the total topic number) in the second layer. This node is composed of two parts, representing texts and images respectively.

Text Representation. The text part of each topic is represented by an *AND-node*, and its three components encode the knowledge of "who", "where" and "what." These three components describe the people involved, related places, and what happened respectively.

They are the three key aspects in the journalism's five W's [Har96, HLJ81] for describing news events and topics. More details will be provided in Section 3.2.2.

The "who", "where" and "what" components are all represented by *OR-nodes* (nodes O_k^{who}, O_k^{where}, and O_k^{what} in Fig. 3.2). All of these nodes can describe a set of possible words for the corresponding components. A certain news story may trigger a subset of these words. The words are represented by *TERMINAL-nodes* in the last layer. We also embed the contextual relations between the three components in the AOG. They are described using information from two aspects: (1) the co-occurrences of words from different components (such as those marked by the dashed red lines in Fig. 3.2); and (2) the ratios of entity numbers of different components (e.g. some topics have more people involved compared to the related locations).

Image Representation. The image part of each topic is also represented by an *AND-node* as shown in Fig. 3.2 (node A_k^{img}). This node has two components that capture two important visual signals in news: faces and objects. Faces show the main people related to the topic, and objects include other general information about the scene and the event. More details will be shown in Section 3.2.3.

The face and object components are represented by *OR-nodes* (nodes O_k^{face} and O_k^{obj} in Fig. 3.2), which can describe a set of possible entities similar to O_k^{who}, O_k^{where}, and O_k^{what}. Each face/object entity corresponds to one cluster of face/object patches, and we use a *TERMINAL-node* to represent it in the last layer. We also encode the contextual relations between the face and object components in the AOG using co-occurrences of face-object pairs (such as the co-occurrence of politicians and suits).

Joint Image-Text Representation. The relationships between image and text parts are explicitly modeled via the frequencies of pairs of an image patch and a text entity.We model three component pair relations selected from these two parts, namely face-who, face-what and object-what pairs. The face-who and face-what pairs can clearly relate the faces appeared in the video to their names and other co-appearing textual knowledge respectively. The object-what pairs can relate the objects to textual descriptions.

In summary, the proposed MT-AOG jointly models texts and images, and their sub-components in a hierarchical structure. The MT-AOG model strikes a balance between the syntactic representation in natural language processing (too complex to compute) and the simplistic bag-of-words representation (too coarse). It supports the news topic detection and tracking tasks with the appropriate complexity accurately.

3.1.2.2 Task: Detecting and Tracking News Topics

In the massive and continuously updated news data, each news topic evolves over time. We aim to detect topics within short time periods and further discover long-time topic trajectories. Therefore, we can show both detailed descriptions for each topic in different time periods, and how each topic develops over time. It also helps prevent the heavy computation incurred by periodically detecting topics using the entire updated news collection.

For topic detection, we group stories that elaborate the same topics. The proposed MT-AOG explicitly describes components of different topics. Thus based on the MT-AOG, we can effectively group related stories and generate meaningful topics. We solve the grouping problem using cluster sampling methods by maximizing a Bayesian posterior probability. An efficient cluster sampling algorithm introduced in image segmentation, i.e. SWC, is adopted for topic detection.

For topic tracking, we link topics detected in different time periods to generate topic trajectories. The MT-AOG model can represent topic compositions and how such information change over time. So using the MT-AOG, we can effectively track and keep updating the news states. The topics are linked by considering both topic similarities and their temporal relations.

The experimental results in Section 3.5 demonstrate that our method can generate meaningful topics and topic trajectories. It also achieves better performance compared to other state-of-the-art algorithms.

3.1.3 Summary of Contributions

This chapter makes the following contributions:

- We propose a Multimodal Topic And-Or Graph that models the semantic structures of events in multimodal dimensions, which is much more suitable in the TV news domain compared to existing methods only using texts [ACD98, BNJ03, XX13, ZS05, WNL06, GSB05, BB09].
- We solve the topic detection problem using a cluster sampling method, Swendsen-Wang Cuts, which has better performance than commonly used greedy algorithms [BNJ03, XX13, WNL06, WNH08].
- We detect and track topics simultaneously over time, generating both topic summaries in different time periods and long-time topic trajectories. The results provide useful data for further media analyses, which can hardly be fulfilled by traditional topic detection and tracking methods [All02, ACD98].
- We propose a novel TV news dataset for joint image-text topic detection and tracking, and the ground-truth annotations for topics.

3.2 Topic Representation

In this section, we define our Multimodal Topic And-Or Graph (MT-AOG) for topic representation.

3.2.1 Overall Representation

A **MT-AOG** can be defined by a three-tuple $\mathcal{G} = (V, E, \Theta)$. The node set V consists of three subsets of nodes: **AND-nodes** V_{AND}, **OR-nodes** V_{OR} and **TERMINAL-nodes** V_{T}, i.e. $V = V_{\text{AND}} \cup V_{\text{OR}} \cup V_{\text{T}}$. E denotes the edge set in the graph. Θ represents the MT-AOG model parameters. We have $\Theta = \{K, \theta_1, ..., \theta_K\}$ where K is the total topic number, and

$\theta_1, ..., \theta_K$ represent the model parameters for these K topics respectively. Fig. 3.2 illustrates the proposed MT-AOG topic representation.

A **parse graph** pg is an instantiation of the MT-AOG by selecting children nodes at OR-nodes (according to the scoring functions defined below in this Section and Section 3.2.2, 3.2.3, and 3.2.4). The green lines in Fig. 3.2 shows one example of the parse graph.

As shown in Fig. 3.2, the MT-AOG has five layers. Nodes in each layer are explained as follows:

1) **Root OR-node** $O^{\mathrm{r}} \in V_{\mathrm{OR}}$ in the first layer of MT-AOG represents different topic configurations and their mutual contextual information. Each topic k ($k = 1, ..., K$) is represented by an AND-node A_k^{topic} in the second layer with the model parameter θ_k.

News stories are reports of topics, i.e. topic instances, from various TV news networks. To find the optimal pg for one news story, i.e. the optimal topic instantiation of the MT-AOG, we define a series of scoring functions at different nodes below. We denote a news story by $\mathbf{d}_i$. For a story $\mathbf{d}_i$, the scoring function at root OR-node O^{r} is defined as:

$$s^{\mathrm{root}}(\mathbf{d}_i; \Theta) = \max_{\theta_k \in \Theta} s^{\mathrm{topic}}(\mathbf{d}_i; \theta_k), \tag{3.1}$$

where $s^{\mathrm{topic}}(\mathbf{d}_i; \theta_k)$ is the scoring function at A_k^{topic}, which will be introduced later. In the following sections, we omit the story index, i, for simplicity.

2) **Topic AND-node** $A_k^{\mathrm{topic}} \in V_{\mathrm{AND}}$ represents one topic configuration. One topic is composed of the text part and the image part. So A_k^{topic} has two children AND-nodes, i.e. text AND-node A_k^{txt} and image AND-node A_k^{img}. Considering both text and image parts and their contextual relations, we define the scoring function at A_k^{topic} as:

$$\begin{aligned} s^{\mathrm{topic}}(\mathbf{d}; \theta_k) = s^{\mathrm{txt}}(\mathbf{d}^{\mathrm{txt}}; \theta_k) + s^{\mathrm{img}}(\mathbf{d}^{\mathrm{img}}; \theta_k) \\ + s^{\mathrm{joint}}(\mathbf{d}^{\mathrm{joint}}; \theta_k) + g(f_{A_k^{\mathrm{topic}}}), \end{aligned} \tag{3.2}$$

where $\mathbf{d}^{\mathrm{txt}}$, $\mathbf{d}^{\mathrm{img}}$ and $\mathbf{d}^{\mathrm{joint}}$ denote the text part, the image part and their joint information of the story $\mathbf{d}$ respectively ($\mathbf{d} = \mathbf{d}^{\mathrm{txt}} \cup \mathbf{d}^{\mathrm{img}} \cup \mathbf{d}^{\mathrm{joint}}$). The two terms $s^{\mathrm{txt}}(\mathbf{d}^{\mathrm{txt}}; \theta_k)$ and $s^{\mathrm{img}}(\mathbf{d}^{\mathrm{img}}; \theta_k)$ are scoring functions at A_k^{txt} and A_k^{img} respectively. The term $s^{\mathrm{joint}}(\mathbf{d}^{\mathrm{joint}}; \theta_k)$

describes the contextual relations between the text part and the image part. These three terms will be explained later. To take the prior of choosing A_k^{topic} at root node O^{r} into consideration, we also add function $g(f_{A_k^{\text{topic}}})$ in the scoring function, where $f_{A_k^{\text{topic}}} \in \theta_k$ is the branching frequency. We observed that in broadcast news, dominant topics with a large amount of coverage are only a small part of all the topics, and sizes of most topics are small. Accordingly, we assume that branching frequencies at O^{r} follow a power law distribution[2] (the verification of our observation will be shown in Section 3.2.5).

3.2.2 Text Representation

For a news story $\mathbf{d}$, its *text part* $\mathbf{d}^{\text{txt}}$ contains the "who" component $\mathbf{d}^{\text{who}}$, the "where" component $\mathbf{d}^{\text{where}}$, and the "what" component $\mathbf{d}^{\text{what}}$. These three components describe the people involved, related places, and what happened respectively. We extract words for different components by performing named entity extraction using the Stanford Named Entity Recognizer [FGM05]. Thus each of these three components can be represented by a list of words (word duplication is allowed in the list), e.g. $\mathbf{d}^{\text{who}} = (w_1, ..., w_{M^{\text{who}}})$ where M^{who} is the total word number in the "who" component in story $\mathbf{d}$. The total numbers of words in "where" and "what" components are denoted by M^{where} and M^{what} respectively.

We extract co-occurring word pairs from the three components in the text part. We consider a pair of words as one co-occurring pair if they belong to two different components, and are extracted from the same sentence. The list of co-occurring word pairs in $\mathbf{d}$ is denoted by $\mathbf{d}_{\text{txt}}^{\text{pair}} = [(w_1, w_2)|w_1 \in \mathbf{d}^{\text{who}}, w_2 \in \mathbf{d}^{\text{where}}] \cup [(w_1, w_2)|w_1 \in \mathbf{d}^{\text{who}}, w_2 \in \mathbf{d}^{\text{what}}] \cup [(w_1, w_2)|w_1 \in \mathbf{d}^{\text{where}}, w_2 \in \mathbf{d}^{\text{what}}]$.

Text AND-node A_k^{txt} in the MT-AOG has three children OR-nodes, i.e. O_k^{who}, O_k^{where}, and O_k^{what}, which represent "who", "where" and "what" components in the text part of topic k respectively. To model these three components jointly, the scoring function at A_k^{txt} (i.e. $s^{\text{txt}}(\mathbf{d}^{\text{txt}}; \theta_k)$ in Eq. 3.2) is defined as:

[2]In the experiments, for the function $g(\cdot)$, we use the Zipf's law probability distribution, i.e. $g(f) = \frac{f^{-\rho}}{\zeta(\rho)}$ and set the parameter ρ that describes the distribution's exponent as $\rho = 1.75$ (ζ is the Riemann Zeta function).

$$s^{\text{txt}}(\mathbf{d}^{\text{txt}};\theta_k) = \sum_{c\in\{\text{who,where,what}\}} s^{\text{c}}(\mathbf{d}^{c};\theta_k) + s^{\text{pair}}(\mathbf{d}^{\text{pair}}_{\text{txt}};\theta_k), \tag{3.3}$$

where the variable c represents the component type $c \in \{\text{who}, \text{where}, \text{what}\}$. $s^{\text{c}}(\mathbf{d}^{c};\theta_k)$ represents the scoring function at the OR-node for one component $O^c_k \in \{O^{\text{who}}_k, O^{\text{where}}_k, O^{\text{what}}_k\}$. $s^{\text{pair}}(\mathbf{d}^{\text{pair}}_{\text{txt}};\theta_k)$ describes the contextual relations between components in the text part:

$$s^{\text{pair}}(\mathbf{d}^{\text{pair}}_{\text{txt}};\theta_k) = \sum_{c_1,c_2} \mathcal{N}(\frac{M^{c_1}}{M^{c_2}};\mu^{c_1c_2}_k,\sigma^{c_1c_2}_k) + \sum_{(w_1,w_2)\in\mathbf{d}^{\text{pair}}_{\text{txt}}} \log(f^{(w_1,w_2)}_k+1), \tag{3.4}$$

where we have the component types $c_1 \in \{\text{who}, \text{where}\}$, $c_2 \in \{\text{where}, \text{what}\}$, $c_1 \neq c_2$. $\frac{M^{c_1}}{M^{c_2}}$ represents the ratio of word numbers from two different components. The three ratios, namely $\frac{M^{\text{who}}}{M^{\text{where}}}$, $\frac{M^{\text{who}}}{M^{\text{what}}}$, and $\frac{M^{\text{where}}}{M^{\text{what}}}$ are assumed to follow Gaussian distributions $\mathcal{N}(\frac{M^{c_1}}{M^{c_2}};\mu^{c_1c_2}_k,\sigma^{c_1c_2}_k)$. $\mu^{c_1c_2}_k, \sigma^{c_1c_2}_k \in \theta_k$ are parameters for corresponding Gaussian distributions. The parameter $f^{(w_1,w_2)}_k \in \theta_k$ is the frequency of co-occurring word pair (w_1, w_2) in topic k.

Three children OR-nodes of $A^{\mathbf{txt}}_k$ in the fourth layer, namely O^{who}_k, O^{where}_k, and O^{what}_k, describe a set of possible words for the corresponding components. A certain news story may trigger a subset of these words. The words are represented by TERMINAL-nodes in the last layer. The scoring functions at these OR-nodes are defined as:

$$s^{\text{c}}(\mathbf{d}^{c};\theta_k) = \sum_{w\in\mathbf{d}^c} \log(f^w_k+1) \tag{3.5}$$

where the component type $c \in \{\text{who}, \text{where}, \text{what}\}$. The parameter $f^w_k \in \theta_k$ is the frequency of word w in topic k.

3.2.3 Image Representation

The story's *image part* $\mathbf{d}^{\text{img}}$ contains the face component $\mathbf{d}^{\text{face}}$, and the object component $\mathbf{d}^{\text{obj}}$, i.e. $\mathbf{d}^{\text{face}}, \mathbf{d}^{\text{obj}} \in \mathbf{d}^{\text{img}}$. Each entity in the face/object component corresponds to one cluster of face/object patches.

To obtain the face component, we first perform face detection using the Viola-Jones face detector [VJ01] and extract face features based on Local Binary Pattern [AHP06] and Local Gabor Binary Pattern Histogram Sequence [ZSG05], and then use the k-means algorithm to cluster faces into groups.

To get the object components, we first extract patches from images using Selective Search [USG13] which generates object proposals. We then extract a 4096-dimensional feature vector for each patch from the fc7 layer of AlexNet [KSH12b] trained on ImageNet data [RDS15]. We use a pretrained model in Caffe [JSD14]. Then we cluster these patches by k-means algorithm.

Fig. 3.3 illustrates how one image can be parsed based on the obtained face and object clusters. Each face/object patch can be represented by its corresponding cluster membership. Then the face and object components of one story $\mathbf{d}$ can also be represented by a list of visual words, e.g. $\mathbf{d}^{\text{face}} = (w_1, ..., w_{M^{\text{face}}})$ where each word $w_j \in \mathbf{d}^{\text{face}}$ represent one face patch's cluster membership. M^{face} is the total number of face patches in $\mathbf{d}^{\text{img}}$ and the total number of object patches is denoted by M^{obj}.

We extract co-occurring word pairs from the face and object components of the image part. A pair of visual words is considered as one co-occurring pair if the two words are from the face and object components respectively, and they both appear in one short time period in the news video. We denote the list of co-occurring pairs extracted from the image part by $\mathbf{d}^{\text{pair}}_{\text{img}} = [(w_1, w_2)|w_1 \in \mathbf{d}^{\text{face}}, w_2 \in \mathbf{d}^{\text{obj}}]$.

Image AND-node A^{img}_k in the MT-AOG represents the image part of topic k. It has two children OR-nodes, i.e. O^{face}_k and O^{obj}_k, which represent face and object components respectively. The scoring function at A^{img}_k is defined in a similar way to the one at A^{txt}_k:

$$\begin{aligned} s^{\text{img}}(\mathbf{d}^{\text{img}}; \theta_k) = & \sum_{c \in \{\text{face,obj}\}} s^{\text{c}}(\mathbf{d}^c; \theta_k) \\ & + s^{\text{pair}}(\mathbf{d}^{\text{pair}}_{\text{img}}; \theta_k), \end{aligned} \tag{3.6}$$

where the component type $c \in \{\text{face}, \text{obj}\}$. The term $s^{\text{c}}(\mathbf{d}^c; \theta_k)$ represents the scoring function at OR-node for one component $O^c_k \in \{O^{\text{face}}_k, O^{\text{obj}}_k\}$.

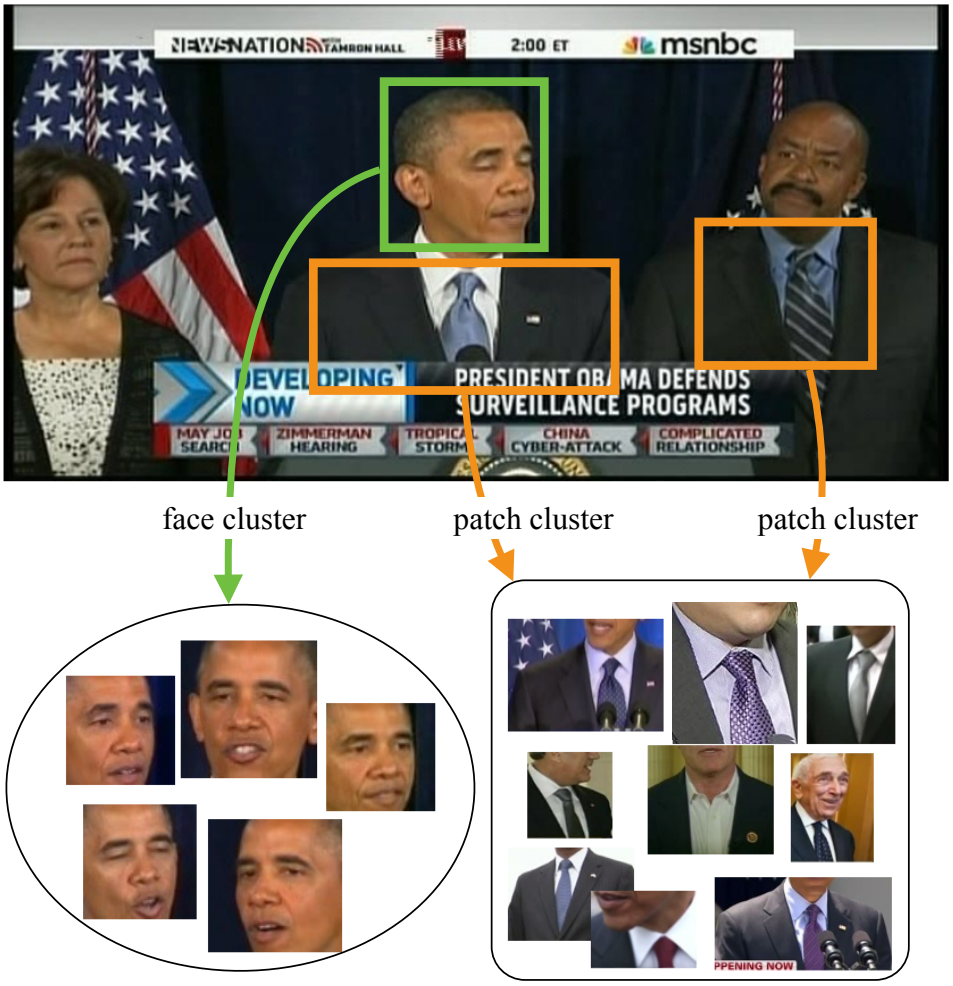


Figure 3.3: A common example pair of a face and a object cluster discovered by our algorithm.

The term $s^{\text{pair}}(\mathbf{d}^{\text{pair}}_{\text{img}}; \theta_k)$ describes contextual relations between face and object components and we define it as:

$$s^{\text{pair}}(\mathbf{d}^{\text{pair}}_{\text{img}}; \theta_k) = \sum_{(w_1,w_2)\in \mathbf{d}^{\text{pair}}_{\text{img}}} \log(f_k^{(w_1,w_2)} + 1), \tag{3.7}$$

where $f_k^{(w_1,w_2)} \in \theta_k$ is the frequency of co-occurring visual word pair (w_1, w_2) in the topic k.

Two children OR-nodes of A_k^{img}, namely O_k^{face}, and O_k^{obj}, can describe a set of alternative visual words. These words are represented by TERMINAL-nodes in the last layer. Scoring functions at these OR-nodes, i.e. $s^c(\mathbf{d}^c; \theta_k), c \in \{\text{face}, \text{obj}\}$, are defined in the same way as those at O_k^{who}, O_k^{where} and O_k^{what} (Eq. 3.5).

3.2.4 Joint Image-Text Representation

To jointly model the topic text and image parts, we extract their co-occurring word pairs. Three kinds of pairs, namely the face-who, face-what, and object-what pairs, are obtained for each news story $\mathbf{d}$. The words in each co-occurring pair appear in one short time period. These image-text co-occurring word pairs are denoted by $\mathbf{d}^{\text{joint}} = [(w_1, w_2)|w_1 \in \mathbf{d}^{\text{who}}, w_2 \in \mathbf{d}^{\text{face}}] \cup [(w_1, w_2)|w_1 \in \mathbf{d}^{\text{what}}, w_2 \in \mathbf{d}^{\text{face}}] \cup [(w_1, w_2)|w_1 \in \mathbf{d}^{\text{what}}, w_2 \in \mathbf{d}^{\text{obj}}]$. We use M^{txt} and M^{img} to denote the total entity numbers of the text part and the image part in $\mathbf{d}$ respectively. So we have $M^{\text{txt}} = M^{\text{who}} + M^{\text{where}} + M^{\text{what}}$, and $M^{\text{img}} = M^{\text{face}} + M^{\text{obj}}$.

The score function $s^{\text{joint}}(\mathbf{d}^{\text{joint}}; \theta_k)$ in Eq. 3.2 is defined as:

$$\begin{aligned} s^{\text{joint}}(\mathbf{d}^{\text{joint}}; \theta_k) = \mathcal{N}(\frac{M^{\text{txt}}}{M^{\text{img}}}; \mu_k^{\text{joint}}, \sigma_k^{\text{joint}}) \\ + \sum_{(w_1,w_2)\in\mathbf{d}^{\text{joint}}} \log(f_k^{(w_1,w_2)} + 1). \end{aligned} \tag{3.8}$$

We assume that the ratio between the total entity numbers of the text part and the image part, i.e. $\frac{M^{\text{txt}}}{M^{\text{img}}}$, follows Gaussian distribution $\mathcal{N}(\frac{M^{\text{txt}}}{M^{\text{img}}}; \mu_k^{\text{joint}}, \sigma_k^{\text{joint}})$ with the parameters $\mu_k^{\text{joint}}, \sigma_k^{\text{joint}} \in \theta_k$. The parameter $f_k^{(w_1,w_2)} \in \theta_k$ is the frequency of the word pair (w_1, w_2) in topic k.

Based on the previous scoring functions, we can select the best children nodes at OR-nodes and find the optimal parse graph pg^* for the story $\mathbf{d}$ by calculating $s^{\text{root}}(\mathbf{d}; \Theta)$.

3.2.5 Empirical Evaluations of Assumptions in MT-AOG

In the MT-AOG representation, we make the assumption that branching frequencies at root OR-node O^{r} follow the power law distribution. To verify our assumption, we collected a news corpus that contains 1,853 news stories during a period of seven days. Annotators were asked to group the stories according to their topics and we collected 355 topics in total after annotation.

To verify that the branching frequencies at O^{r} follow the power law distribution, using the collected corpus, we fit the empirical distribution of the story numbers in the topics to the power law distribution. The p-value (at the 5% significance level) is 0.9984. Fig. 3.4

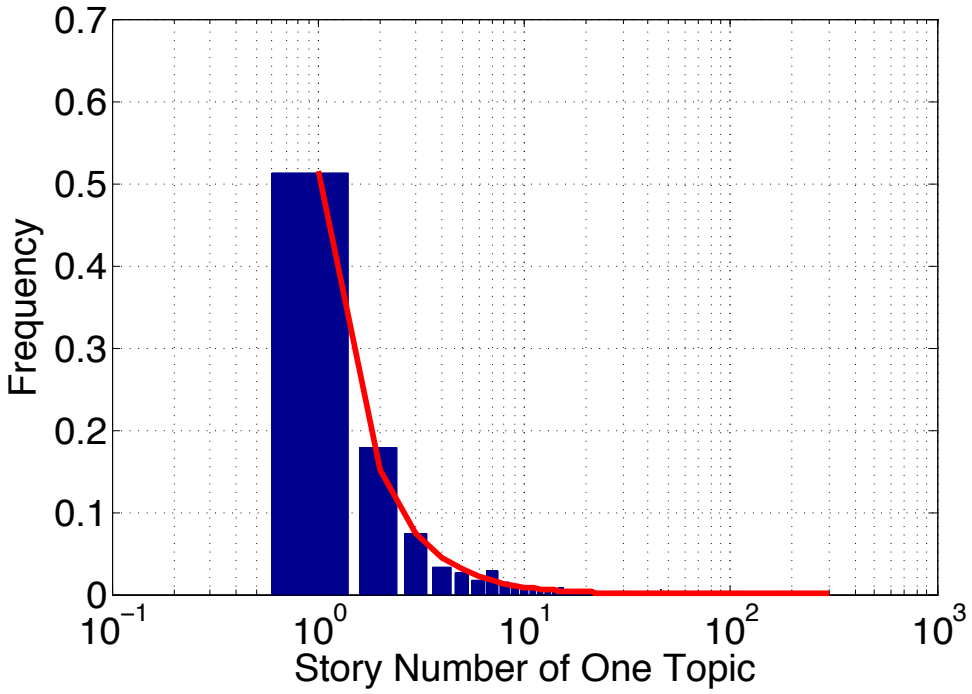


Figure 3.4: The histogram of the number of stories in each topic and the fitting result (the red curve).

shows the empirical distribution and the fitted curve (red line).

3.3 Topic Detection

In this section, we present our formulation of the topic detection problem, and the algorithm for optimizing a Bayesian posterior probability for the problem.

3.3.1 Problem Formulation

With the MT-AOG topic representation, our goal of topic detection is to cluster news stories that describe the same topics and obtain the MT-AOG model parameters Θ for the topics. We pose this clustering problem as a graph partitioning problem in which news stories, as vertices in the adjacency graph, are partitioned into coherent groups. We show one example of the adjacency graph in Fig. 3.5. Edges in the adjacency graph are associated with certain weights corresponding to related story similarities. Partitions can be obtained by dividing the

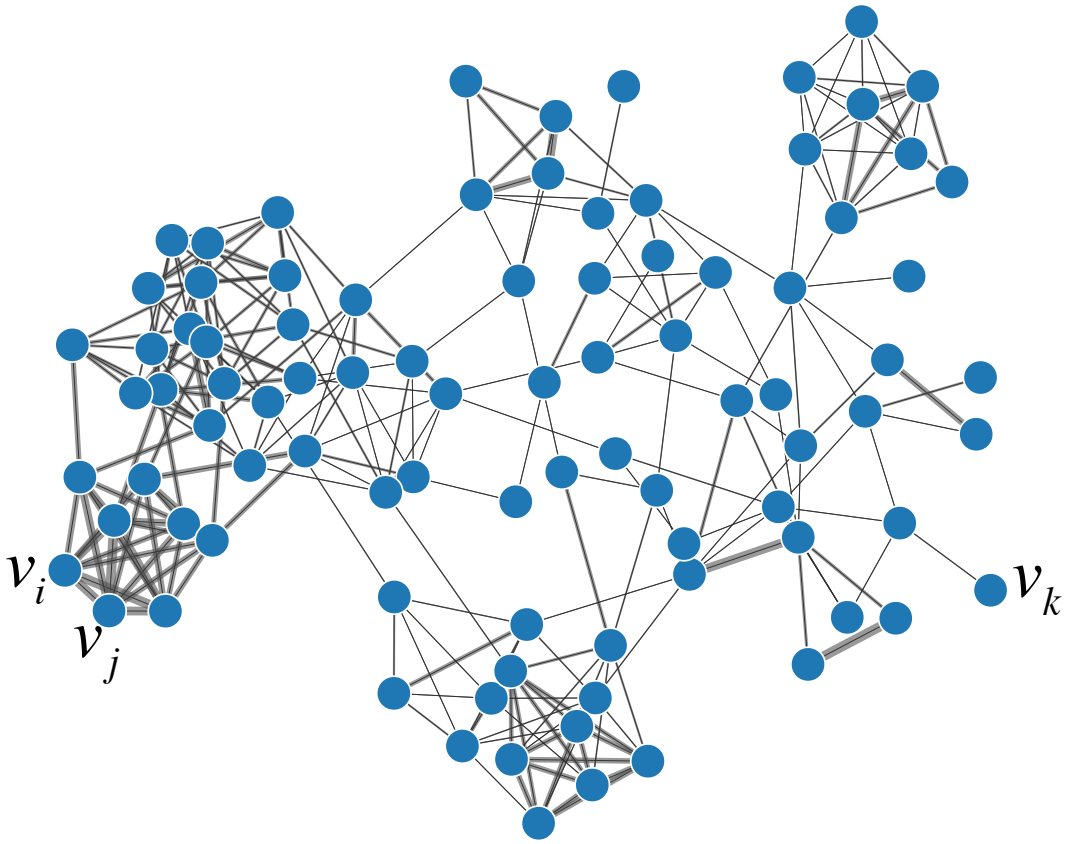


Figure 3.5: One adjacency graph. Each vertex in the graph corresponds to one news story. Edges are associated with weights corresponding to the story similarities (the edge thickness shows the story similarity). The vertices v_i and v_j both talk about the Oklahoma tornado topic and they are adjacent to each other in the graph. The other vertex v_k which is far away from v_i and v_j talks about the California High-Speed Rail project.

vertices into groups with specific properties and also keeping the number of edges between separated components small. Graph partitioning can help the news topic detection since even though news stories from one topic develop over time and drift the topic, they can still be grouped together through the connections between temporally adjacent stories with less changes and more similarities.

Formally, we are given a news story corpus that contains N news stories, i.e. $D = \{\mathbf{d}_i; i = 1, \ldots, N\}$. The adjacency graph is defined as $\mathcal{G}_{\text{ADJ}} = (V_{\text{ADJ}}, E_{\text{ADJ}})$ where V_{ADJ} is a set of vertices and each vertex $v_i \in V_{\text{ADJ}}$ corresponds to one news story $\mathbf{d}_i$. E_{ADJ} is a set of edges

between vertices. The clustering/partition W we are trying to find given D is defined as:

$$W = (K, \pi_K, \Theta), \tag{3.9}$$

where K is determined automatically while solving the partitioning problem and π_K represents the $K-$partition of the adjacency graph. π_K is defined as:

$$\pi_K = (V_1, ..., V_K), \bigcup_{k=1}^{K} V_k = V_{\text{ADJ}}, V_k \cap V_j = \emptyset, \forall i \neq j. \tag{3.10}$$

This becomes an optimization problem which can be solved by maximizing a Bayesian posterior probability:

$$W^* = \arg\max_{W \in \Omega} p(W|D) = \arg\max_{W \in \Omega} p(D|W)p(W), \tag{3.11}$$

where Ω is the solution space. The likelihood probability $p(D|W)$ is formulated as:

$$p(D|W) = \prod_{i=1}^{N} p(\mathbf{d}_i; \Theta) \propto \exp\{\sum_{d_i \in D} s^{\text{root}}(\mathbf{d}_i; \Theta)\}. \tag{3.12}$$

The prior probability $p(W)$ penalizes the partition number K in W and we formulate it as:

$$p(W) \propto \exp\{-\alpha NK\}. \tag{3.13}$$

α is a positive parameter which acts as a threshold for grouping stories into topics. This prior helps us combine close partitions to get dense results.

3.3.2 Inference by Swendsen-Wang Cuts

To solve the topic detection problem formulated above, we adopt a cluster sampling method Swendsen-Wang Cuts (SWC) [BZ05]. It is a Markov Chain Monte Carlo method which samples the solution space Ω efficiently. An alternative method will be the expectation-maximization (EM) algorithm. But in [PTZ15], SWC is shown to be more effective than EM which finds only a local minimum.

SWC changes the labels of a group of vertices at the same time. It thus solves the coupling problem of Gibbs sampler (which flips a single vertex) by quickly jumping between local minima. SWC starts with an initial partition π, which can be the one which sets

all stories to be in the same group, or can be set randomly. We denote the set of edges whose related two vertices belong to the same group under the partition π by $E(\pi)$. The optimal clustering W^* can be obtained by performing the following steps iteratively until convergence.

(1) *Determining edge status.* Each edge $e =< v_i, v_j >\in E(\pi)$ is associated with a Bernoulli random variable $u_e \in \{0, 1\}$:

$$u_e \sim \text{Bernoulli}(q_e \cdot 1(x_i = x_j)), \tag{3.14}$$

which indicates the edge's on/off status and a turn-on probability q_e. x_i and x_j are state variables for vertices v_i and v_j respectively, and they take values from a finite number of labels. We define:

$$q_e = e^{-\mathcal{D}(e)/T}, \tag{3.15}$$

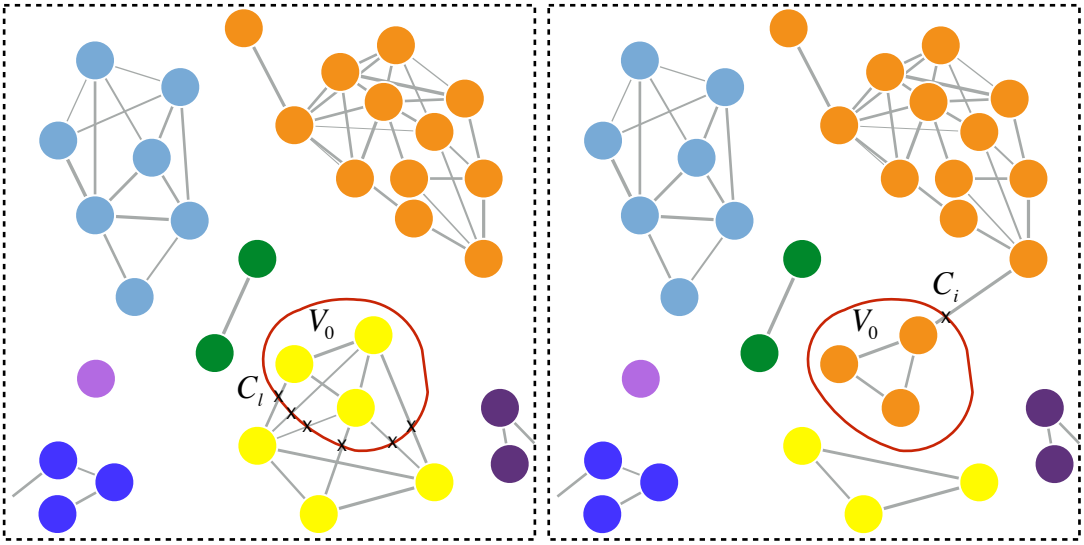
where T is the temperature used in the simulated annealing procedure and it is slowly decreased according to a cooling schedule. $\mathcal{D}(e)$ is the distance of these two vertices obtained based on the Kullback-Leibler (KL) divergence:

$$\mathcal{D}(e) = \sum_{F\in\mathcal{F}} \lambda_F \cdot \frac{KL(F(v_i)||F(v_j)) + KL(F(v_j)||F(v_i))}{2}, \tag{3.16}$$

where $F(\cdot)$ denotes one type of feature of the vertex and λ_F is the weight for feature F. Here we use the distributions for the five components in the text and image parts (i.e. who, where, what, face and object) to construct the feature set $\mathcal{F}$. Moreover, since KL divergence is non-symmetric, we average the KL divergence of $F(v_i)$ given $F(v_j)$ and the KL divergence of $F(v_j)$ given $F(v_i)$ to get a symmetric distance measure for vertices v_i and v_j. Based on these definitions, in this step, we set $u_e = 0$ (i.e, turn e off) with probability $1 - q_e$ for all $e \in E(\pi)$.

(2) *Computing connected components.* Once the states u_e is determined for each edge $e \in E(\pi)$, the graph G is partitioned into a set of connected components, each of which contains vertices that belong to the same group.

(3) *Selecting a component and flipping it.* Among all the connected components formed in (2), we can randomly select one component V_0 to flip. We show one example of V_0 in Fig.



(a) Before flipping. (b) After flipping.

Figure 3.6: SWC flips the selected component V_0. The cuts are marked with crosses.

3.6a. The target label for V_0 can be a new one that has not been used yet or just the same as any other connected components, thus allowing reversible jumps in the solution space. The current partition number is denoted as K'. Then the number of possible new labels for the selected component is $K'+1$. Assuming that $V_0 \subseteq V_l$ in the current partition π, we denote a series of sets

$$S_1 = V_1, S_2 = V_2, ..., S_l = V_l \backslash V_0, S_{K'} = V_{K'}, S_{K'+1} = \emptyset \tag{3.17}$$

that V_0 can be merged with. Then V_0 can be flipped by drawing a random sample l' with probability

$$p(l'(V_0) = i | V_0, \pi) = \frac{\gamma_i p(\pi_i | D)}{\sum_{j=1}^{K'+1} \gamma_j p(\pi_j | D)}, \tag{3.18}$$

where π_i is the partition after assigning the label of the component V_0 to be i and keeping other components' labels the same as in π. We also have

$$\gamma_i = \prod_{e \in \mathcal{C}_i} (1 - q_e), \tag{3.19}$$

where $\mathcal{C}_i$ is the cuts between V_0 and S_i, i.e. $\mathcal{C}_i = \mathcal{C}(V_0, S_i) = \{< s, t >: s \in V_0, t \in S_i\}$. Two examples of the cuts are shown in Fig. 3.6, which are marked by the crosses. Theorem 3

in [BZ05] proved that the acceptance rate will be 1 by choosing the new label of V_0 by Eq. 3.18.

Another thing to be noted is that when generating the adjacency graph, we can use a complete graph of N vertices since each pair of news stories can be related. But this may cause problems since a complete graph of N vertices has $\binom{N}{2} = O(N^2)$ edges and the number of all possible solutions is exponential in the number of edges, i.e. $O(2^{N^2})$, which requires a long convergence time. By investigating the data, however, one may observe that some story pairs have few similarities in terms of contents. Such pair of stories shall never be grouped together. So graph pruning can be performed on the adjacency graph before SWC . We define a threshold τ, and cut all edges e whose $\mathcal{D}(e) \geq \tau$ deterministically.

3.4 Topic Tracking

In this section, we describe our method for tracking topics detected in certain continuous time periods. We link all detected topics in different time periods to form topic trajectories over time.

We divide the whole news data collection into several sub-collections which consist of news stories in different time periods. Topic detection is performed within each sub-collection separately. The sub-collection set of the news corpus D is denoted by $\{C_1, ..., C_M\}$ where $C_1 \cup ... \cup C_M = D$ and M is the number of sub-collections. Each sub-collection contains news documents from one specific time span t_i. Topics extracted within each sub-collection C_i are denoted by $\{\Theta_i^1, ..., \Theta_i^{K_i}\}$, where K_i is the obtained topic number.

For topic tracking, as shown in Fig. 3.7, we link topics detected in the sub-collections to generate topic trajectories. One optional method for solving the linking problem is to do another clustering on the detected topics using SWC. But to fast obtain the topic links, we choose to measure the similarities between topics by considering both the topic content similarities and their temporal distances, and use a threshold to decide whether they can be linked. Formally, the similarity measurement to decide whether two topics can be linked is

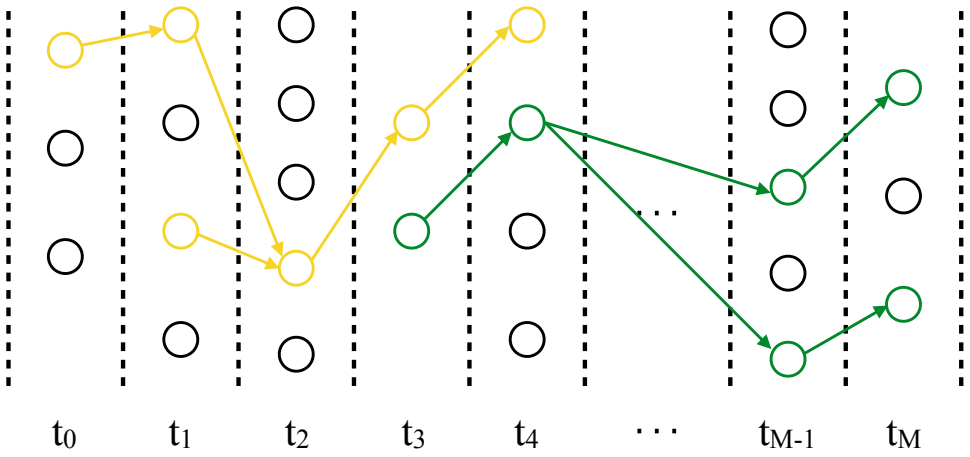


Figure 3.7: One example of the topic tracking trajectory.

calculated as:

$$Sim(\Theta_{i_1}^{k_1}, \Theta_{i_2}^{k_2}) = \alpha_{sim} \exp\{-\beta_{kl}[KL(\Theta_{i_1}^{k_1}||\Theta_{i_2}^{k_2}) + KL(\Theta_{i_2}^{k_2}||\Theta_{i_1}^{k_1})]\} + (1 - \alpha_{sim}) \exp\{-|t_{i_1} - t_{i_2}|\}, \tag{3.20}$$

where $i_1 \neq i_2$, and α_{sim} and β_{kl} are positive parameters. Note that using the proposed topic representation, each topic is composed of the image part and the text part, and they can be further divided into the "who", "where" and "what" components, and the face and object components respectively. Thus we have five components in total. Each component is represented using one model. The KL divergence of one topic given another is therefore averaged over these models:

$$KL(\Theta_{i_1}^{k_1}||\Theta_{i_2}^{k_2}) = \sum_{j=1}^{5} \lambda_j KL(\Theta_{m_1}^{k_1,j}||\Theta_{i_2}^{k_2,j}), \tag{3.21}$$

where λ_j is the corresponding weight, and $\Theta_{i_1}^{k_1,j}$, $\Theta_{m_2}^{k_2,j}$ are the histograms of word frequencies for the j-th component. After calculating the topic similarities using Eq. 3.20, a threshold τ_{link} can be used for pruning the links between topics to get the final topic trajectories.

3.5 Experiment

3.5.1 Datasets

Two datasets are used in our experiment:

1) **Reuters-21578.** Reuters-21578 dataset[3] is a publicly available collection of news stories from Reuters newswire. It is widely used for the evaluation of clustering and classification methods. The dataset contains 21,758 stories which belong to 135 clusters/categories. The clusters/categories are annotated manually. Only textual information is contained in the dataset.

2) **UCLA Broadcast News Dataset.** We collected a multimedia broadcast news dataset from UCLA Library Broadcast NewsScape. Five US networks are included in the dataset: CNN, MSNBC, FOX, ABC, and CBS. It contains 379 news videos broadcasted in the time period from June 1, 2013 to June 14, 2013. The total length of the videos is about 362 hours. Several programs from each news network are included in the dataset, such as "CNN Newsroom", "MSNBC News Live", "FOX Morning News", "ABC Nightline", "CBS News", etc.

Annotation: We annotate the UCLA Broadcast News Dataset for topic detection and tracking.

One annotation choice can be letting annotators manually group news stories based on their related topics [WNL06, ACD98]. However, this will be a hard task and the annotation results may not be accurate since there can be hundreds of news topics even in one week and the annotators can hardly remember all the previously found topics during annotation.

So instead of this, we choose to build the ground-truth by letting annotators decide whether a pair of stories belong to the same topic or not. The topic granularity is chosen to be at the event level, like the definition in the TDT system [ACD98]. In other words, two stories talking about the same event (or closely related ones) belong to the same topic. Since it takes a long time to annotate all story pairs, we choose to annotate a subset selected from the whole story pair collection. We first compute the cosine distances between the two stories in the same pair, and then select 10,000 story pairs to be annotated randomly from the pair set where all pair distances are within the range $[0.6, 0.9]$. This specific range is chosen for the reason that the corresponding story relations are ambiguous compared to

[3]Reuters-21578 dataset can be downloaded at `http://www.daviddlewis.com/resources/testcollections/reuters21578/`.

other ranges. Three annotators are involved in the annotation and for each story pair we treat the relation that most annotators agree as the ground-truth relation.

This dataset is mainly used for quantitative evaluation. To show how our method work on large-scale datasets qualitatively, we also apply it to more news data from the UCLA Library Broadcast NewsScape.

3.5.2 Experiment I: Topic Detection

In this experiment, we conduct topic detection experiments on both the Reuters-21578 dataset and the UCLA Broadcast News Dataset.

3.5.2.1 Results on Reuters-21578

We compare the proposed topic detection method with other story/document clustering methods on the news dataset Reuters-21578 (only texts are available). Stories with multiple cluster labels are discarded and for the remaining stories, only those from the largest 10 clusters are selected [XX13].

Evaluation Protocol. On Reuters-21578, we follow the evaluation protocol in [XX13, CHH11]. Two metrics are used to evaluate the clustering performance, i.e. accuracy, and normalized mutual information. To compute the accuracy, the obtained clusters are mapped to the ground-truth clusters in the dataset. The clustering accuracy is then defined as the percentage of documents that have the correct cluster labels after mapping. In detail, for a document d_i, let l_i and l'_i denote the obtained cluster label and the provided ground-truth cluster label respectively. Then the clustering accuracy AC can be calculated by the following equation:

$$\mathrm{AC} = \frac{\sum_{i=1}^{N} \delta(l_i, map(l'_i))}{N} \tag{3.22}$$

where N is the total number of stories/documents. $\delta(x, y)$ is the delta function:

$$\delta(x, y) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise} \end{cases} \tag{3.23}$$

The function $map(l_i)$ is the permutation mapping function that maps the obtained clusters to the ground-truth clusters in the dataset.

The mutual information measures the mutual dependence of the ground-truth cluster assignments and the obtained clustering assignments of the documents:

$$\mathrm{MI}(C, C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \cdot \log_2 \frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_j)} \tag{3.24}$$

where C and C' denotes the ground-truth clusters and the obtained clusters respectively. $p(c_i)$ and $p(c'_j)$ are the probabilities that one randomly selected document is from the cluster c_i and c'_j respectively. $p(c_i, c'_j)$ is the probability that randomly selected document belongs to both cluster c_i ad cluster c'_j. The normalized mutual information NMI is defined as:

$$\mathrm{NMI}(C, C') = \frac{\mathrm{MI}(C, C')}{\max(H(C), H(C'))} \tag{3.25}$$

where $H(C)$ and $H(C')$ are the entropies of C and C' respectively. We have $\mathrm{NMI}(C, C') \in [0, 1]$, and it equals 0 (or 1) when the clusters C and C' are independent (or identical).

Other Methods. Several other methods are included in the comparison, namely:

(1) K-means and Normalized Cuts (NC) [SM00], which are widely used clustering and graph partitioning algorithms;

(2) Nonnegative-Matrix-Factorization (NMF) based clustering [XLG03], Latent Semantic Indexing (LSI) [Hof99], and Locally Consistent Concept Factorization (LCCF) [CHH11], which are factorization based methods that are effective for document clustering;

(3) LDA + K-means [XX13] and LDA + Naive [XX13] (both of them use LDA to learn topics and topic distributions for documents, and LDA + K-means then clusters documents using K-Means based on these distributions while LDA + Naive treats the label of the most dominant topic as the cluster label for each document);

(4) Multi-grain clustering topic model (MGCTM) [XX13] which has the best clustering result on Reuters-21578 so far.

The inputs of these methods in the comparison are the documents' tf-idf vectors [XX13, CHH11]. These methods all require the cluster number to be specified in the input. Thus

Table 3.1: Clustering Performance of different methods on Reuters-21578.

	Clustering Accuracy(%)	Normalized Mutual Information(%)
K-Means	35.02	35.76
NC [SM00]	26.22	27.40
NMF [XLG03]	49.85	35.89
LSI [Hof99]	42.00	37.14
LCCF [CHH11]	33.07	30.45
LDA + K-means [XX13]	29.73	36.00
LDA + Naive [XX13]	54.88	48.00
MGCTM [XX13]	56.01	50.10
Our method	**67.19**	**51.97**

for these methods, we set the cluster number $K = 10$ in the experiment, which equals the ground-truth cluster number in the dataset. Please refer to [XX13] for other detailed settings of these algorithms.

Parameter Settings of Our Method. To compare with the other algorithms, in our method, we add a Gaussian prior term with the mean $\mu = 10$ and variance $\sigma^2 = 0.5$ to Eq. 3.13 to make the sampling process converge to the state where the cluster number equals 10. The parameter α in Eq. 3.13 is set as $\alpha = 0.2$. The weights $\{\lambda_F, F \in \mathcal{F}\}$ in Eq. 3.16 are set as: $\lambda_{F_{who}} = 0.1$, $\lambda_{F_{where}} = 0.1$, $\lambda_{F_{what}} = 0.4$, $\lambda_{F_{face}} = 0.1$ and $\lambda_{F_{object}} = 0.3$. The threshold τ used for graph pruning is set as $\tau = 160$.

Comparison Results. Table 3.1 shows the results of different methods on Reuters-21578. It can be seen from the results that our approach is better than other methods in terms of both the clustering accuracy and the normalized mutual information. This is because our method uses the Multimodal Topic And-Or Graph representation which organizes topics in a hierarchical way and embeds contexts between different components. The cluster sampling

method Swendsen-Wang Cuts also plays an important role in optimizing the solution.

Among the other methods, topic modeling based methods, i.e. LDA + K-Means, LDA + Naive, and MGCTM generally perform better than K-Means, normalized cuts, and factorization based methods (NMF, LSI and LCCF), which shows that topic modeling can help get better similarity measures compared to the tf-idf vectors. However, they still use the basic word distributions, which are not sufficient for representing news events. Our approach models the semantic structures of news events and the event relationships, which helps represent each event cluster's semantics and distinguish between different clusters. Moreover, most of the solutions these comparison methods get are locally optimal, while in our approach, SWC helps sampling the solution space more effectively, resulting in better clustering performance.

3.5.2.2 Results on UCLA Broadcast News Dataset

We conduct both qualitative and quantitative evaluations of our topic detection method on the UCLA Broadcast News Dataset.

Preprocessing. We preprocess news videos and closed captions to obtain texts and key frames used in the experiment. In detail, we utilize texts from both video frames and closed captions (CC). Text extraction on video frames is performed using optical character recognition (OCR) based on Google OCR engine Tesseract[4], and the results are further refined using the spatial-temporal relations between frames. News CC consist of several stories in one single continuous text stream. Accordingly story segmentation needs to be performed to divide the CC into stories before the topic detection and tracking process. In CC, some special markers are used as the indicators of story boundaries, such as ">>>". Moreover, many news programs insert commercials between stories with special formats of letter cases and indentations. Thus we also do commercial detection based on these special formats. Using the special boundary markers and commercial detection results, most stories boundaries are determined. For the remaining boundaries, we train a classifier using Support Vector Machine to decide the boundary locations based on features including the

[4]Available online at `https://code.google.com/p/tesseract-ocr/`

boundary key words (such as "coming up", "still ahead"), and similarities of sentences near the boundaries.

For the news videos, we extract the keyframes by removing the commercial frames, redundant frames and anchor frames. Commercial frames can be specified using the aforementioned commercial detection results from CC (each CC line has corresponding start and ending time information). Redundant frames are those perceived to be similar to the previous frames. They appear frequently in the news videos especially when the scene is unchanged and people don't move frequently. To detect the redundant frames, we use the frame histograms to decide whether one incoming frame is similar to the previously detected non-redundant frame. One frame is kept only when its similarity with the previous non-redundant frame is low enough. After removing the commercial and redundant frames, we further detect anchor frames among the remaining frames. Anchor frames are those containing the news anchors, especially when they are reporting in the studios. The anchor frames usually appear repeatedly in the video. Hence we detect the anchor frames by exploiting features from two aspects: anchor frames' backgrounds (they usually show the news studios and thus are similar in the videos), and anchors' faces (they also appear repeatedly in the video). Similar backgrounds and faces are grouped by K-means clustering. We can then check the clusters' time distribution and decide whether the corresponding frames are anchor frames or not.

Fig. 3.8 illustrates the results that can be obtained in the previous preprocessing procedure.

Preprocessing visual frames such as face detection or running a convolutional neural network is in fact more time-consuming part in our system. It takes about 5s to preprocess one visual frame. In practice, we use a distributed computing system to extract the features.

After preprocessing, we have 3,633 news stories including 577,721 words and 36,810 keyframes. The whole collection contains 24,036 unique word terms.

1) Qualitative Evaluation. We conduct the topic detection experiment on the whole dataset.

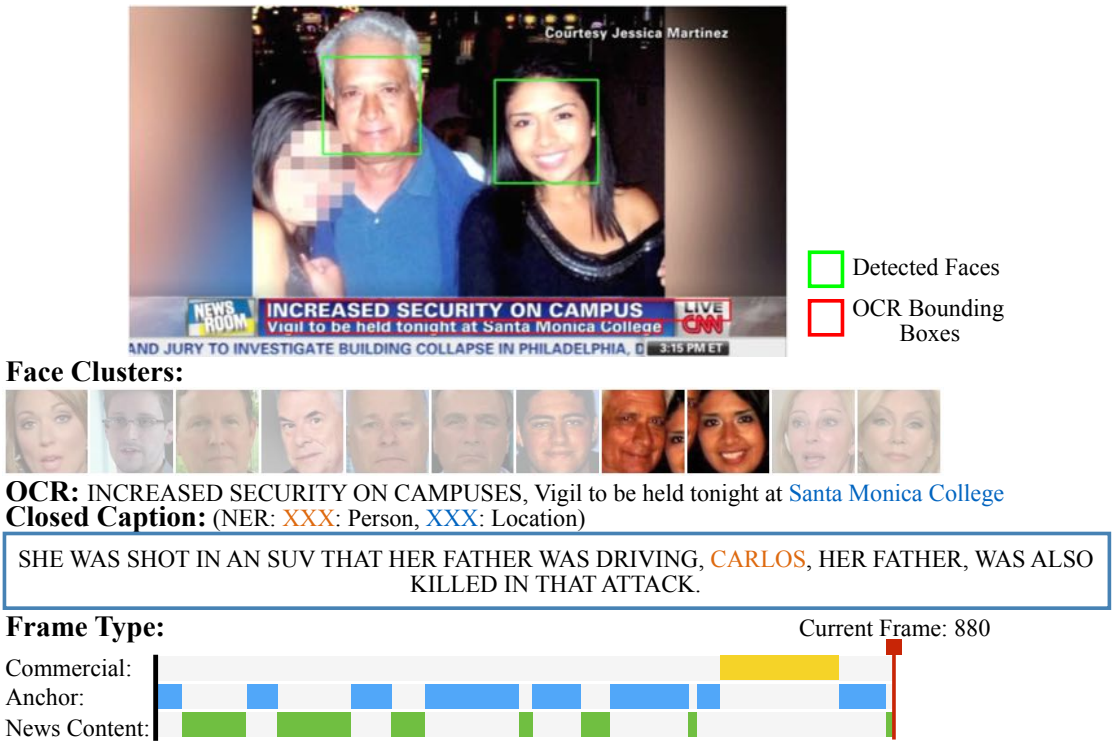


Figure 3.8: An example showing the preprocessing results, including those for OCR, NER, face detection and clustering, commercial detection (marked by "Commercial" label in the "Frame Type" part), and anchor frame detection (marked by the "Anchor" label).

Parameter Settings. The parameter α in Eq. 3.13 is set as $\alpha = 10$. The "fast" mode of Selective Search is used to generate the possible object patches (please refer to [USG13] for more detailed settings of the "fast" mode). The cluster numbers for grouping the faces and object patches in Chapter 3 are set as $1,000$ and $1,500$ respectively. We also delete clusters with a small number of patches. The remaining cluster numbers for face and object are 708 and $1,316$ respectively. Other parameter settings are the same as those in Section 3.5.2.1.

Topic Detection Results. We show the detected top five topics in Fig. 3.9. Topic 1 talks about the news that Edward Snowden leaked information from National Security Agency (NSA). Topic 2 is about the IRS scandal, including the discussion on the misuse of taxpayers' money and the related hearing. Topic 3 mainly talks about the Oklahoma tornado, including its development, the damage it caused, and the storm chasers' stories.

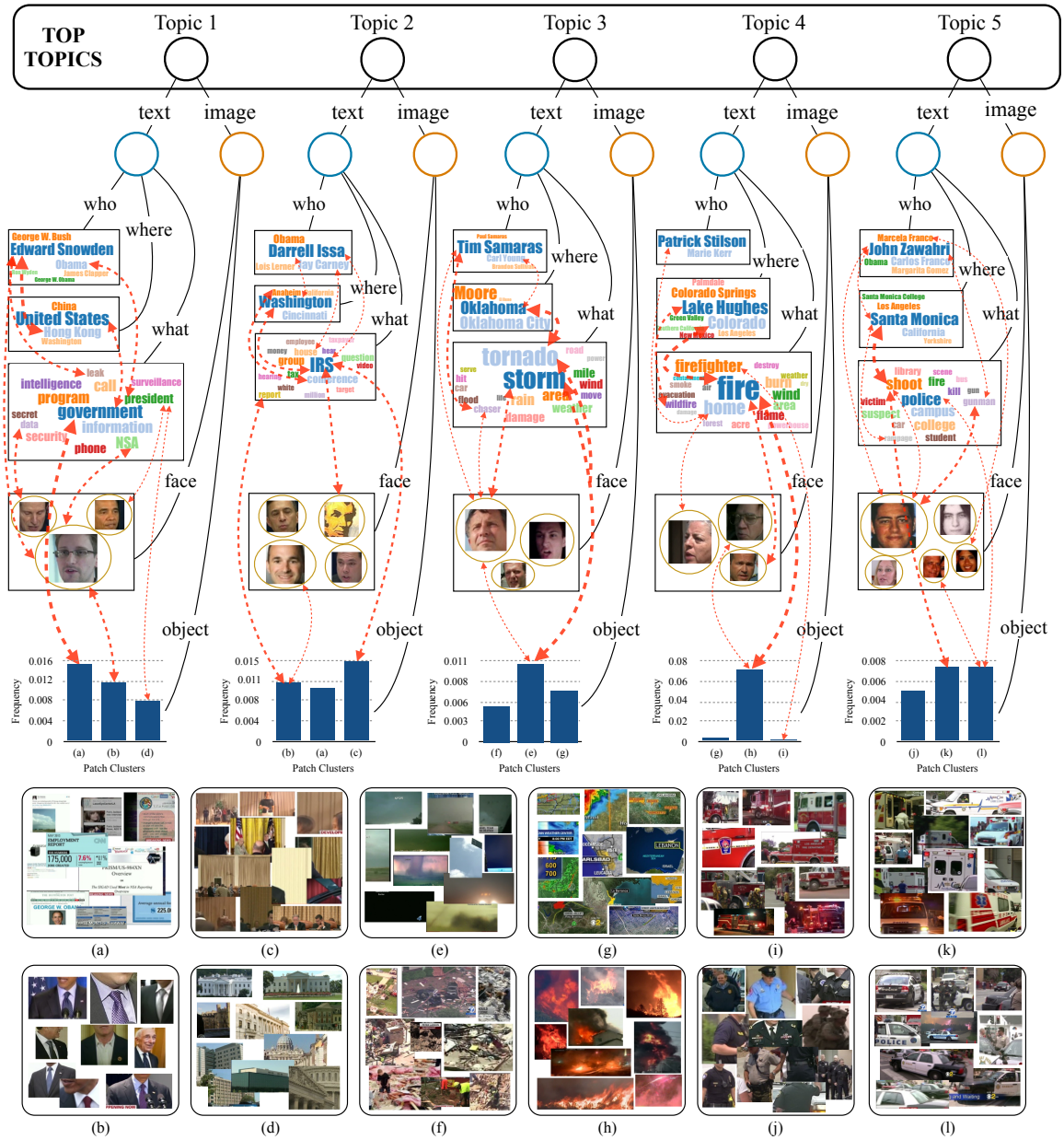


Figure 3.9: Top five topics detected in the dataset we collected. The top words for who, where and what components are shown with their sizes proportional to their frequencies. The top faces and objects are also shown. The face sizes are also proportional to their frequencies. Object clusters are shown in squares at the bottom part of the figure. The objects' frequencies in the topic are shown by the curves above the squares. The dashed red lines show the top co-occurring pairs between different components and the thickness of each line is proportional to the related pair frequency.

Topic 4 is about the wildfires, which also includes the fire development and the related damages. Topic 5 is about the Santa Monica College shooting rampage, and the related gunman and victims' stories are also included.

We can see from the figure that the obtained structured results can clearly describe the related topics. The involved people's names and face patches, related locations, key objects, descriptions about the event, as well as the co-occurrence relations between them (represented by the dashed red lines) are all shown in the structure. And as shown in Fig. 3.9, the topic components and their co-occurrence relations are closely related to the detected topic. For example, in Topic 1 which is about the NSA leaking, the "who" component contains the main people related to the topic, such as "Edward Snowden" and the US president "Obama". The "where" component shows the related locations, such as the main location "United States", and "Hong Kong" where Edward Snowden stayed. Top words in the "what" components, such as "leak", "NSA", "government", "information", etc., further describe the topic in detail. The face component show the faces of the main characters, e.g. Snowden and Obama's faces. The object component describe the objects related to the topic, e.g. the suits and the white house, which also shows that this is a topic related to the US government and politics. The top co-occurrence relations obtained during the topic detection process can tell some common knowledge related to the topic, e.g. Obama is the president and how his face looks like.

2) Quantitative Evaluation.

Evaluation Protocol. Using the annotated story pairs, we draw precision-recall curves for different topic detection methods in the evaluation. The precision is calculated as the fraction of story pairs that actually belong to one topic out of those that are computed to be. The recall is the fraction of story pairs that are computed to belong to one topic out of those that actually do. In detail, let tp, fp, tn, fn denote the number of true positives (i.e. the number of stories correctly labeled as belonging to the same topic), false positives (i.e. the number of stories wrongly labeled as belonging to the same topic), true negatives (i.e. the number of stories correctly labeled as belonging to different topics), and false negatives (i.e. the number of stories wrongly labeled as belonging to different topics), respectively. Then the

precision is defined as precision $= tp/(tp + fp)$, and recall is defined as recall $= tp/(tp + fn)$.

Other Methods. Among the comparison methods used in 3.5.2.1, we select two methods with better performance, i.e. LDA + Naive and MGCTM [XX13]. We also include the widely used k-means algorithm. These algorithms are all unimodal, so their inputs in the experiment are still the stories' textual information, i.e. the stories' tf-idf vectors. Two multimodal methods are also included in the comparison, including the multimodal co-clustering method in [WNL06], and the multimodality graph with topic recovery method (MMG+TR) in [CZL14]. For these method, we set a sequence of cluster numbers in the experiment to generate the precision-recall curves. Other settings of these methods are kept the same as those in [XX13], [WNL06] and [CZL14].

Parameter Settings of our method. To generate the precision-recall curve, for our method, we vary the parameter α_K in Eq. 3.13. Other parameter settings are the same as those in the qualitative experiment. To compare with the unimodal methods, we also conduct experiments where only texts are included in the experiment (i.e. only the text part in the MT-AOG model is used, and the image part is ignored).

Comparison Results. Fig. 3.10 shows the precision-recall curves for different methods. As we can see from the figure, similar to the comparison results on Reuters-21578, based on merely texts, our method has better performance than the other unimodal methods. This shows quantitatively that the proposed Multimodal Topic And-Or Graph (MT-AOG) and the clustering sampling method we use can help generate better topics. The comparison results of k-means, LDA + Naive, and MGCTM are similar to those in Reuters-21578, which again shows that topic modeling based clustering methods, e.g. LDA + Naive and MGCTM, perform generally better than k-means.

With the visual cues added, our performance gets further improved, showing the effectiveness of our method which jointly models texts and images. Multimodal methods also perform better than the other unimodal methods in general, which shows the necessity of using visuals in topic detection. However, the other multimodal methods, i.e. the multimodal co-clustering method and MMG+TR, are not capable of modeling the decomposition of tex-

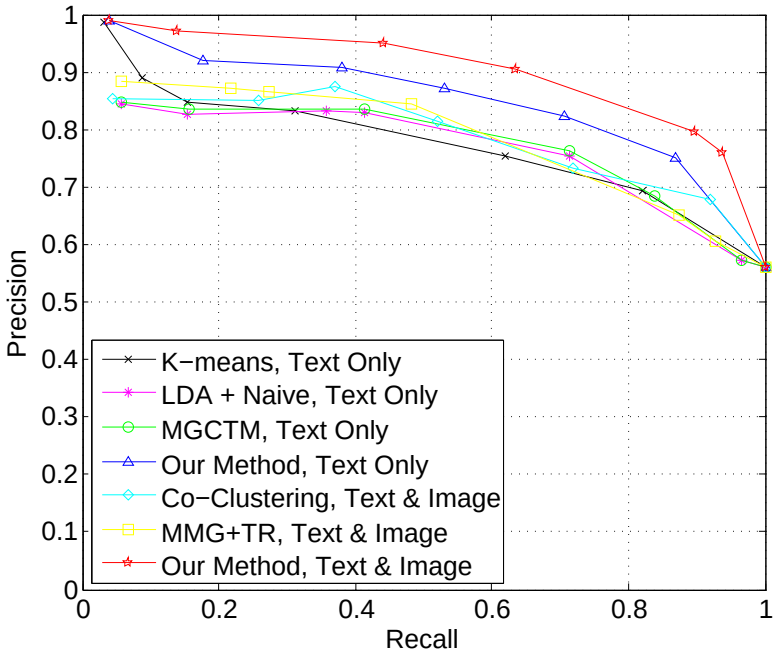


Figure 3.10: Precision-recall curves of our topic detection method and comparisons with other methods on UCLA Broadcast News Dataset.

tual and visual parts in news topics. The bag-of-word representation for texts, and color histograms or near-duplicate keyframes based representation for visuals cannot model the semantics in either texts or visuals. Our method can model not only the semantic structures of both texts and images, but also their contextual relations.

Evaluation of Contextual Relations. To demonstrate the effectiveness of contextual relations in MT-AOG separately, we conducted an ablation study. The contextual relations in the text part (Sec. 3.2.2), in the image part (Sec. 3.2.3), and between these two parts (Sec. 3.2.4) were tested in the experiment.

Note that there are 303 stories in our dataset (8.34%) which do not have any image elements (e.g., only anchor's comments without field footage). These stories were treated as individual clusters in "image-only" cases.

The results are shown in Fig. 3.11. As expected, incorporating contextual relations is critical for achieving a better clustering performance in all cases, which justifies our model choice. In addition, this result also reinforces that using multimodal cues together is better

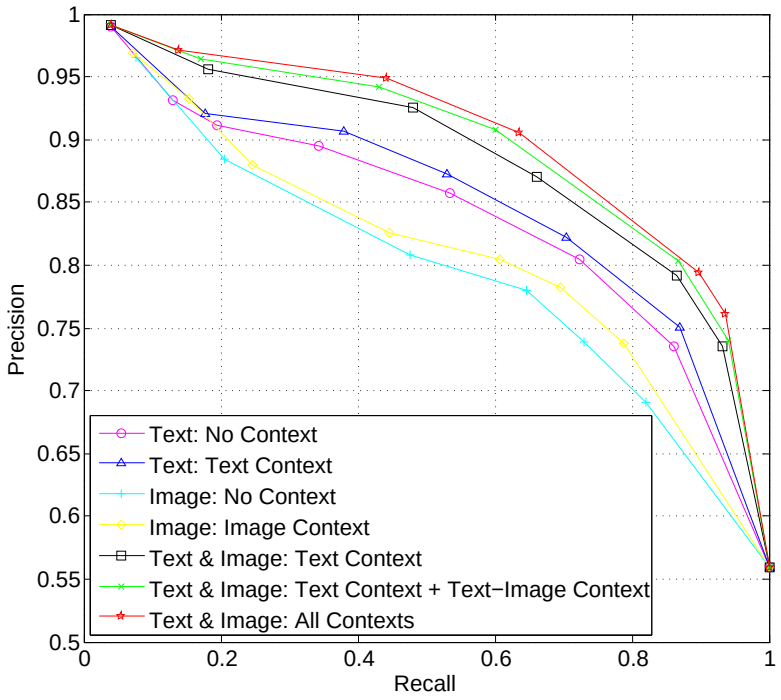


Figure 3.11: Precision-recall curves of our topic detection methods with/without different contextual relations.

than using a single channel.

3.5.3 Experiment II: Topic Tracking

In this experiment, we conduct topic tracking experiments on the UCLA Broadcast News Dataset. Both qualitative and quantitative evaluations of our method are included in the experiment.

1) Qualitative Evaluation. To show that our topic tracking method can generate meaningful topic trajectories, we conduct the qualitative evaluation experiment.

Parameter Settings. To track topics over time, we divide the dataset into 14 sub-collections, each of which contains news stories from one day. The average number of news stories per day is 260, and these stories on average contain 41,266 words and 2,629 keyframes. Topic detection is firstly performed within each sub-collection. Then given the detected topics, we do topic tracking, which links topics over time and generates topic trajectories.

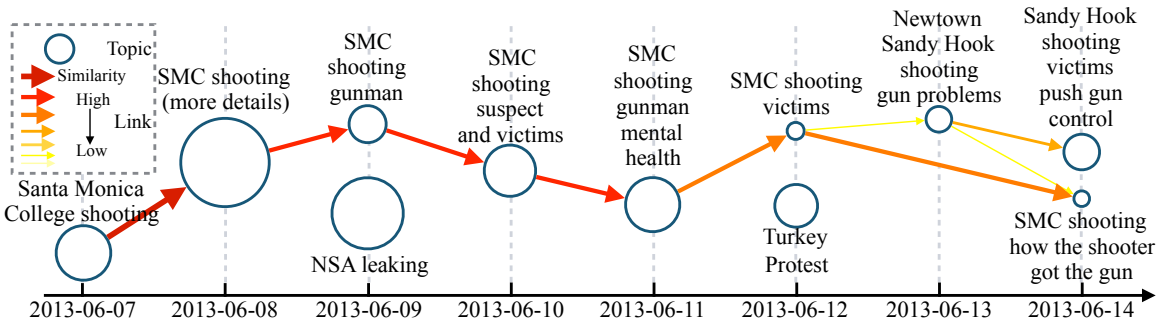


Figure 3.12: Topic tracking result of the event Santa Monica Shooting. Each circle represents one topic and the circle size is proportional to the size of the topic, i.e. the volume of corresponding news stories. Thicker links represent greater similarities between topics.

The parameter α_{sim} and β_{kl} in Eq. 3.20 are set as $\alpha_{sim} = 0.8$ and $\beta_{kl} = 0.005$ respectively. The weights $\{\lambda_i; i = 1, ..., 5\}$ in Eq. 3.21 are set as $\lambda_1 = 0.1, \lambda_2 = 0.1, \lambda_3 = 0.4, \lambda_4 = 0.1, \lambda_5 = 0.3\}$. The threshold τ_{link} for selecting links between topics is set as $\tau_{link} = 0.7$.

Topic Tracking Results. One topic tracking trajectory about the Santa Monica College shooting is shown in Fig. 3.12. The topics are summarized in several words here for space constraints. The descriptions of the text part and the image part for the corresponding topics in the trajectory are shown in Fig. 3.13 and Fig. 3.14 respectively. The probabilities of the top textual and visual words over time are shown in the figure.

As shown in these figures, when the event happened, news reports in the first two days are generally about the shooting scenes and details. The text parts mainly describe the shooting event and victims, and top objects in the image parts are mainly ambulance and police cars. As the topic developed, the suspect was confirmed by the police, so more information about him was covered in the news, as shown in the who component ("John Zawahri"), the face component (the third and fifth faces are from the suspect). Victims' stories were also reported. Later the news talked about why the suspect made the shooting, as shown by the text part where his mental health problem was discussed. At last, this shooting event also lead to the discussion about the gun control problem.

Based on the tracking result, we also analyze the emotional changes as the topic developed. The NRC Emotion Lexicon [MT13] is used for the emotion analysis. Three emotional

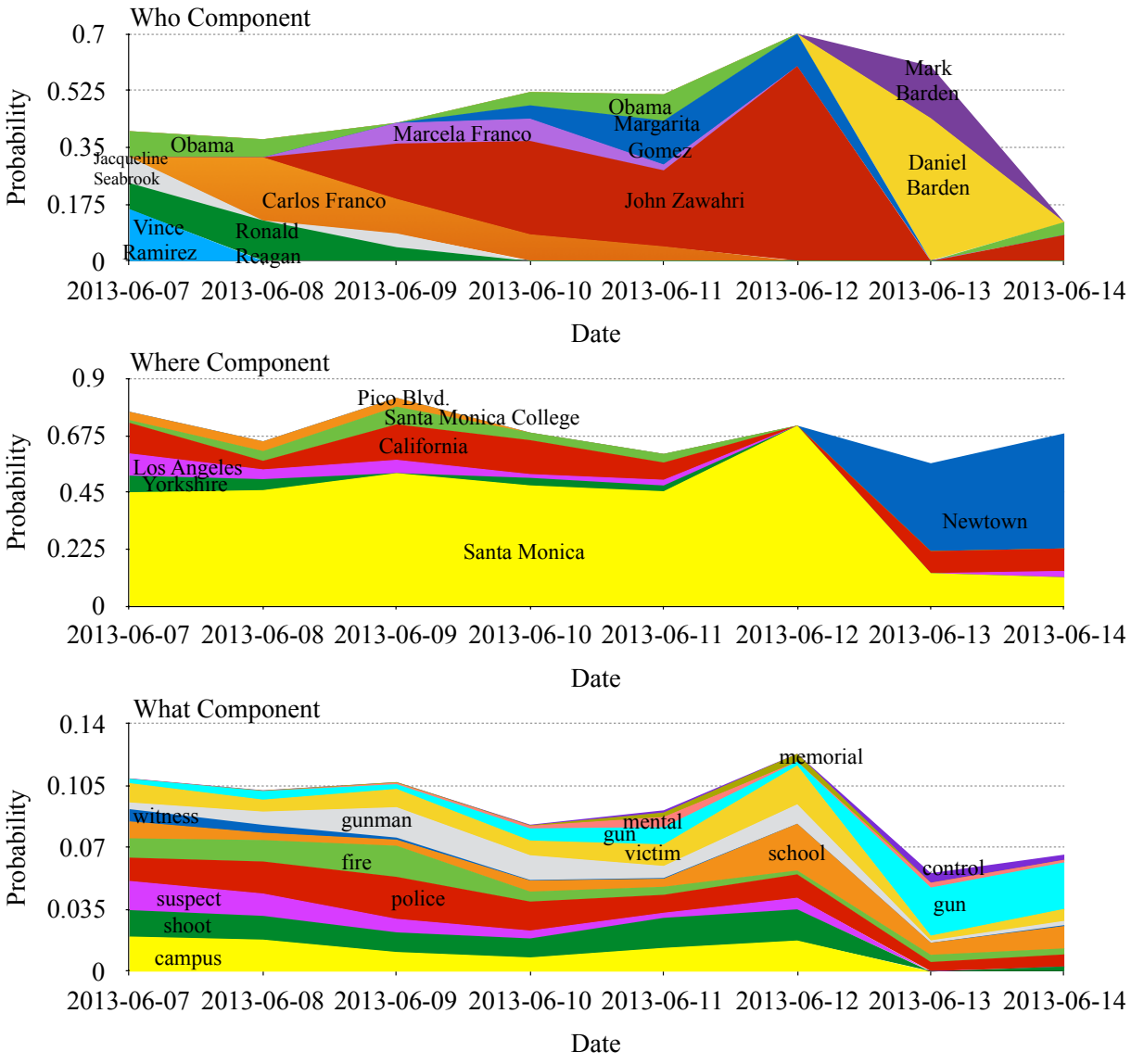


Figure 3.13: The text part of the topics corresponding to the trajectory shown in Fig. 3.12. The probabilities of top words along the time span are shown.

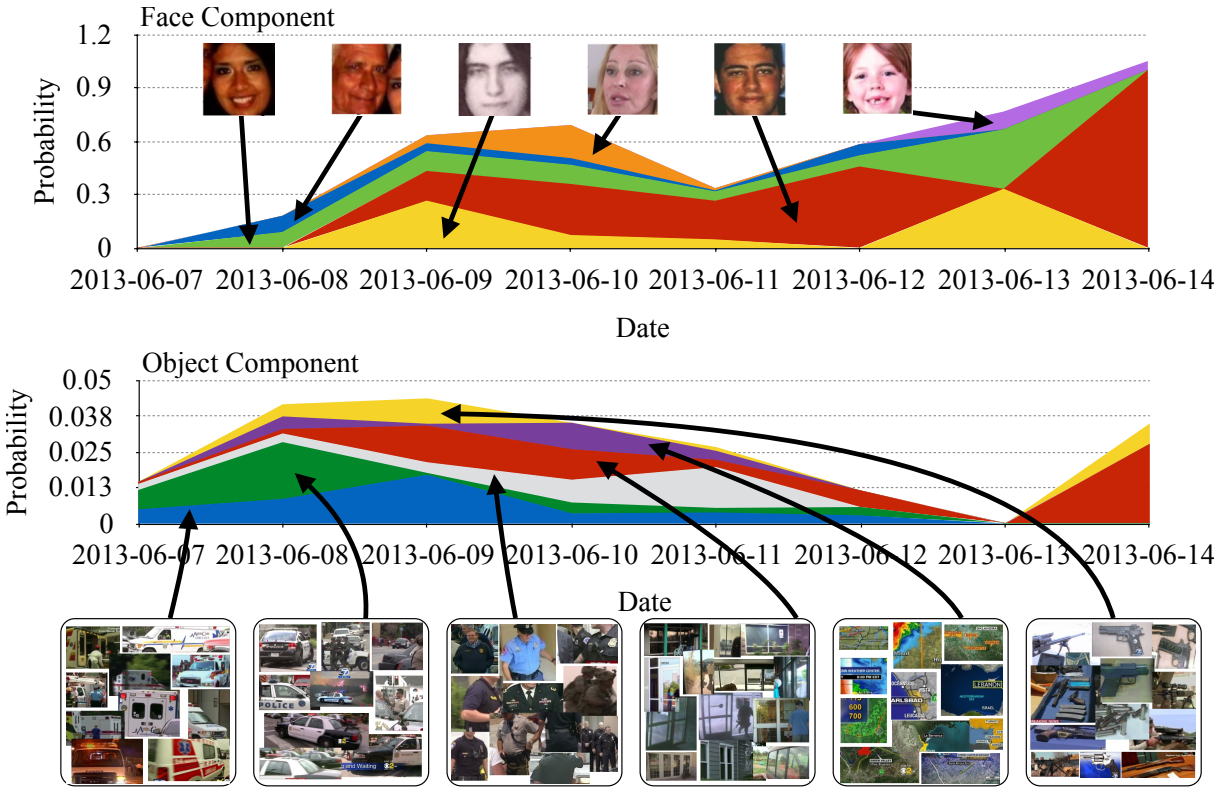


Figure 3.14: The image part of the topics corresponding to the trajectory shown in Fig. 3.12. The probabilities of top faces/objects along the time span are shown.

variables, i.e. fear, anger and sadness, are included in the analysis and the ternary plot on these variables is shown in Fig. 3.15. From these figures we can see that at the beginning, when the shooting happened, news stories mainly describe the shooting scenario and expressed people's fear mostly. Later when the suspect was found, more anger is shown in the news stories. When victims' stories were told later, sadness became dominant. From these results, we can clearly see how mass media reported the event and what emotions they want to express. They also show that our tracking method can generate meaningful tracking trajectories.

Our experiment is conducted on a computer with 3.6 GHz CPU and 16G RAM. The average time for topic detection for one day's stories is 7.16s, and the average time for topic tracking is 2.41s. So our method can deal with news streams efficiently.

2) Quantitative Evaluation. We also conduct quantitative evaluation on our topic

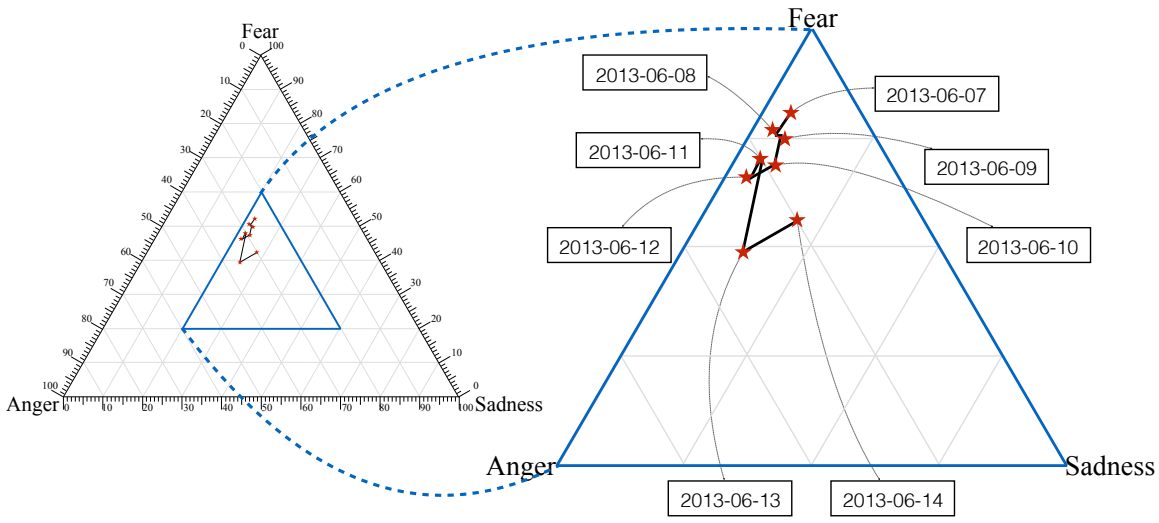


Figure 3.15: Emotion analysis for the Santa Monica Shooting event. The ternary plot on three emotional variables (fear, anger and sadness) shows the emotional changes in the news stories as the topic goes on.

tracking method.

Evaluation Protocol. For topic tracking, we also use the precision-recall curves to compare different methods.

Other Methods. We include three methods in the comparison, namely: (1) Dynamic topic model (DTM) [BL06] which models topic changes over time; (2) topic chain method [KO11] which generates topics in different time periods using LDA and links these topics to form topic chains; and (3) temporal Dirichlet mixture model (TDPM) [AX08] for evolutionary clustering.

These three methods are all unimodal, so we use the tf-idf vector to represent each news story. For DTM, we set different topic numbers to generate its precision-recall curve. For the topic chain method, we set the topic number in each time period as 50 and use a sequence of similarity threshold when building the topic chains. For TDPM, we vary the concentration parameter to obtain its precision-recall curve.

Parameter Settings of our method. We vary the parameter τ_{link} to generate the precision-recall curve.

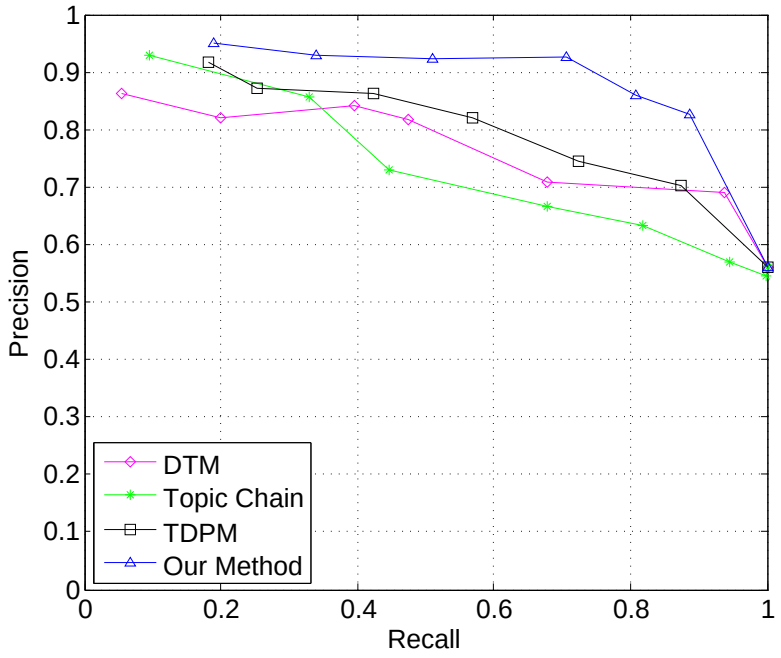


Figure 3.16: Precision-recall curves of our topic tracking method and comparisons with other methods on UCLA Broadcast News Dataset.

Comparison Results. Fig. 3.16 shows the precision-recall curves for our tracking method and the other two methods. Our method outperforms the other two methods since both texts and images are included. Moreover, our topic detection method can generate meaningful topics, which is also an important factor for the topic tracking performance.

3.5.4 Experiment III: Large-Scale Topic Detection and Tracking

To show that our method can work effectively on large-scale datasets qualitatively, we conduct topic detection and tracking experiment using the CNN news data in 2012. We firstly detect topics within each month, and then do topic tracking among the detected topics.

The obtained trajectories are shown in Fig. 3.17. Due to the space limit we only show some top topics and the text parts of topics in the trajectory. The top part of the figure shows the trajectory of George Zimmerman's case and some other related shooting cases such as the Chardon shooting in February and the Colorado theater shooting in July. The middle part of the figure is mainly about topics closely related to the 2012 US election, such as the

health care, the immigration problem, the economy and the debates. The Syria problem, which is another factor related to the election, is shown in the bottom part of the figure. We also get some short trajectories such as the one about Olympic shown in the lower half part of the figure. The "who", "where", and "what" components shown in the figure are highly relevant to the corresponding topics.

From the topic trajectories, we can clearly see how these topics develop over time and how they relate to each other. For instance, from the trajectory of George Zimmerman's case, we can see that the shooting happened in Feb. 2012, followed by the bond hearing and defense process, and the final trial in July 2012. This case is related to other shooting cases shown by the links between them. And as expected, the links within the same event are thicker compared to those between different events.

We also conduct another experiment to track gun shooting events. We use the CNN news data in 2015. For three gun shooting events we detect, we analyze the emotional changes in their coverage. The emotion analysis results are shown in Fig. 3.18. As we can see from the results, even though the emotional changes are not exactly the same, they all follow similar patterns as that in the Santa Monica Shooting event: when the shooting happened, news stories expressed more fear; later they showed more anger; and at last sadness became dominant in the news reports.

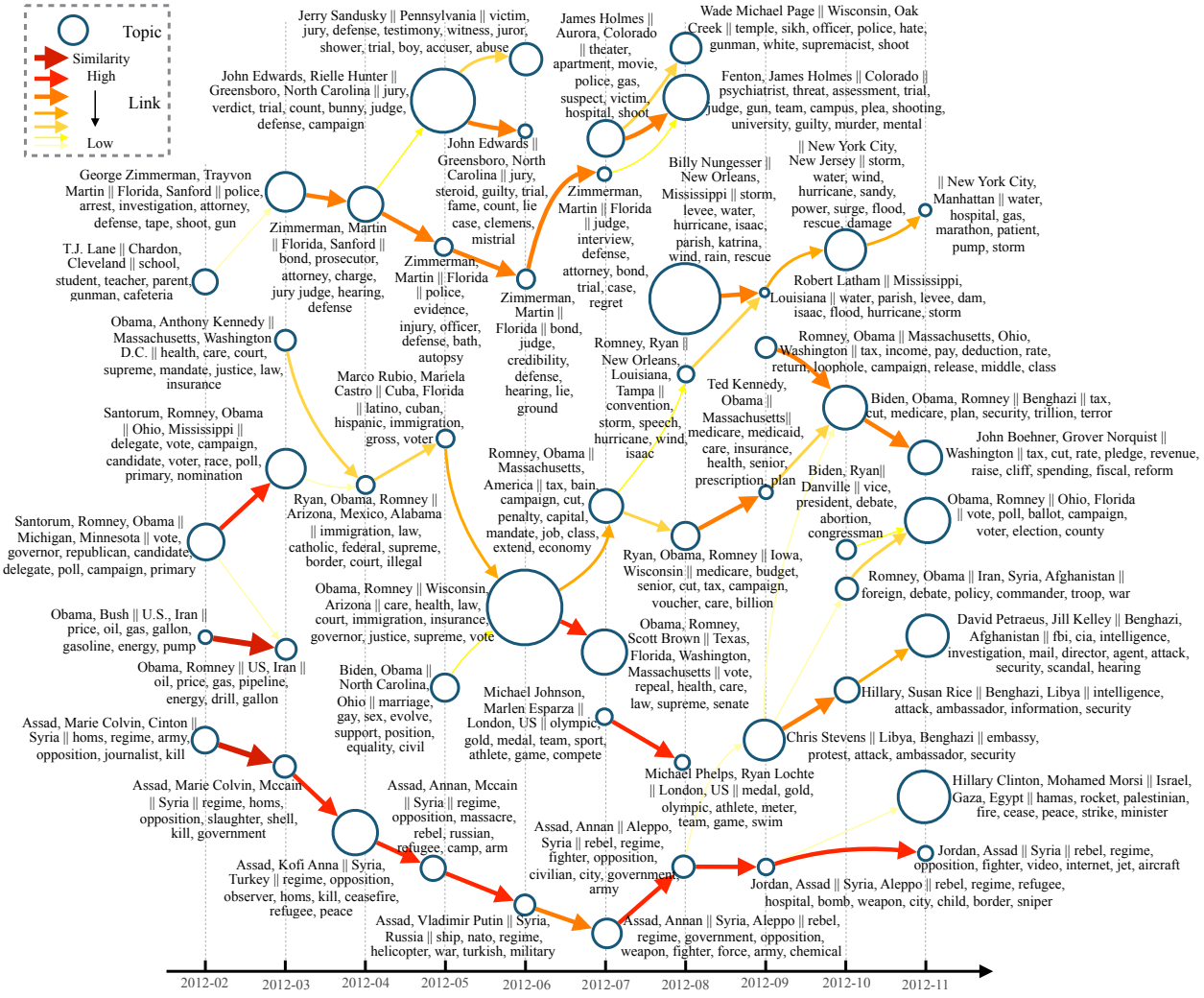


Figure 3.17: Topic trajectories for 2012 CNN news. Each circle represents one detected topic and the circle size is proportional to the topic size, i.e. the volume of corresponding news stories. Thicker links correspond to greater similarities between topics. The who, where and what parts of the topic are separated by the symbol "||".

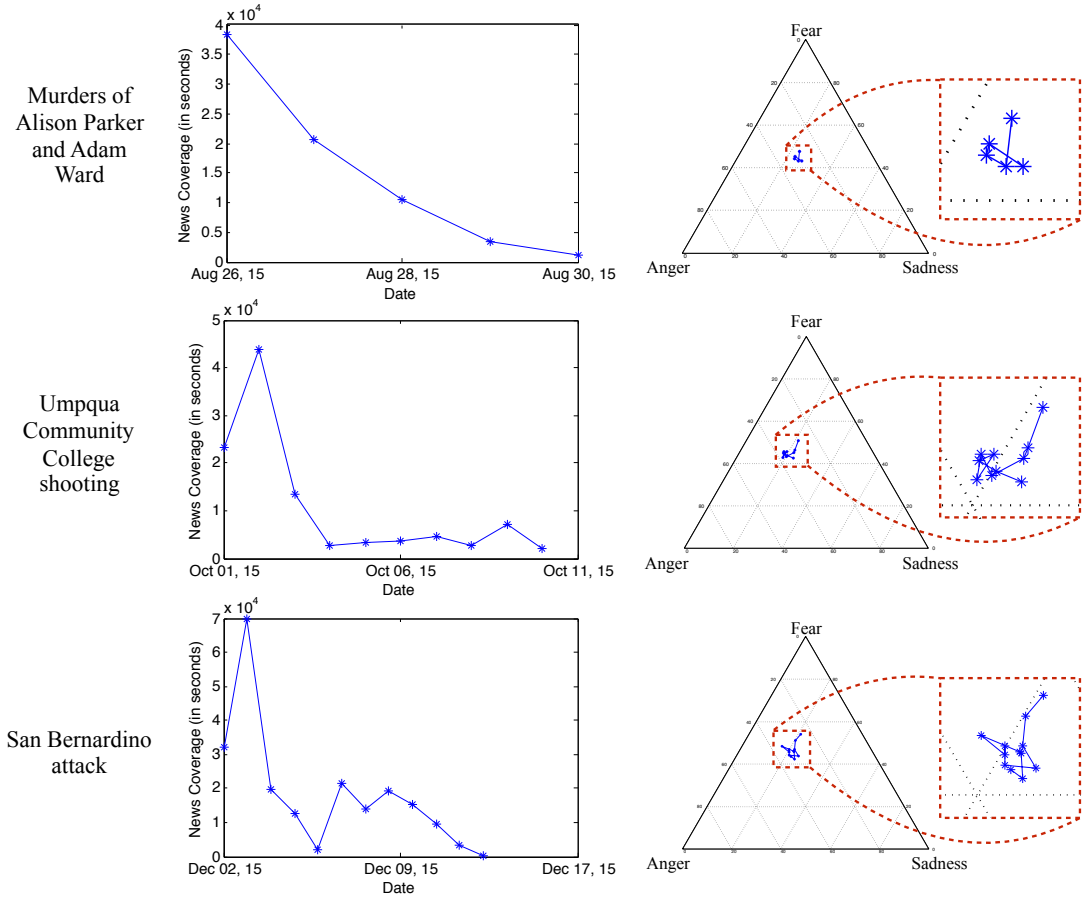


Figure 3.18: Emotion analysis for three gun shooting events, including the Murders of Alison Parker and Adam Ward, Umpqua Community College shooting, and the San Bernardino attack. The curves show the news coverage (in seconds) of events each day. The ternary plots on three emotional variables (fear, anger and sadness) show the emotional changes in the news stories as the topic goes on.

CHAPTER 4

Visualizing the US Presidential Elections

4.1 Introduction

Scholars studying campaign communication may long for the simplicity of the days when a few "boys on the bus" dominated coverage of presidential elections [Cro73] as current campaign discourse threatens to overwhelm scholars' ability to truly understand the information upon which voters act in the voting booth. The core challenge in a quantitative content-based analysis of election communication is the sheer amount of data to annotate (or code), especially due to the development of mass media. Traditionally, social and political science researchers recruit human coders who can provide annotations on given data and then statistically analyze the obtained annotations. In instances where scholars have used automated methods to expand their analysis of news content, they have generally chosen to focus purely on text and ignore the rich visual dimension of communication because of its complexity.

Our study explores several important aspects of campaign communication, but does so using new tools that might rescue scholars from this overwhelming wave of information. Computer vision and machine learning are branches in computer science and statistics aimed at automatically processing digital visual input such as images or videos to recognize and represent visual contents and knowledge. The techniques in these fields have largely grown in the last decade from small-scale experimental and exploratory studies to practically applicable and matured systems. This major improvement has been made possible mainly by enormous sizes of datasets available for model training and enhanced quantity and quality of computing power.

In this work, we apply fully automated coding to a massive collection of news and other

campaign information to track:

- which candidates are discussed on Twitter and shown in traditional television news coverage,
- what topics are being discussed in relation to the candidates (and by which news outlets), and
- which candidates were treated most favorably across news outlets and media (including particularly the images that were selected to represent them on television news).

Our analysis draws on a unique and massive corpus of textual and visual data. It also builds on cutting-edge machine learning methods to address novel research questions in such a massive collection of information. While much of the analysis presented here is preliminary and undergoing active development, we are excited to be working with tools that have the possibility of transforming how research on these topics can be accomplished.

4.2 Research Questions

In presidential elections, the press has often been assigned a role as the "great mentioner" who can make some candidates relevant and others sink without a trace during the so-called "invisible primary." For that reason, we begin by asking the degree to which some candidates were mentioned more than others.

R1: Which candidates receive the most coverage?

Within that overall question, we further sharpen the question to break down our question depending on the medium or outlet through which the information is conveyed.

R1A: Do different media outlets pay attention to candidates similarly? (In particular, do more partisan outlets cover candidates in a systematically different way?) and

R1B: Do some candidates receive disproportionate attention on social media?

Finally, we start to address issues closer to traditional conceptions of media bias research. We begin by asking generally,

R2: How does the content of coverage candidates receive differ across media and outlets?

Which we then focus to two specific, related questions:

R2A: How do issues and topics covered in the campaign differ across candidates and outlets?, and

R2B: Do some candidates receive more favorable coverage than others on Twitter or in the choice of visuals in traditional news?

4.3 Data

4.3.1 Television News Content

The main data source for this project is the UCLA Communication Studies News Archive, founded by Prof. Paul Rosenthal in 1974 with a mandate to preserve ephemeral television news content that would otherwise be lost. The Archive began recording a full daily schedule of all local and national television news available in Los Angeles starting in 1979, and began recording a full daily schedule straight to digital files beginning in October 2006. The Archive collects the full video file and time-coded closed captioning text for each news broadcast, and then collects a variety of other metadata (current metadata includes optical character recognition (OCR) of onscreen text, automated commercial break detection, and thumbnail images collected at regular intervals, among other items). See Table 4.1 for some information on the collection's holdings.

The Archive currently contains news programs recorded during the 2008, 2012, and 2016 presidential election years. Our analysis uses two different subsets of the Archive's data for

[1]Data as of 8/28/16. Includes straight-to-digital files recorded beginning with a pilot project in 2005. Series tally include non-regularly-scheduled programs. The Archive is currently working to digitize our analog back catalog, which started recording in the 1970s. For copyright reasons, research access to the collection is limited to the UCLA campus and to members of the inter-campus Distributed Little Red Hen Lab research consortium.

Table 4.1: UCLA Communication Studies News Archive descriptive statistics[1].

Start of Daily Digital Recording	2006
Networks	46
Series	2,525
Total video files	383,550
Duration in hours	297,596
Closed caption files	383,739
Words in caption files	2,419,185,351
OCR files	371,426
Words in OCR files	825,662,597
Total thumbnail images	107,134,425
Storage	106.93 terabytes
Limited public access link	`tvnews.library.ucla.edu`

this project:

- For our analysis focusing specifically on the 2016 election cycle, we limit our analysis to shows originating from CNN, Fox News, MSNBC, and the three broadcast networks and their local Los Angeles affiliates (190 shows in total, and full list of these shows available upon request).We analyze shows for one year starting on August 1, 2015 and continuing through July 31, 2016.

- For our comparative analysis of the 2008, 2012, and 2016 presidential elections, we chose to limit our analysis to shows appearing on CNN, Fox News, and MSNBC (including special election coverage beyond their regularly-scheduled programs). Because of the longer time frame of this analysis and the inclusion of some special and/or cancelled programs, this dataset actually includes more shows (275 shows, full list available upon request) than the 2015-16 dataset.

4.3.2 Twitter Data

To better understand the flow of campaign information through less-mediated channels, we collected data from Twitter during the same time period as our 2016 television news dataset (August 2015-July 2016). To collect our data, we keyword-searched tweets for candidate names using the API provided by Twitter. We assume the tweet numbers the API returns for different candidates are proportional to the actual distribution in the (unobserved) complete dataset.

4.3.3 Candidates

For our comparative presidential study of 2008, 2012, and 2016, we restrict our analysis here to the eventual Republican and Democratic nominees. For our in-depth analysis of the 2016 nomination battle, we restrict our analysis to the "final four" candidates (Donald Trump (R), former Secretary of State Hillary Clinton (D), Sen. Bernie Sanders (I-VT), and Sen. Ted Cruz (R-TX)). However, it should be noted that we did collect equivalent data on every announced candidate for president as part of our larger Viz2016 initiative (see Figure 4.1 for an example of the Viz2016.com data), and will extend our analysis to these candidates at a later date.

4.4 Methods

Our method in content analysis is based on a coding scheme similar to prior manual studies ([GB09], etc) but replaces all manual hand-coding with automated measures by computer vision. In addition, we developed a fully automated computational processing pipeline which further eliminates the necessity of any manual labors such as sample selection, commercial detection, or story segmentation.

Our news dataset consists of videos and corresponding closed caption text files. Each news video is first preprocessed by frame extraction at the rate of 1 frame per second. This generates 3,600 frame images for an one hour-long video. We also parse the caption text and

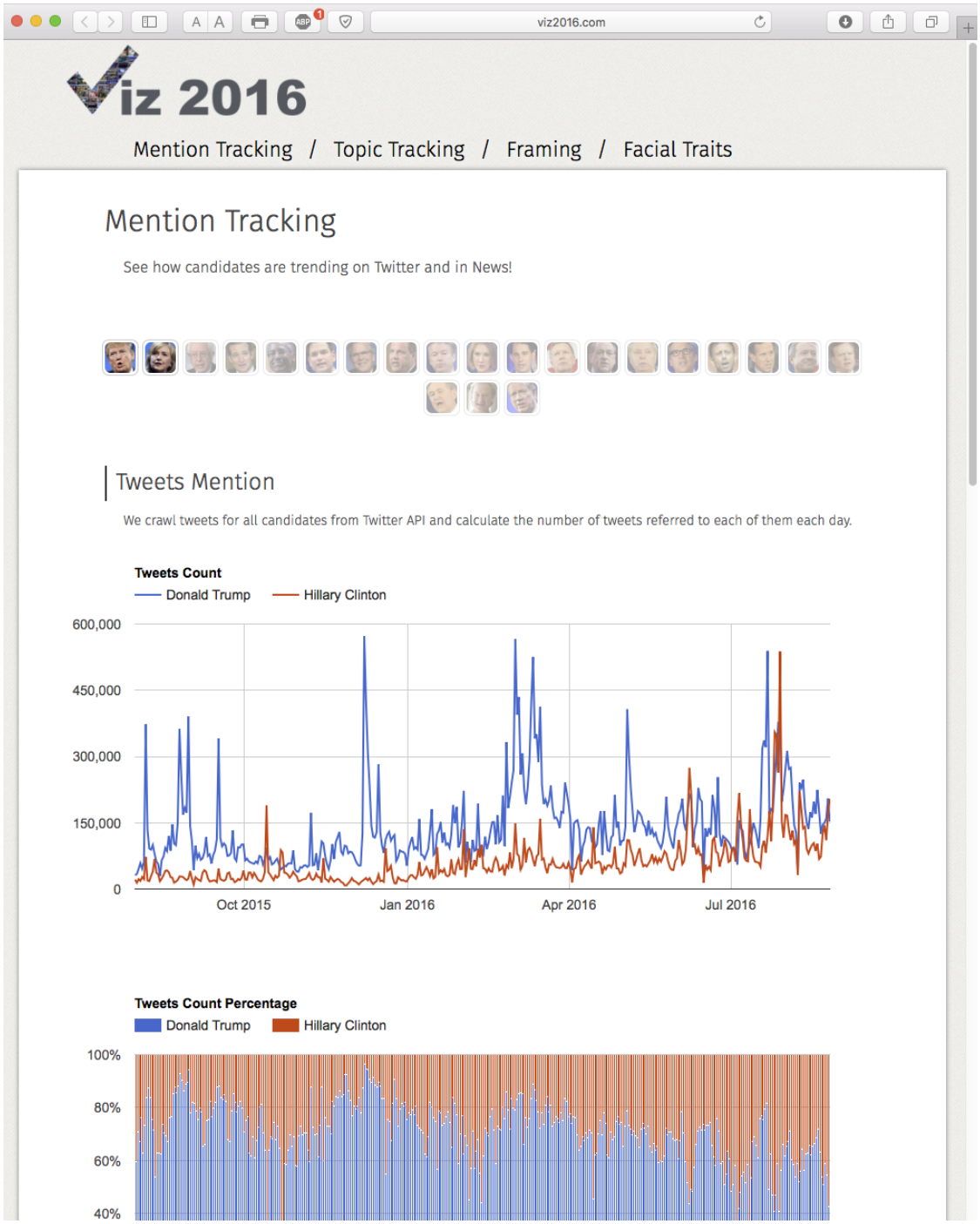


Figure 4.1: Candidate tracking, Viz2016.com site.

automatically detect the commercial parts so that the equivalent parts of the video can be discarded from the analysis.

Excluding these commercial parts, we apply a Viola-Jones face detector [VJ01] to detect any present faces in each frame. This detector, based on local contrast features and trained by adaboost, has been the standard face detector in the field. Then to make subsequent processing more robust and accurate, we further align the facial image region by facial landmark detection. These are all done by public software packages developed for general purposes without any modification to our use.

We further process the aligned face for recognition and smile classification by Convolutional Neural Networks [LBB98]. Convolutional Neural Networks are a model class of machine learning, which takes advantage of complex and deep structures of neuron connectivity and non-linearity. Its effectiveness and accuracy are well known for a wide range of applications including image classification [KSH12a] and face verification [TYR14]. These models are applied to each face sequentially and generate the outputs for the identify of the person and her facial expression as real number confidence values.

For the tasks of face recognition and smile classification, we trained our own models with the facial image data and the annotations made on the set in order to maximize the accuracy of classifiers. We primarily use the implementation and the model architecture introduced by an open source project, OpenFace [SKP15] and fine-tune the model with our own data. For recognition, we collected the facial images of 7 politicians in our work by web search while ensuring each person has at least 1,000 facial images. We also used generic negative images outside of these 7 politicians from public datasets to train the model by a discriminative loss. For smile classification, we took advantage of existing public facial attribute datasets, LFW [HRB07] and CelebA [LLW15].

We verified the accuracies of our models through human verification on a sample dataset. We first sampled 240 facial images for each politician from our news videos according to their smiling classification scores. We divided the whole score range into 12 evenly-spaced bins and randomly selected 20 faces from each group. These 240 faces are randomly permuted

and given to human annotators who were asked to select the actual present facial expression among 4 different categories: smile, frown, neither, smile and frown . In this work, we only focus on smile classification and so we simply merged the two groups of "smile" and "smile and frown" into the "smile" group.

4.5 Results

4.5.1 Which Candidates Received the Most Coverage?

Our first research question (R1) asks which candidate receives the most coverage.

Somewhat unsurprisingly to those who are currently living through the 2016 election, the answer is Donald Trump, almost always/everywhere, although there are some interesting patterns in attention paid to the other three candidates.

4.5.1.1 Twitter Mentions

Figure 4.2 shows our Twitter data for mentions of our "final four" presidential aspirants from August 2015-July 2016. As mentioned above, Trump dominates discourse on Twitter in every month of our year-long sample. In a majority of months, he actually is mentioned in more tweets than Clinton, Sanders, and Cruz *combined*.

So, while we have yet to discuss the tone of all of that attention, it seems clear that the answer to Research Question 1b ("Do some candidates receive disproportionate attention on social media?") is a resounding "Trump yes."

4.5.1.2 News Mentions

In this section, we attempt to answer Research Question 1a. "Do different media outlets pay attention to candidates similarly? (In particular, do more partisan outlets cover candidates in a systematically different way?)" We do so using two different measures of candidate attention in our Communication Studies News Archive programs: mentions and images.

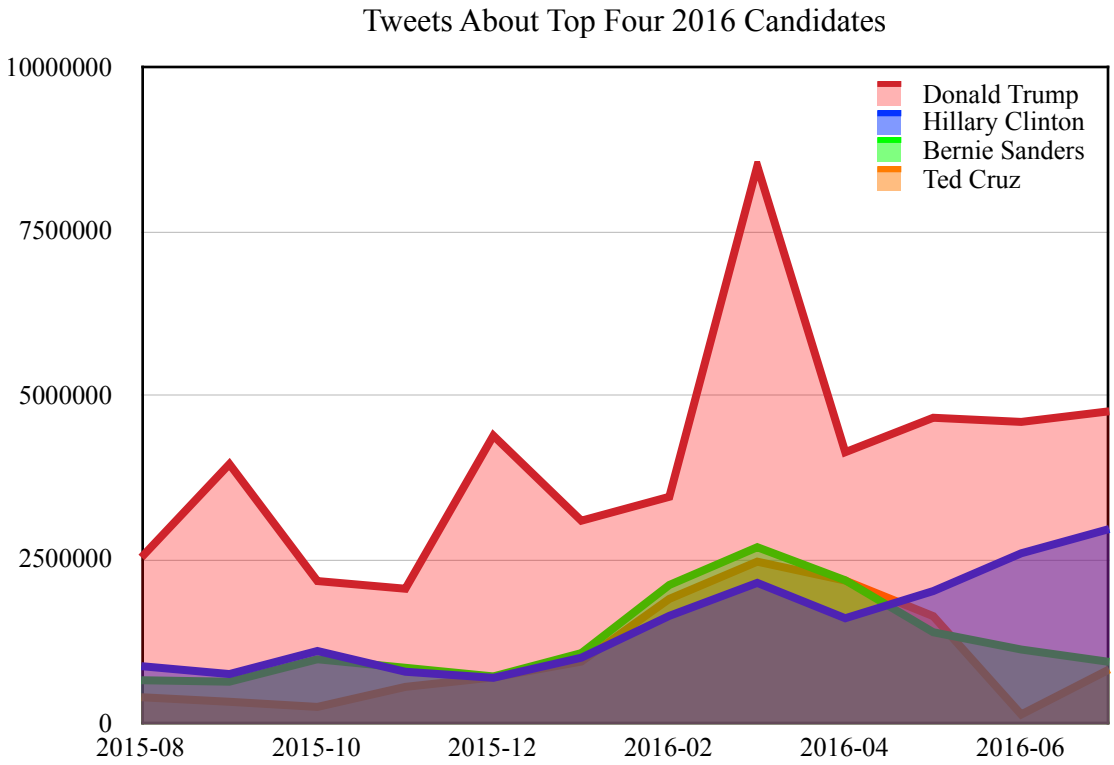


Figure 4.2: Volume of tweets about final four 2016 presidential candidates.

4.5.1.3 Aggregate Mentions

For our 2016 election programs, we searched closed captions for any mentions of each of our final four candidates. We then used pronoun matching based on coreference resolution implemented in the Beautiful Anaphora Resolution Toolkit [VMP08] to determine whether subsequent sentences also mention that specific candidate. We then use the timecodes in the closed captioning to calculate total mention time for sentences that referred to each candidate.

Table 4.2 continues to show the degree to which Trump has dominated public attention in the 2016 election cycle. Across the media outlets we examined, sentences discussing Trump accounted for more airtime than for any of the other three candidates. MSNBC discussed Trump the most; their corporate partner NBC discussed him the least. However, it should be noted that MSNBC discussed every candidate the most, and NBC discussed every candidate the least. Overall, the three cable news networks made up a larger proportion

Table 4.2: Percent of time devoted to discussing candidates (8/15-7/16), by outlet.

Source	Total Hours	Donald Trump	Hillary Clinton	Bernie Sanders	Ted Cruz
All	21660.3	5.7%	3.2%	1.6%	3.7%
CNN	5763.7	8.4%	4.5%	2.4%	5.1%
FOX	3930.6	6.2%	3.9%	1.4%	4.4%
MSNBC	4168.6	9.5%	5.2%	3.1%	6.5%
ABC	1021.3	2.4%	1.4%	0.6%	1.3%
CBS	994.2	2.6%	1.5%	0.7%	1.8%
NBC	1198.3	1.7%	1.1%	0.5%	1.0%
Local News	4556.4	3.6%	2.1%	1.2%	2.6%

of the hours in our sample, but also spent proportionately more of that time discussing the four candidates than their broadcast network peers. Somewhat surprisingly, local news shows spent proportionately more time discussing the candidates than their national broadcast network peers.

Figure 4.3 takes the same underlying data as Table 4.2, but displays it in a monthly series similar to the Twitter data in Figure 4.2.

Figure 4.3 shows a much more nuanced account of the ebb and flow of media attention prior to the general election. Although Trump clearly dominates attention overall, we find several points where Trump received equivalent or less coverage than other candidates. For Hillary Clinton, all outlets show her receiving more discussion than Trump in October 2015 (chiefly corresponding to her testimony in the Benghazi congressional hearings), and then later when she had dispatched the challenge from Bernie Sanders. For his part, Sanders goes from near total obscurity to a serious rival to Clinton for the media's attention, peaking with near-parity with her in early 2016 (especially on MSNBC and CNN). Somewhat unexpectedly, Cruz actually dominates both Sanders and Clinton and rivals Trump for attention on every subcategory of media from Nov. 2015-April 2016 before rapidly trailing off in May

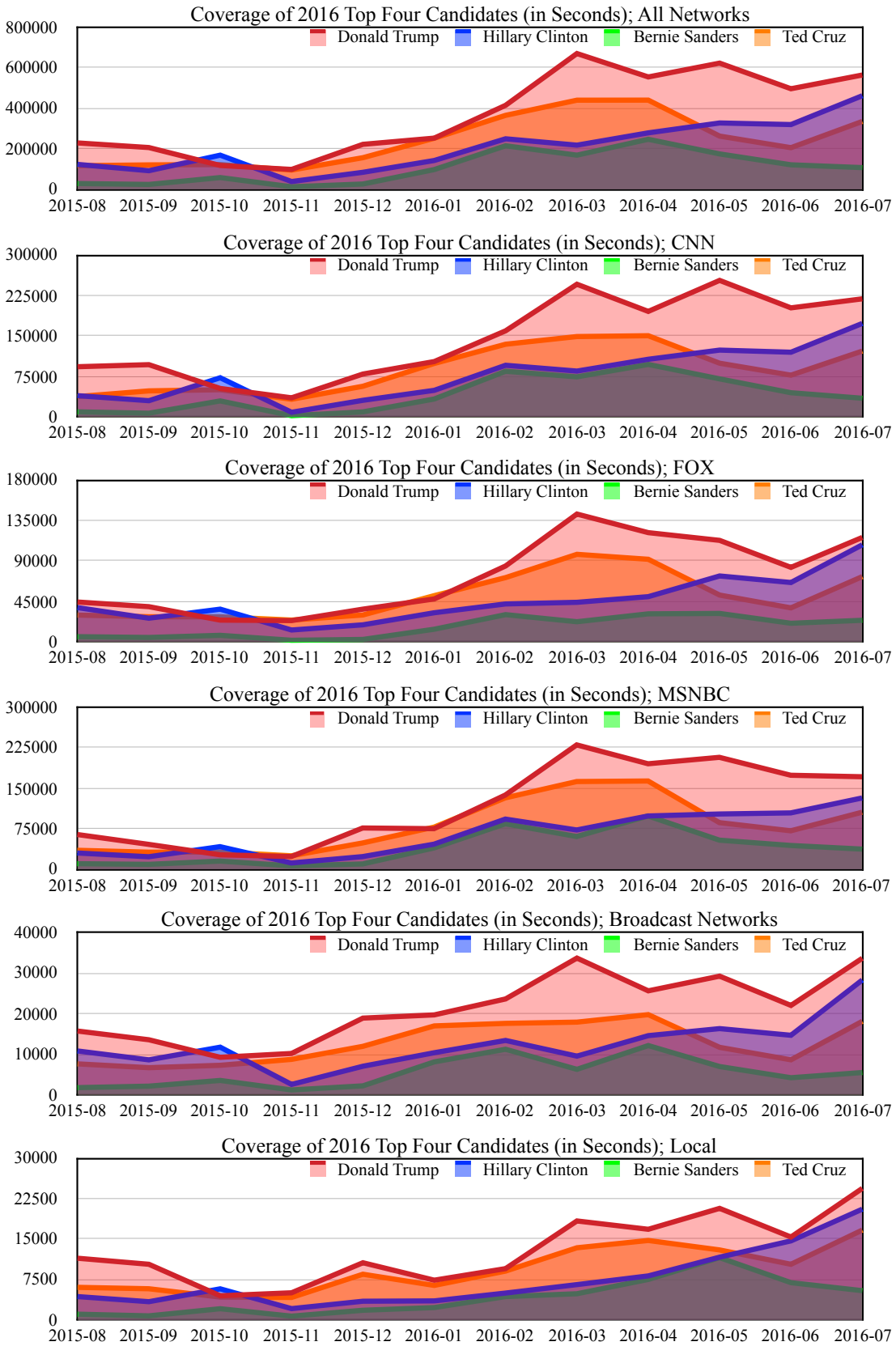


Figure 4.3: Monthly total seconds of candidate coverage, by outlet (8/15-7/16).

2016.

4.5.1.4 Candidate Still Photos

Next, we use an analysis of still images in video to test for trends in candidate attention. Because the use of a still photo of a candidate implies a higher level of attention—and planning—than simply mentioning a candidate's name in passing, the presence of a candidate's photo on-screen should be an especially strong signal of a news program's focus on the candidate at that moment during a show. Table 4.3 and Figure 4.4 summarize the data for this part of our analysis.

While Table 4.3 and Figure 4.4 show that Donald Trump has appeared in still photos proportionately and absolutely more often than Hillary Clinton, the historical data make that achievement somewhat less impressive. In particular, MSNBC apparently used still photos more often for every candidate in 2008 and 2012. In terms of candidates, Obama seems to have received similar amounts of attention in 2008 (and to a lesser degree in 2012), making Trump's level somewhat less distinctive.

4.5.2 What Topics Are being Discussed?

We turn next to Research Question 2a: "How do issues and topics covered in the campaign differ across candidates and outlets?"

4.5.2.1 Topic Trajectories

To answer this question, use analyzed the news programs in our Communication Studies News Archive to automatically track what stories appeared in the news, and how those news stories evolved over time. To do so, we detected news topics by clustering news stories every day, and then linked the detected topics to generate topic tracking trajectories using our joint image-text topic detection and tracking method introduced in chapter 3. We then track the top weekly topics (filterable by news organization) according to their total time, and also show the words, people and places associated with each topic cluster. Figure 4.5 shows the

Table 4.3: Uses of candidate images, by year, outlet, and candidate.

Number of Hours Searched per Network				
		CNN	FOX	MSNBC
2016		3402	2225	2681
2012		4283	2250	1707
2008		4148	2078	1402
Total,Number of Faces in Still Shots				
	Candidates	CNN	FOX	MSNBC
2016	Clinton	88821	41530	52634
	Trump	119753	100924	64268
2012	Obama	95321	61628	92041
	Romney	77876	37264	110236
2008	Obama	132178	66679	79675
	McCain	76928	47509	34790
Average,number of still shots per hour				
	Candidates	CNN	FOX	MSNBC
2016	Clinton	26.1095	18.6691	19.6302
	Trump	35.2022	45.3686	23.9692
2012	Obama	22.2542	27.3941	53.9302
	Romney	18.1814	16.5641	64.5913
2008	Obama	31.8633	32.0942	56.8459
	McCain	18.5445	22.8672	24.8217

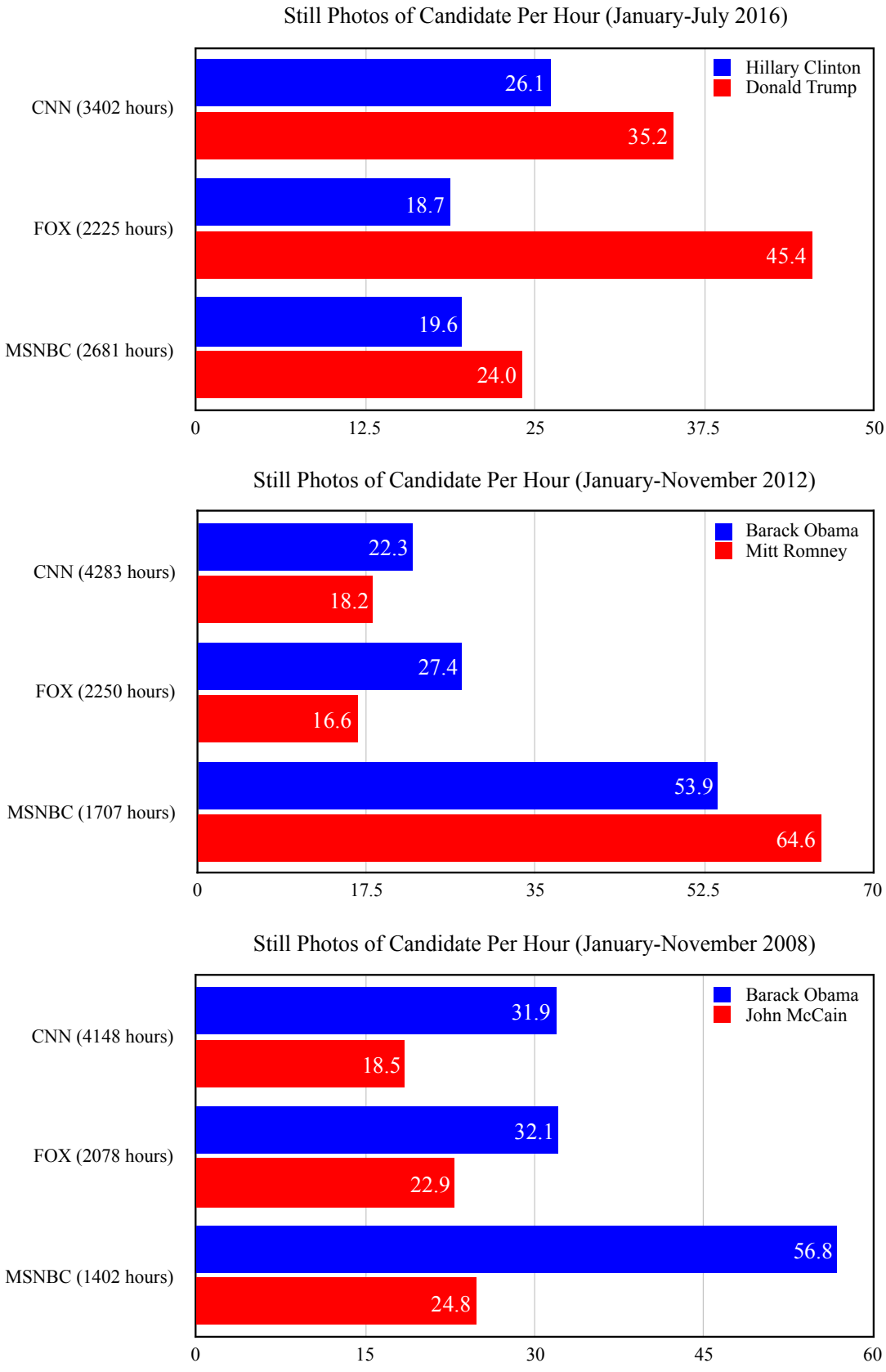


Figure 4.4: Use of candidate still images per hour, by outlet and election year.

online interface for our Topic Tracking tool.

In addition to the weekly summaries, we chart the daily stories and their relationship with prior and subsequent days' news. In these dynamic charts, each circle represents one topic on a given day. Topics in the same topic trajectory are given the same color and connected with lines to show their relationship. The greater the thickness of the line, the more connected the top nodes are.

Although scarcity of time and ink preclude us from completing an in-depth analysis of these data for this work, a brief search of the daily news nodes reveals patterns consistent with our prior findings. In the topic nodes contained in the 392 days from August 1, 2015 to August 27, 2016, Donald Trump is a top-level match on 488 of them. Hillary Clinton is a top-level match in 340. In contrast, Bernie Sanders is a top-level match in 160 nodes, and Ted Cruz is a top-level match in only 27 nodes (suggesting that much of his coverage occurred in topics where another figure was the main focus of the story).

4.5.2.2 Outlet Differences in Framing Coverage of Candidates

Next, we turn to research question 1a, "Do different media outlets pay attention to candidates similarly? (In particular, do more partisan outlets cover candidates in a systematically different way)?" Again, rather than presenting a static analysis here, we have used the Viz2016 site to present our data in an interactive and (hopefully) compelling fashion. By going to `http://viz2016.com/agenda-setting/`, one can select three different news organizations and a presidential candidate, and then see how they differed in their coverage of that candidate.

In these charts, we ranked words by their total occurrence associated with the three vertices of the triangle (news organizations or candidates, depending on the viewer's selection). Proximity to the vertices is an analog to how closely that vertex is associated with the word: If a bubble appears at the center of the triangle, it signifies that it occurs in roughly equal proportions; if it's close to a vertex, it is much more strongly associated with that organization or person than to the other two. The bubble's size conveys the normalized occurrence

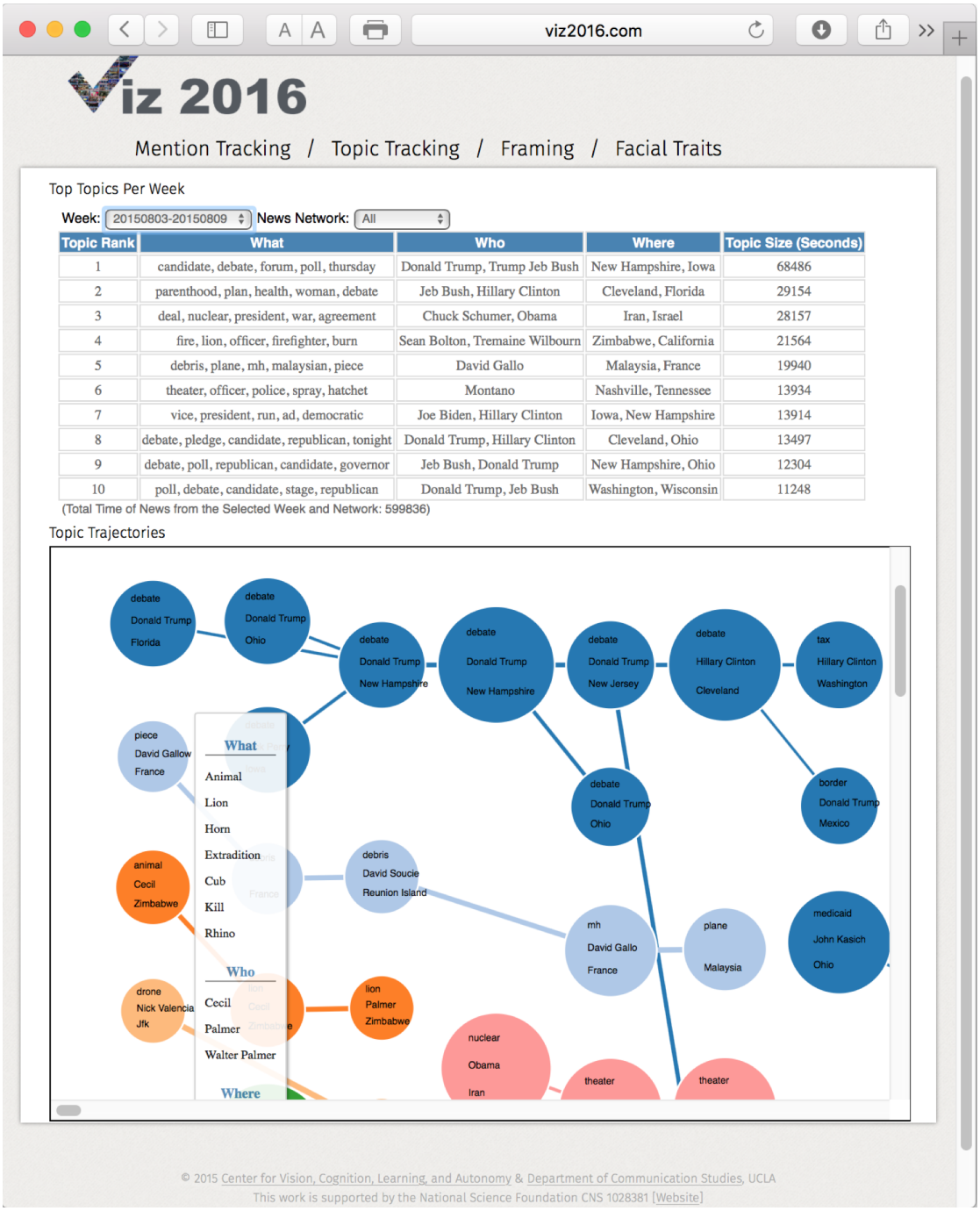


Figure 4.5: Viz2016.com topic tracking.

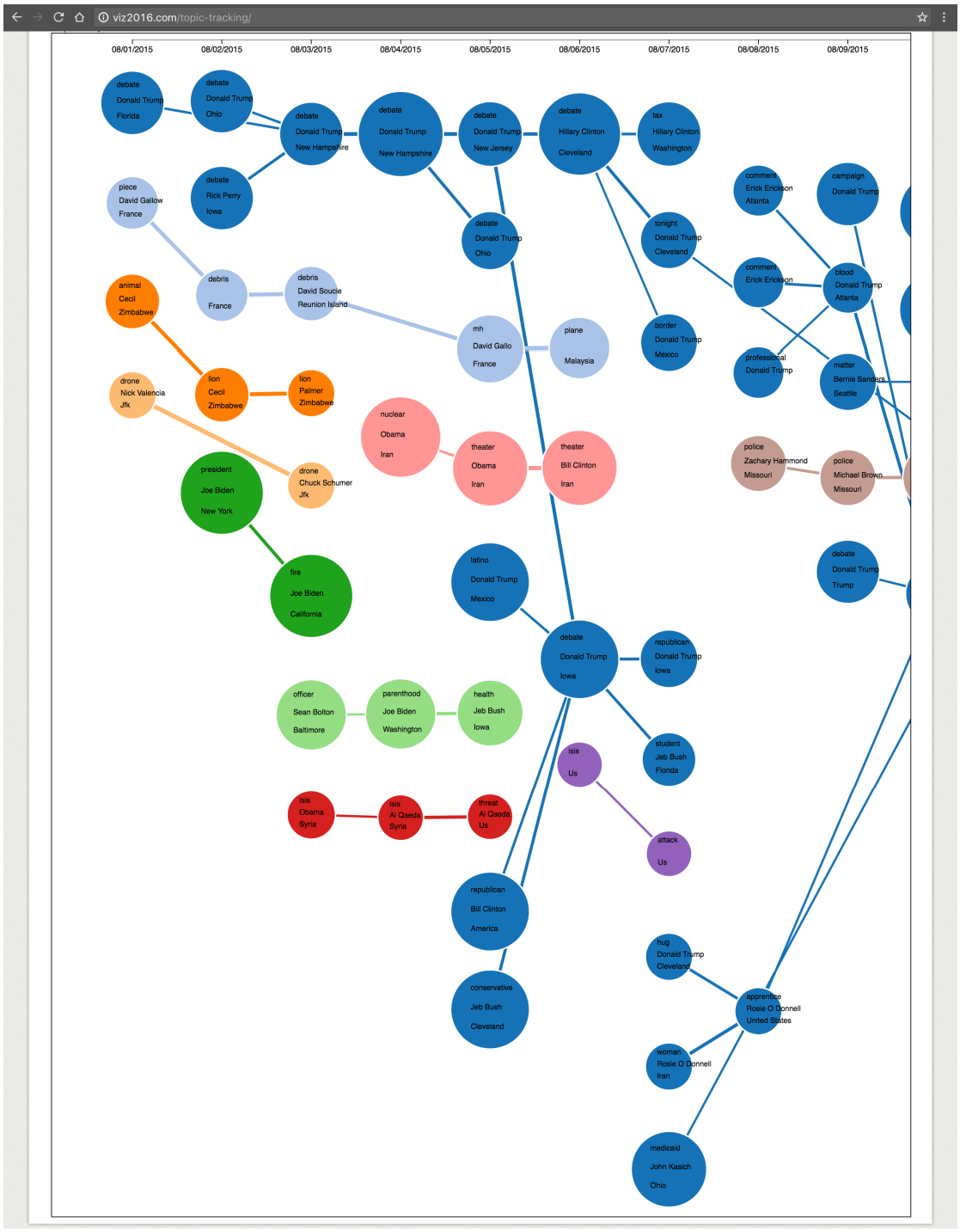


Figure 4.6: Viz2016.com topic trajectory, detailed daily view.

of that word. Colors represent the issue area of the word (note that the same word might show up in two different-colored bubbles if it is associated with different issues).

4.5.3 How Favorably Are the Candidates Being Portrayed?

For our final research question (R2b), we move beyond considering who is in the news, or the news topics with which they are associated. This section tackles a more challenging question: Do some candidates receive more favorable coverage than others?

As was discussed in the literature review, researchers have had a difficult time assessing whether news content favors one political actor or side over another. An important obstacle to this pursuit has been the subjectivity of those consuming and coding the news content. As Campbell and Stanley [CS71] observed, "problems of subjectivity—reliability, validity, and weighting—point to the same moral: even the most ingenious content analysis has difficulty warding off methodological critiques that threaten to impugn its results". Perhaps the most famous example of such subjectivity is the hostile media phenomenon, in which identical stories were perceived as having diametrically opposed biases depending on who viewed them [VRL85]. Particularly in the case of media bias, differing perceptions of story content might arise from differences in the content of news, or might instead reflect prior attitudes regarding the bias of that source. Because of well-documented cognitive biases—such as confirmation and disconfirmation biases, selective perception, anchoring, attention bias, the clustering illusion, and selective perception, among others—partisans might sincerely perceive news as being biased against their preferred stance, even when it is actually unbiased (see [HC54] and [DBH98]).

To answer our question while avoiding problems of subjectivity, we take a closer look at two of our data sources: our Twitter candidate mentions from the 2016 election dataset, and our candidate still photo data from 2008, 2012, and 2016.

4.5.3.1 Twitter Tone

To determine the sentiment of our candidate-related tweets, we used the VADER sentiment analysis tools [HG14]. Using this tool for tweets has several advantages over using it for longer-form news stories, in which the target for the sentiment in a sentence is often ambiguous [BS09]. In particular, the inherent content limits on a tweet mean that there is a clearer match between the sentiment being expressed and its intended target. If there is negative sentiment expressed in a news story that also mentions Donald Trump, it is quite possible that the sentiment might be expressed by Trump, or possibly from one of his supporters. In a tweet, there are fewer opportunities to layer in complexity, so the candidate mentioned in a given tweet is also likely the target of the positive or negative sentiment it contains.

Figure 4.7 shows the volume and sentiment of tweets about our "final four" candidates in the 2016 election data: Trump, Clinton, Sanders, and Cruz. Because we have already discussed Twitter volume earlier, we will focus here on sentiment.

Figure 4.7 shows that—although Trump is sometimes called "the Twitter candidate"—the tone of the tweets that mention him is not that much more positive than those mentioning Clinton or Cruz. Indeed, looking at Figure 4.7, the candidate who clearly seems to be suffering the least at the hands of the Twitterati is Bernie Sanders, whose mentions are more positive than negative every month of the election period.

4.5.3.2 News Visuals (Smiles)

Finally, we turn to our most challenging computational task: assessing the favorability of still images drawn from millions of frames of television news. While we are still developing this analysis, we decided to start with a relatively straightforward indicator of image favorability: smiling. Joo et al. [JLS14] found that a smiling face of a politician in a photograph is highly correlated with a number of positive perceptual dimensions such as favorability, competence, or happiness. In addition, the human face is a very well studied topic in computer vision and machine learning and automated classification of facial attributes is much more reliable than detection of other complex visual cues.

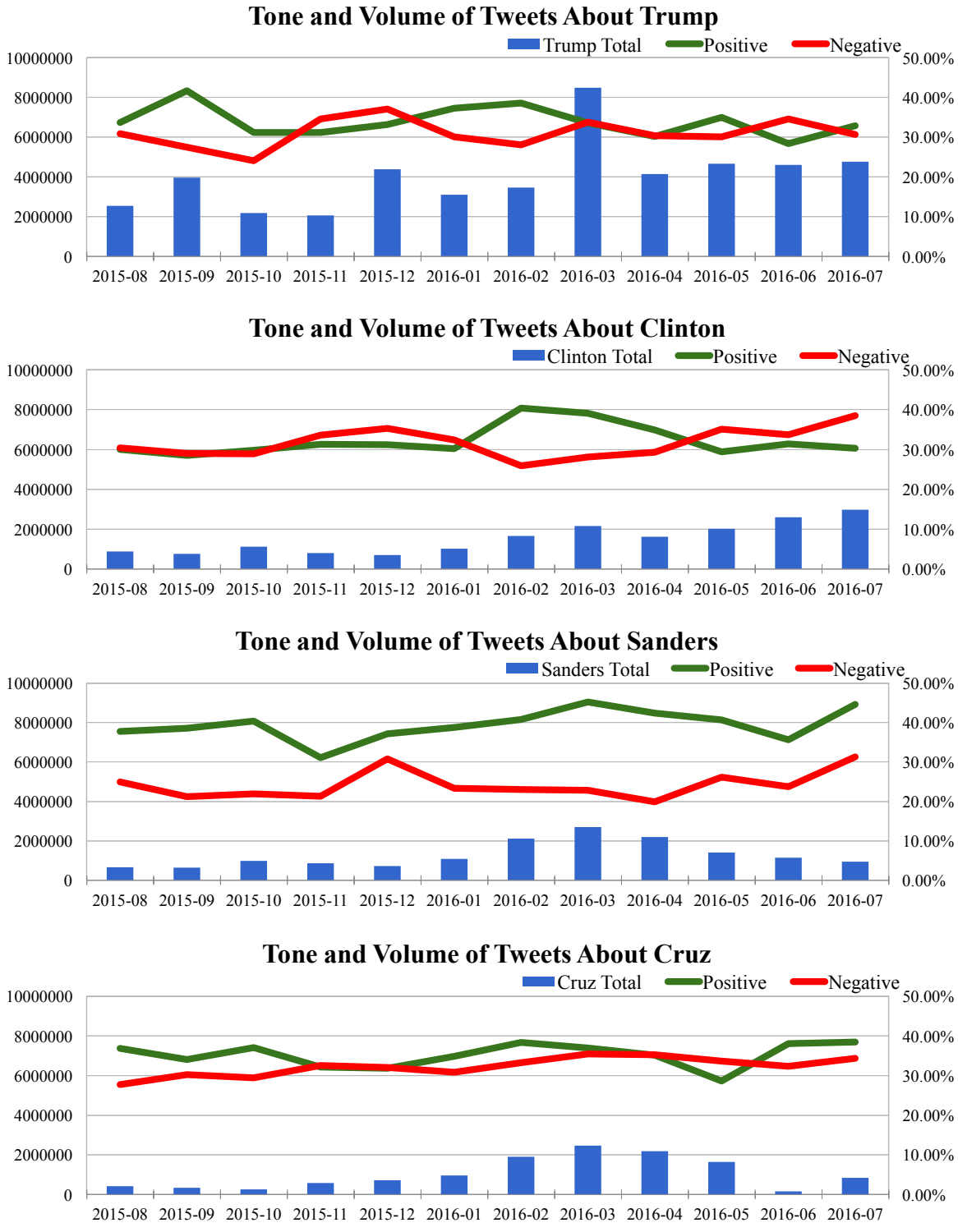


Figure 4.7: Volume and tone of tweets about Trump, Clinton, Sanders and Cruz.

Having a common and fair baseline score for smile classification across different individuals is a difficult task. This is because our perception of a facial expression may be affected by the inherent facial appearance of each individual [LKP08, JSZ15]. Furthermore, the public datasets used to train our model for smile classification may also introduce additional bias. e.g., correlation between gender and expression, which may result in unintended and unbalanced prediction score distributions between subpopulations and across candidates in our analysis. To alleviate this problem, we leverage the annotations obtained from the human coders and probabilistically count smiles by sampling such that the faces of candidates showing the equivalent degree of smiling be classified with an equal probability as a post-hoc calibration. However, our annotation set is relatively small and the comparisons of the raw smile scores between candidates must be noted with a caution. In contrast, it is straightforward to interpret the scores of the same candidate across different networks because all the potential dataset bias would equally apply.

We examine the visual sentiment of our news data by examining the proportion of candidate images that are favorable to the candidate. In this case, we operationalize favorability by assuming smiling images are more favorable than neutral or frowning images. Figure 4.8 presents the result of that analysis, showing how the percentage of smiling still images varies across news outlets for each recent presidential candidate. Since we use automated classification which may not be perfect, we estimate the population mean and confidence intervals by bootstrapping 2,000 faces and sampling their unknown, true expressions (smile or non-smile) according to the empirical distribution obtained by annotation (repeating 1,000 times).

As was noted above, caution should be exercised when comparing results across individuals; however, we have greater confidence when comparing the results for the same candidate across media outlets. Beginning with Hillary Clinton, we find exceptionally high levels of smiling in her still photos across outlets. Somewhat surprisingly, Fox News (generally regarded as conservative) actually presents more images of her smiling than MSNBC (second most) or CNN.

For Donald Trump, the results are much more in line with conventional expectations,

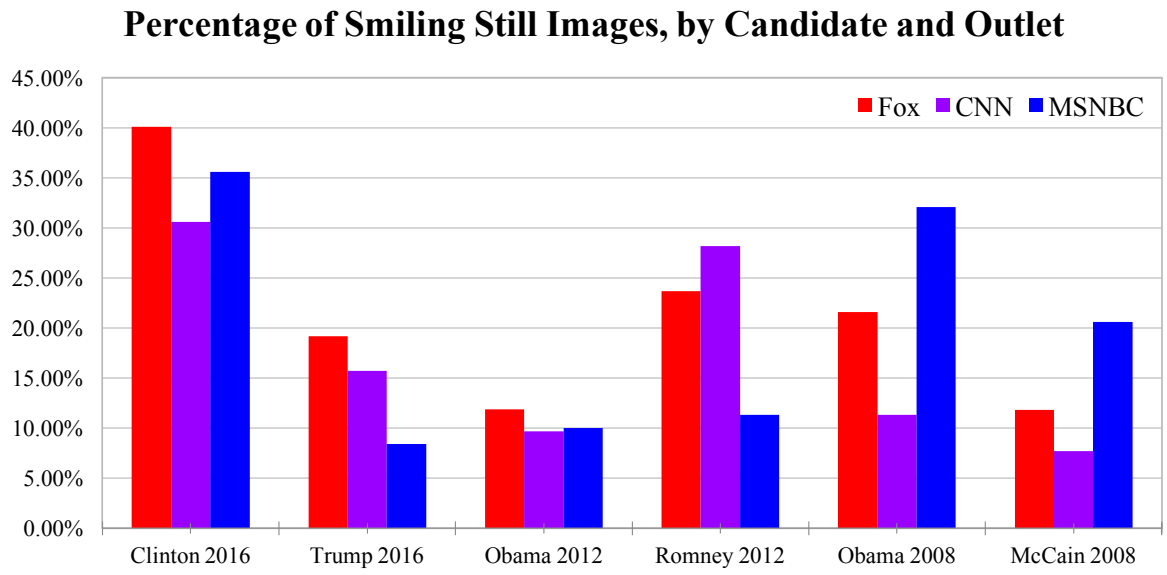


Figure 4.8: Percent of still images containing smiles, by candidate and outlet.

with Fox News showing Trump smiling about three times as often as does MSNBC (CNN falls in between, but is closer to Fox than MSNBC).

Turning to 2012, the uniformly low amounts of smiling for Barack Obama are surprising, as is Fox's lead in likelihood to use smiling pictures. Again, the Republican candidate falls much more in line with expectations, with Fox showing Romney smiling about twice as frequently as MSNBC (CNN actually shows more smiling Romney than Fox, which is also surprising).

Finally, our charts from 2008 show MSNBC selecting smiling still photos around a third of the time. Fox used smiling Obama pictures around one fifth of the time; somewhat surprisingly, CNN used smiling Obama pictures only about 11% of the time. For McCain, the patterns are similar, albeit at a lower level of smiling. MSNBC showed him smiling about a fifth of the time; Fox was lower at around 12% of the time, while CNN reached the lowest percentage we've seen thus far at only 8% smiling.

CHAPTER 5

Conclusions

In this dissertation we present methods for automatic and systematic analysis of the large and continuously updated news collection.

Firstly, we present a joint image-text news topic detection and tracking method which can summarize and organize the news collection into events. We propose a Multimodal Topic And-Or Graph (MT-AOG) for structured topic representation. The MT-AOG embeds a context sensitive grammar that can model the topic's hierarchical decomposition of the text part, the image part, and their subcomponents about related people, locations, faces, objects, and what happened. The contextual relationships between elements in the topic hierarchy are also modeled in the MT-AOG. For topic detection, with the MT-AOG, we cluster news stories into coherent groups about the same events and obtain the MT-AOG model parameters for topics at the same time. We pose this as a graph partitioning problem and solve it using the Swendsen-Wang Cuts cluster sampling method, which can efficiently sample the space defined by a Bayesian posterior probability to get the optimal solution. In topic tracking, the topics detected in different time periods are linked to form topic trajectories. In this way, we can not only deal with the continuous updates of news streams, but also obtain both short-time topic summaries and long-time topic trajectories to show how topics evolve over time. To link topics, we measure the topic similarities by considering both the textual and visual content similarities and the topics' temporal distances. The qualitative experimental results show that our method can explicitly describe the textual and visual data in news videos and produce meaningful topic trajectories. The quantitative evaluation results show that our method outperforms existing methods on Reuters-21578 and our novel dataset, UCLA Broadcast News Dataset.

We further expand our topic detection and tracking work to concrete media analysis for social and political science research. We conduct several large-scale quantitative analyses of the campaign communication to visualize the US presidential elections. We examine what topics are discussed about different candidates in news outlets based on our proposed topic detection and tracking methods. We investigate which candidates are mentioned more in traditional news media and social media by analyzing the TV news and Twitter data. We also measure how different news outlets favored different candidates visually. We have attempted to better understand presidential election communication through the application of innovative machine-learning techniques to a unique, big-data collection of news texts and videos.

References

[ACC15] André Araujo, Jason Chaves, David Chen, Roland Angst, and Bernd Girod. "Stanford I2V: a news video dataset for query-by-image experiments." In *MMSys*, 2015.

[ACD98] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. "Topic Detection and Tracking Pilot Study: Final Report." In *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pp. 194–218, 1998.

[AHP06] T. Ahonen, A. Hadid, and M. Pietikainen. "Face Description with Local Binary Patterns: Application to Face Recognition." *TPAMI*, **28**(12):2037–2041, 2006.

[All02] James Allan. *Topic Detection and Tracking: Event-based Information Organization*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[AX08] Amr Ahmed and Eric P Xing. "Dynamic Non-Parametric Mixture Models and the Recurrent Chinese Restaurant Process: with Applications to Evolutionary Clustering." In *SDM*, pp. 219–230, 2008.

[AZ12] CharuC. Aggarwal and ChengXiang Zhai. "A Survey of Text Clustering Algorithms." In *Mining Text Data*, pp. 77–128. Springer US, 2012.

[BB05] A.W. Barrett and L.W. Barrington. "Bias in newspaper photograph selection." *Political Research Quarterly*, **58**:609–618, 2005.

[BB09] Jordan Boyd-graber and David Blei. "Syntactic Topic Models." In *NIPS*, 2009.

[BC09] S. Banning and R. Coleman. "Louder than words: A content analysis of presidential candidates' televised nonverbal communication." *Visual Communication Quarterly*, **16**:4–7, 2009.

[BH14] Cosmin Adrian Bejan and Sanda Harabagiu. "Unsupervised event coreference resolution." *Computational Linguistics*, **40**(2):311–347, 2014.

[BJ03] David M. Blei and Michael I. Jordan. "Modeling Annotated Data." In *SIGIR*, pp. 127–134, 2003.

[BJC04] I.V. Blair, C. M. Judd, and K.M. Chapleau. "The influence of afrocentric facial features in criminal sentencing." *Psychological Science*, **15**(10):674–679, 2004.

[BL06] David M. Blei and John D. Lafferty. "Dynamic Topic Models." In *ICML*, 2006.

[Ble12] David M. Blei. "Probabilistic Topic Models." *Commun. ACM*, **55**(4):77–84, 2012.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." *J. Mach. Learn. Res.*, **3**:993–1022, 2003.

[BP07] A.W. Barrett and J.S. Peake. "When the president comes to town." *American Politics Research*, **35**:3–31, 2007.

[BS09] A. Balahur and R. Steinberger. "Rethinking Sentiment Analysis in the News: from Theory to Practice and back." In *Proceedings of the first workshop on Opinion Mining and Sentiment Analysis*, pp. 1–9, 2009.

[BZ05] Adrian Barbu and Song-Chun Zhu. "Generalizing Swendsen-Wang to Sampling Arbitrary Posterior Probabilities." *TPAMI*, **27**(8):1239–1253, 2005.

[CB09] Jonathan Chang and David M Blei. "Relational Topic Models for Document Networks." In *AIStats*, volume 9, pp. 81–88, 2009.

[CHH11] Deng Cai, Xiaofei He, and Jiawei Han. "Locally Consistent Concept Factorization for Document Clustering." *TKDE*, **23**(6):902–913, 2011.

[Cro73] T. Crouse. *The boys on the bus*. Ballantine Books, New York, 1973.

[CS71] Donald T Campbell and Julian C Stanley. *Experimental and Quasi-Exprimental Designs for Research*, volume 4. Rand McNally, 1971.

[CSH15] Tao Chen, Hany M SalahEldeen, Xiangnan He, Min-Yen Kan, and Dongyuan Lu. "VELDA: Relating an Image Tweet's Text and Images." In *AAAI*, pp. 30–36, 2015.

[CW07] T.J.A. Covert and P.C. Wasburn. "Measuring media bias: A content analysis of time and newsweek coverage of domestic social issues." *Social Science Quarterly*, **88**(690):1975–2000, 2007.

[CYL15] Hongyun Cai, Yang Yang, Xuefei Li, and Zi Huang. "What are Popular: Exploring Twitter Features for Event Detection, Tracking and Visualization." In *MM*, pp. 89–98, 2015.

[CZL14] L. Chu, Y. Zhang, G. Li, S. Wang, W. Zhang, and Q. Huang. "Effective multimodality fusion framework for cross-media topic detection." *TCSVT*, **PP**(99):1–1, 2014.

[DBH98] Russell J Dalton, Paul A Beck, and Robert Huckfeldt. "Partisan cues and the media: Information flows in the 1992 presidential election." *American Political Science Review*, **92**(01):111–126, 1998.

[DVC13] M. Daneshi, P. Vajda, D.M. Chen, S.S. Tsai, M.C. Yu, A.F. Araujo, H. Chen, and B. Girod. "Eigennews: Generating and delivering personalized news video." In *ICMEW*, 2013.

[FGM05] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling." In *ACL*, pp. 363–370, 2005.

[GB09] M.E. Grabe and E.P. Bucy. *Image Bite Politics: News and the Visual Framing of Elections*. Oxford University Press, Oxford, UK, 2009.

[GCM12] Aude Giraudel, Matthieu Carré, Valérie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard. "The REPERE Corpus: a multimodal corpus for person recognition." In *LREC*, 2012.

[GG84] Stuart Geman and Donald Geman. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *TPAMI*, **6**(6):721–741, 1984.

[Gro13] T. Groeling. "Media Bias by the Numbers: Challenges and Opportunities in the Empirical Study of Partisan News." *Annual Review of Political Science*, **16**:129–151, 2013.

[GS04] T. L. Griffiths and M. Steyvers. "Finding scientific topics." *PNAS*, **101**:5228–5235, 2004.

[GS10] M. Gentzkow and J.M. Shapiro. "What drives media slant? Evidence from U.S. daily newspapers." *Econometrica*, **78**:35–71, 2010.

[GSB05] Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. "Integrating topics and syntax." In *NIPS*, 2005.

[Ham06] J.T. Hamilton. *All the news that's fit to sell: How the market transforms information into news*. Princeton University Press, New Jersey, 2006.

[Har96] Geoff Hart. "The Five W's: An Old Tool for the New Task of Task Analysis." *Technical Communication*, **43**(2):139–145, 1996.

[HC54] Albert H Hastorf and Hadley Cantril. "They saw a game; a case study." *The Journal of Abnormal and Social Psychology*, **49**(1):129, 1954.

[HC06] W.H. Hsu and Shih-Fu Chang. "Topic Tracking Across Broadcast News Videos with Visual Duplicates and Semantic Concepts." In *ICIP*, 2006.

[HG14] C.J. Hutto and E.E. Gilbert. "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text." In *Eighth International Conference on Weblogs and Social Media*, 2014.

[HGH12] E. Hehman, E.C. Graber, L.H. Hoffman, and S.L. Gaertner. "Warmth and competence: A content analysis of photographs depicting american presidents." *Psychology of Popular Media Culture*, **1**:46–52, 2012.

[HLJ81] Julian Harriss, Kelly Leiter, and Stanley P Johnson. *The Complete Reporter: Fundamentals of News Gathering, Writing, and Editing, Complete with Exercises*. MacMillan Publishing Company, 1981.

[Hof76] C.R. Hofstetter. *Bias in the news: Network television coverage of the 1972 election campaign*. Ohio State University Press, Ohio, 1976.

[Hof99] Thomas Hofmann. "Probabilistic Latent Semantic Indexing." In *SIGIR*, pp. 50–57, 1999.

[HQ08] D.E. Ho and K.M. Quinn. "Measuring explicit political positions of media." *Q. J. Polit. Sci.*, **3**:353–377, 2008.

[HRB07] G.B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. "Labeled faces in the wild: A database for studying face recognition in unconstrained environments." *Technical Report, University of Massachusetts, Amherst*, **1**(2):7–49, 2007.

[JG11] Heng Ji and Ralph Grishman. "Knowledge base population: Successful approaches and challenges." In *ACL*, pp. 1148–1158, 2011.

[JGD10] Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. "Overview of the TAC 2010 knowledge base population track." In *TAC*, 2010.

[JLE13] Brendan Jou, Hongzhi Li, Joseph G. Ellis, Daniel Morozoff-Abegauz, and Shih-Fu Chang. "Structured Exploration of Who, What, when, and Where in Heterogeneous Multimedia News Sources." In *MM*, 2013.

[JLS14] Jungseock Joo, Weixin Li, F.F. Steen, and Song-Chun Zhu. "Visual Persuasion: Inferring Communicative Intents of Images." In *CVPR*, pp. 216–223, 2014.

[JSD11] Yangqing Jia, Mathieu Salzmann, and Trevor Darrell. "Learning cross-modality similarity for multinomial data." In *ICCV*, pp. 2407–2414, 2011.

[JSD14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. "Caffe: Convolutional Architecture for Fast Feature Embedding." In *MM*, pp. 675–678, 2014.

[JSZ15] Jungseock Joo, F.F. Steen, and Song-Chun Zhu. "Automated Facial Trait Judgment and Election Outcome Prediction: Social Dimensions of Face." In *ICCV*, pp. 3712–3720, 2015.

[JWZ12] Jungseock Joo, Shuo Wang, and Song-Chun Zhu. "Hierarchical Organization by And-Or Tree." In Johan Wagemans, editor, *Book chapter in Handbook of Perceptual Organization*. Springer, 2012.

[KO11] Dongwoo Kim and Alice Oh. "Topic Chains for Understanding a News Corpus." In *International Conference on Computational Linguistics and Intelligent Text Processing*, 2011.

[KSH12a] A. Krizhevsky, I. Sutskever, and G.E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[KSH12b] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." In *NIPS*, 2012.

[LBB98] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE*, **86**(11):2278–2324, 1998.

[LJE13] Hongzhi Li, Brendan Jou, Joseph G. Ellis, Daniel Morozoff, and Shih-Fu Chang. "News Rover: Exploring Topical Structures and Serendipity in Heterogeneous Multimedia News." In *MM*, 2013.

[LJQ17] Weixin Li, Jungseock Joo, Hang Qi, and Song-Chun Zhu. "Joint Image-Text News Topic Detection and Tracking by Multimodal Topic And-Or Graph." *TMM*, **19**(2):367–381, 2017.

[LKP08] E. Lee, J.I. Kang, I.H. Park, J.J. Kim, and S.K. An. "Is a neutral face really evaluated as being emotionally neutral?" *Psychiatry Research*, **157**(1):77–85, 2008.

[LLW15] Z. Liu, P. Luo, X. Wang, and X. Tang. "Deep learning face attributes in the wild." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

[LWL05] Zhiwei Li, Bin Wang, Mingjing Li, and Wei-Ying Ma. "A Probabilistic Model for Retrospective News Event Detection." In *SIGIR*, pp. 106–113, 2005.

[MAS04] Juha Makkonen, Helena Ahonen-Myka, and Marko Salmenkivi. "Simple Semantics in Topic Detection and Tracking." *Inf. Retr.*, **7**(3-4):347–368, 2004.

[MT13] Saif M. Mohammad and Peter D. Turney. "Crowdsourcing a Word-Emotion Association Lexicon." *Computational Intelligence*, **29**(3):436–465, 2013.

[MZ05] Qiaozhu Mei and ChengXiang Zhai. "Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining." In *SIGKDD*, 2005.

[NCS06] David Newman, Chaitanya Chemudugunta, and Padhraic Smyth. "Statistical Entity-topic Models." In *SIGKDD*, pp. 680–686, 2006.

[NHG14] Zhenxing Niu, Gang Hua, Xinbo Gao, and Qi Tian. "Semi-supervised relational topic model for weakly annotated image recognition in social media." In *CVPR*, pp. 4233–4240, 2014.

[PAN10] D. Putthividhy, H.T. Attias, and S.S. Nagarajan. "Topic regression multi-modal Latent Dirichlet Allocation for image annotation." In *CVPR*, pp. 3408–3415, 2010.

[PTZ15] Maria Pavlovskaia, Kewei Tu, and Song-Chun Zhu. "Mapping the Energy Landscape of Non-convex Optimization Problems." In *EMMCVPR*, pp. 421–435, 2015.

[QZX16] Shengsheng Qian, Tianzhu Zhang, Changsheng Xu, and Jie Shao. "Multi-Modal Event Topic Model for Social Event Analysis." *TMM*, **18**(2):233–246, 2016.

[RDS15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge." *IJCV*, pp. 1–42, 2015.

[RMT11] M. Rojas, D. Masip, A. Todorov, and J. Vitria. "Automatic prediction of facial trait judgments: Appearance vs. structural models." *PloS one*, **6**(8), 2011.

[Sch06] A.J. Schiffer. "Assessing partisan bias in political news: The case(s) of local senate election coverage." *Political Communication*, **23**:23–39, 2006.

[SKK00] Michael Steinbach, George Karypis, and Vipin Kumar. "A comparison of document clustering techniques." In *In KDD Workshop on Text Mining*, 2000.

[SKP15] F. Schroff, D. Kalenichenko, and J. Philbin. "Facenet: A unified embedding for face recognition and clustering." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

[SM00] Jianbo Shi and Jitendra Malik. "Normalized Cuts and Image Segmentation." *TPAMI*, **22**(8):888–905, 2000.

[SOK06] Alan F. Smeaton, Paul Over, and Wessel Kraaij. "Evaluation campaigns and TRECVid." In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pp. 321–330. ACM Press, 2006.

[Sto88] J.G. Stovall. "Coverage of 1984 presidential campaign." *Journalism and Mass Communication Quarterly*, **65**:443–449, 1988.

[SW87] Robert H. Swendsen and Jian-Sheng Wang. "Nonuniversal critical dynamics in Monte Carlo simulations." *Phys. Rev. Lett.*, **58**:86–88, 1987.

[TJB06] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. "Hierarchical Dirichlet processes." *JASA*, **101**(476):1566–1581, 2006.

[TMG05] A. Todorov, A.N. Mandisodza, A. Goren, and C.C. Hall. "Inferences of competence from faces predict election outcomes." *Science*, **308**(5728):1623–1626, 2005.

[TYR14] Y. Taigman, M. Yang, M.A. Ranzato, and L. Wolf. "Deepface: Closing the gap to human-level performance in face verification." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701–1708, 2014.

[USG13] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. "Selective Search for Object Recognition." *IJCV*, **104**(2):154–171, 2013.

[VJ01] Paul Viola and Michael Jones. "Rapid object detection using a boosted cascade of simple features." In *CVPR*, 2001.

[VMP08] Y. Versley, A. Moschitti, M. Poesio, and X. Yang. "Coreference Systems based on Kernel Methods." In *Proceedings of the 22nd International Conference on Computational Linguistics*, 2008.

[VRL85] Robert P Vallone, Lee Ross, and Mark R Lepper. "The hostile media phenomenon: biased perception and perceptions of media bias in coverage of the Beirut massacre." *Journal of personality and social psychology*, **49**(3):577, 1985.

[VSY14] R.J. Vernon, C.A. Sutherland, A.W. Young, and T. Hartley. "Modeling rst impressions from highly variable facial images." *Proceedings of the National Academy of Sciences*, **111**(32):E3353–E3361, 2014.

[Wal06] Hanna M. Wallach. "Topic Modeling: Beyond Bag-of-words." In *ICML*, 2006.

[WB11] Chong Wang and David M. Blei. "Collaborative Topic Modeling for Recommending Scientific Articles." In *SIGKDD*, pp. 448–456, 2011.

[WD98] Paul Waldman and James Devitt. "Newspaper photographs and the 1996 presidential election: The question of bias." *Journalism & Mass Communication Quarterly*, **75**(2):302–311, 1998.

[WNH08] Xiao Wu, Chong-Wah Ngo, and A.G. Hauptmann. "Multimodal News Story Clustering With Pairwise Visual Near-Duplicate Constraint." *TMM*, **10**(2):188–199, 2008.

[WNL06] Xiao Wu, Chong-Wah Ngo, and Qing Li. "Threading and autodocumenting news videos: a promising solution to rapidly browse news topics." *IEEE Signal Processing Magazine*, **23**(2):59–68, 2006.

[XLG03] Wei Xu, Xin Liu, and Yihong Gong. "Document Clustering Based on Non-negative Matrix Factorization." In *SIGIR*, pp. 267–273, 2003.

[XX13] Pengtao Xie and Eric P. Xing. "Integrating Document Clustering and Topic Modeling." In *UAI*, pp. 694–703, 2013.

[YCG99] J. P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. Van Mulbregt. "Topic Tracking in a News Stream." In *In Proceedings of DARPA Broadcast News Workshop*, pp. 133–136, 1999.

[YNL14] Benjamin Z Yao, Bruce X Nie, Zicheng Liu, and Song-Chun Zhu. "Animated pose templates for modeling and detecting human actions." *TPAMI*, **36**(3):436–452, 2014.

[YVC13] Matt C. Yu, Peter Vajda, David M. Chen, Sam S. Tsai, Maryam Daneshi, Andre F. Araujo, Huizhong Chen, and Bernd Girod. "EigenNews: A personalized news video delivery platform." In *MM*, 2013.

[ZFM03] L. A. Zebrowitz, J. M. Fellous, A. Mignault, and C. Andreoletti. "Trait impressions as overgeneralized responses to adaptively significant facial qualities: Evidence from connectionist modeling." *Personality and social psychology review*, **7**(3):194–215, 2003.

[ZL12] Youjie Zhou and Jiebo Luo. "Geo-location Inference on News Articles via Multimodal pLSA." In *MM*, pp. 741–744, 2012.

[ZLJ15] Tongtao Zhang, Hongzhi Li, Heng Ji, and Shih-Fu Chang. "Cross-document event coreference resolution based on cross-media features." In *EMNLP*, 2015.

[ZLY13] J. Zhu, J. Luo, Q. You, and J. R. Smith. "Towards understanding the effectiveness of election related images in social media." In *IEEE International Conference on Data Mining Workshops*, pp. 421–425, 2013.

[ZM06] Song-Chun Zhu and David Mumford. "A Stochastic Grammar of Images." *Found. Trends. Comput. Graph. Vis.*, **2**(4):259–362, 2006.

[ZS05] Yun Zhai and Mubarak Shah. "Tracking News Stories Across Different Sources." In *MM*, pp. 2–10, 2005.

[ZSG05] Wenchao Zhang, Shiguang Shan, Wen Gao, Xilin Chen, and Hongming Zhang. "Local Gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition." In *ICCV*, volume 1, pp. 786–791, 2005.