GEO-TEXTUAL DATA ANALYTICS
EXPLORING PLACES AND THEIR CONNECTIONS

by

Xiaoyi Yuan
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Computational Social Science

Committee:

_____	Dr. Andrew Crooks, Committee Chair
_____	Dr. William G. Kennedy, Committee Member
_____	Dr. Andreas Züfle, Committee Member
_____	Dr. Arie Croitoru, Committee Member
_____	Dr. Jason Kinser, Department Chair
_____	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Fernando Miralles-Wilhelm, Dean, College of Science

Date: _____ Summer Semester 2020
George Mason University
Fairfax, VA

Geo-Textual Data Analytics: Exploring Places and Their Connections

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Xiaoyi Yuan
Master of Arts
Georgetown University, 2015
Bachelor of Arts
China University of Political Science and Law, 2013

Director: Andrew Crooks, Professor
Department of Computational and Data Sciences

Summer Semester 2020
George Mason University
Fairfax, VA

Dedication

To John Asbaghi.

Acknowledgments

Seven years ago I arrived at an apartment building in DC with two giant suitcases and a goal of getting a PhD degree. Never in my wildest dream did I expect meeting so many inspiring people who would have helped me to be who I am today.

First and foremost, I want to say thank you to my advisor and mentor Dr. Andrew Crooks for his guidance. He never holds back from providing the most honest and detailed feedback and yet he always gives me the freedom to be creative. What I have learnt from the five papers we have published together is invaluable. I also want to thank Dr. Crooks for the encouragement and stimulating conversations, especially when I got stuck in my research feeling frustrated. Overcoming frustration and learning to redirect research for me is a tremendously useful skill for becoming an independent researcher.

I would like to thank my committee members, Dr. William Kennedy, Dr, Andreas Züfle, and Dr. Arie Croitoru for their support. Dr. Kennedy, thank you for the GRA opportunities, without which I would not be able to finish my PhD and learn how to write grant proposals. I also very much appreciated Dr. Züfle and Dr. Croitoru for their valuable feedback and the time they spent video chatting during the COVID-19 epidemic.

I would also like to express my gratitude for my parents who have sacrificed too much for the betterment of my education.

Last but not the least, I must thank my husband, one of the most intelligent people that I have ever met. John, I truly enjoyed the extremely helpful discussions we had on my dissertation research and thank you for all the laundries you did over the years.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Global Positioning System ..... GPS
Computational Social Science ..... CSS
Volunteered Geographic Information ..... VGI
Ambient Geographical Information ..... AGI
Natural Language Processing ..... NLP
Aspect Based Sentiment Analysis ..... ABSA
Convolutional Neural Network ..... CNN
Conditional Random Field ..... CRF
Latent Dirichlet Allocation ..... LDA
Rectifed Linear Units ..... ReLU
inside-outside-beginning ..... IOB
Nevada ..... NV
Arizona ..... AZ
North Carolina ..... NC
Pennsylvania ..... PA
Wisconsin ..... WI
Illinois IL
Application Programming Interface ..... API
American Community Survey ..... ACS
Situational Awareness ..... SA
Named Entity Recognition ..... NER
Geographic Information System ..... GIS

Abstract

GEO-TEXTUAL DATA ANALYTICS: EXPLORING PLACES AND THEIR CONNECTIONS

Xiaoyi Yuan, PhD

George Mason University, 2020

Dissertation Director: Andrew Crooks

Place is defined by physical, social, and economic activities and processes. Understanding the complexity of socially constructed places is a fundamental question in geography, sociology, and many other social sciences. Meanwhile, the growing amount of user volunteered geographic information (VGI) leads us to study place through a new perspective. For instance, Flickr users report local activities in various geographic locations that capture individualistic experiences and impressions of the locations. Many previous studies utilizing non-textual VGI have focused primarily on analyzing geographical footprints of places, which separated place from its meaning. This dissertation argues that the textual part of VGI provides us with unprecedented opportunities for deriving patterns of place meanings on an individual level. More specifically, three research questions are pursued in this dissertation. First, how to quantify placeness (i.e., place identities) that has been traditionally studied via theoretical and qualitative methods? Second, as place being innately interconnected, how can we assess connections between places in networks so that we can apply network science to analyze complex connections between places? Third, as geo-textual data can also reveal social events, how to trace critical events across places using geo-textual data? In order to answer these research questions, this dissertation leverages advances in

machine learning, natural language processing and network analysis techniques on geo-textual data. By doing so this dissertation is able to build foundations for geo-textual data analytics and thus providing a new lens to study places and the connections between them from the bottom up. Overall, this dissertation showcases an interdisciplinary effort in computational social science research that combines computational textual data analytics and social scientific theories including human geography and sociology.

Chapter 1: Introduction

1.1 Background

Place is a meaningful site that combines location, locale, and place meanings, which is different from space, grounded in the concept of location. It is one of the most important concepts in geography and the idea of meaning of places has been central to human geography [2]. The "meaning" is what distinguishes place from other concepts such as location and space [3]. It is hard to study place because it is practiced through a complex time-space process and thus the meaning is never fixed [2,4]. The nature of place lies in its dynamic and interconnected characteristics [5]. Place is also an important factor in other social science research, such as sociological research on gentrification, immigration, social inequality [6–8].

Studying places as complex and dynamic systems is challenging because of its bottom-up nature. The meaning is generated by people's interaction with others and our physical environment. To understand this complex and nuanced meaning-making process, previous studies have conducted surveys, interviews and ethnographies [9,10]. Comparing to computational methods, these conventional social science methods have a long history and have been refined after generations of researchers [11]. Studies using surveys take representativeness into consideration (e.g., sample and population distributions of age and education), unlike social media studies with unknown demographics. However, these previous studies have several limitations for studying urban places. One of the most important limitations is that they are hard to scale as every survey has the problem of reply rate [12] and the purpose of ethnographic studies and interview is not to scale. To understand the nuances of place as a bottom-up process, the constantly updated large-scale online crowdsourced data (e.g., social media data) and the modern computing power gives us an unprecedented opportunity from a new perspective. This new perspective comes from both its large volume

and individualized micro-level nature. Another important aspect of online data is oftentimes contains geographic information and also unstructured natural language data [13]. In the last decade, there has been a growing body of research studying place using online "geo-referenced" data and "geo-textual" data from social media platforms such as Twitter, Foursquare, and Flickr [14–17]. There is no unified term that describes the research area that uses crowdsourced natural language data for geographic research. For example, Hu [18] defines geo-textual data as datasets containing links between geographic locations and natural language. Caquard [19] defines it as travel blogs or interviews. Other research uses "digital narrative" to refer online text reviews (e.g. [20]). A few researchers have also used the term "geo-narrative" to refer to online, large amount, crowdsourced textual data (e.g. [21, 22]). Here, I define geo-textual data as:

> ... natural language data with geo-references that can be crowdsourced and harvested through computational approaches.

Using volunteered geographic data to study places is not new. The so-called "geo-tagged" data is part of the concept of "Volunteered Geographic Information" (VGI), which was coined by Goodchild [23]. Later on Stefanidis et al. [15] proposed the concept of "Ambient Geographical Information" (AGI). Even though both AGI (e.g., Twitter) and VGI (e.g., OpenStreetMaps) refer to geographic information volunteered by users, a major difference between them is that users contributing to AGI are not necessarily aware that the content is intended to be used as geographic information. Some other similar terms are "crowdsourced graphical information" [24] and "urban computing" [25]. Liu et al. [26] proposed the concept "social sensing", as a complement to remote sensing and a new way to understand spatial interactions and place semantics/sentiments. While agreeing with the importance of social sensing, MacEachren [27] argues that not all data types share the same level of importance considering the state of art of the "big data". He points out that unstructured language data, in particular, enables us to understand the special aspect of places, beyond determining *where*, but also *what* and *why*.

1.2 Previous Research on Place

Place has been studied traditionally using both qualitative methods (such as theoretical analysis, interviews and participatory approach) and quantitative methods (such as surveys). Lynch's book Image of the City [28] showed how urban form impacts people's perception (mental map) of places. Lynch's work created an original way of exploring urban form by breaking a city down to nodes, paths, districts, edges, and landmarks. Another widely cited research on meaning of places was by Gustafson [9]. Gustafson [9] interviewed 14 residents in Western Sweden and from their in-depth interviews, he constructed a framework for understanding meaning of places based on the interaction between self, others, and the environment. Another classic research in urban places is Höflich's [10] ethnographic observation in a square located at city Udine, Italy. The study was conducted in the early stages of prevalent use of mobile phones. He sketched people's movement patterns (with speed and direction). Because of the nature of the ethnographic methodology, it was able to get highly detailed information about people's activities such as how people turned their back to others when they talk on the phone, which indicates the transition from public space to private space. In addition to ethnography, Whyte [29] studied New York City using interviews and observations to interpret why certain places are more popular and sociable than others. Sociologist Haffner [30] studied aerial photography and examined the development of urban social spaces to criticize of capitalistic society. Up until now, urban places have been an active researching area centered around the concept of "sense of place", with application in urban studies, psychology, and ecosystem management [31–37]. Sense of place is often interchangeable with place attachment or identity of places. While it is widely used and applied, sense of place has mostly been studied using non-positivistic methods [38]. Thus, one of the disadvantages of these methods is that they lack systematic analysis [3]. In the last two decades, there have been more and more survey-based research studying the sense of place (e.g., [39–45]). Surveys can cover a larger scale compared to ethnographic observation and interviews, but it is usually conducted to measure a sense of

place in one or a few specific areas. Computational methods enable a large scale of studies and more importantly, lead to the emergence of new research agendas.

Urban places have been studied using geo-tagged social media data extensively based on both Global Positioning System (GPS) locations and derived geo-location content in the past decade by researchers from multiple disciplines such as computer science, information science, and geography. One of the advantages of using geo-tagged social media data is that it captures the ever-changing dynamics of urban places. The common data sources are geo-tagged Twitter data and check-in data from Foursquare and Flickr. Geo-tagged social media data provide an unprecedented amount of data with reference to locations. In addition to its large volume, it has many important features that make geo-tagged data important for studying urban places. First, geo-tagged social media data is also time-stamped. This allows us to understand the process of how places and neighborhoods change overtime. Second, geo-tagged social media data is "bottom up". Crowdsourced social media data provides us a way of understanding city's form and function from first-hand information instead of a traditional aggregated one [46]. For instance, the notion of neighborhood aggregated from crowdsourced social media data can be different from an authoritative version [47]. Third, geo-tagged social media data is often not only "geo-data" but always contextual. When harvesting social media data, its metadata is often included as well. Sometimes, the geographic information is imbedded in the content and the users are not explicitly volunteering to share their locations. Stefanidis et al. [15] call it "Ambient Geographical Information" (AGI). Unlike check-in data, Twitter and Yelp data also contains texts (Tweets/reviews) and images. The extra information that comes with geo-tags differentiates itself from trajectory or footprint data. One of the main themes of research using geo-tagged social media data, which explains and instantiates the three characteristics is landmark and point of interest (POI) identification. Flickr is a social media site with one of the largest image-sharing volume. It was widely used for research on making sense of collective notion of landmarks. In practice, the difficulty lays in conflation, which means merging multiple locations as one. Often, the combination of metadata analysis and unsupervised learning on images is

used to get a ranked cluster on how representative the images are for certain location from Flickr (e.g., [48, 49]). Visual analysis of places, coupled with temporal features and textual features, is also used for predicting locations [50]. In addition to using Flickr locations as separate data source, user's trajectories are also mapped in the project The Geo-taggers' World Atlas [51]. In addition, trajectory patterns are studied for the purpose of predicting future movement and location recommendation. The assumptions of these studies were that people goes to places that either they think is interesting or for specific intentions based on their past visiting histories [52, 53]. Trajectory pattern analysis has also been applied to understanding travelers' behaviors [25, 54, 55]. Other than POI and trajectory analysis, another theme of research is to challenges the very idea of pre-established concept of places and neighborhoods by using geo-tagged social media data. One of the most quoted research is the Livehoods project, where they replace "neighborhood" with their own term "livehood" [47]. Hochman and Schewartz [56], however, believe that current research trend in understanding urban dynamics is too much in favor of aggregation—the highly visible (most frequently visited) areas and the commonality of places but excludes the particularity of certain places. They argue that the big advantage of using social media data is that it facilitates a much more nuanced understanding of physical places. In summary, despite the different voices on how to use geo-tagged social media data, they all share the common ground of recognizing the significance of social media data for understanding the emergence nature of urban places. Crooks et al. [46] elaborated the reasons of such significance. They point out that comparing to traditional authoritative data collection, crowdsourced social media data is a big transition for harvesting urban form and function information because it enables us to study the constantly evolving urban landscape.

As discussed, one of the three characteristics about geo-tagged social media data is that it is contextual. Geo-located Tweets, Yelp reviews, or TripAdvisor reviews all have rich contextual data including the content (Tweets, review, or hashtags), user metadata and online social connections, and temporal information. What has not been explored as extensive as others is the *content* of the contextual data, namely geo-textual data. Topics of

current research for geo-textual analysis can be classified into 1) places inference, 2) thematic place analysis and topic modeling, and 3) sentiment analysis. First, geo-textual data has been used for place inference. Since Twitter and many other social media platforms allow user to create hashtags, these tags usually reveal the topic of corresponding social media posts. Ratternbury and Naaman [57] analyzed Flickr tags and demonstrated that location metadata (place tags) enables the extraction of place semantics that improves image search. Tomaszewski et al. [58] created a system for geo-referenced natural language search engine, SensePlace. With an input containing, geo-information, it returns a list of main events going on in that region. However, they did not use social media data but traditional media news articles as data source. The second research category, thematic place analysis, is the most well researched one within geo-textual studies. With topic modeling, places can be categorized into several "topics" based on social media posts. For example, Jenkins et al. [22] categorized different regions in Singapore, London, Los Angles and the New York City and compared the result of categorization between Twitter data (crowdsourced sense of place) and Wikipedia data ("objective" sense of place). In terms of sentiment analysis using geo-textuald data, Sekar et al. [20] used online review data from Yelp and TripAdvisor and investigated sense of places through digital narratives and classified sentiments of places into multiple dimensions other than naïve positive/negative: dissatisfaction, attachment, dependence, identity, aesthetic, and social/cultural for sentiment analysis. Nelson et al. [59] proposed geo-visual analytics for exploring public political discourse with respect to geographic regions (which region is talking about what theme).

In summary, previous studies using geo-textual data has taken advantages of these advances in natural language processing (NLP) methods with an emphasis on unsupervised learning, especially topic modeling [60–63]. In the last decade, deep learning methods have proven to be more effective and accurate in many NLP tasks. The research presented in this dissertation utilizes various NLP and machine learning algorithms. Section 1.3 will introduce them in the context of three research questions that this dissertation aims to answer.

1.3 Research Themes and Questions

The main theme of this dissertation is to demonstrate how geo-textual data analytics can be used for exploring places and place connections. This overaching theme is illustrated in Figure 1.1. As discussed in Section 1.1, place is a dynamic, nuanced, and bottom-up concept and the meaning of place is what differentiate "place" and "locations" or "space". This dissertation aims to argue that fine-grained and constantly updating geo-textual data harvested online can be utilized to understand place. While the concept of place is rooted in human geography, analyzing a large amount of geo-textual data requires computational techniques. Therefore, Figure 1.1 also depicted how computational social science (CSS) acts as the underpinning of this dissertation. CSS can be broadly defined as the study of social science using computational techniques [64]. This dissertation showcases and demonstrates how a large amount of textual data can be processed and analyzed computationally for answering questions that pertain to human geography.

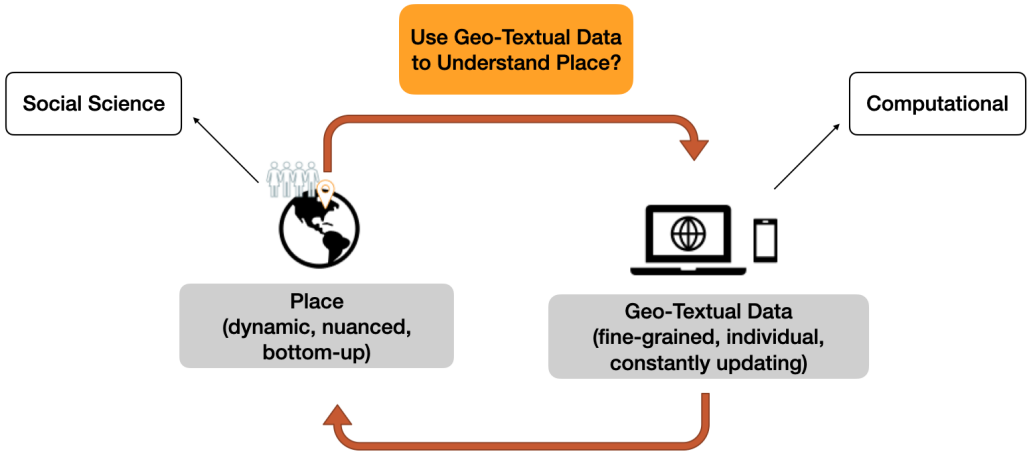


Figure 1.1: An illustration of the overarching research question of this dissertation.

What does it mean to understand place? To answer this question, we need to trace back to the definition of place. As discussed in Section 1.1, meaning is what differentiate

between place and location. To understand place is to understand the meaning of place constructed by individuals, such as people's perceptions of place. This dissertation argues that despite that place has been studied by various methods, geo-textual data provides us unprecedented opportunity to study place from a new perspective. For instance, a classic critical lens of studying urban place is on the problem of place identities [65] and most of the work on such topic is through theoretical and qualitative analysis [66–69]. Geo-textual data enables us to quantify the level of placeness (i.e., place identity) across different regions. The first research presented in Chapter 2 showcases how to use geo-textual data to assess placeness of several urban areas. Moreover, as Agnew [5] argues, the nature of place lies in its dynamic and interconnected characteristics, the second theme of this dissertation aims to build and quantify place connections through people's perceptions and reviews of place. In addition, as geo-textual data is defined as "natural language data with geo-references that can be crowdsourced and harvested through computational approaches", the "textual" part of geo-textual data is not limited to be used for studying place perceptions. For instance, geo-located tweets are not just about places, but could also be expressions on anything that users are willing to share. During emergency and critical events, geolocated social media provides us timely information on current situations [70]. Therefore, the third research included in this dissertation examines how geo-textual data can be used for understanding critical events across places. The remainder of this section introduces the three research questions in detail.

- RQ1: How to quantify "placeness" of urban places using geo-textual data?

The first research question arises from various urban local policies stating that urban places are becoming placeless and identical because of the growth of chain stores. To answer this question, I analyze geo-textual data in the form of Yelp reviews of independent and chain stores. More specifically, the question is answered by identifying opinion aspects through a deep neural network, a kind of machine learning model. The research intents to discover how people perceive different places and which aspects of the places people care about the most through analyzing the opinion aspects.

- RQ2: How to examine place connections in networks derived from geo-textual data?

Places are never in isolation. This research question aims to quantify place connections by linking places in networks. The network is built upon the thematic similarities of places, derived from topics (i.e., themes) of texts harvested with geolocations of these places. The research question is answered using three different geo-textual data sources, i.e., TripAdvisor restaurant reviews, TripAdvisor attraction reviews, and Twitter data. By comparing networks derived from three data sources, the research aims to get a well-rounded perspective on place connections from geo-textual data as well as understanding the implications of different commonly used geo-textual data sources for studying places and their connections.

- RQ3: In what way can we use geo-textual data to achieve situational awareness of events across places?

This research question reflects a more specific usage of geo-textual data analytics with an application in achieving situational awareness of events. This research question is answered in a specific context of drug cartel events. Place and place connections are explored in the context of notable events of drug cartels detected from geolocated Twitter data. Through extracting named entities from the tweets and identifying spatiotemporal patterns of the named entities, this research aims to make an initial effort on achieving situational awareness of these transnational crime organizations that poses great threat to communities world wide.

1.4 Dissertation Outline

This dissertation is organized into 5 chapters. Following this chapter (Chapter 1) that introduces the background and research questions of the dissertation, Chapter 2, 3, and 4 answer the three research questions presented in Section 1.3 respectively. Chapter 5 summarizes findings and results of the three studies and discusses contributions to research areas of geo-textual data and place and contributions from a broader computational social science

perspective. In addition, Chapter 5 also addresses the limitations of this dissertation and future work.

Research presented in Chapter 2 publishes in the Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery [71] and research from Chapter 3 publishes at ISPRS International Journal of Geo-Information [72].

Chapter 2: Assessing the Placeness of Locations Through User-Contributed Content[1]

Previous research has argued that urban places are becoming "placeless" and inauthentic. Many local policies have also proposed to encourage more independent stores in order to restore urban identity. Others argue, however, that chain stores provide affordable merchandise and different locations of the same chain may have different meanings to an individual. The research presented in this chapter uses a Convolutional Neural Networks model to extract opinion aspects from more than 3 million user-contributed Yelp restaurant reviews. The results show high homogeneity among cities in terms of the average proportions of aspects in restaurant reviews. In addition, for fast food chains, "location" is the only aspect category reviewed proportionally higher than independent fast food restaurants. An analysis of the co-occurrences of "location" indicates that the identity of chain restaurants stems from the comparison between the same chain of different locations whereas the identity of the independent restaurants is more diverse, implying the intricacies of placeness of urban stores. This research demonstrates that fine-grained sentiment analysis (i.e., opinion aspect extraction and analysis) with geo-tagged text data is fruitful for studying nuanced place perceptions on a large scale.

2.1 Introduction

Many studies have argued that urban places are becoming inauthentic because of urban commodification and standardization [67,73,74]. As restaurants occupy an increasingly important place in urban culture and the economy [75,76], the research presented in Chapter 2

[1]This chapter is based on: Yuan X. and Crooks A.T. (2019), Assessing the Placeness of Locations through User-contributed Content, in Gao, S., Newsam, S., Zhao, L., Lunga, D., Hu, Y., Martins, B., Zhou, X. and Chen, F. (eds.), *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (GeoAI)*, Chicago, IL. pp. 15-23.

aims to understand the relationship between restaurants and urban identities by extracting and analyzing the key information from Yelp restaurant reviews using a technique called "opinion aspect extraction." By analyzing the extracted aspects, the concept of placeness is examined through two research questions: first, to what extent cities share similarities or differences in their Yelp restaurant reviews; second, how opinion aspects are mentioned differently in urban chain and independent restaurants.

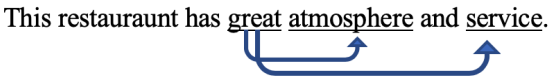


Figure 2.1: An example of opinion aspects.

User-contributed review data enables us to understand urban places and restaurant culture through detailed descriptions and evaluations of restaurants, written voluntarily by users on a large scale. The technique aspect extraction used in this research is a key subtask of aspect-based sentiment analysis (ABSA) [77]. An example is shown above 2.1, in which "great" is the opinion and "atmosphere" and "service" are the opinion aspects. ABSA is a branch of sentiment analysis that has been a widely used computational method to analyze public opinions or customer satisfactions [78]. With the increasing amount of review data and social media data available, ABSA has gained in popularity as a more fine-grained approach than sentiment analysis [79] and has been applied to analyzing movie reviews, product reviews, and Twitter data to understand the opinions towards not only movies, products, or other subjects as a whole, but also their various aspects [80, 81].

While opinions vary along with the reviewer's taste, opinion aspects indicate the most important information of the reviews. Instead of focusing on the sentiments, this research aims to extract only the opinion aspects to understand *what reviewers care about the most*. Previous research using review data to study urban places often used topic modeling, which is a technique that extracts the latent topics of documents (reviews) [82]. Other research

has also used topic modeling for aspect extraction [83]. Topic modeling, however, can only find some general aspects, not the precise aspects [77]. Therefore, a deep learning method of aspect extraction is used in this research to extract the exact, fine-grained aspects. More specifically, this research uses a convolutional neural network (CNN) model.

More than 3 million Yelp reviews from 37,000 restaurants were processed using a CNN. An aggregation rule was then applied to aggregate aspects to 18 categories for analysis. The average proportion of each aspect category mentioned in the reviews was then compared among cities and between chain and independent places to explore the relationship between urban restaurants and sense of place as many petitions were made to restore urban culture through encouraging independent stores. San Francisco enacted the Formula Business Policy in 2004 requiring any retail store to notify the community when they move in [84]. In 2015, The East Village Community Coalition in New York City released the report for informing policy makers arguing chain businesses can be detrimental to community character and local economics [85]. In the remainder of Chapter 2, I will first discuss related literature, which sets the scene for our research (Section 2.2). I will then describe the data in Section 2.3 and introduce the methodology in Section 2.4. Finally, I will present results in Section 2.5 before a discussion of the implications of the research (Section 2.6).

2.2 Related Work

This research intends to examine placeness of location from the perspective of urban stores by analyzing extracted aspects from Yelp reviews. In this section, I will first discuss previous works on urban places and stores from both computational and non-computational perspectives. Then I will examine the previous works on extracting opinion aspects and explain why a CNN model is used for this research.

2.2.1 Urban Places and Stores

The theories of (urban) placeness was pioneered by Relph, in which he differentiates "place" from "space" [74]. Relph argues that as place affects human intentions, experiences, and actions spatially, our understanding and perception of space is related to the places where we reside, and thus the modern society of highly commodified place lacks authenticity [74, 86]. Similarly, Agnew (2011) points out that "Strip malls and chain stores replace the elemental variety that once characterized the landscape" [87]. Other studies on urban places also consider stores as an important part of urban identity, urban culture, and place authenticity [88, 89].

Different from the previous studies on urban places using non-computational methods, research using computational methods to analyze crowdsourced or user-contributed data is often on a larger scale, including research using textual data from Twitter and Wikipedia [82,90] and image data from Google Street View [91,92]. Apart from the large scale, studying places using computational methods on crowdsourced and user-contributed data (especially geo-tagged text data) enables us to study urban places in fundamentally new ways by understanding the dynamic complexity of places [93]. While many studies have argued that chain stores contribute to place uniformity, the nuances of chain store, independent store, and store locations are explored more often from a business marketing perspective such as store placement [94]. Chapter 2 explores people's perceptions towards urban restaurants and thus closes the research gap between restaurant stores and urban identity.

2.2.2 Extracting Opinion Aspects

Many methods of aspect extraction have been proposed by natural language processing (NLP) researchers in the past decade for aspect extraction, including frequency-based methods, syntax-based methods, supervised machine learning models, and unsupervised machine learning models [95]. Earlier methods extract aspects by extracting the most frequent nouns because most of the aspects are nouns and noun phrases. However, a major shortcoming of this frequency-based method is the high false positive rate even when a part-of-speech

pattern filter is applied [96, 97]. Syntax-based rules were then designed to reduce the low recall of frequency-based methods using a dependency parser [98,99]. Before deep learning, Conditional Random Field (CRF) was a common supervised learning method for aspect extraction. Features including the words,their part-of-speech, dependency relation, and distance between the word and sentiment are used to train CRF models [100]. For unsupervised learning, many researchers have utilized the Latent Dirichlet Allocation (LDA) based topic modeling approach for aspect extraction. Using document-level LDA for fine-grained tasks such as aspect extraction can be problematic because the topics that LDA finds are too global to be defined as "aspects"[77,95]. However, topic modeling approach is still useful for analyzing geo-tagged social media texts and for getting rough topics to study people's perceptions of urban places [82, 101]. In the past few years, deep learning has emerged as a powerful technique that produces state-of-art results in sentiment analysis and aspect extraction [102–105]. Among many other deep learning methods, deep CNN models achieved high performances for aspect extraction [106, 107]. In terms of studying urban places, CNN has been primarily applied to analyze image data such as OpenStreetMap data [108].

2.3 Data

The data used in Chapter 2 was downloaded in November 2018 from the Yelp dataset challenge (`https://www.yelp.com/dataset/challenge`).

Table 2.1: Vocabularies used to include and exclude business categories

Included	**Excluded**
"Restaurants'	"Butcher", "CSA", "Convenience Stores", "Custom Cakes"
"Restaurant"	"Farmers Market", "Food Delivery Services", "Food Trucks"
"Food"	"Grocery", "Honey", "Kombucha","International Grocery"
	"Meaderies", "Specialty Food", "Vendors", "Wineries"
	"Water Stores"

In total, the dataset has 192,609 businesses, from which I selected restaurant businesses using the field "category" in its business data. Keywords used to select restaurant categories are "Restaurant", "Restaurants", and "Food". Although using "Food" as a keyword ensures a more complete selection, it also selects other non-restaurant businesses such as grocery stores. Therefore, several keywords were also used to exclude the irrelevant business types (Table 2.1). Finally, restaurants were further selected based on their location. In the Yelp dataset, most of the states only have data on less than 10 businesses. To avoid over-representation, these states are excluded from the study. Eventually, in total, there are 37,818 businesses and 3,399,993 reviews are used for this research from 6 states (NV, AZ, NC, PA, WI, and IL). On average, there are 89.9 reviews for each restaurant.

2.4 Methodology

2.4.1 Opinion Aspects Extraction

To extract and analyze the aspects of the reviews, there are two main tasks. The first is to extract the opinion aspects from each review sentence and the second is to aggregate the aspects to categories for further analysis. A CNN model is used for the first task, aspect extraction. This section will first provide some background on CNN and how it is applied in the task of aspect extraction, followed by explanation of the method used to aggregate the aspects.

Background on CNN and word embedding

CNN has been widely used for computer vision before it inspired NLP researchers to adopt it to many NLP tasks [109]. CNN, as a deep learning model, has achieved state-of-the-art performance for various tasks in NLP, including part-of-speech tagging, named entity recognition, and sentiment analysis [110, 111].

Representing word as vectors ("word embeddings" or "word vectors") is crucial for making deep learning in NLP possible [112]. There are two common methods of training

word vectors: "Word2Vec" and "GloVe" [113, 114]. Both of the methods assign each word in the corpus to a corresponding vector in the space so that words sharing common contexts are located close to one another. Since training word embeddings requires a corpus with billions of vocabularies, pre-trained embeddings are often used in deep learning models. In this research, I used a GloVe embeddings pre-trained with 6 billion tokens and 400 thousand vocabularies (`https://nlp.stanford.edu/projects/glove/`). As many other pre-trained embeddings, tokens in GloVe are mapped to a 300-dimensional space.

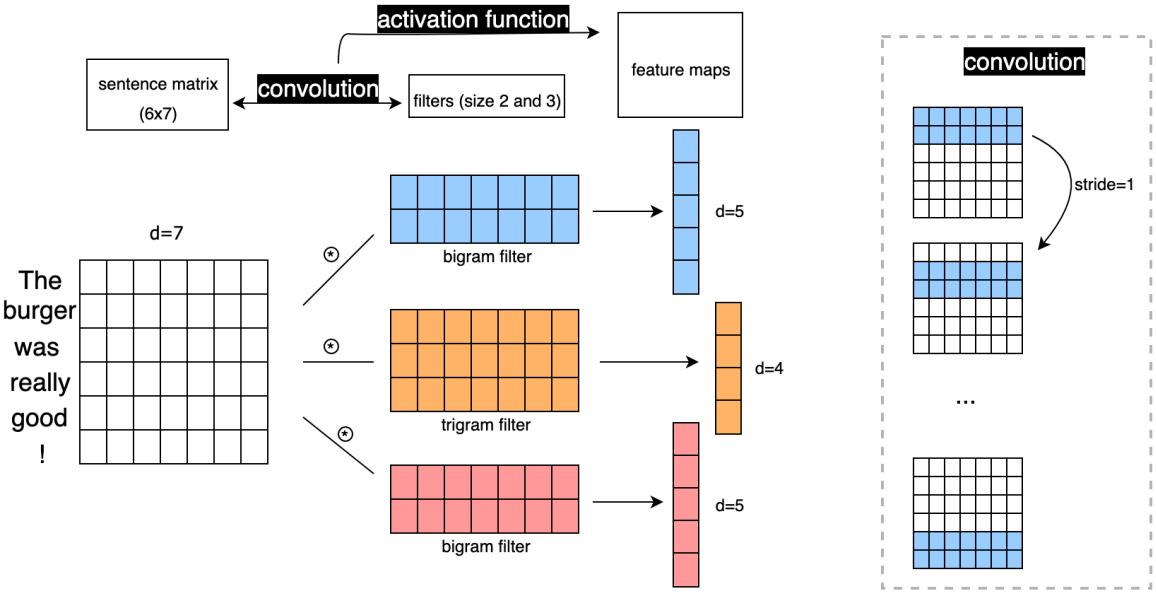


Figure 2.2: Illustration of an example of a CNN layer.

With word embeddings, each sentence can be represented as a matrix. Figure 2.2 shows a simplified example demonstrating how a convolution layer works. For better readability, the word vectors are in a 7-d space instead of 300-d and thus the input sentence is represented by a 6 by 7 matrix. As shown in the example, multiple filters containing the weights are applied to the sentence matrix with an activation function to produce feature maps. The power of deep learning neural networks comes from its non-linearity. So the result of the convolution is applied as the input of a non-linear function. The process of convolution is

shown in detail in the right panel of Figure 2.2. When a filter is on top of the sentence matrix, it means applying the activation function to the element wise multiplication of the overlapped area, which produces the first element of the vector in the feature maps (i.e. a neuron):

$$c_i = f(w^T X_{i:i+h-1} + b) \tag{2.1}$$

$$c = [c_1, c2, ..., c_{n-h+1}] \tag{2.2}$$

where f is the activation function and h is the size of the filter. Biases (b) are not depicted in Figure 2.2 for simplification because biases are calculated along with weights in filters. The filter slides one row down at each step and then repeats the same process until it reaches to the bottom. The number of rows a filter slides down the matrix each step is called "stride." If the sentence matrix is not padded (adding columns and rows to the four sides of the matrix), the top and bottom rows of the matrix are calculated once while other rows are calculated twice. Therefore, the sentence matrix is sometimes padded to make sure each row is calculated the same number of times. In this example, with stride 1 and padding 0, the feature maps from the bi-gram filter is 5-dimensional and that from the tri-gram filter is 4-dimensional. Followed by a convolutional layer, there is often a max-pooling layer $\hat{c} = max(c)$ [115]. After max-pool, the weights are trained so that they pick up the most important bigrams/trigrams, the definition of importance depending on the task. Stride, padding, number and size of filters are all hyper-parameters of CNN models. Eventually a fully connected layer is applied to the vector result from max-pool. A fully connected layer ensures the output matches the number of classes of the training data. For instance, if the task is a 3-classes classification, then the model will produce a matrix of 3 columns, corresponding to the prediction of each class.

Using CNN to extract opinion aspects

As mentioned in Section 2.2 and Section 2.4.1, CNN uses a shared weights strategy so that the number of weights that need to be updated is much smaller than in a fully connected

neural network. The hyper-parameters of a CNN model, however, still need to be tuned. It is computationally expensive to train CNN models. Therefore, researchers often start with a pre-existing model structure based on previous studies [116]. Previous research using CNN to extract aspects often have 2-3 CNN layers, followed by max-pool and a fully connected layer [106, 107]. With pre-existing CNN models and experiments through cross-validation, the final model structures with the highest F1 score are: 5 convolutional layers (256 filers with size 5, stride 1, padding 1), each followed by a dropout layer, and finally a fully connected layer. The activation function is Rectified Linear Units (ReLU), defined as $f(x) = max(0, x)$. After the experiments, the best model on the Yelp dataset has an average F1 score of 68.79% with 5 runs.

Aspect extraction, in other words, is to label each word of the sentence to be "aspect words" or "not aspect words", which is "sequence labeling" in NLP. A standard method for sequence tagging is inside-outside-beginning (IOB) [117]. As depicted in Table 2.2, "B" labels the beginning word of an aspect word and "I" is for non-beginning words. Non-aspect words in a sentence were labeled as "O".

Table 2.2: An example of IOB sequence labeling

I	do	not	enjoy	the	chicken	soup	at	all	.
O	O	O	O	O	B	I	O	O	O

Gaining training data

Training the CNN model requires IOB labeled training data. For aspect extraction, previous literatures have been using the training data from SemEval-14 (`alt.qcri.org/semeval2014/task4`) and SemEval-16 (`alt.qcri.org/semeval2016/task5`). However, after training the model with SemEval restaurant data, the model does not perform well on Yelp test data (F1 score 63%) but performs much better on SemEval's own test data (F1 score 77%). Therefore, the model was trained using 3,000 hand-labeled sentences randomly

selected from the Yelp dataset. The labeling was done by two labelers with an 5% overlap.

2.4.2 Opinion Aspects Aggregation

After the model was trained, aspects of each sentences were extracted. Analyzing these aspects requires aspect aggregation. Using word embeddings, we can measure semantic similarities between words. Through trial and error, aspects were assigned to the 18 aspect categories. Those categories are "chef", "ingredient", "quality", "portion", "texture", "taste", "atmosphere", "experience", "reservation", "crowd", "menu", "location", "food", "drink", "dessert", "place", "interior", and "price". The assignment involves two aggregation rules (Table 2.3), a dictionary-based, and a semantic similarity-based rule (Table 2.4). For each aspect, if the word can be found under "aspects" in Table 2.3, it is assigned to the corresponding "aspect category". If not, the word will be assigned to the corresponding aspect category that has the highest cosine similarity between this word to the "aspects" in Table 2.4. For aspects that has more than one word, the aspect vector is calculated by adding vectors of all the words in that aspects.

Table 2.3: Dictionary-based aspect aggregation rules

Categories	**Aspects**
"chef"	"chef"
"ingredient"	"ingredient", "ingredients"
"quality"	"quality"
"portion"	"portion"
"texture"	"texture", "textures"
"taste"	"taste", "tastes", "flavor", "flavors"
"atmosphere"	"atmosphere", "vibe", "vibes", "ambience"
"experience"	"experience", "experiences"
"reservation"	"reservation", "reservations"
"crowd"	"crowded", "wait", "waiting"
"menu"	"menu", "options", "option", "choices"
"location"	"location", "locations"

Table 2.4: Semantic similarity-based aspect aggregation rules

Categories	**Aspects**
"food"	"food Food foods", "food dish meal", "entree", "seafood" "vegetable vegetables", "meat beef steak chicken poultry", "pasta" "pizza", "soup", "cheese", "sauce", "rice beans", "salad salads" "appetizers appetizer", "burger sandwich sandwiches"
"dessert"	"dessert desserts sweet sugar"
"drink"	"drink wine tea alcohol soda"
"place"	"place restaurant"
"interior"	"seating room table layout area"
"price"	"price pricing prices priced cost costs"

2.5 Results

2.5.1 Statistical Summaries

For each restaurant, there are three attributes that will be used as control factors for aspect analysis. As seen in Figure 2.3, a majority of the reviews are from businesses of price range 1 or 2. Moreover, the median of stars for businesses is 4 stars while there are also many reviews from businesses with 3 and 4.5 stars. Unlike price range and stars, the number of reviews distribution is long-tailed with a few businesses with a large number of reviews and most others having much less.

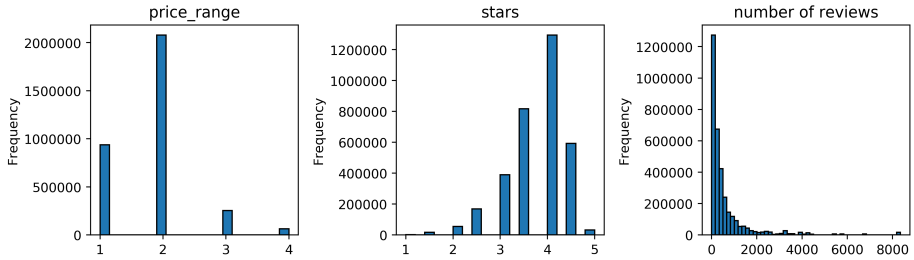
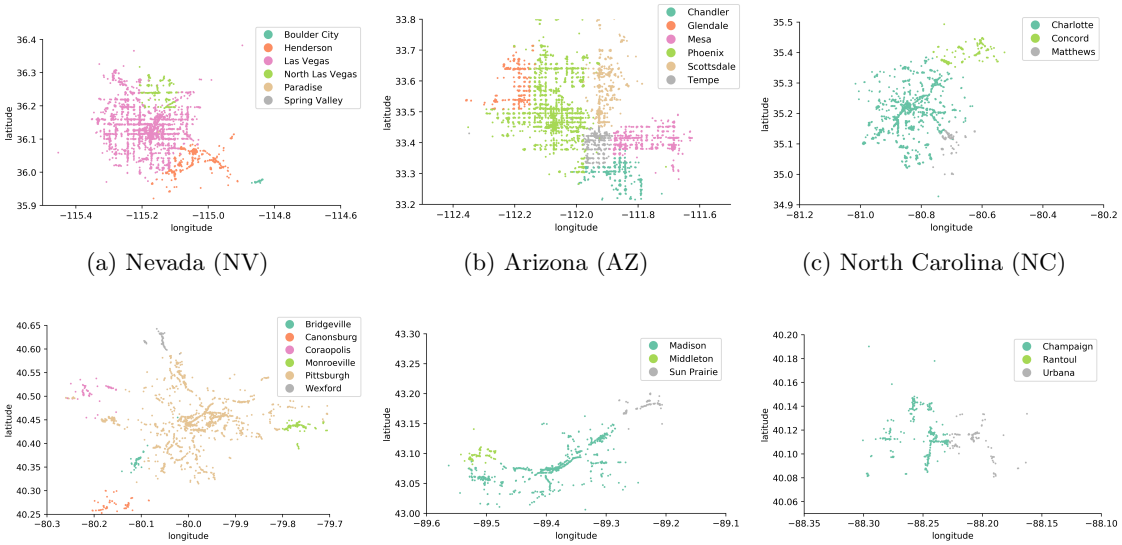


Figure 2.3: Frequency distributions of restaurant price range, stars, and number of reviews.



(a) Nevada (NV) (b) Arizona (AZ) (c) North Carolina (NC)

(d) Pennsylvania (PA) (e) Wisconsin (WI) (f) Illinois (IL)

Figure 2.4: Mapping restaurants in NV, AZ, PA, NC, WI, IL. Not all cities are shown in each state. Only cities have data that accounts for the majority of the restaurants in that state are mapped, for the sake of visual clarity.

2.5.2 Homogeneity Among Cities

Six states are selected from the original Yelp dataset (Section 2.3). Meanwhile, reviews in these states are shown to be surrounded by one major city in that state (Figure 2.4). Therefore, cities in the cluster centers are picked for analysis. These cities are Las Vegas (NV), Phoenix (AZ), Charlotte (NC), Pittsburgh (PA), Madison (WI), and Champaign (IL).

Figure 2.5 shows the average proportion of each aspect category from the study cities. Many aspect categories have relatively smaller proportion. It should not be taken for granted that these do not matter for the analysis. The reason is that reviewers tend to list out the food they ordered and write a detailed description of the service they received, which leads to the result that "food" and "service" constitute the majority of the total amount. For those aspect categories with higher proportions (the bottom two rows in Figure 2.5), the biggest differences between cities are below 1%. For the aspect categories with proportions that

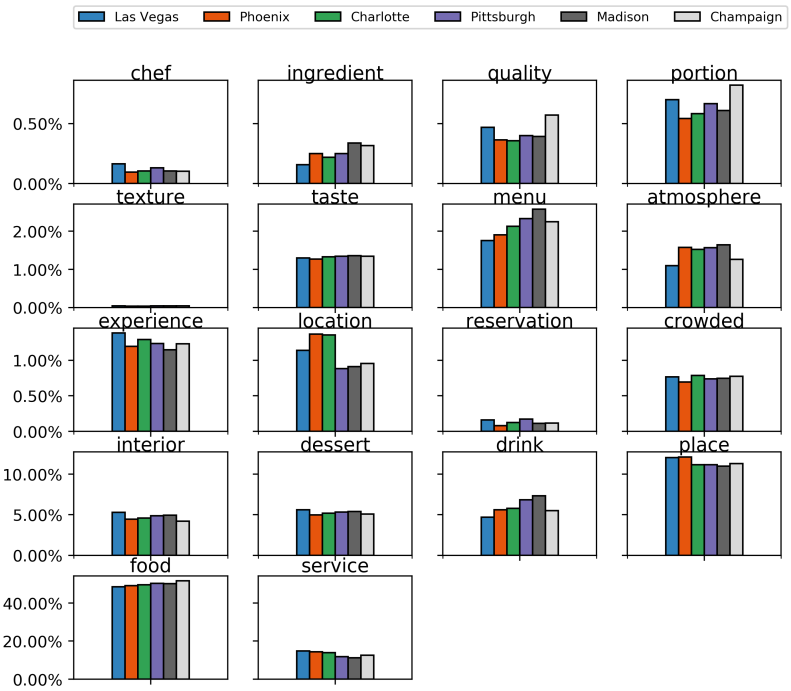


Figure 2.5: Average proportions of aspect categories in a review. Each column represents a city.

are significantly smaller (top three rows of Figure 2.5), "texture", "taste", and "crowded" show extremely high similarities among cities. For other aspect categories, as an example, reviews for Las Vegas restaurants emphasize on "chef" and "experience" slightly more and "atmosphere" slightly less than that of other cities. However, overall, figure 2.5 shows high homogeneity among cities. This result echos several theories concerning the "placelessness" of urban places [68].

At the same time, policies concerning restoring city "personalities" argue that chain stores are the "ones to blame" and the "war" between chain and independent stores is even depicted as a social movement as many anti-chain store organizations seeking tax reform for urban stores [118]. Other people argue, however, that chain stores provide bargained prices, make large range of products available and absorb the characteristics of the local place, and therefore stores belonging to the same corporate chains are actual competitors [119, 120]. The next section, therefore, compares patterns of aspects from independent and chain fast

food restaurant reviews and shows how the aspect "location" plays a role in them.

2.5.3 "Placeness" of Fast Food Chains

To compare independent and chain restaurants using the data from the same cities from Section 2.5.2, I controlled the type of restaurants to be fast food because fast food restaurants are becoming more and more important in urban restaurant culture [121]. The other factor that needs to be controlled was price range since high-end fast food places can be perceived very differently. Information on whether the restaurant is "fast food" can be found in "categories" in the Yelp data. After selecting fast food restaurants, restaurants with "price_range" of 1 and 2 are selected. Eventually, to separate chain and independent restaurants, those with their names appearing in the data once are classified as independents and those that show up multiple times are classified as chains. However, some restaurants classified as independents were actually chains that only appeared once in the dataset. To solve this problem, I manually removed the false positives.

Figure 2.6 shows the differences of aspect categories for independent and chain fast food restaurants. The x-axis is the average proportion of each aspect normalized by dividing the mean. For each aspect, therefore, one bar is over 100% and the other is below 100%. The most prominent feature from Figure 2.6 is that for chains, only the category "location" is higher than independent restaurants. It means that "location" for chains is extremely important while independents restaurants were reviewed in a much more diverse manner. Cities are also compared in terms of independent and chain fast food restaurants and no significant differences between cities were found, which is consistent with findings in Section 2.5.2.

To gain further insights on what causes the differences in the aspect category "location" for chain and independent restaurants, aspects assigned under category "location" are examined. Table 2.5 lists examples of review sentences that contain location-related aspects. The review text was truncated for simplicity and for privacy reasons. From these examples, we can tell that the meaning of "location" varies in different contexts, ranging from the

characteristics of the place that the restaurant is located in, such as "mall location", to the generic features of a restaurant, such as the example 3 under "Independent" in Table 2.5. Location is also being used for chains for comparison. For instance, the "best KFC location I've ever been to."

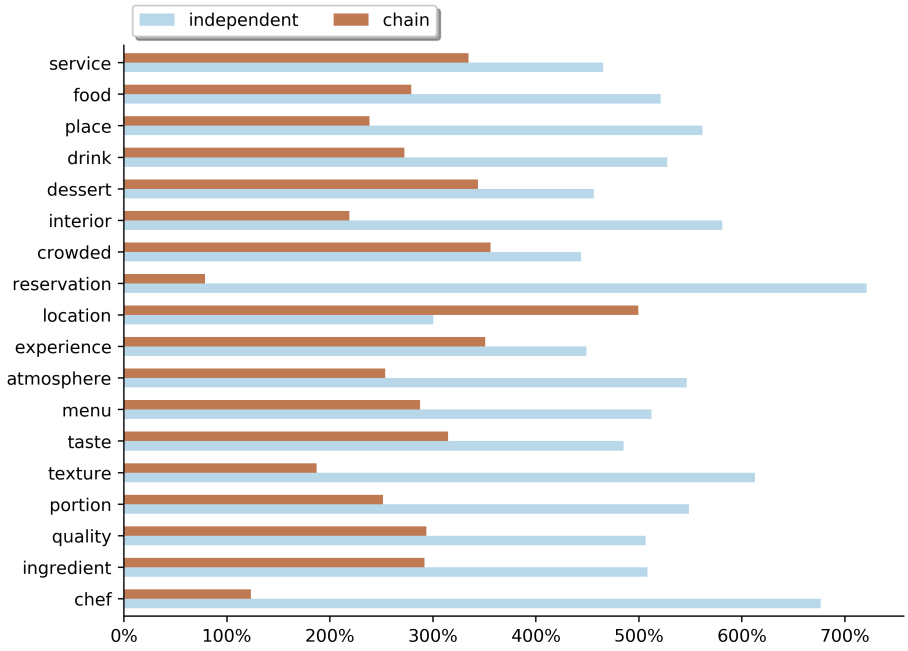


Figure 2.6: Average proportions of aspect categories for chain and independent fast food restaurants, normalized by dividing the mean for comparison.

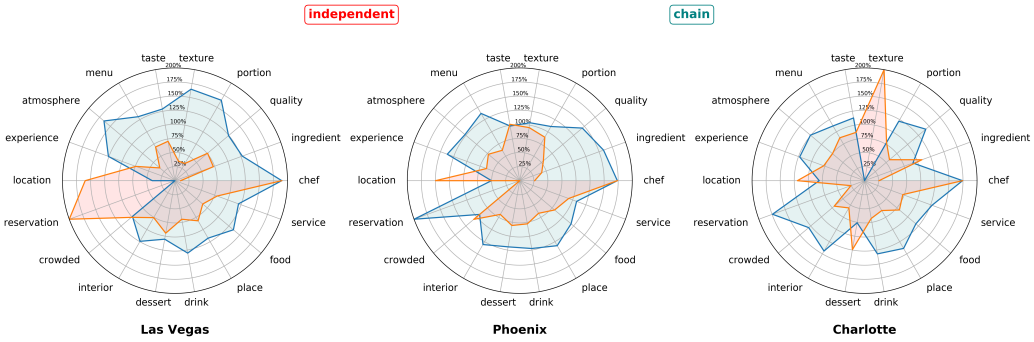
Table 2.5: Examples of reviews that contain aspect category "location" for independent and chain (fast food) restaurants

	Examples
Independent	1. "So, this place might not be the best **veg friendly location**..." 2. "We'll definitely be back soon! Support a **neighborhood location**!!" 3. "The **location** is very convenient as well."
Chain	1. "Haven't had this issue with **other In-n-Out locations** in Phoenix." 2. "We only come here for the **location convenience**." 3. "The **other Madison locations** are not like this. I would not recommend."

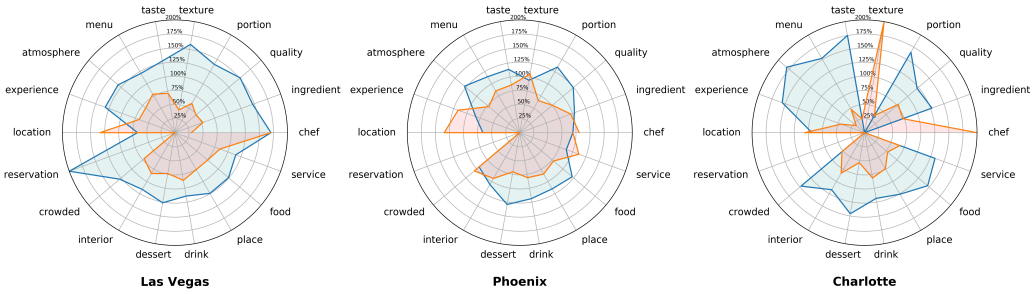
Many location-related aspects are single words (i.e. "location" or "locations") with no co-occurrences. Others have co-occurred words that were extracted along with "location" or "locations", which provides us an opportunity to understand the contextual meaning of them. For independent restaurants, "location" is primarily used to describe the characteristics of the place where the restaurant is situated. For chain restaurants, however, the co-occurrences take many different forms, which can be grouped into three types: co-occur with location characteristics, co-occur with store names, and co-occur with place names (Table 2.6). The grouping is done using k-means with word embeddings. After k-means, the grouping result is then examined manually to ensure a better performance. As same as independent restaurants, some co-occurrences for chains are about restaurant features, such as "drive thru", "24hr", "remodeled", "neighbor", or "airport". What marks the difference between independent and chain "location" is that chain restaurants have co-occurrence with store name or with place name for comparing different restaurant locations. Table 2.6 ("Percentages") shows that two categories (50.98% and 38.73%) account for the main reasons why chain restaurants are reviewed heavily on "location". Comparing the same franchise with different locations is a crucial part of "location" for chains. This pattern does not exist for independent restaurants, the reviews of which emphasize more diverse aspect categories than those of chain stores.

Table 2.6: Unique co-occurrence of "location" and "locations" for independent and chain (fast food) restaurants

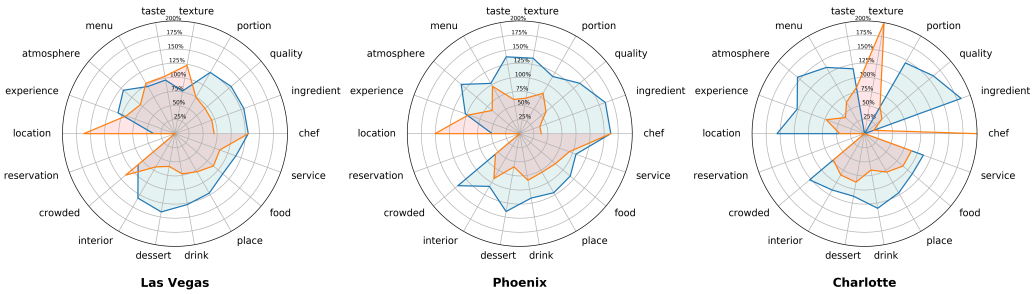
	Types of Co-Occur	**Percent**	**Examples**
Independent	all	100%	"Veg friendly location", "mall location", "downtown location", "neighborhood location"
Chain	location characteristics	10.29%	"location convenience", "24hr location", "drive thru location", "neighbor location" "airport locations"
	franchise names	50.98%	"Chick Fil A locations", "subway locations' "Firehouse Subs location"
	place names	38.73%	"uptown location", "Madison location" "Metrocenter location"



(a) **American** fast food restaurants.



(b) **Mexican** fast food restaurants.



(c) **Asian** fast food restaurants.

Figure 2.7: Average proportions of aspect categories for chain and independent fast food restaurants for three kinds of cuisine (American, Mexican, Asian) in Las Vegas, Phoenix, and Charlotte, normalized by dividing the mean for comparison.

While "location" is reviewed more often for chain restaurants, the results could potentially vary with cities or with restaurants of different cuisine. To examine whether this pattern exists in different cities, cities that have the highest number of reviews (Las Vegas, Phoenix, and Charlotte) and three most reviewed kinds of cuisine (American, Mexican, and Asian) were selected for demonstration (Figure 2.7). Overall, "location" for fast food

chains still plays an important role and independent fast food places still have more diverse aspects. Despite these consistencies with earlier findings, a few exceptions exist when breaking data down to different cuisine and cities. For Charlotte, there is less difference between independent and chain in terms of the average proportion of "location". Meanwhile, for Asian fast food restaurants in Charlotte, "location" for chains is even lower than that of independents. It suggests that although cities share high similarities in terms of all kinds of restaurants (Figure 2.7), the differences between cities exist at a more detailed level, such as different kinds of restaurants (fast food and non-fast food) or restaurants with different kinds of cuisine.

2.6 Discussion and Conclusion

This research characterized people's perception of urban restaurants through analyzing the opinion aspects of Yelp reviews. Chapter 2 discovered that, first, cities show homogeneity in terms of restaurant reviews and, second, "location" is the only category of aspects in which chain restaurants are higher than independent restaurants. Meanwhile, the co-occurrences of "location" for chain restaurants are primarily restaurant names and place names. It implies that the contextual meaning of "location" varies for chain and independent restaurant reviews. For chain restaurants, "location" often emphasizes the differences between different stores of the same chain whereas for independent restaurant reviews, the aspect "location" reflects the characteristics of the places the restaurants are situated. Reviews of different cuisine in various cities are also examined to demonstrate the potential of using fine-grained analysis of geo-tagged textural data to study urban places. In summary, this research has three major contributions:

1. It demonstrated that the Aspect Based Sentiment Analysis (ABSA) with deep learning method can be applied for analyzing public perceptions on review data. Very little research on ABSA has focused on analyzing opinion aspects, especially in the area of urban social sensing [122]. Recent research has applied ABSA using topic

modeling to study neighborhood perception from neighborhood review data [123]. ABSA, therefore, shows potential in various applications because of its fine-grained characteristics. Analyzing aspects of sentiment digs deeper into people's opinions and perceptions that would be missed otherwise. In the example of this research, chain and independent restaurants were compared from various aspects and "location" was discovered as a unique aspect that differs between them.

2. This research also showcased how geo-tagged review data analysis can be applied to study place perceptions from the perspective of urban restaurants. People's perception of urban places is individualized. Therefore, comprehending patterns of place meanings requires a large amount of data. Compared to conventional interview and survey data, crowdsourced geo-tagged textual data is on a much bigger scale [60].

3. It found that the most important factor for chain fast food is the comparison between different locations of the same chain while other factors such as food quality, and service are not mentioned as often compared to independent fast food places. For independent fast food places, "location" is about store surroundings, such as "neighborhood location". These results show that although urban stores (especially chain stores) are blamed as the cause of lack of place identity [66], the meaning of location of urban stores is nuanced.

There are, however, limitations to this research. First, it is unknown how representative the restaurant data used in this research is of the studied regions. Yelp users have higher proportion in high income, high level of education, and a younger population than the US demography (`https://www.yelp.com/factsheet`). Second, the aspect aggregation rule applied in this research could be improved by a machine learning method to produce more accurate results. For future research, co-occurrence analysis could also be applied to other aspect categories and non-fast food restaurants could also be analyzed and compared to gain further insights into placeness. Another direction for future research is to study restaurant perception using other methodologies such as surveys or interviews to examine

whether different demographics and restaurants outside our study area show consistent results. Furthermore, future research could also use other review data such as TripAdvisor to study perceptions of places other than restaurants. With this being said, the research in presented in Chapter 2 provided a way to study placeness at a fine-grained level through user-contributed text data over wide geographical area.

Chapter 3: A Thematic Similarity Network Approach for Analysis of Places Using Volunteered Geographic Information [1]

The research presented in Chapter 3 proposes a thematic network approach to explore rich relationships between places. I connect places in networks through their thematic similarities by applying topic modeling to the textual volunteered geographic information (VGI) pertaining to the places. The network approach enhances previous research involving place clustering using geo-textual information, which often simplifies relationships between places to be either in-cluster or out-of-cluster. To demonstrate our approach, I use a case study in Manhattan (New York) that compares networks constructed from three different geo-textural data sources—TripAdvisor attraction reviews, TripAdvisor restaurant reviews, and Twitter data. The results showcase how the thematic similarity network approach enables us to conduct clustering analysis as well as node-to-node and node-to-cluster analysis, which is fruitful for understanding how places are connected through individuals' experiences. Furthermore, by enriching the networks with geodemographic information as node attributes, I discovered that some low-income communities in Manhattan have distinctive restaurant cultures. Even though geolocated tweets are not always related to place they are posted from, our case study demonstrates that topic modeling is an efficient method to filter out the place-irrelevant tweets and therefore refining how of places can be studied.

[1] This chapter is based on: Yuan X., Crooks, A.T. and Züfle, A. (2020), A Thematic Similarity Network Approach for Analysis of Places Using Volunteered Geographic Information, *ISPRS International Journal of Geo-Information*, 9(6), 385, https://doi.org/10.3390/ijgi9060385.

3.1 Introduction

Place and space are among the most fundamental concepts of geography [124, 125]. Space is often considered to be points of locations represented by coordinates. Place, on the other hand, is an "experience-based dynamic construct" [126]. Compared to space, the concept of place emphasizes on the meaning-making process that is complex, dynamic, and individualistic [127]. In Chapter 3, I study how different places are semantically similar, based on textual topics that appear in Volunteered Geographic Information (VGI) in these places. Our goal is to create a thematic similarity network that connects places of similar topics regardless of their physical distance. By applying a network clustering algorithm, I find groups of semantically similar places and analyze their topics and spatial autocorrelation qualitatively and quantitatively.

Analyzing and theorizing about places from a variety of perspectives, has a long history in geographical analysis—from social area analysis [128–131] to more recent geodemographic analysis that derives collective behaviors and characteristics from demographic data of geographic regions [132, 133]. In the past decade, studies on place have taken advantage of a new data source, that of VGI [134]. VGI comes in many different forms, from that of maps created by users to text from Wikipedia which has a geographic component (e.g., place names) [21]. The research presented in Chapter 3 uses the textual form of VGI, specifically crowdsourced reviews from TripAdvisor and geolocated Twitter data (see Section 3.3). Such platforms provide large amounts of textual data with either explicit or implicit geographic information contributed by users [135, 136]. Leveraging this unstructured geographical information found in such texts allows us to comprehend the complexity of places at scale [137–139].

Generally speaking, the most common method utilized by prior research to analyze geotextual data is to structure the unstructured texts into themes (i.e., topics) through topic models (e.g., [140]). This is then often followed by applying clustering algorithms (e.g., k-means) to expose the underlying patterns of sentiments, experiences, or activities captured

in the text (e.g., [21, 141–147]). When places are clustered for further analysis, however, those in the same cluster are assumed to be carrying similar characteristics. Relationships between places are reduced to being either in-cluster or out-of-cluster. However, I would argue that the connectedness and relationships of places, in reality, are more complex. For instance, when connecting places in a network, places at the edge of their own clusters still have relatively weak out-of-cluster connections. The network approach presented in Chapter 3, recognizes them as places with both in-cluster and out-of-cluster connections. Thus, this approach does not limit us to only perform network-level place clustering, but also to discover unique places based on their positions in the networks. To highlight this I use a case study to demonstrate the approach in the context of Manhattan, New York. In the remainder of Chapter 3, I will first discuss related research pertaining to topic modeling and thematic similarity network analysis (Section 3.2). This is followed by introducing the data (Section 3.3) and the methodology (Section 3.4) that I applied to our case study. The results are then presented in Section 3.5 and finally, the implications and conclusions of our research are presented in Section 3.6.

3.2 Related Work

The approach proposed in Chapter 3 involves two major steps, topic modeling using geo-textual data and thematic similarity network analysis. In what follows I review related work with respect to these steps. For step one, topic modeling is a widely used language model for understanding large amounts of unstructured textual data. Previous research has adopted generic topic modeling algorithms (e.g., [140]) to ones that incorporate geographic information (e.g., [148, 149]). Utilizing these geographical topic models, studies have been able to derive the topics from travel blogs and Flickr tags to specific geographical units, such as states [148, 149]. Other work has analyzed the relationships between topics and countries from online news articles and blogs [150], or generated activity patterns from

check-in data [146]. In addition, topic modeling has been used to recommend travel destinations using travel blogs [151, 152], create location related question-answering systems using Twitter and blogs [153], and predict the future distribution of topics [154]. Despite research on innovating geographic topic models, many researchers often chose to use generic topic models (such as Latent Dirichlet Allocation, LDA [140]) to analyze geo-textual data. In such instances, the geolocational information in the text does not contribute to the results of the topic models but is used only after applying the model. For example, Adams et al. [143] explored the temporal themes related to places using travel blogs and applied a similarity score between places based on the topics. Jenkins et al. [63] compared themes of geographic areas from Twitter and Wikipedia.

In addition, previous work has defined "place" at various levels of aggregation including countries, cities, neighborhoods, buildings, but such aggregations artificially split geographical areas. For example, at the neighborhood level, Cranshaw et al. [47,155] detected boundaries of neighborhoods using check-in data and Foursquare venue descriptions in order to show that crowdsourced and official neighborhood definitions differed. At a more aggregated level, Preoţiuc-Pietro [61] viewed cities as collections of Foursquare venues and clustered cities hierarchically using venue descriptions to show that similarities between cities can be captured through crowdsourced data. Since Foursquare venue data also provides venue categories, Noulas et al. [156] clustered both geographic areas (in terms of 625×625 square meters) and users based on their visit history in order to enhance recommendation systems for different users. In another work, Crooks et al. [21] proposed a multi-level (individual building, streets, and neighborhoods level) approach for discovering social functions through mining place topics. Clustering at different aggregations also allows us to find places where people share similar experiences [143, 157] along with places with similar functions [21]. When applying clustering, however, the relationship between places becomes binary, being similar or not similar, and thus the relationships between places in *different clusters* are often ignored.

Turing to work pertaining to thematic similarity network analysis (i.e., the second step

of our approach), previous studies have analyzed place similarities but rarely used a network approach in the context of geo-textual data. For example, Janowicz et al. [158] used semantic similarity for developing geographic information retrieval applications. While Yan et al. [159] trained word embeddings for place types that was then used for exploring similarity and relatedness between point-of-interests types. In terms of using a similarity network based approach, Quercini and Samet [160] created graph-based similarity measures to address spatial relatedness of a concept to a location using Wikipedia articles. In another work, Hu et al. [161] placed cities into networks based on their semantic relatedness (i.e., number of news articles which contain the co-occurrences of the two cities). Similarity networks, however, have seen much wider applications in domains outside of geography, ranging from analyzing protein sequences and structures [162], genome data [163] to that of hospital patients [164, 165]. Methodologically, such studies have demonstrated that one of the most important analysis for similarity networks is clustering (i.e., community detection), which captures groups of nodes that are most similar to each other. Although place clustering does not require connecting places in networks, one of the advantages of conducting network-based clustering is that it enables for downstream node level analysis in relation to clusters. For example, Valavanis et al. [162] discovered structural similarities of protein folds and classes in the downstream analysis after carrying out network clustering. Similarly, in the case study presented throughout the rest of Chapter 3, I will apply clustering to the similarity network as well as identifying special nodes (i.e., places) based on their positions in the network.

3.3 Data

3.3.1 Data Collection

To apply our methodology (see Section 3.4) to showcase how a network approach can be used to study place, data was needed to be collected. In this study I used two geo-textual data sources: TripAdvisor and Twitter. The rationale for choosing these data sources

are twofold. First, they are open source that have been widely used by previous research (as discussed in Section 3.2). Therefore, future studies could use these data sources to extend the research presented in this chapter. Secondly, most prior research using geo-textual data often choose only one data source. In the research presented in this chapter, we aim to provide a thematic similarity network approach which can compare multiple geo-textual data sources. The TripAdvisor data was collected in September 2019 which included reviews for attractions and restaurants in New York City. For each attraction and restaurant, the addresses, neighborhood, and reviews were retrieved. An example of this is shown in Figure 3.1, in which I highlight content that was used in our analysis (i.e., locational information and reviews). With respect to Twitter, I was only interested in tweets that had a precise geographical coordinate. The Twitter data that was collected from January 1, 2015 to December 31, 2015 with a bounding box of latitude ranging from 40.481867 to 40.9325 North and longitude between -74.2721 and -73.626201 West, which includes the New York City.

3.3.2 Data Pre-processing and Aggregation

As the locational information from TripAdvisor attractions and TripAdvisor restaurants was in the format of addresses, the first step of data pre-processing was to geocode TripAdvisor data addresses using Google Maps Geocoding application programming interface (API) [166]. After all the data was geocoded, the second step was to define what a place is. As was discussed in Section 3.2, previous studies have treated places at various levels of aggregation. For this case study, a place is a census tract defined by the United States Census [167]. Although the aggregation may result in the modifiable areal unit problem that statistical summaries of the aggregated area are influenced by the shape and size of the area [168], the reason of using census tract in this research was to incorporate the census demographic data into the analysis. The rationale for this was to be able to explore the connection between the patterns found in crowdsourced reviews (or tweets) and the underlying geodemographics of an area. Furthermore, by using census data, while not only

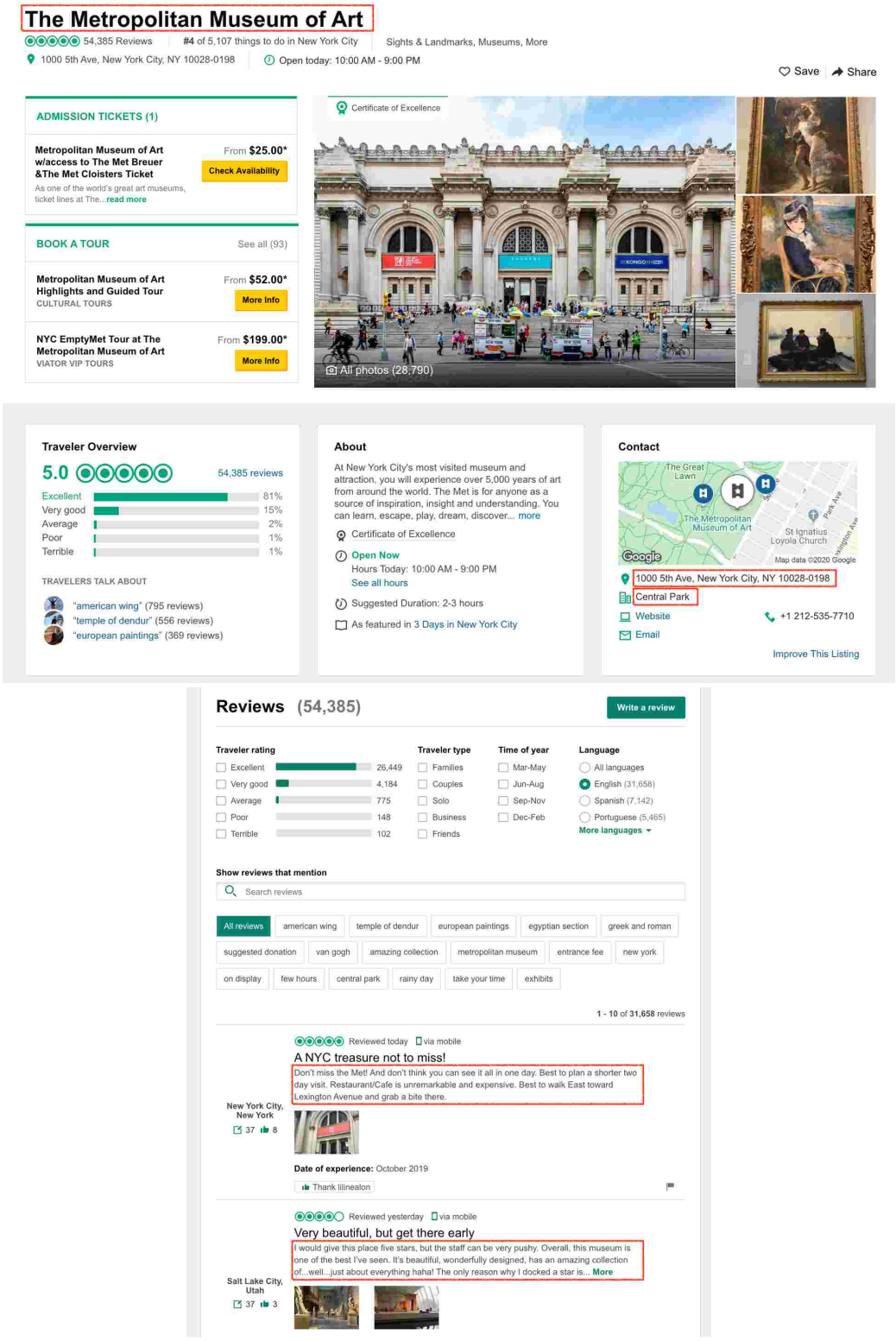


Figure 3.1: An example of a TripAdvisor page and the highlights are the information scraped from the page.

demonstrating how our case study allows for a novel approach to studying places through thematic similarity networks, but it also allows for others to use it in different areas within the United States or in other countries where census data is available (e.g., as in the United Kingdom). It should be noted however, if readers are not interested in comparing the geo-textual data to census data, our approach could be applied to other levels of aggregation such as grids, road segments etc. (see [21]).

After aggregating data to the census tract level, the final step was to select only tracts that appeared in all three datasets to make them comparable. It was found that the majority of the attractions within New York City from TripAdvisor were located in Manhattan, and thus for the analysis in Chapter 3, the tracts only in Manhattan were analyzed. Table 3.1 shows the number of restaurants/attractions, reviews/tweets, and census tracts after restricting the study to Manhattan. Furthermore, the texts (i.e., reviews and tweets) were filtered to only be those that were in English. Although special characters such as "@", emojis, and stop words may contribute to the meaning of the text [169], we do not consider them in this work as is common in text pre-processing (e.g., [170]). Next all words were converted into lower case in order to treat all words with the same text the same. Finally in order to reduce the number of vocabularies (e.g., words with the same meaning such as walking, walk, walked), a stemmer (i.e., Porter Stemmer) was applied and only the stems of the words were retained [171].

Table 3.1: Statistical summaries of three datasets after pre-processing.

Dataset	**Count**	**Number of Reviews**	**Number of Tracts**
TripAdvisor attractions	956 (attractions)	446,747	210
TripAdvisor restaurants	7,946 (restaurants)	865,055	210
Twitter	268,224 (users)	2,009,498	210

3.4 Methodology

In this section, we will first introduce how the topic model was trained on each dataset and how the thematic similarity network is constructed based on the similarities between the derived topics (Section 3.4.1). Figure 3.2 illustrates the work flow from data input to thematic similarity network output including data collection and prepossessing which was described in Section 3.3. After the thematic similarity networks are constructed, we carried out the network community detection on these networks (Section 3.4.2) and node level network analysis (Section 3.4.3). Finally the algorithm and implementation is presented in Section 3.4.4.

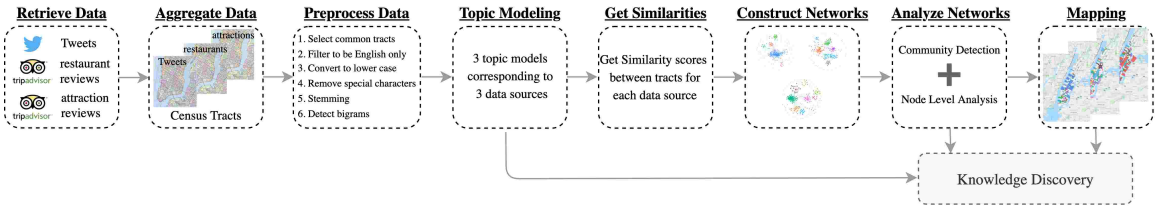


Figure 3.2: Work flow from data input to the construction of the thematic similarity network and analysis (i.e., community detection and unique nodes discovery).

3.4.1 Topic Modeling and Thematic Similarity Networks

As noted in Section 3.2, The first major step towards gaining a meaning from large collections of text is topic modeling. Topic modeling is a type of statistical model that discovers latent topics in documents. For example, when writing a review, the reviewer might not be thinking of specific topics, but topic modeling assumes that there are underlying topics, which are known as latent topics. One of the most widely used topic models is Latent Dirichlet Allocation (LDA) [140], which is a generative probabilistic model that treats each document as a distribution of latent topics and each topic as a distribution of words. A document can be a news article, a review or a social media post. In this research, each document is comprised of all the texts for each census tract from one dataset. For example, for the TripAdvisor restaurant dataset, tract "36061000700" had 57 restaurants, which had

total of 1951 reviews. The 1951 reviews were treated as one document in the LDA model.

One LDA model is trained for each of the three datasets. To train LDA models, I need to tune the hyper-parameters using the datasets to obtain models that have the best performance. A LDA model can predict the words in each topic and the topics are in each document. Therefore the hyper-parameters that need to be tuned include include the a-priori probability vector α that maps each topic to a probability, and the a-priori probability vector β that maps each word to a probability. Moreover, the total number of topics (K) need to be learnt from the data as well. The model was implemented using gensim LDA library, in which alpha and beta was set as "auto" so that both hyper-parameters can be learned from the data [172]. To ensure a better model interpretability, I detected the bi-grams in the texts, after which the corpus became a mixture of uni-grams and bi-grams. In addition, the vocabularies in the corpus was truncated since otherwise many of the most frequent words that bears less concrete meanings in the context of our tasks, such as "I" and "is", would become the top words of the topics. However, the threshold of the word frequency (top_n) to be truncated is a parameter of the model that needs to be tuned during training as well. To tune K and top_n, experiments on each dataset were carried out. Since there is no ground truth for topic models, the common model evaluation metrics are perplexity and coherence [173]. However, the experiments showed that optimizing coherence or perplexity scores in all three datasets did not generate models with better topic interpretability (code available at: `https://bitbucket.org/xiaoyiyuan/network_vgi/src/master/script/topic_model_results.ipynb?viewer=nbviewer`). As a result, I adopted interpretability and manual observations as the metrics for evaluating the topic model quality that can be found. For instance, when k is too high, the model produces topics that have many common words, meaning that new topics are not contributing to generating new knowledge about the data. When top_n is too high, more documents have only one or two topics, which makes it hard to comprehend topic meanings. Table 3.2 shows the parameters that produces the most interpretable model for each dataset. Each of the experiments and the results with various values for the hyper-parameters can be found in

the shared source code.

Table 3.2: Parameters of the trained LDA models on the datasets.

dataset	K	top_n
TripAdvisor attractions	30	100
TripAdvisor restaurants	40	500
Twitter	70	700

When the LDA model is trained, each document is represented by a distribution of topics. The square root of Jensen-Shannon divergence is a commonly used metric of measuring distance between discrete distributions. The Jensen-Shannon distance between two (topic) probability distributions P and Q is defined formally as:

$$JSD(P||Q) = \sqrt{\frac{D_{KL}(P||M) + D_{KL}(Q||M)}{2}}, \text{ where } M = \frac{1}{2}(P + Q).$$

The Jensen-Shannon divergence is symmetrical (i.e., $JSD(P||Q) = JSD(Q||P)$). As a result, the edges of the similarity networks are not directed but weighted and the weights are the similarity scores. Using the same method, three similarity networks were constructed from the three datasets. Since there is always a similarity score between each pair of tracts, there is always an edge between them in the networks as well, making the networks fully connected.

3.4.2 Community Detection

Discovering communities of a fully connected network requires network sparsification [174]. Network sparsification reduces number of edges while preserving structural and statistical properties of interest. Thus, the principle is to reduce the network size by retaining only the important edges and in a similarity networks, the edge weight (i.e., the similarity score) is an indicator of edge importance [175]. The cut-off value for edge weights to sparsify

the networks depends on the data and the clustering algorithm. In this work, I used the Girvan-Newman algorithm [176] to conduct network community detection (i.e., clustering) on the sparsified networks. The Girvan-Newman algorithm is a hierarchical method of detecting communities in complex networks, which can also be applied to weighted networks [176]. Within each step of the Girvan-Newman community detection algorithm, it progressively removes edges with highest edge betweenness centrality (i.e., edges with highest number of shortest path passing through them) and recalculates edge betweenness after each iteration of removal. By removing these high betweenness edges, the communities are separated from each other and consequently, the underlying community structure of the network is revealed. In a weighted network, the Girvan-Newman algorithm calculates the edge betweenness as described above, ignoring the edge weights. Then it divides the edge betweenness by the weight of the corresponding edge. As with unweighted networks, the algorithm then removes edges of highest betweenness. The result of the algorithm is a dendrogram which repeats the steps until no edges can be removed or the most ideal communities have achieved (i.e., the highest modularity of clusters). Therefore, I need to cross validate two parameters, the edge weight cut-off threshold for sparsification and the number of iterations for the Girvan-Newman algorithm. The process of sparsifying fully-connected network to community detection is shown in a stylized network in Figure 3.3.

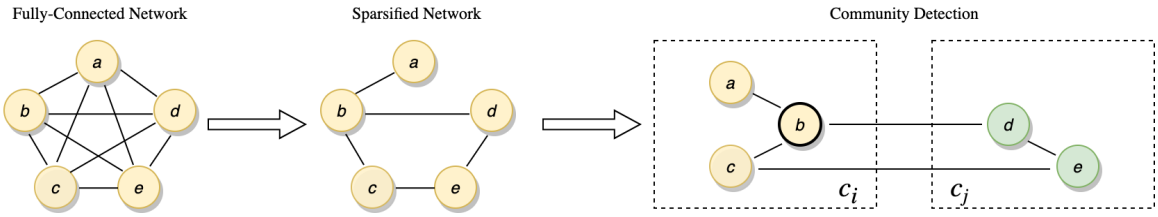


Figure 3.3: A stylized network demonstrating the process of community detection from a fully-connected similarity network.

A common evaluation metric for network community detection quality is modularity, which is a measure of the strength of division of a network into communities [177]. High

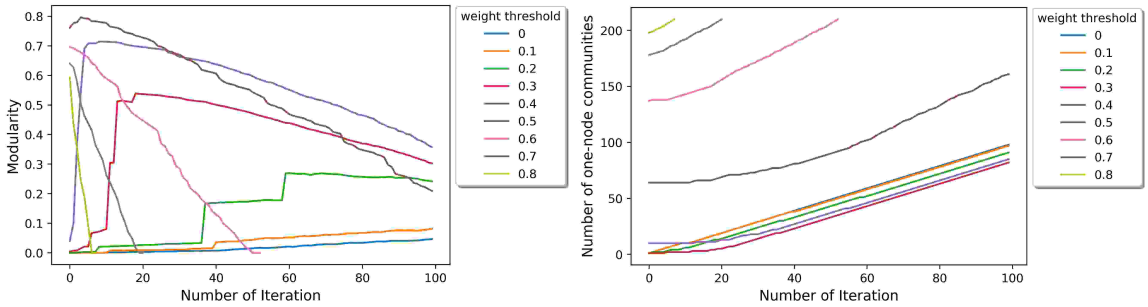
modularity means that within each community detected, there are dense connections within the community and sparse connections between nodes in different communities. However, using modularity as the sole metric for the Girvan-Newman algorithm is not sufficient for our task—Figure 3.4 shows that when modularity is at its highest value, the network has become too scattered with a large number of one-node communities, which hinders the downstream analysis of interactions between communities and relationships between nodes and communities. To mitigate this problem, I selected the set of parameters that has the highest modularity without generating a large number of one-node communities. For instance, the highest modularity for TripAdvisor attraction network (Figure 3.4a, left, brown) is 0.8 at iteration 8 but it produced more than 50 one-node communities (Figure 3.4a, right, brown). However, choosing a model with a slightly lower modularity value, e.g., 0.7 (Figure 3.4a, left, purple) significantly reduced the number of one-node communities (Figure 3.4a, right, purple). Using the same heuristics, the best parameters for each network are presented in Table 3.3.

Table 3.3: Parameters utilized for community detection in the three networks.

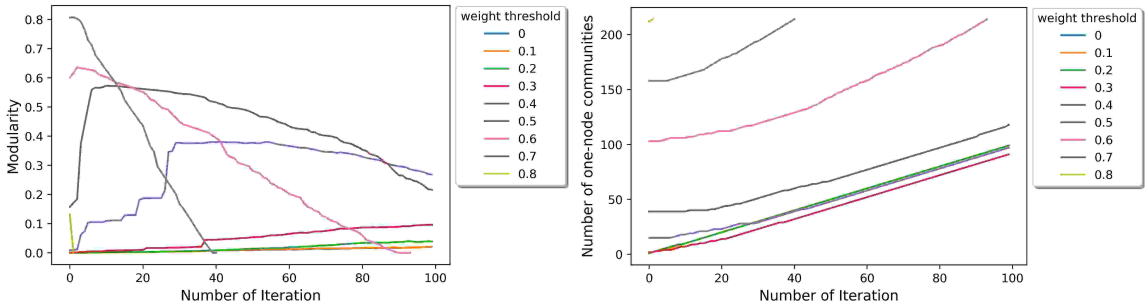
Network	**Iteration**	**Weight Threshold**	**Modularity**
TripAdvisor attractions	8	0.4	0.714
TripAdvisor restaurants	10	0.5	0.573
Twitter	19	0.7	0.365

3.4.3 Discovering Unique Nodes

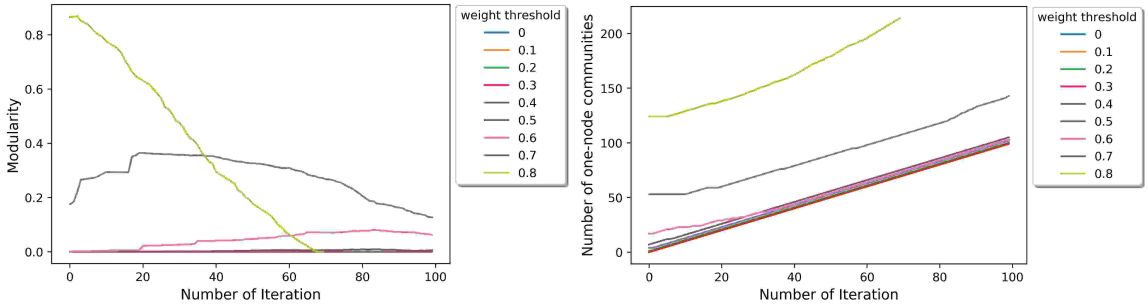
Since the edge weights represent similarity, a node (i.e., a place) with a low degree centrality value means that it bears low similarity to other nodes in the network. It is straightforward, therefore, to discover the ends of the uniqueness spectrum—on the one end, the highly unique places are nodes with a low degree centrality while at the other end high degree centrality nodes are the least unique places. Other than these two extremes, there are nodes that act as bridges between communities that carry their unique characteristics, i.e.,



(a) TripAdvisor attractions thematic similarity network.



(b) TripAdvisor restaurants thematic similarity network.



(c) Twitter thematic similarity network.

Figure 3.4: Cross validation results for community detection in three networks, modularity (left) and number of one-node communities (right).

the community boundary nodes. The concept of community boundaries is rarely applied in similarity network analysis but is often used in social network analysis. In social networks, the community boundaries are the people who convey outside information to those in the community with no out-of-community connections [178, 179]. I adopted and modified the definition of community boundaries from the social network analysis by Guerra et al. [178].

This modified definition of a community boundary is that of a node v that is a boundary node of community C_i for community C_j when:

1. node $v \in C_i$ has at least one edge connecting to community C_j and
2. all the neighborhoods of v have no edge connecting to community C_j.

The community boundaries are identified in the three sparsified networks instead of the original full-connected ones because identifying boundary nodes relies on the community structures detected in the sparsified network. Finally, the last step in Figure 3.3 illustrates a stylized network with community boundaries. For community C_i, node b and node c both have edges connecting to the outside community C_j. Node b qualifies as a boundary node for community C_i because it has a neighbor (node a) having *no* edges connecting to community C_j. Since node c does not have a neighbor meeting this requirement, node c does not count as a community C_i's boundary node to C_j. In social networks, the definition of boundary nodes guarantees that node b is the only node that brings outside information to node a. In the context of place similarity networks, a node having no connection with the outside community (i.e., node a) indicates it has characteristics that are unique to its own community and the boundary nodes (e.g., node b) are the ones that connect the uniqueness of the communities.

3.4.4 Algorithm and Implementation

To summarize what has been discussed above with respect to our methodology, the pseudo-code for it is described in Algorithm 1. The algorithm takes one data source (e.g., Twitter corpus or TripAdvisor) as an input. The loop in Lines 1- 6 constructs topic models from the input texts and Lines 7- 11 calculate similarities between each document of the input. Line 12 constructs a thematic similarity network (which was explained in Section 3.4.1). Finally, Lines 12-14 detect communities in the network (Section 3.4.2) and the loop from Lines 15 to 22 discover boundary nodes (Section 3.4.3). The complete Python code and

information pertaining to the software versions is available at `https://bitbucket.org/xiaoyiyuan/network_vgi`.

Algorithm 1: Network Construction, Community Detection and Boundary Node Detection

Input: Corpus split by their geolocated census tracts $D = d_1, d_2, ..., d_n$

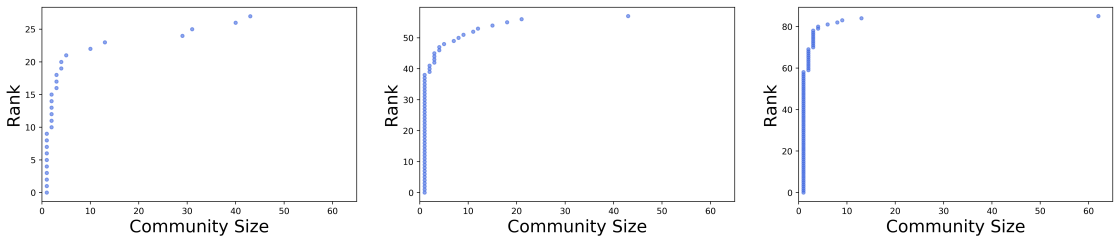
foreach d *in* D **do**
  $pairs \leftarrow [\,]$
  $pair\_similarities \leftarrow [\,]$
  /* TM maps topic ID t and words w (from document D) to a probability */
  $TM(t, w) \leftarrow topic\_model(D)$
  $d.topics = topic\_model(d)$
end
foreach $d \neq d'$ *where* $1 \leq d, d' \leq n$ **do**
  $pairs$.insert($[d, d']$)
  /* Jensen-Shannon Distance */
  $distance \leftarrow$ JSDistance($d.topics, d'.topics$)
  $pair\_similarities$.insert($distance$)
end
$G = (D, D \times D, pair\_similarities)$
/* Sparsify the graph by pruning edges having low similarity */
$G \leftarrow$ sparsify(G)
/* Girvan-Newman Community Detection */
$C \leftarrow$ girvan_newman(G) /* $C = c_1, ..., c_{|C|}$ */
$boundary\_nodes \leftarrow [\,]$
foreach d_i, d_j *in* D **do**
  /* d_i and d_j are from different communities and are connected */
  $condition1 = d_i \in c_i \wedge d_j \in c_j \wedge d_i \neq d_j \wedge d_i.has\_edge(d_j)$
  $condition2 = d_i neighbors.has\_no\_edge(c_j)$
  if $condition1 \wedge condition2$ **then**
    $boundary\_nodes$.insert(d_i)
  end
end
return $G, C, boundary\_nodes$

3.5 Results

Building upon our methodology, in this section, I will present the results for the thematic similarity network analysis of places in Manhattan, New York. Specifically, Section 3.5.1 maps and visualizes the major network communities and their topics and presents results from the spatial autocorrelation of these communities using Moran's I measure of spatial autocorrelation(Section 3.5.1). In section 3.5.2, I enrich the network nodes with geodemographic data and finally in Section 3.5.3 I identify and analyze nodes by their degrees of uniqueness.

3.5.1 Major Network Communities and Their Topics

In this section, I evaluate the clusters found using our proposed community detection approach described in Section 3.4.2. For this purpose, I first visualize and qualitatively analyze the community clusters. Then to ensure that the communities that I found are clustered significantly, I test each community for spatial autocorrelation using Moran's I. The sizes of the communities are show in Figure 3.5. Even though I lowered the number of one-node communities in the community detection (as discussed in Section 3.4.2), the distributions of the community sizes still appear to be long-tailed.



(a) TripAdvisor attractions (b) TripAdvisor restaurants (c) Twitter

Figure 3.5: The sizes of communities from the community detection results of the three networks.

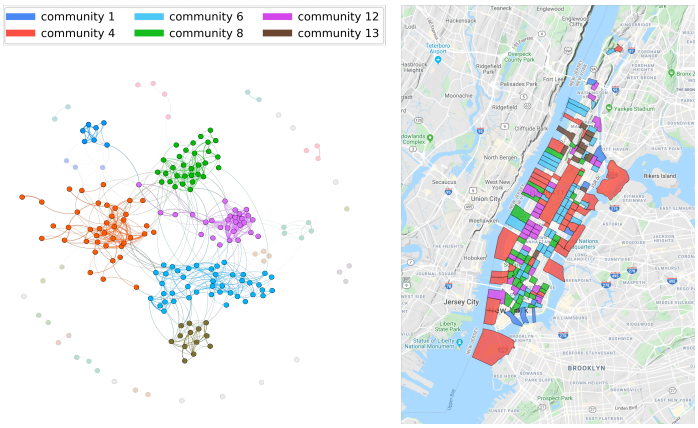
For the sake of clear visualization, only the major communities (i.e., communities with a size equal or larger than 5 nodes) from the community detection are presented for each

network. The topic modeling results for all communities are available online at `https://bitbucket.org/xiaoyiyuan/network_vgi/src/master/script/topic_model_results.ipynb?viewer=nbviewer`.

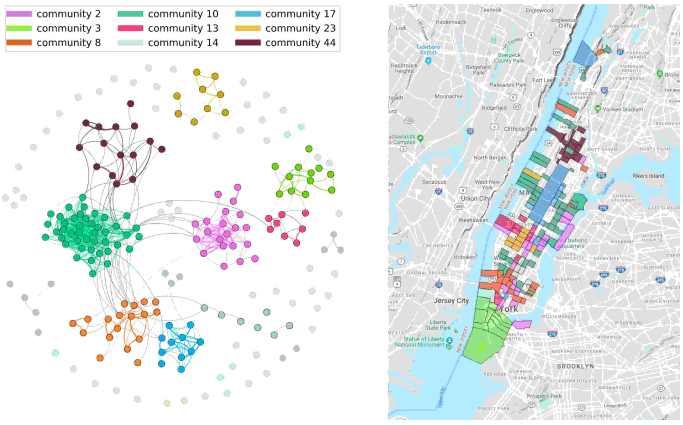
Network Visualization and Mapping

Figures 3.6a and 3.7a shows the visualizations of networks, maps, and topics from the community detection results of the TripAdvisor attractions thematic similarity network. Major communities are highlighted and tracts of the major communities are mapped in the same colors. In Figure 3.7a, dominant topics (i.e., topics with coefficients equal or higher than 0.1) of the major communities are shown.

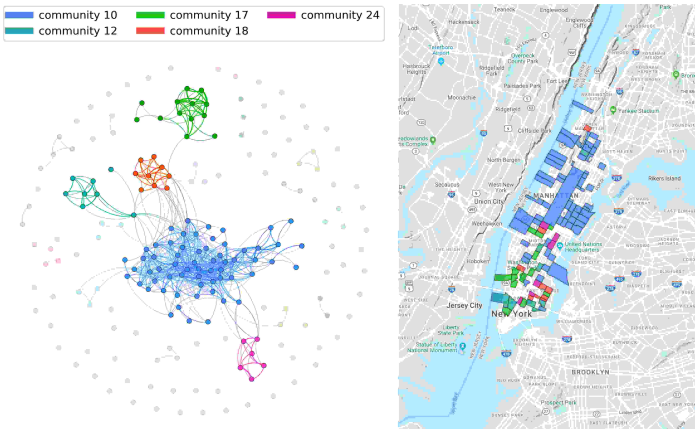
Based on the words in the topics, communities can be characterized into categories such as church tracts in Harlem (Community 13), restaurant tracts that includes two famous restaurant areas in Chelsea and Chinatown (Community 8), bridge tracts (Community 1), and theater tracts (Community 12) while other communities have more hybrid characteristics (i.e. topics). Observing the combination of the topics and their locations on the map (Figure 3.6a), some communities have tracts which are visually close to each other and their topics reflect the main characteristics of the attractions in these geographic regions. For instance, topics of Community 12 are about "broadway", "theater", "concert", and "venu" (venue) and most of these tracts are clustered around the Broadway theater district. Furthermore, as shown on the map (Figure 3.6a), not all communities are not clustered perfectly in a geographic region and some of the tracts of a community are in the same region. For example, even though most of the tracts of Community 12 are located Midtown, the rest of the tracts are scattered around the Downtown area. The reason is that the topics of Community 12 include not only Broadway but also more broadly "concert", "game", and "venu" (venue) (Figure 3.7a). A similar example is that of Community 13 that has a dominant topic with keywords "harlem", "church", and "theater" and most of the tracts of Community 13 are located in Harlem and tracts that are not in Harlem have church related attractions such as The New York Mosque in Midtown Manhattan and



(a) TripAdvisor attractions



(b) TripAdvisor restaurants

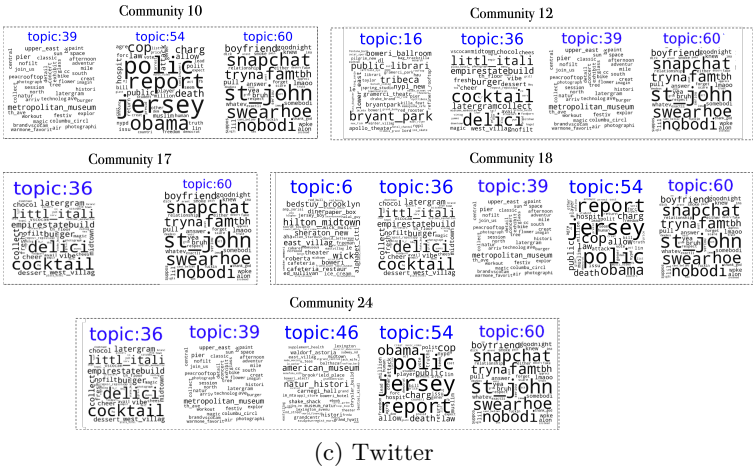


(c) Twitter

Figure 3.6: Network visualization of all communities from the thematic similarity networks using Gephi [1] with major communities highlighted. Only the major communities are shown on the map for the sake of clarity. Major communities in Network visualization and mapping for each network are colored the same and thus the legend applies for both.

(a) TripAdvisor attractions

(b) TripAdvisor restaurants



(c) Twitter

Figure 3.7: Dominant topics of all major communities in each thematic similarity network. Dominant topics are topics with coefficients equal or higher than 0.1.

Mariners' Temple Baptist Church in Downtown Manhattan. Such findings indicate that the network communities are reflections of people's similar experiences of various attractions as they are mined from a large amount of crowdsourced reviews from individuals. For the restaurant thematic similarity network, communities show higher level of spatial proximity (Figures 3.6b and 3.7b). One of the most prominent of such is that of Community 3, which is shown in Figure 3.6b clustered in Downtown Manhattan. Primary topics of community 3 (Figure 3.7b) are "pub" and "eatali" (Eataly food market), and "financi district" (financial district). In addition, tracts of Community 8 have close geographic proximity as well. This is evident from the map of Figure 3.7b, where most of the tracts in Community 8 are located between Downtown and Midtown Manhattan. Community 8 has Topics 14 and 32 featuring word stems such as "greenwich_villag" (Greenwich Village), "west_villag" (West Village), "japanes" (Japanese), and "bagel". Similarly, Community 17 has Topic 36 that can be interpreted as Central Park related even though it has the common Topic 32 that shows up across many other communities (e.g., Community 2, 8, 10, 14, and 17). Interestingly, communities from TripAdvisor attractions network have counterparts from the restaurants network communities. For example, Community 13 from attraction network and Community 44 from restaurant network are about Harlem, which can be seen from the geographic clusters on the map and their dominant topics. A similar finding is for the theater district, which appears in both Community 12 of attraction network and Community 23 of the restaurant network. This suggests that people's dining experiences can be intertwined with the characteristics of the surrounding attractions or vice versa.

Turning to the results of the Twitter thematic similarity network, one of the most noticeable pattern is that of Community 10 (i.e., the blue community in Figure 3.6c) which dominates this network. Unlike the communities in Trip Advisor attractions and restaurants where there is a more even distribution of community sizes. Furthermore, communities from the Twitter dataset have more diverse topics than that of restaurant and attraction networks from TripAdvisor (Figure 3.7c). This could partly due to the distinction between Twitter and TripAdvisor as data sources for studying places. In that TripAdvisor reviews

are directly about places but this is not necessarily the case for Twitter, which is a more generic social media platform where users can contribute a whole variety of topics [135]. Therefore, some topics (e.g., Topic 54 and Topic 60 in Figure 3.7c) from Twitter are not about places but relate to news or social and political discussions. This indicates that although geolocated Tweets can be used to study people's perceptions and experiences about places, it needs to be used with awareness that the texts may need to be filtered. The results in Figure 3.7c shows that it is viable to use topic modeling to filter out the non-related topics (e.g., Topic 54 which relates to police reporting and New Jersey). The reason could be that tweets pertaining to social discussions often use different vocabularies than texts directly about places. Since topic modeling is a bag-of-words approach, the model is sensitive to vocabularies and thus is able to "tell them apart" as separate topics.

Quantitative Test for Spatial Autocorrelation of Communities

In the previous section (Section 3.5.1), I discussed network communities and whether the communities have geographically proximate tracts. In this section, I will present the results of Moran's I measure of spatial autocorrelation to quantify geographical proximity of the major communities in Table 3.4 and the Moran's I results for all communities are found in Table A.1 in the appendix. Moran's I is a measure for spatial autocorrelation that is often applied on continuous data. To measure each community's autocorrelation level, I therefore encoded tracts of a specific community as 1 and all the other tracts as 0. I defined neighborhood using Queen's contiguity, that is any polygons (i.e., tracts) that shares a point-length border are neighbors.

Among the three networks, all of the major communities from the TripAdvisor restaurant network have statistical significance in their spatial autocorrelation results (Table 3.4). Communities 3, 8, 10, 23, and 44 of the restaurant network have spatial autocorrelation at the 99% confidence interval, which are generally about pubs, west village, vegetarian Indian, theater, and Harlem respectively. These results inform us that these geographic

Table 3.4: Moran's I spatial autocorrelation of major communities in each network.

Network	Community ID	Moran's I	P value
TripAdvisor attraction	1	-0.022740	.413
	4	0.120743	.014*
	6	0.142676	.006**
	8	0.002285	.410
	12	0.107151	.014*
	13	0.106869	.034*
TripAdvisor restaurants	2	0.120561	.017*
	3	0.682913	.001***
	8	0.223211	.001***
	10	0.174987	.001***
	13	0.112362	.026*
	14	0.167896	.010**
	17	0.255313	.002**
	23	0.345028	.001***
	44	0.572042	.001***
Twitter	10	0.328338	.001***
	12	0.056188	.085
	17	0.118175	.020*
	18	0.194717	.007**
	23	0.019150	.258

*Significant at $p \leq 0.05$; **Significant at $p \leq 0.01$; ***Significant at $p \leq 0.001$.

clusters in Manhattan have their own restaurant culture, manifested by the topical summaries from TripAdvisor reviews. Other major communities from the restaurant network (i.e., Communities 2, 13, 14, and 17) also show statistical significance (at a confidence interval of over 95%) in their spatial autocorrelation results. These communities have relatively lower scores from Moran's I test, which can be observed on the map as they are more spread out over Manhattan. For the attraction network, Communities 4, 6, 12, and 13 have spatial autocorrelation and have topics pertaining to zoo/High Line/kid, cathedral/gallery, Broadway/seat/venue, and Harlem/Broadway, which can summarized from word clouds in Figure 3.7a. Therefore, attractions from these communities are more of a mixture of many different topics, which also explained the reason of the Moran's I for the attraction network being relatively lower than that from the restaurant network. Similarly, for the Twitter

thematic network, besides the biggest community (i.e., Community 10), the others have low Moran's I values which suggests that the topics discussed on Twitter are less correlated with their geographic locations.

3.5.2 Enriching Network Communities with Geodemographic Attributes

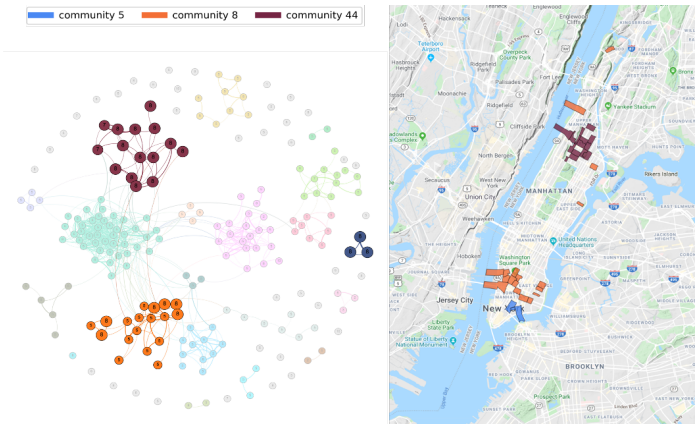
One advantage of examining places as the Census tracts is to combine Census demographic data with the results from the derived networks. If I was to use the demographic data from the US Census such as the American Community Survey (ACS), a tract can be described by multiple variables (e.g., total population, mean household income, education attainment, and marital status). An alternative is that proposed by Spielman and Singleton [180] who took the ACS data and clustered it to generate a single variable description known as a geodemographic classification (e.g., "Hispanic and Kids" and "Wealthy Nuclear Families"). I enriched our node attributes with this geodemographic classification at the tract level in order to explore the relationship between the network communities and their demographics.

Table 3.5: Geodemographic distributions of tracts (i.e., network nodes).

	Type	Type Description	Percentage
High Income	2	"Wealthy Nuclear Families"	1.87%
	5	"Wealthy, urban without Kids"	68.22%
	7	"Wealthy Old Caucasian"	2.80%
Low Income	8	"Low income, mix of minorities"	22.90%
Others	10	"Residential Institutions"	1.40%
	3	"Middle Income, Single Family Home"	0.47%

Note: the geodemographic type and type descriptions are from research by [180]. The descriptions are abbreviated to give the reader a sense of the classification schema and only the types found within our study area are shown. The percentages are based on all tracts across all network communities in our study area and thus are used as baseline for defining if network community is predominantly high-income or low-income.

Based on the results from [180] shown in Table 3.5, most of the tracts that I study are classified as wealthy and the column "Percentage" shows percentages of tracts in that



(a) Network visualization of all communities and mapping of major communities (colored the same as Figure 3.6b). The node label represents their demographic classification.

(b) Word cloud of topics in major communities

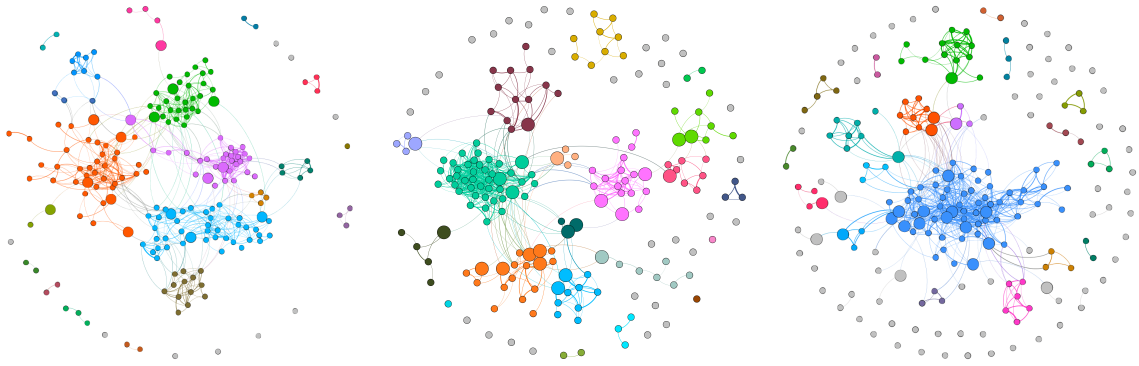
Figure 3.8: Low-income communities highlighted. Node labels represent the geodemographic type. Topics of low income communities are in visualized (b).

demographic classification. I use it as baseline to compare the percentage of each demographic classification for network communities. For instance, if a community has more than 22.90% of Demographic Type 8, based on Table 3.5, I define that community to have a high proportion of low income residents. Using this baseline, I discovered that even though Manhattan tracts are mostly rich and the majority of low-income tracts reside in a few network communities, Communities 5, 8, and 44. From the topics of these communities, two of them are in Chinatown and Harlem, presented in Figure 3.8). This suggests that these low-income areas have a distinctive restaurant culture. When applying the same method to the communities from TripAdvisor attractions and Twitter thematic networks, I do not find communities having high percentages of demographic types. This implies that discussions on Twitter and TripAdvisor attractions in Manhattan do not have patterns that correspond

to the characteristics of its residents.

3.5.3 Identifying Nodes with Degrees of Uniqueness

Besides network level analysis, node level analysis allows us to identify important or interesting places. As discussed in Section 3.4.3, nodes with the lowest weighted centrality are the most unique ones and vice versa. In this Section, I will first examine the central nodes (top 5 highest weighted centrality nodes) and the outliers (top 5 lowest weighted centrality nodes), followed by exploring the community boundary nodes in the networks (i.e., nodes that act as bridges between communities that carry their unique characteristics).



(a) TripAdvisor attraction (b) TripAdvisor restaurant (c) Twitter

Figure 3.9: Visualization of the networks and nodes where large node size represents boundary nodes. Communities are colored the same as Figure 3.6).

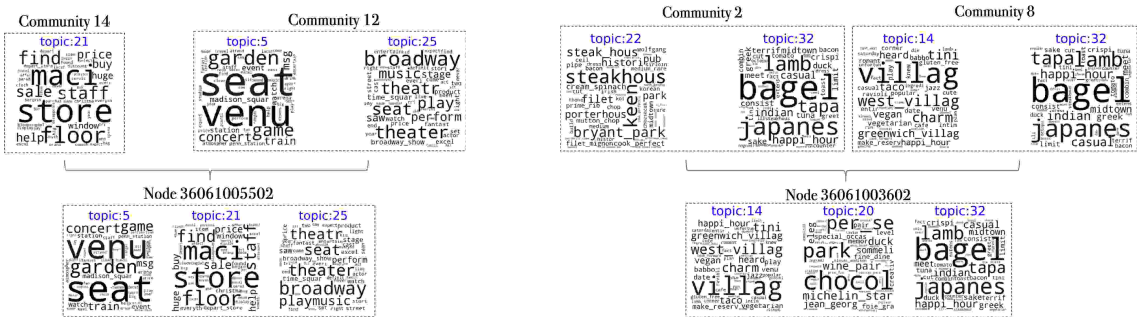
Table 3.6 shows the topics for the central and outlier nodes in the network of TripAdvisor restaurants. Observing the number of topics for the two kinds of nodes, central nodes tend to have more diverse topics than the outliers. The topics of the outlier nodes show that these are the tracts with attractions that are unique to Manhattan, including “Skylin” (Skyline), “rockefel_center” (Rockefeller Center), “time_squar” (Time Square), “grand_central” (Grand Central Station), “Statu” (Statue of Liberty), and “elli” (Ellis Island). Being unique and distinctive, the outlier nodes have very low weighted degree centralities. This pattern

of low degree centraliy nodes having distinctive topics also applies to the thematic similarity network from Twitter and TripAdvisor restaurants. On the contrary, the central nodes have a combination of common topics that enable them to have connections with many other nodes.

Table 3.6: The topics of central nodes and outliers in thematic similarity network of TripAdvisor attractions.

	Central Nodes				Outlier Nodes	
36061012000					36061013700	
36061005000					36061010400	
36061005400					36061007600	
36061016700					36061009200	
36061005502					36061000100	

Figure 3.9 shows the positions of community boundary nodes in the three networks, which to be expected are often at the edges of the communities. Identifying nodes with these special positions facilitates us to identify places with hybrid characteristics from both communities. Community boundary nodes connect the uniqueness between communities. To demonstrate the ways that the topics from community boundary nodes include topics



(a) An example from TripAdvisor attractions.

(b) An example from TripAdvisor restaurants.

Figure 3.10: Two examples of communities with boundary nodes and their respective topics.

from the communities, here I will show two examples from the TripAdvisor networks. First, Figure 3.10a shows an example of the topics and characteristics of a community boundary node and how the boundary node has the topics from two communities (i.e., Communities 14 and 12). However, when two communities have overlapping topics, the topics of the community boundary node are not always the perfect combination of topics from the two communities as shown in Figure 3.10b.

3.6 Conclusion

A place is a geographic location that has individuals' experiences and meaning-making processes from it. As a place has different meanings to different individuals, it is difficult to summarize the collective sense of place (e.g., [127,135,138]). As more and more textual data that describe people's experiences of places are now available online via social media etc., it is possible to study places by crowdsourcing the online textual data from place reviews and geolocated social media. Furthermore, through the utilization of network science and natural language processing of crowdsourced experiences collected from individuals I can also study the connections between places which reveals not only richer information about the place itself and how places relate to each other.

To this end, in Chapter 3 I utilized TripAdvisor reviews and geolocated Twitter data to

understand the characteristics and the connectedness of places. The complex relationships between places were modeled via thematic (i.e., topical) similarity networks. While previous research utilizing crowdsourced data allowed for the clustering of places (e.g., [63]), they do not explore the connections between places and clusters. The network approach developed in Chapter 3 enables us to perform clustering (i.e., network community detection) to discover the network and node level patterns of places. More specifically, the contributions of this research are as follows. First, similar to previous research with place clustering, the network approach of places also allow to perform clustering on places using a network community detection algorithm. The case study in Section 3.5.1 shows that community detected from the thematic similarity network from restaurant reviews tend to have higher Moran's I value (i.e. geographical proximity) than that of attraction reviews and tweets. It suggests that certain geographical clusters correspond to certain restaurant culture in Manhattan (e.g., the bar and pub area at Downtown). Second, by using the network approach (as discussed in Section 3.4), I am able to discover places of interest by exploiting the positions of the places in the network. In the case study shown in Chapter 3, the places of interest are places of different levels of uniqueness (Section 3.5.3). Third, from the TripAdvisor restaurant network results, I found that even though most of the study area in Manhattan is high-income, the low-income communities have a distinctive restaurant culture that the high-income areas do not have (Section 3.5.2). Fourth, I explored implications of using such data for studying places by comparing different datasets, which are Trip Advisor restaurants and attractions reviews and Twitter (Section 3.5.1). TripAdvisor review data represents experiences and perceptions people have directly about places, whereas geolocated Twitter data does not necessarily reflect places. However, as our case study shows, by using topic modeling one can overcome this challenge and filter out place-irrelevant topics, which do not require the time consuming hand labeling process as supervised learning (as shown in Section 3.5.1).

While this study has shown how places are connected through individuals' experiences

and adds to the growing area of geographic data science [181], there are several limitations to this research. First, although the clustering algorithm used in this study (i.e., the Girvan-Newman algorithm [176]), produces deterministic results, it might not be an ideal choice when the networks become larger, say when expanding this research to larger areas. Therefore, researchers who expand this research might want to consider less computationally inexpensive algorithms such as Louvain community detection algorithm [182] which use modularity optimization and has been shown to be scalable [183]. Second, in this study, we define places as census tracts and further analysis is required to test whether some of the results still stand when places are defined otherwise (e.g., zip codes, city blocks etc.). Nonetheless, using census tracts in this research had the advantage of combining textual VGI data with Census data for further analysis (as shown in Section 3.5.2). Turning to future work, other centrality measures (e.g., betweenness centrality, eigenvector centrality) could be explored to discover places of interest other than degree centrality and boundary nodes. Additionally, topic models could be trained by merging data from three datasets so that the topics are comparable across networks. As this work does not take tokens such as emojis into consideration, future work could explore topic models by incorporating them (e.g., [184,185]). The network can also constructed differently with edges representing similarities measured by methods other than topic similarity such as similarity based on users' visit history, which has often been used in collaborative recommender systems [186]. Even with these limitations and potential areas of further work, the research presented in this chapter demonstrates a novel approach of studying places and their connections by combining textual VGI with network analysis.

Chapter 4: Achieving Situational Awareness of Drug Cartels with Geolocated Social Media

Using geolocated tweets to achieve situational awareness is an often researched topic in disaster and emergency management. However, little has been done in the area of drug cartels, which as transnational crime organizations, continue to pose great risk to the stability of our communities. This paper made an initial effort in using geolocated social media via Twitter to achieve situational awareness of drug cartels through spatial and temporal analysis of clusters of named entities. A cluster of named entities contains entities with similar semantic meanings. The results show that detecting peaks in the time series of frequently occurring entity clusters enables us to track important events in regard to drug cartels. Correlations between time series also provide valuable insights into the synchronicity between different events. Further, spatiotemporal analysis allowed us to study the temporal saliency of events in different countries that are being impacted by drug cartel activity. By focusing on prominent peak activity of high frequency tweets, we show important information on events that account for variance in their spatiotemporal patterns. Finally, as previous research using social media to achieve situational awareness argue that one of the challenge of using geolocated social media data is that of working with messy and noisy data, this research also addresses challenges of language ambiguity in the context of achieving situational awareness on drug cartels.

4.1 Introduction

Drug cartel organizations, as transnational crime organizations, are widely recognized as a threat to societies world wide [187]. According to O'Neil [188], over 10,000 drug-related killings between 2006 and 2009 in Mexico alone and from mid-2006 to September 2011,

more than 47,500 people died in drug cartel or gang-related killings [189]. In the 1990s, U.S. law enforcement officials had estimated the annual revenues from cocaine trafficking to be 29 billion dollars a year in just the United States [190]. Therefore, previous research on drug cartels placed a heavy focus on drug related gangs, trade, and corruption in the Americas such as Mexico and Columbia [191–194]. In terms of drug cartel situational analysis, most of the previous research conducts qualitative analysis (e.g., review articles on drug cartel organizational evolutions [195]). Meanwhile, recent studies have also shown that Twitter is one of the most important social media platform that people use to discuss drug cartels and crimes, especially in the countries (e.g., Mexico) where traditional media is censored by drug cartels [196, 197]. The research presented in this chapter aims to propose a pipeline to gain situational awareness using geolocated tweets in regard to drug cartels. Achieving situational awareness (SA) is to gain perception of the elements in the environment through time and space and understand their meaning in order to make decisions for future actions [198]. Based on this widely cited definition from Endsley [198], there are a few key components of achieving SA, which are "elements", "time", and "space". Similar definition of SA in the context of disaster management that organizes information is into place, time, and concept/theme [199]. To adopted these definitions and theories to the context of drug cartel informatics, I define elements of SA for drug cartels to be key events of drug cartels, and achieving SA is to gain perception of the key events through space and time.

Since collecting tweets by keywords often lead to very noisy results [200], I first clustered named entities of tweets using k-means and filtered out the irrelevant clusters of entities. The clusters of relevant entities are therefore, considered as concepts/themes that are analyzed by characterizing spatiotemporal trends of their frequencies. Then to gain better understanding of notable events underlying the spatiotemporal trends, high frequency tweets are queried from peak days of certain geographic regions. In the remainder of Chapter 4, I will first discuss previous related research (Section 4.2) followed by introducing the dataset used in this research (Section 4.3). I will also introduce the methodology used for

named entity recognition, clustering and time series analysis (Section 4.4). Finally I will present the results of querying notable events through identifying peaks of time series of entity cluster frequencies in Section 4.5 and conclude the chapter in Section 4.6.

4.2 Related Work

Previous research on situational awareness, social media, and drug cartels can be categorized into two areas, first is research about drug cartel and social media and the second one is situational awareness using social media data. This section will discuss research in this two area and how these related studies inform the research design presented in Chapter 4.

Research on social media and drug cartels has focused on how citizens and journalists use social media as an alternative for news reporting and information curating in the context of Mexican drug war and censorship posed by Mexican governmental and non-governmental forces (e.g., [196, 197, 201]). While many Mexican journalists choose to write about crimes of drug cartels using more implicit and intentionally ambiguous words, one of the most followed Twitter account that shapes the Mexican public sphere is by "Blog del Narco" that reflect what Mexican citizens discuss about drug cartels [202, 203]. Meanwhile, Roberts et al. [204] proposed that even though social media data as a form of news reporting has many biases, such as internet access and demographic bias, it provides contextual information that helps formal news reporting. Besides journalists, citizens in armed conflict areas also use social media as a form of civic media [197,203] and De Choudhury et al. [205] found that the affective response to armed conflicts showed Twitter users became desensitization to violence reporting. In addition to citizens using social media for acquiring drug cartel information, some studies explored how drug cartels themselves use social media for communication in order to further their criminal activities on Facebook [206]. Womer et al. [207] discovered that self-claimed terrorists, Sureños gang members and Mexican drug cartel members show an organizational culture on YouTube with religious and spiritual tones. For the research presented here, I do not assume that drug cartel members use Twitter nor do I seek to search for information posed by them. However, as previous studies have shown that social media

is used as a crucial data source by citizens and journalists in the context of the Mexican drug war, Twitter has been shown to be a quality data source to achieve situational awareness for drug cartel information [197, 203, 204].

In the area of using social media to enhance situational awareness, there has been a large volume of research carried out in the case of emergency and disaster management [200]. Geolocated social media has been considered to be a form of volunteered geographic information (VGI) [208–210]. VGI has also been shown to be useful for crisis management [70,211]. Previous studies have shown that there is correspondence between real events with the algorithm detected events from Twitter data and linking event detection from Twitter and online news can provide situational awareness [212]. Event detection from Twitter has also been demonstrated to be helpful for overcoming limitations of keyword-based Twitter data collection and thus I can focus on specific pieces of situational information that are valuable for responders [200]. It has been demonstrated that one of the most challenging task within this area of research is on how to extract relevant information from tweets. For example, Vieweg et al. named those who post tweets relevant to situational awareness with geolocations as "high-yield tweeters" [213] and Amirkhanyan et al. [214] aimed to understand which parameters from Twitter data can help us recognize tweets that are providing useful situational awareness information. With identifying the relevant information from Twitter data, various methods have been proposed for knowledge extraction from the tweets, including learning temporal change of emergency situations, how the temporal changes relates to the location of the tweets, and topic discovery to better enhance situation awareness of the emergency and disaster events [215, 216]. Research has also highlighted that the location data from Twitter is an important element of situational awareness, which include geolocation of the tweets (i.e., where the tweets was collected) and geo-reference data (i.e. location and place names mentioned in tweets in the text form) [199, 213]. Other than Twitter data, previous research has shown how other social media platforms, such as Flickr, can inform us on events during crisis [16]. In addition, MacEachren et al. [199] argued that using social

media data for situational assessment in crisis management, the classic approach to organizing information into place, time, and concept components supports foraging and making sense of crisis.

To summarize, even though previous research using Twitter data to achieve situational awareness did not touch on the topic of drug cartels, many studies has also shown that social media data has been a critical channel of communicating cartel related information especially in the case of Mexican censorship on traditional media outlets. In Chapter 4, I use Twitter data to achieve situational awareness of drug cartel related events and information. As related research has shown that the challenges of using Twitter to gain situational awareness is data filtering, in the next section, I will describe the dataset used for this research and the technique and results of filtering tweets that do not contribute to the knowledge of drug cartel situations.

4.3 Data

To gain situational awareness of drug cartels, geolocated tweets were collected using keywords including "sinalo", "cartel", "cartels", "zetas", "ms-13", and "ms13". The geolocation of the tweets mainly comes from three sources—location from Twitter user profile, locations from computer IP address, and locations from the precise GPS coordinates from the tweets and the number of tweets from these categories are 407812, 9236, and 2005 respectively. The data were collected during November 1, 2018 to November 30, 2018. The total number of tweets collected is 828,596, among which more than 95% were from English or Spanish speaking countries. However, minor differences exist between the language spoken by the countries that the tweets were collected from (i.e. Twitter's language setting) and the language that the tweets were written in. I removed languages that are not in English and Spanish speaking countries to focus on the situation of Mexican drug war (Section 4.2). Table 4.1 presents the summaries of the data after removal other languages and the column "Entities" refers to the count of named entities for each language, which will be explained further in Section 4.4. Table 4.1 also shows that many tweets in the data

are not original but retweets from the other users (i.e., number of unique tweets).

Table 4.1: Tweet and entity counts by languages.

Language	Tweets	Unique Tweets	Entities
English	403,742	139,369	456,962
Spanish	384,591	131,790	353,556
Total	788,333	271,159	810,518

4.4 Methodology

In order to mine knowledge and thus achieve situational awareness from the tweets, the methodology used to in this research involves three major steps—named entity recognition (NER), named entity clustering, and spatiotemporal analysis of named entity clusters. The workflow is illustrated in Figure 4.1.

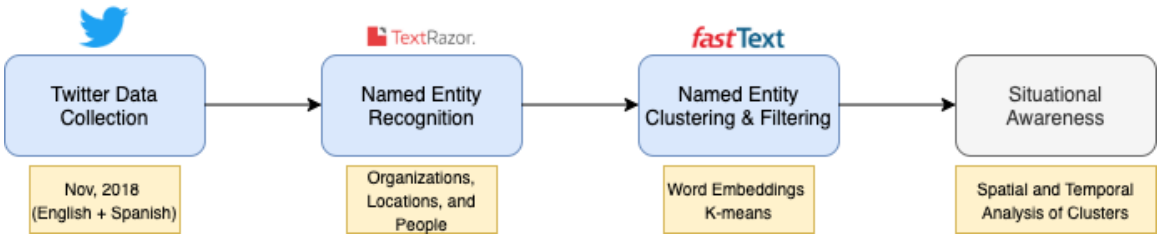


Figure 4.1: The workflow of achieving situational awareness of drug cartels using geolocated tweets.

Named Entity Recognition (NER) is a widely used technique in Natural Language Processing (NLP) to extract structured information from unstructured textual data [217]. Early works of NER aimed to recognize "proper names" in general and overall the most studied types are "people", "locations", and "organizations" [217]. Each of the three entity types can be divided into subcategories such as identifying "politicians" instead of "people" or "city" rather than "locations". As a well-researched area, many algorithms and systems

have been designed for NER tasks. In this research, NER was conducted using an off-the-shelf NER system named TextRazor [218]. Other than the named entities, the results from TextRazor also includes fine-grained entity types based on Freebase [219], a collaborative knowledge base and a Wikipedia link for the entities. I then removed entities that are not essential for enhancing situational awareness such as urls and numbers. Even though some urls may lead to information that is relevant for gaining SA of drug cartels, the research of this chapter does not consider this type of data and Section 4.6 will discuss it further as future research. Eventually, removing non-essential entities results in a total number of 810,518 named entities (Table 4.1).

Keywords-based social media data collection often results in a lot of noise [200]. The named entities extracted from tweets are no exception. The bilingual nature of the dataset also make it more challenging to disambiguate entities that are relevant for gaining SA of drug cartels from irrelevant ones. For instance, one of the keywords used for collecting tweets is "cartel", which in Spanish has meaning of "poster". To filter entities that are irrelevant in order to gain situational awareness of drug cartels, I transformed entities in natural language form into pre-trained word embeddings and used a clustering algorithms to cluster entities into groups and then removed groups of entities that are irrelevant manually. Pre-trained word embeddings map words (i.e., entities) into vectors of numbers so that words with similar semantic meanings are closer to each other in vector space. More specifically, the pre-trained embeddings used in this chapter are from FastText that provides aligned multilingual word vectors [220]. One of the disadvantages of the common pre-trained word embeddings are that they cannot handle homographs, words with different meanings [221]. To mitigate this problem, in the case of clustering tweets by using embeddings of the entities in the tweets, I sum up vectors of all words of all entities in each tweet. For instance, although "cartel" has different meanings, by adding the vector of "drug" and the vector of "cartel", the vector of "drug cartel" represents more accurately the meaning needed for the research presented in Chapter 4. The convention of vector length for pre-trained word embeddings is 300 [114] and thus named entities are transformed into vectors of 300

numbers.

After tokenizing words into vectors, I clustered tweets using K-means clustering algorithm. K-means is an unsupervised learning method to cluster data into k number of topics. It is an iterative algorithm starting with k random centroids and at each iteration, the algorithm then assigns data points to their nearest centroids. After each iteration, it recalculates the centroids of clusters. The algorithm stops when there is no change of centroids. Applying K-means to the entity vectors, the parameter k was decided to be 100 using the elbow method [222]. The elbow method is a common heuristic that determines the optimal number of clusters where the quality of clusters (i.e., the sum of in-cluster distance) does not increase as sharp as the number of clusters increases and thus does not worth the computation cost to split data into higher number of clusters [223]. After clustering, entities in the same clusters have similar semantic meanings. I then assume that each cluster has underlying coherent topics, discussed further in Section 4.5.1.

Finally, spatiotemporal analysis is a common method used for SA research (e.g., [199, 215, 216]). More specifically, the research presented in this chapter examines the temporal and spatial patterns of the clusters (i.e., topics). The purpose of temporal analysis of clusters is to understand the patterns of time series of cluster frequencies throughout Nov, 2018, which is the time frame of the data collection. Patterns of time series involve detecting peaks and peak prominence used widely in signal processing [224]. In the context of the research presented in Chapter 4, detecting peaks and their prominence of entity clusters frequencies enables us to discover important events discussed on Twitter, which are bursts (i.e., sharp in frequency) of certain discussions [225]. In addition, the time series of cluster frequencies are also tested for correlations between each other in order to explore potential concurrence of detected events. While analyzing the detected named entities from all geolocations inform us about the general situations of drug cartel related events, it is also possible that certain events show local peaks in the time series from different places (e.g., countries). Therefore, the spatial patterns of time series peaks were then examined to discover local events and thus achieving a more local situational awareness.

4.5 Results

In this section, I will first present the results from named entity recognition, clustering results of these entities, and keywords in the relevant clusters after cluster filtering in Section 4.5.1. Then I will discuss the results from spatiotemporal analysis of these entity clusters in Sections 4.5.2 and 4.5.3.

4.5.1 Entity Filtering

As discussed in Methodology section, named entities are clustered into 100 clusters. I determined manually which clusters are relevant for achieving situational awareness and removed clusters that are determined to be irrelevant. The complete code and results are available at `https://bitbucket.org/xiaoyiyuan/cartel` and the statistical summaries of the clusters before and after filtering is shown in Table 4.2. Even though only 14 out of 100 clusters are chosen to be relevant, they consisted of 40.53% of all the entities with a total number of 328,519. Table 4.2 also shows that even though 40.53% of the entities are in relevant clusters, the unique entities in these clusters are only consisted of 11.61% of all unique entities. It suggests that the irrelevant entity clusters that are filtered out have lower frequency than the relevant entity clusters. The reason is that the relevant entities are overall high frequency entities.

Table 4.2: Named entity counts before and after K-Means filtering.

	Entities	**Unique Entities**
Before Filtering	810,518	53,586
After Filtering	328,519	6,221
Ratio	40.53%	11.61%

The relevant clusters are also shown in Table 4.3. For each cluster, only 15 entities that have the highest frequencies in the dataset are shown. It is worth noting that some keywords used for data collection (i.e.,"ms-13" and "ms13") are not shown in Table 4.3.

Table 4.3: Clusters of relevant entities and their frequencies.

ID	Entities
2	mexico (10064), mexican (9620), méxico (4509), mexicans (568), mexico presidents (233), mexico presidents (233), mexicanos (212), mexico city (172), peru (103), u.s. mexico (100), ciudad méxico (83), narcos mexico (55), tijuana mexico (51), mex (47), guatemalan (40), ciudad mexico (23)
14	gang member (4618), caravan (4419), white supremacis (1268), gangs (1161), 18th street gang (267), gang violence (256), arctic (232), rush (209), pale (108), massive attack (106), bloods (75), running (75), plastic surgeon (70), diamond (67), terrorist attack (56)
18	us (8918), democrats (3840), law (3735), police (2359), citizens (1812), taxes (1784), congress (1767), democrat (928), mexican police escort (913), haiti (683), republicans (588), democrat party (388), governor (379), republican (332), prosecutors (302)
25	sinaloa (5163), tijuana (2292), jalisco (1534), oaxaca (1009), guadalajara (547), tamaulipas (459), cártel sinaloa (394), veracruz (333), mexicano (203), monterry (186), tecate (158), chilpancingo (143), acapulco (142), coahuila (135), chihuahua (117)
26	trump (9023), clinton (3868), obama administration (3179), obama (2176), president (1059), clinton foundation (698), hillary (623), president trump (601), gop (290), pelosi (285), donald trump (273), bush (147), putin (138), trudeau (106), muslim obama (102)
43	drug cartel (4026), factory new ak-47 cartels (433), trafficking (340), human traffickers (314), tinubus drug cartel (199), rush u.s. border drug cartels find new routes(195), human trafficker (86), traffickers (56), cartels drugs human traffickers (55), drug (51), cartels (48), record cocaine production (23), cartels lobbyists (22), money cartel families (21), founding zetas cartel member killed mexican prison (20)
52	abogado (3658), sindicato (608), partido gobierno (331), u.s. federal (317), político (302), partido (261), ceda (252), partido revolucionario (241), gobierno (210), amlo (207), presidente (186), alcalde (175), congreso (142), gobierno eeuu (118), unión (105)
55	zetas (4285), terrorists (1226), farc (836), fbi (650), dea (597), interpol (456), cártel zetas (325), cia (305), eln (199), mafia (198), terrorist (186), grupo terrorista (145), dea agent (133), swat (132), asesinan (130)
70	peña nieto (653), calderón (492), maldonado (311), damaris reyes rivas (275), diosdado cabello (256), castro (253), cifuentes villa (147), escobar (147), galán (136), cabal (100), medina mora (94), bravo (75), peña (64), yunes linares (47), león (44)

Table 4.3: Clusters of relevant entities and their frequencies (continued).

ID	Entities
73	colombia (4282), venezuela (2156), honduras (2130), guatemala (1431), bogotá (1212), medellín (860), honduran (439), nicaragua (360), medellin (312), ecuador (223), caracas (220), cali (195), perú (168), columbian (160), venez (151)
77	cartel (116228), cartel sinaloa (7829), sinaloa cartel (2036), cartel medellín (1047), zetas cartel (950), tijuana cartel (861), carteles (438), gulf cartel (386), cartel toga (334), sinaloa drug cartel (323), guadalajara cartel (271), cartel medellin (200), jefe cartel soles (176), cartel paz (176), red dragon cartel (146)
78	chapo guzmán nueva york (1270), new york (706), new route (245), new (202), brooklyn (198), neuva york (186), neuva orleans (162), jalisco nueva generación (152), new ms13 (118), new york times (90), new mexico (64), new blood (43), newyork (41), nyc (39), jersey (37)
98	chapo (5193), chapo guzmán (1666), zambada (472) chapo trial (382), abogado chapo (87), chapo guzman (66), joaquín chapo (55), juicio chapo (55), rey zambada (46), julgamento chapo (44), joaquin chapo (40), narcotraficante zambada (34), joaquín el chapo (34), advogados chapo (22), joaquín el chapo (19)
99	border patrol (7301), u.s. border (5262), u.s. border patrol (742), us border (591), customs border protection (565), mexican border (324), united states border (190), united states border patrol (189), border (159), reserach (150), borders (143), earth (118), us border patrol (111), border mexico (77), texas border (74)
$e1$	ms13 (7060), ms-13 (35637)

The reason is that there is a small proportion of entities that are not found in the FastText English and Spanish aligned embeddings. As presented in Table 4.2, the number of unique entities before filtering is 53586, among which those that are found in FastText are 46870. Therefore, around 87% of unique entities have FastText word embeddings and thus were included in the clustering algorithm. Even though the entities that were not included in the clustering algorithm are often low frequency entities because they will not contribute as much to the overall trend of cartel related events as high frequency entities, "ms-13" and "ms13" are exceptions that appear very frequently in the dataset. Therefore, I included the keywords as an extra relevant cluster, listed on the bottom of Table 4.3 as $e1$. As presented in Table 4.3, majority of the clusters have coherent themes/topics based on the top entities.

For instance, Cluster 2 and 25 are city names and place names of Mexico and Cluster 73 is consisted of country names of the Northern Triangle. In addition, Cluster 14 is about gangs and Cluster 43, 77, 78, and 98 are about cartels with different emphasis. Cluster 43 is on cartel and human trafficking and Cluster 77 includes a few world famous drug cartel organizations (e.g., Sinaloa, Medellin, and Zetas and Tijuana cartel). Cluster 78 and Cluster 98 has entities that are referring to a specific event that the leader of Sinaloa cartel El Chapo was sentenced by the U.S. federal court. There are also clusters consisted of entities of US politicians (Cluster 26), US politics (Cluster 18), and U.S. border control (Cluster 99). Similarly, Cluster 70 includes Mexican politicians ("peña nieto" and "calderón"), Former Speaker of the National Assembly of Venezuela ("diosdado cabello"), drug cartel ("cifuentes villa" and "escobar"). In addition, Cluster 77 (Table 4.3) has an outlier entity "cartel" that has a much higher frequencies than the most frequent entities in other cluster. Considering that the word "cartel" alone is too general and will dwarf the patterns of other clusters in comparison, I removed entity "cartel" from the rest of the analysis. It is worth noting that entities that contain the word "cartel" are not removed, such as "cartel sinaloa". To summarize, two goals are achieved in this step of entity clustering, which are filtering out irrelevant clusters of entities and grouping relevant entities into clusters (i.e., topics).

4.5.2 Temporal Analysis of Entity Clusters

The overall frequencies of each cluster appear in the dataset are presented in Figure 4.2. Clusters about MS-13 (Cluster $e1$) and cluster about US and politics (Cluster 18) have the highest frequencies while cluster about drug cartel and human trafficker (Cluster 43) clusters of Central America politicians (Cluster 70) and Chapo Guzmán New York (Cluster 78) are the lowest in frequencies.

Even though the frequencies enable us to gain insights on the relative popularity of clusters, it does not show *when* the clusters reach its peak in frequencies and what the underlying events cause the peaks. To dig deeper into how entity clusters fluctuate along the one-month time frame, Figure 4.3 presents time series of the cluster frequencies. In

the remainder of this section, I will use measures from signal processing to characterize the time series patterns to make sense of the drug cartel situations.

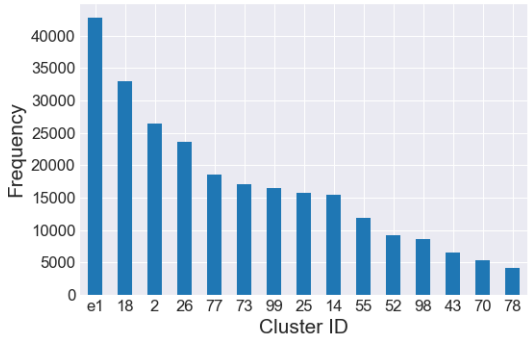


Figure 4.2: Frequencies of entity clusters.

Peak Detection

One of the common ways to characterize time series is to detect peaks and their prominences (i.e., how sharp the peak is) [224]. When the prominence is high, it is easy to visually detect peaks. However, visually detecting peaks relies heavily on data visualization, which is sometimes misleading. For instance, Figure 4.2 shows that Cluster 43 overall has low frequencies compared to other clusters and thus in Figure 4.3, the fluctuation is not as visually obvious as some of the other clusters (e.g., Cluster 2). Therefore, I used a quantitative measure to detect peaks and their prominences [226]. The peaks are detected by finding all local maxima by comparing neighboring values. Figure 4.3 shows the time series of each cluster within a month and the peaks of each time series are marked by red. One of the most distinctive features from Figure 4.3 is that all of the clusters peak have peaks on Day 14 (i.e., Nov 14, 2018) except for Cluster $e1$ ("MS-13"). To find out what events caused the burst of discussions on almost all clusters on Day 14, I queried the original tweets that have the highest frequencies of clustered that peaked on Day 14. Table 4.4 shows the tweets, which features El Chapo and the jury. El Chapo is a former leader of the Sinaloa cartel, which was one of the biggest supplier of drugs to the US. The detected trial event is referring to

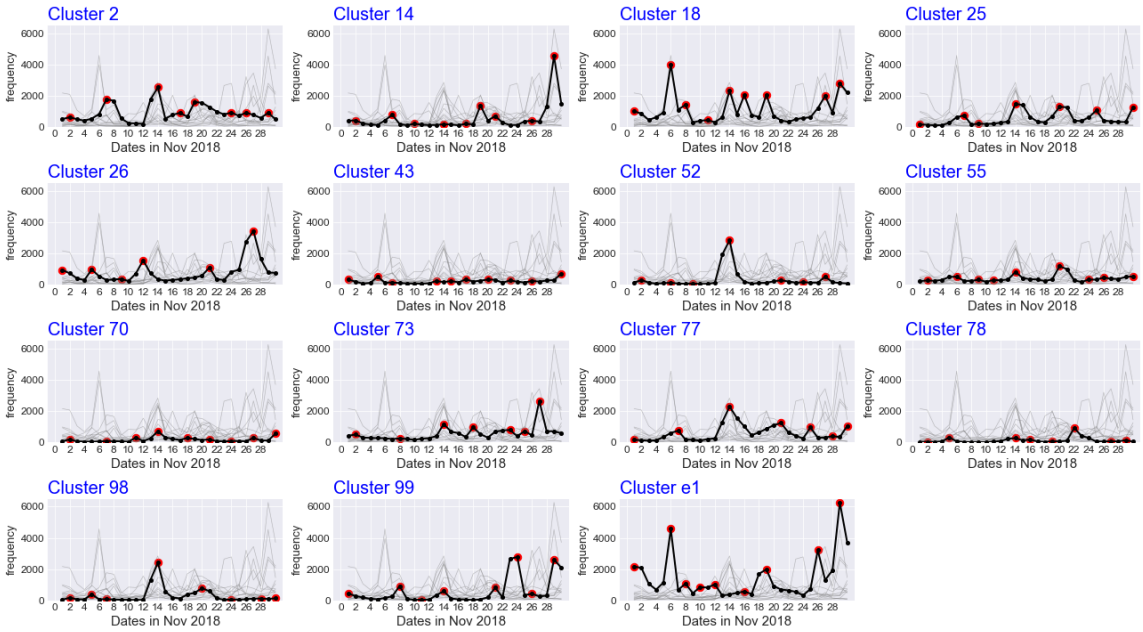


Figure 4.3: Time series of frequencies of entity clusters with peaks highlighted with red.

the trial in New York city where his key associates are expected to testify against him. Day 14 is also the day that cartel member testifies at New York El Chapo trial. In addition, Cluster 98 features the entities such as "chapo" and "chapo trial" has its peak with highest prominences (Cluster 98 in Table 4.3). This example shows that by quantifying peaks and peak prominences, one can gain understanding of current events in regards to drug cartels by querying high frequency tweets on peaks. In the example of Day 14 and El Chapo trial, the high frequency tweets are sufficient to construct a consistent topic of what is happening on Day 14 and the events underlying the peaks of Day 14. Interestingly, the highest frequency tweet from Cluster 18 does not mention El Chapo trial and insteand, it is a poll about border security. Cluster 99 is also about U.S. border, featuring the event of caravan immigrants from Mexico. It suggests two reasons for peaks of various clusters to occur on the same day—same event discussed from different perspective (e.g., El Chapo trial bribe and Sinaloa cartel) and different events (e.g., El Chapo trial and border control).

Table 4.4: Tweets of highest frequency from clusters that peaked on day 14 of Nov 2018

ID	Original Tweets	English Translation*
2	ÚLTIMO MOMENTO: según la agencia AFP, el abogado de El Chapo denunció que el actual Presidente de México y su anteceso recibieron sobornos del cartel de Sinaloa.	LAST MOMENT: according to the AFP agency, El Chapo's lawyer reported that the current President of Mexico and his predecessor received bribes from the Sinaloa cartel.
14	The Disappeared' became a chilling part of Latin America's Cold War vocabulary. Today gang violence is taking an even bigger toll.	N/A
18	#LDTPoll: Do you believe every American should support border security to protect citizens from criminal illegal migrants	N/A
25	Pues algo si es cierto: FelipeCalderon empoderó al Chapo (intencionadamente o no). El cartel de Sinaloa fue el más beneficiado durante su gobierno (intencionadamente o no).	Well, something is true: FelipeCalderon empowered El Chapo (intentionally or not). The Sinaloa cartel was the most benefited during his government (intentionally or not).
52	ÚLTIMO MOMENTO: según la agencia AFP, el abogado de El Chapo denunció que el actual Presidente de México y su anteceso recibieron sobornos del cartel de Sinaloa.	LAST MOMENT: according to the AFP agency, El Chapo's lawyer reported that the current President of Mexico and his predecessor received bribes from the Sinaloa cartel.
55	Cuando esta gente habla de "los organismos internacionales" ¿Se refiere al Cártel de Sinaloa? ¿A los Zetas?¿A los Soles? ¿Al frente Oliver Sinisterra? Qué curiosidad.	When these people talk about "international organizations", do they mean the Sinaloa Cartel? To the Zetas? To the Suns? Oliver Sinisterra up front? How curious.
70	Sinaloa financiaba a la derecha incluyendo a peña nieto y a Calderón. Lo dijo el propio abogado del Chapo Guzmán.	Sinaloa financed the right, including Peña Nieto and Calderón. Chapo Guzmán's own lawyer said it.
73	Palabras de @lopezobrador_"Convertir a México en Venezuela es malo; pero trágico sería convertirla en Colombia, un país que vive una dictadura disfrazada de democracia. Un país donde su clase política autoriza más asesinatos que el Cartel de Sinalo".	To turn Mexico into Venezuela is bad; but tragic would be to turn it into Colombia, a country that lives a dictatorship disguised as democracy. A country where its political class authorizes more murders than the Sinaloa Cartel.
77	Amplía El abogado del capo narco Chapo Guzmán, Jeffrey Lichtman, aseguró al jurado que el cartel de Sinaloa pagó millonarios sobornos al actual presidente de México, Enrique Peña Nieto, y a su antecesor Felipe Calderón.	Expand The lawyer of drug lord Chapo Guzmán, Jeffrey Lichtman, assured the jury that the Sinaloa cartel paid millionaire bribes to the current President of Mexico, Enrique Peña Nieto, and his predecessor Felipe Calderón.
78	#UPDATE Drug baron Joaquin "El Chapo" Guzman's defense told his New York trial that his cartel bribed Mexican presidents.	N/A
98	ÚLTIMO MOMENTO: según la agencia AFP, el abogado de El Chapo denunció que el actual Presidente de México y su anteceso recibieron sobornos del cartel de Sinaloa.	LAST MOMENT: according to the AFP agency, El Chapo's lawyer reported that the current President of Mexico and his predecessor received bribes from the Sinaloa cartel.
99	Mexican Government Partnering with Cartels to Move Migrants to U.S. Border: Mexican Police Escort 400 from Migrant Caravan.	N/A

* The English translation was provided by Google Translation.

This example of Day 14 shows that I can query tweets from clusters based on shared time series characteristics to understand concurrent events (e.g., peaking on the same day). However, the patterns of multiple clusters peaking on the same day was detected by observation. To automatically detect whether these time series sharing similar features (i.e.,

shapes), I calculated the correlations between them, which are presented in the following section.

Correlations Between Temporal Trends

To quantify correlations between the time series of entity clusters in order to extract correspondence of events, I used Pearson's correlation [227]. Table 4.5 shows the correlation coefficients and their P values. Among all pairs of clusters, Clusters 77 and 98 are correlated with most of other clusters, including Cluster 2, 25, 55. Based on Table 4.3, Cluster 77 is about various cartel names and similarly, Cluster 25, 55, and 98 are clusters that include cartel names as well, while Cluster 2 features place names in Mexico and Cluster 70 is about Central American politicians. Meanwhile, it is not surprising to find that Clusters 2, 25, 52, and 55 are correlated with on another as well. Among all the combinations of these clusters, Cluster 52 and Cluster 98 has the highest correlation coefficients, i.e. 0.928 with more than 99.999% confidence interval. Based on previous discussion and Table 4.4, Cluster 52 and 98 share the same high frequency tweets with exactly the same texts, drawing connection between El Chapo (Cluster 98) and Sinaloa cartel (Cluster 52) by claiming that El Chapo's lawyer reported that the president of Mexico and his predecessors received bribes from Sinaloa cartel. Cluster 98 (i.e., El Chapo) is also highly correlated with Clusters 25 (i.e., Mexican place names such as Sinaloa) and 77 (cartels such as Sinaloa and Medellín) with correlation coefficients as 0.859 and 0.817 respectively. Based on the entity clustering results in Table 4.3, the reason that Cluster 25 and Cluster 77 being highly correlated may be that the entities in these two clusters are highly similar to each other. Sinaloa is both a place name in Mexico and a cartel name. Although it is easy for a human to differentiate that Sinaloa in the context of drug cartel is more likely to be referred to be a cartel name, this problem of language ambiguity, specifically homonyms in natural languages is still a difficult problem for computational models [228]. Especially in this research, a generic word embedding was utilized where one word (even with various meanings) is only mapped to one vector. As many researchers have found advanced algorithms to mitigate this problem,

Table 4.5: Correlation coefficients and P values between clusters

	2	14	18	25	26	43	52	55	70	73	77	78	98	99	e1
2	1.000	0.058	0.280*	0.450*	-0.150*	0.000	0.633***	0.430*	0.373*	0.121	0.631***	0.163*	0.715***	0.079	-0.138*
14	0.058	1.000	0.462*	0.047	0.061	0.310*	-0.135*	0.152*	0.042	0.074	-0.024	-0.076	-0.096	0.472	0.763***
18	0.280*	0.462*	1.000	0.238*	0.092	0.191*	0.166*	0.220*	0.308*	0.245*	0.249*	-0.118	0.156	0.174*	0.686***
25	0.450*	0.047	0.238*	1.000	-0.104	0.343*	0.389*	0.698***	0.587***	0.168*	0.859***	0.046	0.535	0.132*	-0.070
26	-0.150*	0.061	0.092	-0.104	1.000	0.119	-0.033	0.094	-0.012	0.579***	-0.203*	-0.138*	-0.143	-0.027	0.217*
43	0.000	0.310*	0.191*	0.343*	0.119	1.000	-0.046	0.363*	0.303*	0.066	0.216*	0.024	0.088	0.360*	0.322*
52	0.633***	-0.135*	0.166*	0.389*	-0.033	-0.046	1.000	0.305*	0.669***	0.298*	0.722***	0.204*	0.928***	-0.043	-0.232*
55	0.430	0.152*	0.220*	0.698***	0.094	0.363*	0.305*	1.000	0.356*	0.100	0.573	-0.013	0.533**	0.040	0.091
70	0.373*	0.042	0.308*	0.587***	-0.012	0.303	0.669***	0.356*	1.000	0.411*	0.719***	-0.005	0.686***	0.065*	-0.019
73	0.121	0.074	0.245*	0.168*	0.579***	0.066	0.298*	0.100	0.411*	1.000	0.221*	0.147*	0.201	0.079	0.020
77	0.631***	-0.024	0.249*	0.859***	-0.203*	0.216*	0.722***	0.573***	0.719***	0.221*	1.000	0.183*	0.817***	-0.032	-0.179*
78	0.163*	-0.076	-0.118	0.046	-0.138*	0.024	0.204*	-0.013	-0.005	0.147*	0.183*	1.000	0.175	0.212*	-0.194*
98	0.715***	-0.096	0.156*	0.535**	-0.143*	0.088	0.928***	0.533**	0.686***	0.201*	0.817***	0.175*	1.000	-0.088	-0.229*
99	0.079	0.472**	0.174*	0.132*	-0.027	0.360*	-0.043	0.040	0.065	0.079	-0.032	0.212*	-0.088	1.000	0.328*
e1	-0.138*	0.763***	0.686***	-0.070	0.217*	0.322*	-0.232*	0.091	-0.019	0.020	-0.179	-0.194*	-0.229	0.328*	1.000

$P < .001$ ***; $P < .01$ **; $P < .05$ *

it is beyond the scope of the research presented in Chapter 4 and it will be addressed as one of the future research directions. Despite the challenge of language ambiguity, it shows that highly correlated clusters inform us about the connection between clusters (e.g., Cluster 98 of El Chapo and Cluster 52 of Sinaloa cartel) and by querying the high frequency tweets from these clusters on their peak days (e.g., "El Chapo's lawyer reported that the current President of Mexico and his predecessor received bribes from the Sinaloa cartel"), I can gain a better understanding in terms of *why* these two clusters are correlated.

In addition, Cluster $e1$ ("MS-13") has the most negative correlation coefficients with other clusters (Clusters 2, 25, 52, 77, 78, and 98) but have high positive correlation coefficients with Clusters 2 and 14 (Table 4.5). While the clusters negatively correlated with "MS-13" are clusters featuring Mexican and Northern Triangle drug cartels and politicians, the clusters it correlates positively with are clusters mostly written in English featuring "gang members", "white supremacies", "US" and US politics. It shows that the discussions of "MS-13" are less as a drug cartel but more as a "gang" in the context of the US. It begs the question that whether the same peaks and time series correlations will be presented if the data is examined by different geolocations (i.e., where the tweets was posted). Therefore, the next section will explore the spatial component of the time series characteristics.

4.5.3 Spatial and Temporal Analysis of Entity Clusters

The twitter data was consisted of tweets written in English and Spanish. Even though there are about 51% English tweets, entities with geolocation of the US is consisted of 70% of them because some tweets in Spanish are posted from a US location (Table 4.1). The top four countries with highest number of entities are the US (168,041), Mexico (19,859), Columbia (11,170), and Venezuela (7,694) and Figure 4.4 is the time series from these four countries for Cluster 98 about El Chapo, which was analyzed in the previous section as an important theme of Nov 2018. Figure 4.4 shows that the time series shape varies significantly from country to country.

In Figure 4.4, some of them have the highest peak on Day 14 (i.e., the US and Mexico)

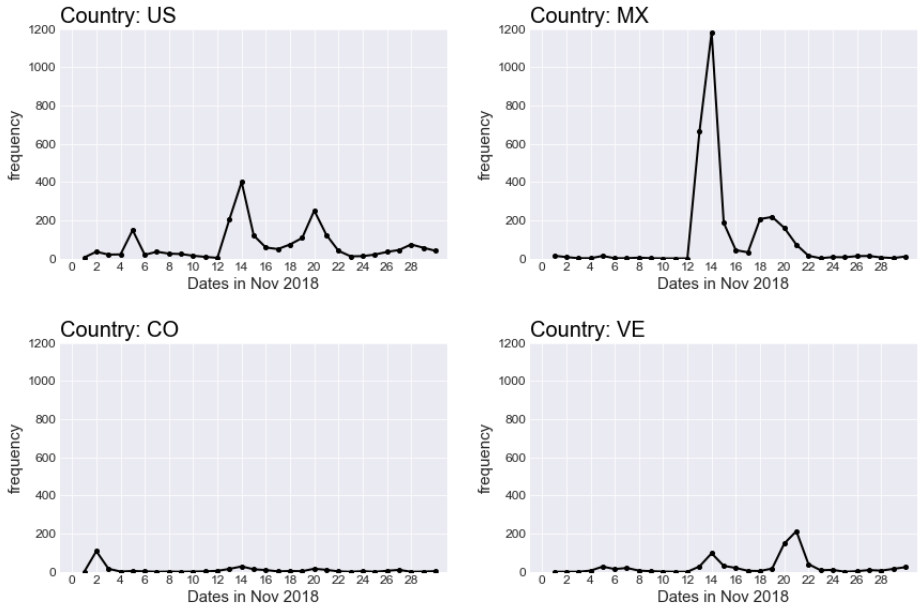
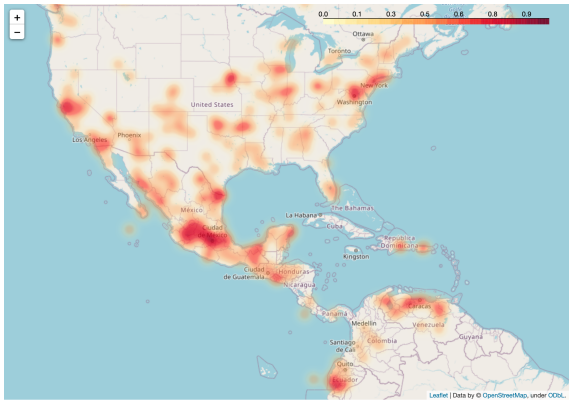
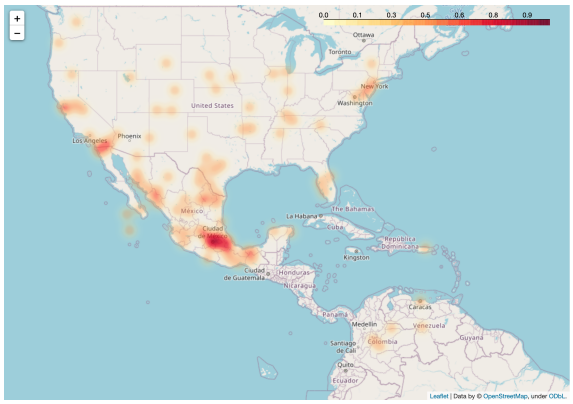


Figure 4.4: Time series of frequencies of entity Cluster 98 in the US, Mexico, Columbia, and Venezuela.

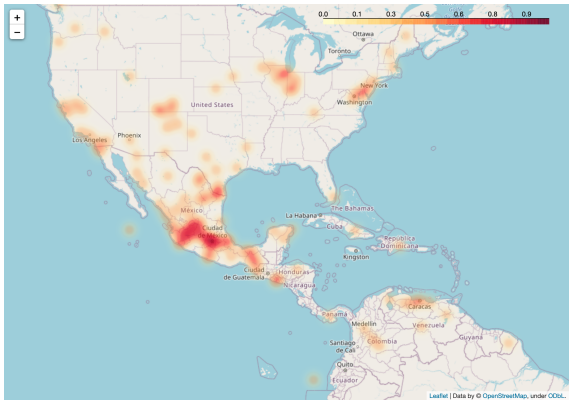
while others peaked the highest on Day 21 (i.e., Venezuela). To make sense of the spatial component of the time series for cluster 18, Figure 4.5 shows heat maps of the tweets on days that peaks appear in these four countries. On Day 14 that many geolocated countries peaked, hot spots appeared on the map in all of the four countries. Additionally, Mexico has the warmest colored spot (around Mexico City) denoting that the area has the highest number of tweets of Cluster 98 on Day 14, which is the day of a large number of El Chapo trial related tweets. Besides Day 14, all of the four countries peaked at around Days from 19-21 (Figure 4.4). The heat map illustrated the peaks spatially as hot spots in Figure 4.5. The hot spots in Mexico from Day 18 to Day 21 (Figure 4.5(b)-(e)) slowly dissipate. At the same time, new hot spots appear from Day 18 to 21 in Ecuador and Venezuela. To find out the reason why these peaks appear in different days and in different countries, I examined the high frequency tweets on Day 14 and Day 18-21 in the four countries. As discussed in Section 4.5.2, the high frequency tweets on Day 14 are about El Chapo trial and bribe and when examining tweets by country, the theme stays the same no matter in which country



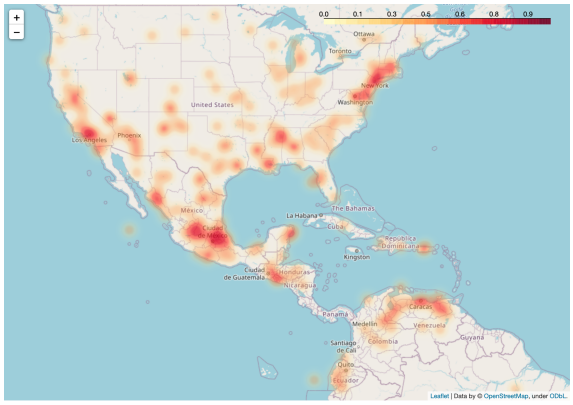
(a) Day 14



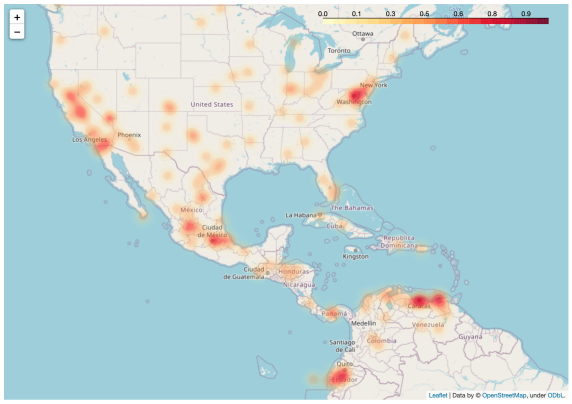
(b) Day 18



(c) Day 19



(d) Day 20



(e) Day 21

Figure 4.5: Heat maps of frequencies of Cluster 98 for Day 14 and Days 18-21. The legend on the upper right of each map denotes the percentage of magnitude.

on Day 14. However, on Day 21, the high frequency tweets are no longer about the same El Chapo bribe. In Venezuela, people are more interested in an event about El Chapo and Venezuela, presented in Table 4.6. The tweet is still relevant to the trial but from the perspective that is Sinaloa cartel fleeing the country in a Venezuela airport. This example showcases the spatiotemporal pattern highlights how people in different locations care about same events (Cluster 98) but from different perspective (e.g., Cluster 98 in Mexico on Day 14 versus Cluster 98 in Venezuela on Day 21).

Table 4.6: An example of tweets of high frequency on peak day in Venezuela

Original Tweets	**English Translation***
Ahora que en el juicio del "Chapo Guzmán" salen a relucir aviones de Aeropostal, es bueno recordar que el cartel de Sinaloa aterrizó y despegó aviones en Maiquetía ...y nada más y nada menos que a escasos metros del hangar presidencial.	Now that in the "Chapo Guzmán" trial Aeropostal planes come to light, it is good to remember that the Sinaloa cartel landed and took off planes in Maiquetía ... and nothing more and nothing less than a few meters from the presidential hangar.

* The English translation was provided by Google Translation.

4.6 Conclusion

Drug cartels are detrimental to the stability of societies across the globe and gaining timely situational awareness of of drug cartels enables us to understand what is the current notable events in regard to these transnational crime organizations. Even though geolocated social media is a useful data source for this task, they are often noisy and messy. In the research presented in Chapter 4, I extracted the named entities of geolocated tweets collected using several drug cartel related key words (i.e., "sinalo", "cartel", "cartels", "zetas", "ms-13", and "ms13"). However, not all of the tweets are relevant to drug cartels because of language ambiguity such as homonym (e.g., "sinalo" as in place name in Mexico or as in Sinalo cartel) and the nature of multilingual tweets (e.g., different meanings of "cartel" in Spanish and English). I transformed the extracted named entities into word vectors using aligned

FastText pre-trained word vectors and clustered them into different groups of entities. The irrelevant entity clusters are then removed and the relevant entity clusters are analyzed as concepts/themes.

The main contributions of this research are three-folded. First, this research proposes a pipeline of achieving drug cartel situational awareness, which detects the temporal trend of entity clusters by borrowing measurements from signal processing, including peak detection and peak prominence. By detecting the peaks of entity cluster frequencies, I can query tweets containing notable events of certain time frame (e.g. a month). In addition, by measuring correlations between these time series, I can detect the concurrent events from tweets. Second, the research demonstrated that by obtaining time series peaks from different regions (e.g., countries), I can detect the nuances of how same events are perceived differently. For instance, tweets geolocated in Mexico presented peaks on Day 14 for events of El Chapo trial in New York in terms of bribe to Mexican presidents whereas tweets geolocated in Venezuela showed peaks on Day 21 for events of El Chapo and Sinalo cartel fleeing from a Venezuela airport. Third, the research in Chapter 4 demonstrated that even though keywords-based Twitter data collection results in noisy results, clustering entity word vectors is useful for removing noises to a considerable degree. There are, however, several limitations to this research. The process of detecting relevant clusters is manual, which could result in delay of achieving timely situation awareness. Even though we could use the results of manual labels for relevant/irrelevant clusters in the future as labeled data to cluster tweets collected in the future (e.g., by using Nearest Neighbor clustering algorithm), it is possible that some vocabularies that are not in our labeled relevant clusters could end up not being recognized as relevant. One part of the future research, therefore, is to develop an approach that could reduce the time for the manual process of identifying relevant clusters (e.g., semi-supervised learning [229]). That being said, this chapter presents a pipeline from Twitter data collection, data filtering and clustering, to spatiotemporal analysis in order to achieve situational awareness for drug cartel related events and thus achieve situational awareness through the detected events.

Chapter 5: Conclusion

5.1 Summary of Dissertation Results

Place is a location associated with meaning. As more VGI and AGI become available, we are able to explore the rich and individualistic meaning of place from bottom up. Specifically, this dissertation utilizes geo-textual data to study places and their connections by answering three research questions.

- RQ1: In what aspects do urban places become placeless based on geo-textual data?

Comparing opinions towards chain stores and independent stores from store reviews, the research presented in Chapter 2 is able to discover that among many other factors (e.g. food quality, atmosphere, taste), *location* is the most important factor for chain stores overall. By digging deeper into what kind of information co-occurs with the aspect "location", the research further finds out that for chain stores, the characteristics of the surrounding areas (e.g., "airport location"), franchise names (e.g., "Subway location"), and place names (e.g., "Madison location") co-occur with "location". The results suggests that even though the standardization of chain stores seems to make urban places look and feel the same, individuals create features, meanings, and identities out of the characteristics of locations (e.g., "The other Madison locations are not like this."), which call into the question of the assumptions underlying policies protecting local business that chains stores are the cause of loss of place identity [118].

- RQ2: How to examine place connections in networks derived from geo-textual data?

While harvesting and analyzing a large amount of geo-textual data is valuable for understanding places, places in reality are never in isolation. We perceive places in relations

as well (e.g., one place reminds us of another) [230]. Chapter 3 proposes a novel approach of analyzing places in networks. The network is constructed by the thematic/topic similarities between places. While many previous geo-textual data analytics for studying places often conduct clustering on geo-texts to explore underlying patterns of sentiments, experiences, or activities of places (e.g., [46, 141, 144–147, 231, 232]), clustering simplifies place relations into either "in-cluster" or "out-of-cluster". Research in Chapter 3 improves the previous geo-textual data analytics on places by structuring places into networks and discovers rich information about place relations using a variety of network statistics. The case study presented in Section 3.5 on Manhattan (New York) is able to discover places by their degrees of uniqueness by the position of nodes (i.e., places) in the networks. The case study also examines the correlation between themes/topic derived from the geo-textual data and the corresponding geodemographics and reveals that a few low income areas in Manhattan (e.g., Harlem and Chinatown) have distinctive restaurant themes and culture.

- RQ3: In what way do geo-textual data enable us to achieve situational awareness of drug cartels?

The impact of events is an application of geo-textual data analytics that has been well researched by many previous studies [60, 233]. The research presented in Chapter 4 attempted to an initial effort in discovering notable events in order to achieve situational awareness of drug cartels. To extract meaningful information out of unstructured geolocated tweets, this research utilizes named entity recognition to identify key players of the tweets (e.g., organizations, locations, or person). Events are then discovered by peaks in the spatiotemporal trend (i.e., time series) of named entities. By conducting spatiotemporal analysis of the named entities, the research also discovered that notable events show nuances as how they are manifested in different places (i.e., countries). For instance, tweets from Mexico and the US showed great interest in the events of the trial of El Chapo from the perspective of Sinalo cartel's bribe to Mexican presidents whereas tweets from Venezuela were focued on Sinalo cartel fleeing from a Venezuela airport. The research in Chapter 4

presents a pipeline from data collection, data filtering and clustering, to the discovery of notable events from spatiotemporal analysis.

In summary, this dissertation examined approaches and applications of geo-textual analytics for studying places, connections between places and place related event detection. The next section discusses how these results contribute to improving the state-of-the-art geo-textual analytics for studying places.

5.2 Contributions

The major contributions of this dissertation are: first, it identified and demonstrated the importance of geo-textual data for comprehending places in various levels of nuances. Research presented in Chapter 2 and Chapter 4 has built analytics on the basis of extracting key elements (i.e., opinion aspects and named entities) of the geo-texual data whereas Chapter 3 extracted latent topics out of whole reviews and tweets. Extracting key elements facilitates us to conduct fine-grained analytics that highlights specific aspects mentioned in geo-textual data, which would be potentially buried and ignored otherwise. For instance, the specific aspects would not stand out in frequency-based topic modeling approach because these aspects are not necessarily repeatedly mentioned in geo-texual data. However, low frequency does not necessarily mean unimportant. For studying place reviews, aspects associated with certain kind of sentiments inform us a lot of place perceptions (e.g., what people care about the most of places). That being said, topic modeling, as a widely used approach for studying places using geo-textual data, it serves well when the research question does not require understanding low-level and nuanced elements in regard to places (e.g., generic themes of people's activities and perceptions of places).

Second, the dissertation addresses the widely recognized problem of geo-textual data source [18, 199]. The problem is to differentiate geo-textual data *about* and *from* places. The data sources used in this dissertation are Twitter, TripAdvisor restaurant and attraction reviews, and Yelp (restaurant) reviews. The review data is geo-textual data *about* places

so that we can use them directly to study place. However, the problem raises when using Twitter, which is a platform that users can contribute various topics. Section 3.5.1 showed that users tend to use different vocabularies when posting place-relevant and place-irrelevant tweets. Since topic modeling is a vocabulary-sensitive model, it can tease out place-relevant tweets into their own topics without time consuming hand label process. Therefore, for future research using tweets for studying places, it is worthwhile to try as a quick method to pre-process geo-textual data for relevancy. Third, this dissertation also showcases how and why it is fruitful to combine advances in natural language processing and network science with geo-textual data. Some researchers (e.g., [199] have argued that the amount of geo-textual data nowadays allows us to explore places and build analytics that cannot be done before the "big data era". This dissertation further demonstrated that development in methodologies (e.g., natural language processing algorithms) is also a driving force that makes recent geo-textual analytics perform better or even possible. For instance, deep learning model for natural language processing (NLP) has been developed primarily in the last decade [115]. Even though many NLP tasks including aspect-based sentiment analysis (Chapter 3) could be done by machine learning models instead of deep learning models, NLP using deep learning shows better performance (e.g., F1 score) and requires less complicated syntax based feature selection processes [110].

Therefore, this dissertation exemplifies computational social science (CSS) research [64] through combining theories and practices from multiple disciplines as illustrated in Figure 5.1. Grounded in social science and geography, the concept of place has been examined through non-computational methods such as interviews and ethnography as discussed in Chapter 1. Facing the large amount of geo-textual data, place, place connections, and correlations between place and events that this dissertation aims to study call for various of new approaches. These approaches are natural language processing (NLP), geographic information system (GIS), machine learning, and network science. This dissertation, therefore, is a work of both geography and CSS. Meanwhile, according to Torrens, "social network analysis, which is central to computational social science, is largely overlooked by geographers

and geographical forays into this area have largely been missed in computational social science" [234]. This dissertation bridges the gap between between geography and CSS. On one hand, place and their connections, as classic geographical concepts are studied using major CSS approaches including information retrieval and network analysis [64] in Chapters 2 and 3. On the other hand, research in Chapter 4 includes spatial component in the computational analysis of sociopolitical events. Geo-textual data analytics call for interdisciplinary efforts [234] and many future work can be done on the basis of this dissertation that is discussed in the next section.

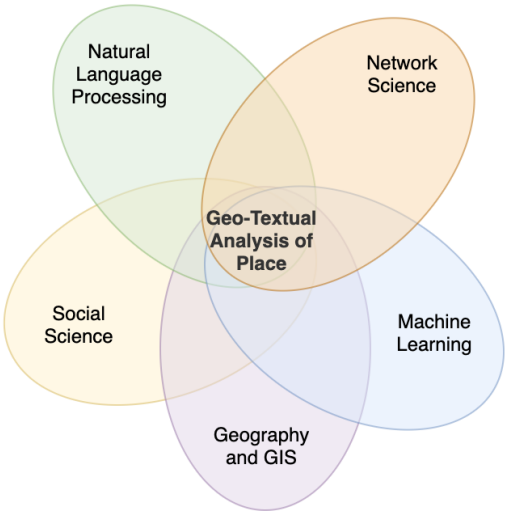


Figure 5.1: A Venn diagram that depicts interdisciplinary characteristics of geo-textual data analytics.

5.3 Limitations and Future Work

As more data and advanced computational methodologies become available, many potential areas of future work can be explored. In this dissertation, place is examined in different levels of aggregation. Chapter 2 defined place as individual stores whereas Chapter 3 treats a place as a Census tract and in Chapter 4, places are geolocations of tweets. Meanwhile, in Chapter 3 Section 3.3.2, I pointed out that defining place in any level would result in

the modifiable areal unit problem that is statistical summaries of the aggregated area being influenced by the shape and size of the area [168]. However, future research could explore the spatiotemporal trend in Chapter 4 Section 4.5.3 state or even city level of geolocated areas instead of country level. Moreover, it is still a relatively new area to investigate connections between places using geo-textual data [27]. Chapter 3 proposed an approach of applying network science to unveil complex relationships between places. Building upon this approach, future research can apply network statistics beyond network community detection, degree centrality and boundary nodes (Chapter 3 Section 3.4). For example, one can construct two networks from places of two cities. By comparing network level statistics such as clustering coefficient, we can grasp the degrees of a city is tightly knit in terms of their restaurant culture if the network is built upon restaurant reviews. In addition, future work can also combine non-textual open source geo-located or geo-referenced data to study place. Chapter 3 Section 3.5.2 incorporated US Census data into the thematic networks of places to discover whether communities detected based on network structure from online geo-texual data has connection with demographics of residents of these places. Besides Census data, there is a large amount of open source mobile activity data available. Future research could incorporate activity data into geo-textual data analytics to explore whether certain place perception is associated with certain activity patterns. In order to help researchers explore these research areas, all the programming code for this dissertation has made available at `https://bitbucket.org/xiaoyiyuan/workspace/projects/PUB`. In summary, this work demonstrates the great potential of geo-textual data for exploring places and their connections, and thus advances research in geography and CSS.

Appendix A: An Appendix

Table A.1: Moran's I for All Communities.

Network	Community ID	Moran's I	P value
TripAdvisor Attractions	1	-0.022740	.413
	2	-0.002260	.105
	3	0.000641	.054
	4	0.120743	.014*
	5	0.162247	.010**
	6	0.142676	.006**
	7	-0.020767	.035*
	8	0.002285	.410
	9	-0.003477	.131
	10	0.057185	.097
	11	0.141883	.015*
	12	0.107151	.014*
	13	0.106869	.034*
	14	0.053348	.111
	15	-0.005520	.326
	16	-0.005640	.280
	17	-0.006208	.157
	18	-0.005532	.345
	19	-0.015263	.430
	20	-0.000060	.073
	21	-0.007165	.067
	22	-0.005024	.492
	23	-0.004145	.208
	24	-0.006130	.140
	25	-0.005303	.389
	26	-0.009513	.381
	27	0.171691	.009**
	28	0.338116	.002**
TripAdvisor Restaurants	1	0.004695	.052
	2	0.120561	.017*
	3	0.682913	.001***
	4	-0.001215	.045*
	5	0.459894	.002**
	6	-0.003185	.100
	7	-0.009306	.014*
	8	0.223211	.001***
	9	-0.005161	.416
	10	0.174987	.001***
	11	-0.006029	.140
	12	-0.025863	.045*
	13	0.112362	.026*

Table A.1: Moran's I for all communities (continued).

Network	Community ID	Moran's I	P value
TripAdvisor Restaurants	14	0.167896	.010**
	15	0.038441	.170
	16	-0.004510	.365
	17	0.255313	.002**
	18	-0.003777	.186
	19	0.116421	.020*
	20	-0.005524	.266
	21	-0.004135	.232
	22	-0.006081	.139
	23	0.345028	.001***
	24	0.085735	.044*
	25	-0.009043	.358
	26	-0.004745	.425
	27	0.102172	.030*
	28	0.076950	.046*
	29	-0.006169	.111
	30	-0.005366	.317
	31	-0.004373	.352
	32	-0.004645	.415
	33	-0.011160	.188
	34	-0.004428	.350
	35	-0.005358	.343
	36	-0.004536	.124
	37	-0.008849	.015*
	38	-0.004531	.391
	39	-0.005106	.442
	40	-0.004121	.251
	41	-0.004531	.390
	42	-0.004252	.282
	43	-0.005566	.301
	44	0.572042	.001***
	45	-0.004745	.444
	46	-0.004745	.429
	47	-0.005653	.277
	48	-0.005114	.443
	49	-0.005894	.195
	50	-0.002693	.085
	51	0.004695	.038*
	52	-0.004088	.232
	53	-0.005155	.420
	54	-0.002693	.075
	55	-0.007618	.024*
	56	0.004695	.053
	57	0.004695	.043*
	58	0.004695	.042*

Table A.1: Moran's I for all communities (continued).

Network	Community ID	Moran's I	P value
Twitter	1	0.004695	.046*
	2	-0.003338	.086
	3	-0.006387	.161
	4	0.004695	.046*
	5	-0.001215	.048*
	6	0.004695	.046*
	7	-0.010278	.404
	8	-0.002693	.095
	9	-0.004061	.221
	10	0.328338	.001***
	11	-0.012091	.006**
	12	0.056188	.085
	13	-0.010706	.293
	14	-0.010190	.414
	15	-0.003404	.113
	16	-0.007665	.031*
	17	0.118175	.016*
	18	0.194717	.013*
	19	-0.004334	.347
	20	-0.005818	.224
	21	-0.015831	.304
	22	-0.006734	.076
	23	-0.009470	.010*
	24	0.019150	.246
	25	0.196924	.005**
	26	0.077334	.063
	27	-0.003777	.169
	28	-0.009687	.451

*Significant at $p \leq 0.05$; **Significant at $p \leq 0.01$; ***Significant at $p \leq 0.001$.

Bibliography

[1] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks," in *Third International AAAI Conference on Weblogs and Social Media*, 2009.

[2] T. Cresswell, *Place: An Introduction*. John Wiley & Sons, 2014-11-24.

[3] Y.-F. Tuan, "Space and place: Humanistic perspective," in *Philosophy in Geography*. Springer, 1979, pp. 387–427.

[4] A. Pred, "Place as historically contingent process: Structuration and the time-geography of becoming places," *Annals of the association of american geographers*, vol. 74, no. 2, pp. 279–297, 1984.

[5] J. Agnew, "Space and place," *Handbook of geographical knowledge*, vol. 2011, pp. 316–331, 2011.

[6] B. C. Eaton and R. G. Lipsey, "An economic theory of central places," *The Economic Journal*, vol. 92, no. 365, pp. 56–72, 1982.

[7] R. Waldinger, "Immigration and urban change," *Annual Review of Sociology*, vol. 15, no. 1, pp. 211–232, 1989.

[8] N. Smith, "New globalism, new urbanism: Gentrification as global urban strategy," *Antipode*, vol. 34, no. 3, pp. 427–450, 2002.

[9] P. Gustafson, "Meanings of place: Everyday experience and theoretical conceptualizations," *Journal of environmental psychology*, vol. 21, no. 1, pp. 5–16, 2001.

[10] J. R. Höflich, "A certain sense of place," *A sense of place—the global and the local in mobile communication, Passagen Verlag*, pp. 159–168, 2005.

[11] R. M. Groves, F. J. Fowler Jr, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau, *Survey Methodology*. John Wiley & Sons, 2011, vol. 561.

[12] Y. Zhang, "Using the Internet for survey research: A case study," *Journal of the American society for information science*, vol. 51, no. 1, pp. 57–68, 2000.

[13] S. Andriole. (2015) Unstructured Data: The Other Side of Analytics. [Online]. Available: https://www.forbes.com/sites/steveandriole/2015/03/05/the-other-side-of-analytics/

[14] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis, "Discovering geographical topics in the twitter stream," in *Proceedings of the 21st International Conference on World Wide Web*, ser. WWW '12. Association for Computing Machinery, 2012-04-16, pp. 769–778. [Online]. Available: https://doi.org/10.1145/2187836.2187940

[15] A. Stefanidis, A. Crooks, and J. Radzikowski, "Harvesting ambient geospatial information from social media feeds," *GeoJournal*, vol. 78, no. 2, pp. 319–338, 2013. [Online]. Available: https://doi.org/10.1007/s10708-011-9438-2

[16] G. Panteras, S. Wise, X. Lu, A. Croitoru, A. Crooks, and A. Stefanidis, "Triangulating Social Multimedia Content for Event Localization using Flickr and Twitter," *Transactions in GIS*, vol. 19, no. 5, pp. 694–715, 2015. [Online]. Available: http://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12122

[17] D. Tasse, "How Geotagged Social Media Can Inform Modern Travelers," 2017.

[18] Y. Hu, "Geo-text data and data-driven geospatial semantics," *Geography Compass*, vol. 12, no. 11, p. e12404, 2018.

[19] S. Caquard, "Cartography I: Mapping narrative cartography," *Progress in Human Geography*, vol. 37, no. 1, pp. 135–144, 2013.

[20] A. Sekar, R. B. Chen, A. Cruzat, and M. Nagappan, "Digital narratives of place: Learning about neighborhood sense of place and travel through online responses," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2666, pp. 10–18, 2017.

[21] A. Crooks, D. Pfoser, A. Jenkins, A. Croitoru, A. Stefanidis, D. Smith, S. Karagiorgou, A. Efentakis, and G. Lamprianidis, "Crowdsourcing urban form and function," *International Journal of Geographical Information Science*, vol. 29, no. 5, pp. 720–741, 2015-05-04. [Online]. Available: https://doi.org/10.1080/13658816.2014.977905

[22] A. Jenkins, A. Croitoru, A. T. Crooks, and A. Stefanidis, "Crowdsourcing a collective sense of place," *PloS one*, vol. 11, no. 4, p. e0152932, 2016.

[23] M. F. Goodchild, "Citizens as sensors: The world of volunteered geography," *GeoJournal; Dordrecht*, vol. 69, no. 4, pp. 211–221, 2007. [Online]. Available: http://search.proquest.com/docview/223670705/abstract/DA38993995644F4DPQ/1

[24] M. F. Goodchild and J. A. Glennon, "Crowdsourcing geographic information for disaster response: A research frontier," *International Journal of Digital Earth*, vol. 3, no. 3, pp. 231–241, 2010.

[25] Y.-T. Zheng, Z.-J. Zha, and T.-S. Chua, "Mining travel patterns from geotagged photos," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 3, p. 56, 2012.

[26] Y. Liu, X. Liu, S. Gao, L. Gong, C. Kang, Y. Zhi, G. Chi, and L. Shi, "Social Sensing: A New Approach to Understanding Our Socioeconomic Environments," *Annals of the Association of American Geographers*, vol. 105, no. 3, pp. 512–530, 2015. [Online]. Available: https://doi.org/10.1080/00045608.2015.1018773

[27] A. M. MacEachren, "Leveraging Big (Geo) Data with (Geo) Visual Analytics: Place as the Next Frontier," in *Spatial Data Handling in Big Data Era: Select Papers from the 17th IGU Spatial Data Handling Symposium 2016*, ser. Advances in Geographic Information Science, C. Zhou, F. Su, F. Harvey, and J. Xu, Eds. Springer Singapore, 2017, pp. 139–155. [Online]. Available: https://doi.org/10.1007/978-981-10-4424-3_10

[28] K. Lynch, *The Image of the City*. The MIT Press, 1960.

[29] W. H. Whyte, *The Social Life of Small Urban Spaces*. Conservation Foundation, 1980.

[30] J. Haffner, *The View from Above: The Science of Social Space*. MIT Press, 2013.

[31] A. Buttimer, "Home, reach, and the sense of place," *The human experience of space and place*, vol. 3, pp. 166–87, 1980.

[32] J. Meyrowitz, *No Sense of Place: The Impact of Electronic Media on Social Behavior*. Oxford University Press, 1986.

[33] E. Relph, "Sense of place," *Ten geographic ideas that changed the world*, pp. 205–226, 1997.

[34] D. R. Williams and S. I. Stewart, "Sense of place: An elusive concept that is finding a home in ecosystem management," *Journal of forestry*, vol. 96, no. 5, pp. 18–23, 1998.

[35] B. S. Jorgensen and R. C. Stedman, "Sense of place as an attitude: Lakeshore owners attitudes toward their properties," *Journal of environmental psychology*, vol. 21, no. 3, pp. 233–248, 2001.

[36] D. Massey, *A Global Sense of Place*. Aughty. org, 2010.

[37] N. Ujang, "Place attachment and continuity of urban place identity," *Procedia-Social and Behavioral Sciences*, vol. 49, pp. 156–167, 2012.

[38] S. Shamai, "Sense of place: An empirical measurement," *Geoforum*, vol. 22, no. 3, pp. 347–358, 1991. [Online]. Available: http://www.sciencedirect.com/science/article/pii/001671859190017K

[39] R. C. Stedman, "Toward a social psychology of place: Predicting behavior from place-based cognitions, attitude, and identity," *Environment and behavior*, vol. 34, no. 5, pp. 561–581, 2002.

[40] ——, "Is It Really Just a Social Construction?: The Contribution of the Physical Environment to Sense of Place," *Society & Natural Resources*, vol. 16, no. 8, pp. 671–685, 2003. [Online]. Available: https://doi.org/10.1080/08941920309189

[41] D. R. Williams and J. J. Vaske, "The measurement of place attachment: Validity and generalizability of a psychometric approach," *Forest science*, vol. 49, no. 6, pp. 830–840, 2003.

[42] J. Mohan and L. Twigg, "Sense of place, quality of life and local socioeconomic context: Evidence from the survey of English housing, 2002/03," *Urban studies*, vol. 44, no. 10, pp. 2029–2045, 2007.

[43] S. Semken and C. B. Freeman, "Sense of place in the practice and assessment of place-based science teaching," *Science Education*, vol. 92, no. 6, pp. 1042–1057, 2008.

[44] M. J. Stebleton, R. L. Huesman Jr, and A. Kuzhabekova, "Do I Belong Here? Exploring Immigrant College Student Responses on the SERU Survey Sense of Belonging/Satisfaction Factor. SERU Consortium Research Paper. Research & Occasional Paper Series: CSHE. 13.10." *Center for studies in higher education*, 2010.

[45] A. Kudryavtsev, R. C. Stedman, and M. E. Krasny, "Sense of place in environmental education," *Environmental education research*, vol. 18, no. 2, pp. 229–250, 2012.

[46] A. Crooks, D. Pfoser, A. Jenkins, A. Croitoru, A. Stefanidis, D. Smith, S. Karagiorgou, A. Efentakis, and G. Lamprianidis, "Crowdsourcing urban form and function," *International Journal of Geographical Information Science*, vol. 29, no. 5, pp. 720–741, 2015.

[47] J. Cranshaw, R. Schwartz, J. Hong, and N. Sadeh, "The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City," in *Sixth International AAAI Conference on Weblogs and Social Media*, 2012-05-20. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4682

[48] I. Simon, N. Snavely, and S. M. Seitz, "Scene summarization for online image collections," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference On*. IEEE, 2007, pp. 1–8.

[49] L. S. Kennedy and M. Naaman, "Generating diverse and representative image search results for landmarks," in *Proceedings of the 17th International Conference on World Wide Web*. ACM, 2008, pp. 297–306.

[50] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Proceedings of the 18th International Conference on World Wide Web*. ACM, 2009, pp. 761–770.

[51] E. Fischer, *The Geotaggers' World Atlas*. Flickr, 2013.

[52] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W.-Y. Ma, "Recommending friends and locations based on individual location history," *ACM Transactions on the Web (TWEB)*, vol. 5, no. 1, p. 5, 2011.

[53] J. J.-C. Ying, W.-C. Lee, and V. S. Tseng, "Mining geographic-temporal-semantic patterns in trajectories for location prediction," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 1, p. 2, 2013.

[54] S. Kisilevich, M. Krstajic, D. Keim, N. Andrienko, and G. Andrienko, "Event-based analysis of people's activities and behavior using Flickr and Panoramio geotagged photo collections," in *Information Visualisation (IV), 2010 14th International Conference*. IEEE, 2010, pp. 289–296.

[55] H. Q. Vu, G. Li, R. Law, and B. H. Ye, "Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos," *Tourism Management*, vol. 46, pp. 222–232, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0261517714001356

[56] N. Hochman and R. Schwartz, "The social media life of public spaces: Reading places through the lens of geotagged data," in *Locative Media*. Routledge, 2014, pp. 68–81.

[57] T. Rattenbury and M. Naaman, "Methods for extracting place semantics from Flickr tags," *ACM Transactions on the Web (TWEB)*, vol. 3, no. 1, p. 1, 2009.

[58] B. Tomaszewski, J. Blanford, K. Ross, S. Pezanowski, and A. M. MacEachren, "Supporting geographically-aware web document foraging and sensemaking," *Computers, Environment and Urban Systems*, vol. 35, no. 3, pp. 192–207, 2011.

[59] J. K. Nelson, S. Quinn, B. Swedberg, W. Chu, and A. M. MacEachren, "Geovisual Analytics Approach to Exploring Public Political Discourse on Twitter," *ISPRS International Journal of Geo-Information*, vol. 4, no. 1, pp. 337–366, 2015. [Online]. Available: https://www.mdpi.com/2220-9964/4/1/337

[60] Y. Hu, "Geo-text data and data-driven geospatial semantics," *Geography Compass*, vol. 12, no. 11, p. e12404, 2018.

[61] D. Preoţiuc-Pietro, J. Cranshaw, and T. Yano, "Exploring venue-based city-to-city similarity measures," in *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, ser. UrbComp '13. Association for Computing Machinery, 2013, pp. 1–4. [Online]. Available: https://doi.org/10.1145/2505821.2505832

[62] S. Gao, K. Janowicz, D. R. Montello, Y. Hu, J.-A. Yang, G. McKenzie, Y. Ju, L. Gong, B. Adams, and B. Yan, "A data-synthesis-driven method for detecting and extracting vague cognitive regions," *International Journal of Geographical Information Science*, vol. 31, no. 6, pp. 1245–1271, 2017.

[63] A. Jenkins, A. Croitoru, A. T. Crooks, and A. Stefanidis, "Crowdsourcing a Collective Sense of Place," *PLOS ONE*, vol. 11, no. 4, p. e0152932, 2016-04-06. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152932

[64] C. Cioffi-Revilla, "Introduction," in *Introduction to Computational Social Science: Principles and Applications*, ser. Texts in Computer Science, C. Cioffi-Revilla, Ed. Springer International Publishing, 2017, pp. 1–33. [Online]. Available: https://doi.org/10.1007/978-3-319-50131-4_1

[65] E. Relph, *Place and Placelessness*. Pion, 1976.

[66] R. Freestone and E. Liu, *Place and Placelessness Revisited*. Routledge, 2016.

[67] M. Arefi, "Non-place and placelessness as narratives of loss: Rethinking the notion of place," *Journal of urban design*, vol. 4, no. 2, pp. 179–193, 1999.

[68] D. Seamon and J. Sowers, "Place and placelessness (1976): Edward relph," *Key texts in human geography*, pp. 43–52, 2008.

[69] C. Shim and C. A. Santos, "Tourism, place and placelessness in the phenomenological experience of shopping malls in Seoul," *Tourism Management*, vol. 45, pp. 106–114, 2014-12-01. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0261517714000569

[70] K. Glasgow, A. Ebaugh, and C. Fink, "# Londonsburning: Integrating Geographic, Topical and Social Information during Crisis," in *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[71] X. Yuan and A. Crooks, "Assessing the placeness of locations through user-contributed content," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 2019, pp. 15–23.

[72] X. Yuan, A. Crooks, and A. ZÃŒfle, "A Thematic Similarity Network Approach for Analysis of Places Using Volunteered Geographic Information," *ISPRS International Journal of Geo-Information*, vol. 9, no. 6, p. 385, 2020. [Online]. Available: https://www.mdpi.com/2220-9964/9/6/385

[73] J. A. Agnew, "Devaluing place:"people prosperity versus place prosperity" and regional planning," *Environment and Planning D: Society and Space*, vol. 2, no. 1, pp. 35–45, 1984.

[74] E. C. Relph, *Place and Placelessness*. Pion Limited, 1976.

[75] A. Latham, "Urbanity, lifestyle and making sense of the new urban cultural economy: Notes from Auckland, New Zealand," *Urban Studies*, vol. 40, no. 9, pp. 1699–1724, 2003.

[76] S. Zukin, "Urban lifestyles: Diversity and standardisation in spaces of consumption," *Urban Studies*, vol. 35, no. 5-6, pp. 825–839, 1998.

[77] L. Zhang and B. Liu, "Aspect and entity extraction for opinion mining," in *Data Mining and Knowledge Discovery for Big Data*. Springer, 2014, pp. 1–40.

[78] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[79] S. Ruder, P. Ghaffari, and J. G. Breslin, "A hierarchical model of reviews for aspect-based sentiment analysis," *arXiv preprint arXiv:1609.02745*, 2016.

[80] T. T. Thet, J.-C. Na, and C. S. Khoo, "Aspect-based sentiment analysis of movie reviews on discussion boards," *Journal of information science*, vol. 36, no. 6, pp. 823–848, 2010.

[81] W. Zhang, H. Xu, and W. Wan, "Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis," *Expert Systems with Applications*, vol. 39, no. 11, pp. 10 283–10 291, 2012.

[82] A. Jenkins, A. Croitoru, A. T. Crooks, and A. Stefanidis, "Crowdsourcing a Collective Sense of Place," *PLOS ONE*, vol. 11, no. 4, p. e0152932, 2016. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152932

[83] S. Wang, Z. Chen, and B. Liu, "Mining aspect-specific opinion using a holistic lifelong topic model," in *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 167–176.

[84] S. Mitchell. (2012) How San Francisco is Dealing With Chains. [Online]. Available: https://ilsr.org/san-francisco-dealing-chains/

[85] EVCC. (2015) REPORT: Preserving Local, Independent Retail. [Online]. Available: http://evccnyc.org/blog/2015/06/12/preserving-local-independent-retail/

[86] C. W. Withers, "Place and the" Spatial Turn" in Geography and in History," *Journal of the History of Ideas*, vol. 70, no. 4, pp. 637–658, 2009.

[87] J. Agnew, "Space and place," *Handbook of geographical knowledge*, pp. 316–331, 2011.

[88] L. G. Rivlin, "Group membership and place meanings in an urban neighborhood," *Journal of Social Issues*, vol. 38, no. 3, pp. 75–93, 1982.

[89] N. Ujang, "Place attachment and continuity of urban place identity," *Procedia-Social and Behavioral Sciences*, vol. 49, pp. 156–167, 2012.

[90] K. M. Jang and Y. Kim, "Crowd-sourced cognitive mapping: A new way of displaying people's cognitive perception of urban space," *PLOS ONE*, vol. 14, no. 6, p. e0218590, 2019. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0218590

[91] A. Dubey, N. Naik, D. Parikh, R. Raskar, and C. A. Hidalgo, "Deep learning the city: Quantifying urban perception at a global scale," in *European Conference on Computer Vision*. Springer, 2016, pp. 196–212.

[92] D. Quercia, J. P. Pesce, V. Almeida, and J. Crowcroft, "Psychological maps 2.0: A web engagement enterprise starting in London," in *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 2013, pp. 1065–1076.

[93] A. M. MacEachren, "Leveraging big (geo) data with (geo) visual analytics: Place as the next frontier," in *Spatial Data Handling in Big Data Era*. Springer, 2017, pp. 139–155.

[94] W. Applebaum, "Can store location research be a science?" *Economic Geography*, vol. 41, no. 3, pp. 234–237, 1965.

[95] K. Schouten and F. Frasincar, "Survey on aspect-level sentiment analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 3, pp. 813–830, 2015.

[96] M. Hu and B. Liu, "Mining opinion features in customer reviews," in *AAAI*, vol. 4, 2004, pp. 755–760.

[97] Z. Li, M. Zhang, S. Ma, B. Zhou, and Y. Sun, "Automatic extraction for product feature words from comments on the web," in *Asia Information Retrieval Symposium*. Springer, 2009, pp. 112–123.

[98] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion word expansion and target extraction through double propagation," *Computational linguistics*, vol. 37, no. 1, pp. 9–27, 2011.

[99] B. Wang and H. Wang, "Bootstrapping both product features and opinion words from chinese customer reviews with cross-inducing," in *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, 2008.

[100] N. Jakob and I. Gurevych, "Extracting opinion targets in a single-and cross-domain setting with conditional random fields," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 1035–1045.

[101] B. Adams and G. McKenzie, "Inferring Thematic Places from Spatially Referenced Natural Language Descriptions," in *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*, D. Sui, S. Elwood, and M. Goodchild, Eds. Springer Netherlands, 2013, pp. 201–221. [Online]. Available: https://doi.org/10.1007/978-94-007-4587-2_12

[102] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.

[103] W. Wang, S. J. Pan, D. Dahlmeier, and X. Xiao, "Coupled multi-layer attentions for co-extraction of aspect and opinion terms," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[104] H. Dai and Y. Song, "Neural Aspect and Opinion Term Extraction with Mined Rules as Weak Supervision," *arXiv preprint arXiv:1907.03750*, 2019.

[105] D. Ma, S. Li, F. Wu, X. Xie, and H. Wang, "Exploring Sequence-to-Sequence Learning in Aspect Term Extraction," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3538–3547.

[106] S. Poria, E. Cambria, and A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," *Knowledge-Based Systems*, vol. 108, pp. 42–49, 2016.

[107] H. Xu, B. Liu, L. Shu, and S. Y. Philip, "Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 592–598.

[108] J. E. Vargas-Muñoz, S. Lobry, A. X. Falcão, and D. Tuia, "Correcting rural building annotations in OpenStreetMap using convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 147, pp. 283–293, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S092427161830306X

[109] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[110] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[111] X. Li and W. Lam, "Deep multi-task learning for aspect term extraction with memory interaction," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2886–2892.

[112] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119. [Online]. Available: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[113] M. Naili, A. H. Chaibi, and H. H. Ben Ghezala, "Comparative study of word embedding methods in topic segmentation," *Procedia Computer Science*, vol. 112, pp. 340–349, 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877050917313480

[114] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[115] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing [Review Article]," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018-08, conference Name: IEEE Computational Intelligence Magazine.

[116] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," 2015.

[117] L. Ramshaw and M. Marcus, "Text Chunking using Transformation-Based Learning," in *Third Workshop on Very Large Corpora*, 1995. [Online]. Available: https://www.aclweb.org/anthology/W95-0107

[118] P. Ingram and H. Rao, "Store wars: The enactment and repeal of anti-chain-store legislation in America," *American Journal of Sociology*, vol. 110, no. 2, pp. 446–487, 2004.

[119] S. B. Cohen, "Location Research Programming for Voluntary Food Chains," *Economic Geography*, vol. 37, no. 1, pp. 1–11, 1961.

[120] V. Postrel. (2006) In Praise of Chain Stores. [Online]. Available: https://www.theatlantic.com/magazine/archive/2006/12/in-praise-of-chain-stores/305400/

[121] M. D. Jekanowski, "Causes and Consequences of Fast Food Sales Growth," *Food Review: The Magazine of Food Economics*, vol. 22, no. 1, 1999.

[122] C. C. Aggarwal and T. Abdelzaher, "Social sensing," in *Managing and Mining Sensor Data*. Springer, 2013, pp. 237–297.

[123] Y. Hu, C. Deng, and Z. Zhou, "A Semantic and Sentiment Analysis on Online Neighborhood Reviews for Understanding the Perceptions of People toward Their Living Environments," *Annals of the American Association of Geographers*, vol. 109, no. 4, pp. 1052–1073, 2019. [Online]. Available: https://doi.org/10.1080/24694452.2018.1535886

[124] M. F. Goodchild, "Formalizing Place in Geographic Information Systems," in *Communities, Neighborhoods, and Health: Expanding the Boundaries of Place*, ser. Social Disparities in Health and Health Care, L. M. Burton, S. A. Matthews, M. Leung, S. P. Kemp, and D. T. Takeuchi, Eds. Springer, 2011, pp. 21–33. [Online]. Available: https://doi.org/10.1007/978-1-4419-7482-2_2

[125] Y.-F. Tuan, *Space and Place: The Perspective of Experience*. U of Minnesota Press, 1977.

[126] J. Agnew, "Space and place," *The SAGE Handbook of Geographical knowledge*, pp. 316–330, 2011.

[127] T. Cresswell, *Place: An Introduction*. John Wiley & Sons, 2014-12-03.

[128] E. Shevky and W. Bell, *Social Area Analysis; Theory, Illustrative Application and Computational Procedures*, ser. Social Area Analysis; Theory, Illustrative Application and Computational Procedures. Stanford University Press, 1955.

[129] T. R. Anderson and J. A. Egeland, "Spatial aspects of social area analysis," *American Sociological Review*, pp. 392–398, 1961.

[130] S. E. Spielman and J.-C. Thill, "Social area analysis, data mining, and GIS," *Computers, Environment and Urban Systems*, vol. 32, no. 2, pp. 110–122, 2008.

[131] P. Spicker, "Charles Booth: The examination of poverty," *Social Policy & Administration*, vol. 24, no. 1, pp. 21–38, 1990. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9515.1990.tb00322.x

[132] R. Webber, "Papers: Designing geodemographic classifications to meet contemporary business needs," *Interactive Marketing*, vol. 5, no. 3, pp. 219–237, 2004. [Online]. Available: https://doi.org/10.1057/palgrave.im.4340240

[133] A. D. Singleton and P. A. Longley, "Creating open source geodemographics: Refining a national classification of census output areas for applications in higher education," *Papers in Regional Science*, vol. 88, no. 3, pp. 643–666, 2009. [Online]. Available: https://rsaiconnect.onlinelibrary.wiley.com/doi/abs/10.1111/j.1435-5957.2008.00197.x

[134] M. F. Goodchild, "Citizens as sensors: The world of volunteered geography," *GeoJournal*, vol. 69, no. 4, pp. 211–221, 2007. [Online]. Available: https://doi.org/10.1007/s10708-007-9111-y

[135] A. Stefanidis, A. Crooks, and J. Radzikowski, "Harvesting ambient geospatial information from social media feeds," *GeoJournal*, vol. 78, no. 2, pp. 319–338, 2013. [Online]. Available: https://doi.org/10.1007/s10708-011-9438-2

[136] D. Sui, S. Elwood, and M. Goodchild, *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*. Springer Science & Business Media, 2012.

[137] A. M. MacEachren, "Leveraging Big (Geo) Data with (Geo) Visual Analytics: Place as the Next Frontier," in *Spatial Data Handling in Big Data Era: Select Papers from the 17th IGU Spatial Data Handling Symposium*, ser. Advances in Geographic Information Science, C. Zhou, F. Su, F. Harvey, and J. Xu, Eds. Springer, 2017, pp. 139–155. [Online]. Available: https://doi.org/10.1007/978-981-10-4424-3_10

[138] Y. Hu, "1.07 - Geospatial Semantics," in *Comprehensive Geographic Information Systems*, B. Huang, Ed. Elsevier, 2018, pp. 80–94. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B978012409548909597X

[139] X. Yuan and A. Crooks, "Assessing the placeness of locations through user-contributed content," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, ser. GeoAI 2019. Association for Computing Machinery, 2019, pp. 15–23. [Online]. Available: https://doi.org/10.1145/3356471.3365231

[140] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[141] E. Drymonas, A. Efentakis, and D. Pfoser, "Opinion mapping travelblogs," in *Proceedings of Terra Cognita Workshop (in Conjunction with the 10th International Semantic Web Conference)*. Citeseer, 2011, pp. 23–36.

[142] A. Ballatore and B. Adams, "Extracting place emotions from travel blogs," in *Proceedings of AGILE*, vol. 2015, 2015, pp. 1–5.

[143] B. Adams and G. McKenzie, "Inferring Thematic Places from Spatially Referenced Natural Language Descriptions," in *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*, D. Sui, S. Elwood, and M. Goodchild, Eds. Springer Netherlands, 2013, pp. 201–221. [Online]. Available: https://doi.org/10.1007/978-94-007-4587-2_12

[144] G. Mai, K. Janowicz, S. Prasad, and B. Yan, "Visualizing the semantic similarity of geographic features," in *Proceedings of the Conference: AGILE, Lund, Sweden*, 2018, pp. 12–15.

[145] Y. Hu, G. McKenzie, K. Janowicz, and S. Gao, "Mining human-place interaction patterns from location-based social networks to enrich place categorization systems," in *Proceedings of the Workshop on Cognitive Engineering for Spatial Information Processes at COSIT 2015*, 2015.

[146] S. Hasan and S. V. Ukkusuri, "Urban activity pattern classification using topic models from online geo-location data," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 363–381, 2014-07-01. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X14000928

[147] S. Gao, K. Janowicz, and H. Couclelis, "Extracting urban functional regions from points of interest and human activities on location-based social networks," *Transactions in GIS*, vol. 21, no. 3, pp. 446–467, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12289

[148] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang, "Geographical topic discovery and comparison," in *Proceedings of the 20th International Conference on World Wide Web*, ser. WWW '11. Association for Computing Machinery, 2011, pp. 247–256. [Online]. Available: https://doi.org/10.1145/1963405.1963443

[149] Q. Mei, C. Liu, H. Su, and C. Zhai, "A probabilistic approach to spatiotemporal theme pattern mining on weblogs," in *Proceedings of the 15th International Conference on World Wide Web*, ser. WWW '06. Association for Computing Machinery, 2006-05-23, pp. 533–542. [Online]. Available: https://doi.org/10.1145/1135777.1135857

[150] C. Wang, J. Wang, X. Xie, and W.-Y. Ma, "Mining geographic knowledge using location aware topic model," in *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval*, ser. GIR '07. Association for Computing Machinery, 2007, pp. 65–70. [Online]. Available: https://doi.org/10.1145/1316948.1316967

[151] Q. Hao, R. Cai, C. Wang, R. Xiao, J.-M. Yang, Y. Pang, and L. Zhang, "Equip tourists with knowledge mined from travelogues," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. Association for Computing Machinery, 2010-04-26, pp. 401–410. [Online]. Available: https://doi.org/10.1145/1772690.1772732

[152] B. Hu and M. Ester, "Spatial topic modeling in online social media for location recommendation," in *Proceedings of the 7th ACM Conference on Recommender Systems*, ser. RecSys '13. Association for Computing Machinery, 2013-10-12, pp. 25–32. [Online]. Available: https://doi.org/10.1145/2507157.2507174

[153] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann, "Who, where, when and what: Discover spatio-temporal topics for twitter users," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '13. Association for Computing Machinery, 2013, pp. 605–613. [Online]. Available: https://doi.org/10.1145/2487575.2487576

[154] K. A. Schmid, A. Züfle, D. Pfoser, A. Crooks, A. Croitoru, and A. Stefanidis, "Predicting the evolution of narratives in social media," in *International Symposium on Spatial and Temporal Databases*. Springer, 2017, pp. 388–392.

[155] J. Cranshaw and T. Yano, "Seeing a home away from the home: Distilling proto-neighborhoods from incidental data with latent topic modeling," in *CSSWC Workshop at NIPS*, vol. 10, 2010.

[156] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks," in *Fifth International AAAI Conference on Weblogs and Social Media*, 2011. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/3845

[157] B. Adams and M. Raubal, "Identifying salient topics for personalized place similarity," *Research@ Locate*, vol. 14, pp. 1–12, 2014.

[158] K. Janowicz, M. Raubal, and W. Kuhn, "The semantics of similarity in geographic information retrieval," *Journal of Spatial Information Science*, vol. 2011, no. 2, pp. 29–57, 2011-05-25. [Online]. Available: http://www.josis.org/index.php/josis/article/view/26

[159] B. Yan, K. Janowicz, G. Mai, and S. Gao, "From ITDL to Place2Vec: Reasoning About Place Type Similarity and Relatedness by Learning Embeddings From Augmented Spatial Contexts," in *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL '17. Association for Computing Machinery, 2017, pp. 1–10. [Online]. Available: https://doi.org/10.1145/3139958.3140054

[160] G. Quercini and H. Samet, "Uncovering the spatial relatedness in Wikipedia," in *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL '14. Association for Computing Machinery, 2014, pp. 153–162. [Online]. Available: https://doi.org/10.1145/2666310.2666398

[161] Y. Hu, X. Ye, and S.-L. Shaw, "Extracting and analyzing semantic relatedness between cities using news articles," *International Journal of Geographical Information Science*, vol. 31, no. 12, pp. 2427–2451, 2017.

[162] I. Valavanis, G. Spyrou, and K. Nikita, "A similarity network approach for the analysis and comparison of protein sequence/structure sets," *Journal of Biomedical Informatics*, vol. 43, no. 2, pp. 257–267, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046410000134

[163] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, vol. 11, no. 3, pp. 333–337, 2014. [Online]. Available: https://www.nature.com/articles/nmeth.2810

[164] S.-A. Brown, "Patient Similarity: Emerging Concepts in Systems and Precision Medicine," *Frontiers in Physiology*, vol. 7, 2016. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fphys.2016.00561/full

[165] S. Pai and G. D. Bader, "Patient Similarity Networks for Precision Medicine," *Journal of Molecular Biology*, vol. 430, pp. 2924–2938, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022283618305321

[166] Google Geocoding API, Available online: https://developers.google.com/maps/documentation/geocoding/start, (accessed February 3, 2020).

[167] US Census, Available online: https://www2.census.gov/geo/pdfs/education/CensusTracts.pdf, (accessed February 3, 2020).

[168] S. Openshaw, "The modifiable areal unit problem," *Quantitative Geography: A British View*, pp. 60–69, 1981.

[169] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!" in *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[170] J. Boyd-Graber, D. Mimno, and D. Newman, "Care and feeding of topic models: Problems, diagnostics, and improvements," *Handbook of Mixed Membership Models and their Applications*, vol. 225255, 2014.

[171] A. Schofield and D. Mimno, "Comparing apples to apple: The effects of stemmers on topic models," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 287–300, 2016.

[172] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[173] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring Topic Coherence over Many Models and Many Topics," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012, pp. 952–961. [Online]. Available: https://www.aclweb.org/anthology/D12-1087

[174] M. A. Serrano, M. Boguná, and A. Vespignani, "Extracting the multiscale backbone of complex weighted networks," *Proceedings of the National Academy of Sciences*, vol. 106, no. 16, pp. 6483–6488, 2009.

[175] G. Lindner, C. L. Staudt, M. Hamann, H. Meyerhenke, and D. Wagner, "Structure-preserving sparsification of social networks," in *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2015, pp. 448–454.

[176] M. E. J. Newman, "Analysis of weighted networks," *Physical Review E*, vol. 70, no. 5, p. 056131, 2004. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.70.056131

[177] ——, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 23, pp. 8577–8582, 2006. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1482622/

[178] P. C. Guerra, W. Meira Jr, C. Cardie, and R. Kleinberg, "A measure of polarization on social media networks based on community boundaries," in *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[179] S. E. Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, no. 1, pp. 27–64, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1574013707000020

[180] S. E. Spielman and A. Singleton, "Studying neighborhoods using uncertain data from the American community survey: A contextual approach," *Annals of the Association of American Geographers*, vol. 105, no. 5, pp. 1003–1025, 2015.

[181] A. Singleton and D. Arribas-Bel, "Geographic data science," *Geographical Analysis*, 2019.

[182] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008. [Online]. Available: https://doi.org/10.1088%2F1742-5468%2F2008%2F10%2Fp10008

[183] X. Que, F. Checconi, F. Petrini, and J. A. Gunnels, "Scalable Community Detection with the Louvain Algorithm," in *2015 IEEE International Parallel and Distributed Processing Symposium*, 2015, pp. 28–37.

[184] M. Swartz and A. Crooks, "Comparison of Emoji Use in Names, Profiles, and Tweets," in *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*. IEEE, 2020, pp. 375–380.

[185] M. Swartz, A. T. Crooks, and W. Kennedy, "Emoji and keyword cues for diversity in social media," in *Proceedings of the 11th International Conference on Social Media and Society*, 2020.

[186] D. Almazro, G. Shahatah, L. Albdulkarim, M. Kherees, R. Martinez, and W. Nzoukou, "A survey paper on recommender systems," 2010.

[187] J. P. Sullivan and R. J. Bunker, "Drug cartels, street gangs, and warlords," *Small Wars and Insurgencies*, vol. 13, no. 2, pp. 40–53, 2002.

[188] S. O'Neil, "The Real War in Mexico: How Democracy Can Defeat the Drug Cartels," *Foreign Affairs*, vol. 88, no. 4, pp. 63–77, 2009. [Online]. Available: https://www.jstor.org/stable/20699622

[189] S. M. Young, "Going nowhere fast (or furious): The nonexistent US firearms trafficking statute and the rise of Mexican drug cartel violence," *U. Mich. JL Reform*, vol. 46, p. 1, 2012.

[190] D. A. Andelman, "The Drug Money Maze," *Foreign Affairs*, vol. 73, no. 4, pp. 94–108, 1994. [Online]. Available: https://www.jstor.org/stable/20046746

[191] S. Longmire, *Cartel: The Coming Invasion of Mexico's Drug Wars*. St. Martin's Press, 2011.

[192] K. J. Durbin, "International Narco-Terrorism and Non-State Actors: The Drug Cartel Global Threat." *Global Security Studies*, vol. 4, no. 1, 2013.

[193] T. Wainwright, *Narconomics: How to Run a Drug Cartel*. PublicAffairs, 2016.

[194] L. Freeman, "State of Siege: Drug-related violence and corruption in Mexico," *WOLA Special Report, Washington Office on Latin America, June*, 2006.

[195] R. J. Bunker and J. P. Sullivan, "Cartel evolution: Potentials and consequences," *Transnational Organized Crime*, vol. 4, no. 2, pp. 55–74, 1998.

[196] C. G. de Bustamante and J. E. Relly, "Journalism in times of violence," *Digital Journalism*, vol. 2, no. 4, pp. 507–523, 2014-10-02, _eprint: https://doi.org/10.1080/21670811.2014.882067. [Online]. Available: https://doi.org/10.1080/21670811.2014.882067

[197] A. Monroy-Hernández, E. Kiciman, D. Boyd, and S. Counts, "Narcotweets: Social media in wartime," in *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[198] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human Factors*, vol. 37, no. 1, pp. 32–64, 1995.

[199] A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford, "Senseplace2: Geotwitter analytics support for situational awareness," in *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2011, pp. 181–190.

[200] H. M. Saleem, Y. Xu, and D. Ruths, "Novel Situational Information in Mass Emergencies: What does Twitter Provide?" *Procedia Engineering*, vol. 78, pp. 155–164, 2014.

[201] R. J. Bunker and J. P. Sullivan, "Criminal Insurgencies in Mexico: Web and Social Media Resources," *CGU Faculty Publications and Research*, 2011.

[202] K. Blandford, R. D'Amour, K. Leasor, A. Terry, and I. V. de Latour, "Citizen Security and Social Media in Mexico's Public Sphere," *Virtualis*, vol. 4, no. 7, pp. 13–40, 2013. [Online]. Available: http://www.revistavirtualis.mx/index.php/virtualis/article/view/68

[203] G. Correa-Cabrera and J. Nava, "Drug wars, social networks and the right to information: The rise of informal media as the freedom of press's lifeline in northern mexico," *Social Science Research Network*, no. ID 1901909, 2011. [Online]. Available: https://papers.ssrn.com/abstract=1901909

[204] T. Roberts and G. Marchais, "Assessing the Role of Social Media and Digital Technology in Violence Reporting," *Contemporary Readings in Law and Social Justice*, vol. 10, no. 2, pp. 9–42, 2018. [Online]. Available: https://heinonline.org/HOL/P?h=hein.journals/conreadlsj10&i=176

[205] M. De Choudhury, A. Monroy-Hernández, and G. Mark, ""Narco" emotions: Affect and desensitization in social media during the mexican drug war," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '14. Association for Computing Machinery, 2014-04-26, pp. 3563–3572. [Online]. Available: https://doi.org/10.1145/2556288.2557197

[206] J. Nix, M. R. Smith, M. Petrocelli, J. Rojek, and V. M. Manjarrez, "The Use of Social Media by Alleged Members of Mexican Cartels and Affiliated Drug Trafficking Organizations," *Journal of Homeland Security and Emergency Management*, vol. 13, no. 3, pp. 395–418, 2016-09-01. [Online]. Available: https://www.degruyter.com/view/journals/jhsem/13/3/article-p395.xml

[207] S. Womer and R. J. Bunker, "Surenos gangs and Mexican cartel use of social networking sites," *Small Wars & Insurgencies*, vol. 21, no. 1, pp. 81–94, 2010.

[208] M. F. Goodchild, "Citizens as sensors: The world of volunteered geography," *GeoJournal*, vol. 69, no. 4, pp. 211–221, 2007.

[209] A. Stefanidis, A. Crooks, and J. Radzikowski, "Harvesting ambient geospatial information from social media feeds," *GeoJournal*, vol. 78, no. 2, pp. 319–338, 2013.

[210] S. Elwood, "Volunteered geographic information: Key questions, concepts and methods to guide emerging research and practice," *GeoJournal*, vol. 72, no. 3-4, pp. 133–135, 2008.

[211] S. Roche, E. Propeck-Zimmermann, and B. Mericskay, "GeoWeb and crisis management: Issues and perspectives of volunteered geographic information," *GeoJournal*, vol. 78, no. 1, pp. 21–40, 2013.

[212] N. Thapen, D. Simmie, and C. Hankin, "The early bird catches the term: Combining twitter and news data for event detection and situational awareness," *Journal of biomedical semantics*, vol. 7, no. 1, p. 61, 2016.

[213] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: What twitter may contribute to situational awareness," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2010, pp. 1079–1088.

[214] A. Amirkhanyan and C. Meinel, "Analysis of the value of public geotagged data from twitter from the perspective of providing situational awareness," in *Conference on E-Business, e-Services and e-Society*. Springer, 2016, pp. 545–556.

[215] Z. Wang, X. Ye, and M.-H. Tsou, "Spatial, temporal, and content analysis of Twitter for wildfire hazards," *Natural Hazards*, vol. 83, no. 1, pp. 523–540, 2016-08-01. [Online]. Available: https://doi.org/10.1007/s11069-016-2329-6

[216] A. Karami, V. Shah, R. Vaezi, and A. Bansal, "Twitter speaks: A case of national disaster situational awareness," *Journal of Information Science*, p. 0165551519828620, 2019.

[217] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[218] TextRazor - The Natural Language Processing API. [Online]. Available: https://www.textrazor.com/documentation

[219] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, 2008, pp. 1247–1250.

[220] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," *arXiv preprint arXiv:1802.06893*, 2018.

[221] J. Camacho-Collados and M. T. Pilehvar, "From word to sense embeddings: A survey on vector representations of meaning," *Journal of Artificial Intelligence Research*, vol. 63, pp. 743–788, 2018.

[222] T. M. Kodinariya and P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *International Journal*, vol. 1, no. 6, pp. 90–95, 2013.

[223] D. J. Ketchen and C. L. Shook, "The application of cluster analysis in strategic management research: An analysis and critique," *Strategic management journal*, vol. 17, no. 6, pp. 441–458, 1996.

[224] G. Palshikar, "Simple algorithms for peak detection in time-series," in *Proc. 1st Int. Conf. Advanced Data Analysis, Business Analytics and Intelligence*, vol. 122, 2009.

[225] J. Ansah, L. Liu, W. Kang, J. Liu, and J. Li, "Leveraging burst in twitter network communities for event detection," *World Wide Web*, 2020. [Online]. Available: https://doi.org/10.1007/s11280-020-00786-y

[226] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, and van Mulbregt, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[227] T. Warren Liao, "Clustering of time series data—a survey," *Pattern Recognition*, vol. 38, no. 11, pp. 1857–1874, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320305001305

[228] E. Navarro, F. Sajous, B. Gaume, L. Prévot, H. ShuKai, K. Tzu-Yi, P. Magistry, and H. Chu-Ren, "Wiktionary and NLP: Improving synonymy networks," in *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, ser. People's Web '09. Association for Computational Linguistics, 2009, pp. 19–27.

[229] M. Peikari, S. Salama, S. Nofech-Mozes, and A. L. Martel, "A Cluster-then-label Semi-supervised Learning Approach for Pathology Image Classification," *Scientific Reports*, vol. 8, no. 1, pp. 1–13, 2018. [Online]. Available: https://www.nature.com/articles/s41598-018-24876-0

[230] S. C. Hirtle and P. B. Heidorn, "Chapter 7 The Structure of Cognitive Maps: Representations and Processes," in *Advances in Psychology*, ser. Behavior and Environment, T. Gärling and R. G. Golledge, Eds. North-Holland, 1993, vol. 96, pp. 170–192. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0166411508600436

[231] A. Ballatore and B. Adams, "Extracting place emotions from travel blogs," in *Proceedings of AGILE*, vol. 2015, 2015, pp. 1–5.

[232] B. Adams and G. McKenzie, "Inferring thematic places from spatially referenced natural language descriptions," in *Crowdsourcing Geographic Knowledge*. Springer, 2013, pp. 201–221.

[233] A. Crooks, A. Croitoru, A. Stefanidis, and J. Radzikowski, "# Earthquake: Twitter as a distributed sensor system," *Transactions in GIS*, vol. 17, no. 1, pp. 124–147, 2013.

[234] P. M. Torrens, "Geography and computational social science," *GeoJournal*, vol. 75, no. 2, pp. 133–148, 2010. [Online]. Available: https://doi.org/10.1007/s10708-010-9361-y

Curriculum Vitae

Xiaoyi Yuan is originally from China. After graduate from Qibao High School in Shanghai in 2009, she moved to Beijing to attend China University of Political Science and Law and graduated in 2013 with a B.A. in journalism. After a few internships as a journalist and an editor, she decided to pursue a research career. She then moved to Washington DC and obtained a M.A. degree in Communication, Culture, and Technology at Georgetown University. During her time at Georgetown, she started programming and grew an interest in applying computational techniques on social science questions. She then started her PhD at George Mason University in 2013.

Over the course of her study at George Mason, she first authored 5 publications with her advisor Dr. Andrew Crooks including 3 conference papers and 2 journal articles. Xiaoyi has been a PhD candidate in Computational Social Science since 2018.