

University of Massachusetts Boston

ScholarWorks at UMass Boston

Graduate Doctoral Dissertations

Doctoral Dissertations and Masters Theses

8-2024

Unraveling Collective Anomalies in Data-Driven Systems: Manifestation, Detection, and Enhancement Implications

Mohammad Bakhsh

Follow this and additional works at: https://scholarworks.umb.edu/doctoral_dissertations



Part of the [Library and Information Science Commons](#)

UNRAVELING COLLECTIVE ANOMALIES IN DATA-DRIVEN SYSTEMS:
MANIFESTATION, DETECTION, AND ENHANCEMENT IMPLICATIONS

A Dissertation Presented
by
MOHAMMAD K. BAKHSH

Submitted to the Office of Graduate Studies,
University of Massachusetts Boston,
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2024

Information Systems for Data Science and Management Program

© 2024 by Mohammad K. Bakhsh
All rights reserved

UNRAVELING COLLECTIVE ANOMALIES IN DATA-DRIVEN SYSTEMS:
MANIFESTATION, DETECTION, AND ENHANCEMENT IMPLICATIONS

A Dissertation Presented
by
MOHAMMAD K. BAKHSH

Approved as to style and content by:

Josephine M. Namayanja, Assistance Professor
Chairperson of Committee

Shan Jiang, Associate Professor
Member

One-Ki Daniel Lee, Associate Professor
Member

Ehsan Elahi, Graduate Program Director
Management Science & Info Sys Program

Peng Xu, Chairperson
Management Science & Info Sys Department

ABSTRACT

UNRAVELING COLLECTIVE ANOMALIES IN DATA-DRIVEN SYSTEMS: MANIFESTATION, DETECTION, AND ENHANCEMENT IMPLICATIONS

August 2024

Mohammad K. Bakhsh, B.Sc., Northumbria University

MBA, San Francisco State University

Ph.D., University of Massachusetts Boston

Directed by Professor Josephine M. Namayanja

This dissertation presents an in-depth investigation into collective anomalies—complex patterns of data that deviate from the norm when considered as a group, rather than individually. It encompasses three pivotal studies that explore the intricacies of identifying and analyzing these anomalies within online customer reviews and urban traffic patterns. The research initially focuses on the subtle shifts in consumer feedback patterns, highlighting the challenges of detecting early signs of collective anomalies. It then advances to the analysis of urban traffic, emphasizing the detection of anomalous trajectory patterns and their implications for urban planning. The final study introduces a spatio-temporal framework to uncover microtransit bottlenecks, aiming to enhance urban mobility. This body of work offers insights into the nuanced manifestation, detection and implications of collective anomalies, providing a significant contribution to the field of data-driven decision-making.

ACKNOWLEDGEMENTS

“If you are grateful, I will certainly give you more” Holy Quran (14:7). My deepest thanks go to Allah Almighty for providing me with the strength, knowledge, and perseverance to complete this dissertation. I am deeply grateful to everyone who offered their invaluable support and guidance. Special thanks to my dissertation advisor, Professor Josephine Namayanja, for her insightful advice, continuous support, and inspiring mentorship. I would also like to extend my heartfelt thanks to my committee advisors, Professor Shan Jiang and Professor One-Ki Daniel Lee, for their critical insights, constructive feedback, and unwavering support. Their expertise has been crucial in the development of this dissertation. My thanks also go to Professor Ehsan Elahi for his invaluable support and encouragement in leading the program. I am deeply grateful to all the professors in the Information Systems for Data Science and Management Program for their expertise, encouragement, and the insightful discussions that have profoundly influenced my research. Finally, I am also grateful to the Saudi Arabian Ministry of Education for its generous scholarship, which has been instrumental in this achievement.

On a personal note, first and foremost, I am grateful to my parents, whose unconditional love and endless support have shaped me into the person I am today. I am deeply thankful to my wife, Nahlah, for her support, love, and faith in me throughout this journey and for the past 20 years. My children, Albaraa, Dana, and Asalah, for bringing boundless love, endless joy, and infinite happiness into my life. Lastly, my sincere thanks go to my siblings, cousins, and friends for their love, support, and for always being a source of strength.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	V
LIST OF FIGURES.....	VIII
LIST OF TABLES	IX
CHAPTER	Page
CHAPTER 1: INTRODUCTION.....	1
CHAPTER 2: BACKGROUND.....	7
2.1 Definition and Types of Anomalies.....	7
2.2 Manifestation and Detection of Collective Anomalies.....	10
2.3 Complexities in Collective Anomaly Identification	16
2.4 References	21
CHAPTER 3: THE MANIFESTATION OF REPORTING BIAS IN ONLINE CUSTOMER REVIEWS.....	23
3.1 Introduction	23
3.2 Related Work.....	26
3.3 Research Model and Hypotheses Development	32
3.4 Research Methods.....	44
3.5 Discussion	52
3.6 References	65
3.7 Appendices	69
CHAPTER 4: DISCOVERING AHP-TDT: ANOMALOUS HOTSPOT PATHS IN TRAJECTORY NETWORKS BASED ON TOTAL DISTANCE TRAVELED	71
4.1 Introduction	71
4.2 Related Work.....	75
4.3 Methodology.....	81
4.4 Research Methods.....	91
4.5 Discussion	101
4.6 References	107
4.7 Appendices	110
CHAPTER 5: TACKLING MICROTRANSIT BOTTLENECKS: A SPATIO- TEMPORAL COLLECTIVE ANOMALY DISCOVERY FRAMEWORK.....	119
5.1 Introduction	119
5.2 Related Work.....	122
5.3 Methodology.....	127

CHAPTER	Page
5.4 Research Methods.....	132
5.5 Discussion	141
5.6 References	148
5.7 Appendices	151
CHAPTER 6: CONCLUSION.....	157
BIOGRAPHICAL SKETCH OF THE AUTHOR.....	162

LIST OF FIGURES

Figure	Page
Figure 2.1: Collective Anomaly Examples.....	8
Figure 3.1: Example Cause of Self-Reporting Bias	28
Figure 3.2: Aggregated Opinion in Ideal and Biased Cases	34
Figure 3.3: Opinion Dispersion and Polarization Matrix	38
Figure 3.4: Descriptive Analysis Results	69
Figure 3.5: Spatial Analysis Using Multidimensional Scaling (MDS).....	70
Figure 4.1: Anomalous Hotspot Paths–Total Distance Traveled Framework.....	82
Figure 4.2: Results of Applying the Anomalous Hotspot Paths Framework	94
Figure 4.3: Changing Weight from Frequency to Total Distance Traveled.....	97
Figure 4.4: Spatial Hotspot Nodes and Edges	98
Figure 5.1: Spatio-Temporal Collective Anomaly Discovery Framework	129
Figure 5.2: Raw and Temporal Decomposition of Total Distance Traveled.....	136
Figure 5.3: Matrix Profile Results.....	137
Figure 5.4: Trajectory Network	138
Figure 5.5: Anomalous Hotspot Paths Within our Trajectory Network.....	141
Figure 5.6: Alignment of ST Path Initiation with São João Festival on June 23 rd	143

LIST OF TABLES

Table	Page
Table 2.1: Summary of Anomaly Types	8
Table 2.2: Overview of Anomaly Manifestation vs. Detection	14
Table 3.1: Ratio of Reviews to Total Users.....	24
Table 3.2: Types of Biases Summary.....	26
Table 3.3: Measurement Items and Sources	45
Table 3.4: Wilcoxon–Mann–Whitney U Test Results	50
Table 3.5: Wilcoxon–Mann–Whitney U Test Robustness Results	51
Table 3.6: Sample Reviews	54
Table 4.1: Summary of Scopes in Hotspot Detection	76
Table 4.2: Summary of Themes in Hotspot Detection.....	80
Table 4.3: Notations of all Variables.....	91
Table 4.4: A Sample of the Raw Data.....	92
Table 4.5: The Top Hotspot Points	97
Table 4.6: The Top Hotspot Edges.....	98
Table 4.7: Results of Experiment 1	110
Table 4.8: Results of Experiment 2.....	114
Table 5.1: Spatial and Temporal Thematic Methods Overview	127
Table 5.2: A Sample of the Raw Data.....	133
Table 5.3: Results of Experiment 1	151
Table 5.4: Results of Experiment 2.....	154

CHAPTER 1: INTRODUCTION

The landscape of data-driven systems is increasingly intricate and multifaceted. While point anomalies are easily distinguishable due to their inherent deviation from the norm, collective anomalies present a more subtle and complex challenge. In the context of collective anomalies, it is not the individual data points that are of primary concern, but rather their combined occurrence within a group. This complexity highlights how individual data points, which might seem normal on their own, can reveal significant irregularities when analyzed together (Chandola et al., 2009).

In data-driven contexts such as online review aggregation and urban traffic pattern analysis, the identification of collective anomalies plays a pivotal role. These anomalies can be a source of both concern and interest, depending on their nature and implications. For instance, in online customer reviews (OCR), an unusual divergence in review characteristics may signal a shift in consumer perception or indicate underlying biases, necessitating a deeper investigation into the factors influencing these changes. Similarly, an unusual high concentration of vehicle trajectory observations in particular spatial location could signal an

unexpected traffic pattern, necessitating interventions such as deploying additional transit services.

This study embarks on a journey to explore the phenomenon of collective anomalies, focusing on the complexities inherent in their manifestation and detection. The research problem centers on the subtle and context-dependent nature of these anomalies as they first appear within datasets. This early manifestation phase demands thorough observation and analysis to identify emerging, yet often inconspicuous, patterns. For example, in OCR, the challenge lies in discerning nuanced differences in review patterns, especially when eliciting feedback from the typically silent majority. Such variations in opinion diversity between solicited and organic reviews, while subtle, can be indicative of reporting bias, underscoring the intricate nature of identifying these anomalies in their early stages.

The detection phase, in contrast, involves a more complex and detailed analytical process. This is particularly evident in spatial contexts such as network or graph data, where the relationships between nodes and edges must be carefully examined. Identifying hotspot paths within trajectory networks exemplifies this challenge, as it entails evaluating multiple points and paths, often resulting in a multitude of combinations and significant computational demands. Furthermore, in spatio-temporal contexts, the complexity of collective anomaly detection escalates due to the need to understand the interrelationships between sequential data points, making the task even more challenging. This study aims to address these complexities, offering insights into the nuanced world of collective anomaly manifestation and detection.

In this dissertation, three distinct yet interconnected papers are presented, each delving into the complexities of collective anomalies in data-driven systems. The first paper explores the manifestation phase of collective anomalies in OCR, focusing on the subtle shifts in review patterns as a response to review solicitation. The second paper shifts the focus to the detection phase, analyzing anomalous patterns in trajectory networks and their implications for urban planning and traffic management. Building on this, the third paper delves further into the detection phase of collective anomalies by introducing a spatio-temporal collective anomaly discovery framework, aiming to enhance urban mobility through the identification of microtransit bottlenecks. Together, these studies provide a comprehensive exploration of collective anomalies, from their manifestation in consumer feedback to their detection in complex urban systems, highlighting their significance and applications in diverse data-driven contexts.

Chapter 2 lays the groundwork for understanding anomalies by diving into their definitions, types, and the specific nuances of collective anomalies. It particularly addresses the manifestation and detection of collective anomalies, highlighting the inherent challenges involved in each phase. The chapter also introduces the enhancement implications that arise from understanding collective anomalies. This exploration provides a critical foundation for the subsequent analytical approaches and applications discussed in the following chapters.

Chapter 3, titled “The Manifestation of Reporting Bias in Online Customer Reviews,” delves into the nuanced world of online customer review (OCR) systems, a vital component of modern business strategy. This chapter addresses the manifestation phase of collective anomalies, particularly focusing on identifying and understanding the subtle differences in

review patterns that emerge when the typically silent majority is prompted to provide feedback through review solicitation. The chapter conducts a thorough examination of how customer feedback dynamics shift between solicited and organic reviews, identifying and interpreting early indications of anomalies. Employing the “Experience Sphere” as a comprehensive theoretical framework, the chapter integrates theories such as herding behavior, spiral of silence, and customer review helpfulness. This framework allows for a dissection of the typical review dynamics and the distinct changes prompted by the solicitation process. The chapter aims to shed light on the silent majority’s response to solicitation, investigating how their participation via solicitation influences OCR biases. This critical exploration into the nuanced shifts in review characteristics is key to understanding the manifestation of collective anomalies and the complexities involved in early-stage anomaly detection. This chapter offers vital insights into the intricate processes of review solicitation, significantly contributing to our understanding of collective anomaly manifestation within OCR systems.

Chapter 4, titled “Discovering AHP–TDT: Anomalous Hotspot Paths in Trajectory Networks Based on Total Distance Traveled,” the emphasis shifts towards the detection phase of collective anomalies. This chapter focuses on identifying and analyzing anomalous hotspot paths within trajectory networks, an approach that extends beyond traditional point- and edge-focused methods. By emphasizing the total distance traveled (TDT), the study uncovers intricate patterns of movement that reveal collective anomalies, offering new insights into urban planning, traffic management, and other applications. In this part of the dissertation, the analysis of trajectory data is employed to discover spatial hotspots and

anomalous paths, which are essential for understanding complex travel patterns in urban environments. The chapter tackles the computational challenges involved in this analysis by employing advanced methods, including weighted connected components, to efficiently process and interpret large-scale trajectory data. Through this exploration, the study seeks to answer key questions about the comprehensive understanding of networks through the lens of collective anomalies and the impact of sophisticated analytical approaches on detecting these anomalies. This chapter contributes to the dissertation by offering a practical example of collective anomaly detection, demonstrating its enhancement implications in real-world scenarios.

Chapter 5, titled “Tackling Microtransit Bottlenecks: A Spatio-Temporal Collective Anomaly Discovery Framework,” delves further into the detection phase of collective anomalies. This chapter introduces a spatio-temporal collective anomaly discovery (STCAD) framework, which is designed to enhance urban mobility by identifying microtransit bottlenecks through the integration of spatial insights from anomalous hotspot paths and temporal nuances. Employing advanced techniques such as the matrix profile, this chapter offers a nuanced approach to anomaly detection, going beyond traditional methods to reveal complex urban mobility patterns. The STCAD framework’s combination of spatial and temporal data analysis provides a detailed understanding of microtransit bottlenecks, leading to enhanced anomaly detection and practical solutions for urban transit challenges, particularly in identifying and addressing systemic issues in microtransit systems. This approach not only enriches theoretical knowledge but also has significant practical implications, marking a substantial contribution to urban mobility studies. Crucially, it

underscores the importance of understanding collective anomalies in data-driven environments, demonstrating how a comprehensive analysis of these anomalies can lead to more effective and informed solutions in urban transit and beyond.

Chapter 6 synthesizes the insights from the preceding studies, providing a cohesive conclusion that encapsulates the multi-dimensional understanding of collective anomalies, their impact, and practical applications across the various data-driven environments.

References

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A Survey. *Computers, Materials and Continua*, 14(1), 1-22. <https://doi.org/10.1145/1541880.1541882>

CHAPTER 2: BACKGROUND

2.1 Definition and Types of Anomalies

Anomalies are also referred to as outliers, abnormalities, discordants, deviants, events, novelties, change points, faults, intrusions, misuses, surprises, and peculiarities in the data mining literature and different application domains (Aggarwal, 2017; Chandola et al., 2009; Gupta et al., 2014). Anomalies represent patterns or observations within data that do not conform to an expected behavior. In data-driven systems, understanding and identifying anomalies is crucial, as they can signal critical, often actionable insights across diverse domains such as finance, healthcare, cybersecurity, online platforms, and urban mobility (Aggarwal, 2017). Therefore, the task of anomaly detection is to identify these unusual patterns that do not align with the expected norm.

Unlike a point anomaly, where an individual data point deviates significantly from the norm, a collective anomaly occurs when an individual data point might not be anomalous by itself, but its occurrence together with other points as a group is anomalous (Ahmed et al., 2016; Chandola et al., 2009). This group could be a collection of points, a sequence in a time series, or a path in a network or a graph, as illustrated in Figure 2.1. While anomalies denote

patterns that deviate from the norm, they are distinct from simple noise in data. Noise, an unwanted phenomenon, can obstruct data analysis. The process of noise removal aims to eliminate these disturbances before any analytical procedure, whereas noise accommodation focuses on adjusting statistical models to be resilient against such disturbances (Chandola et al., 2009). In some contexts, the terms weak outliers and strong outliers are employed to differentiate between noise and genuine anomalies (Aggarwal, 2017). The different types of anomalies are summarized in the table 2.1.

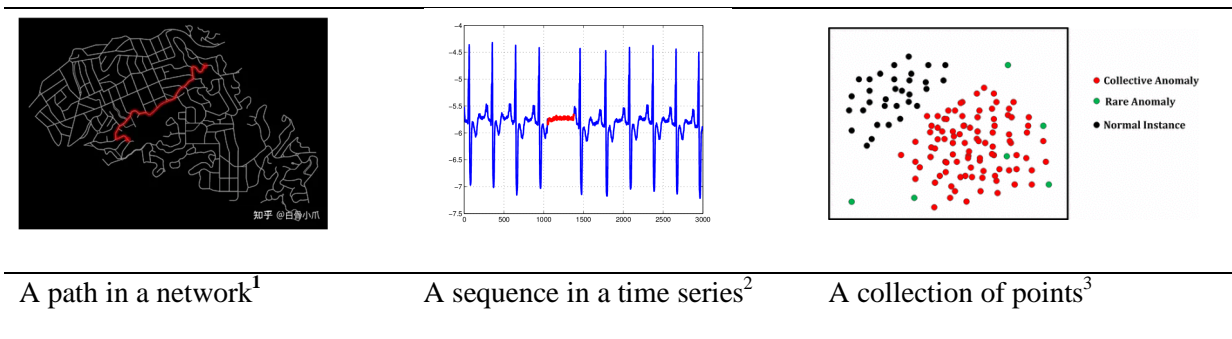


Figure 2.1: Collective Anomaly Examples

Table 2.1: Summary of Anomaly Types

Type of Anomaly	Description	Example
Point Anomalies	Individual data points that are abnormal compared to the rest of the data (Ahmed et al., 2016; Chandola et al., 2009).	A transaction where an extremely high amount of money is withdrawn.
Contextual Anomalies	Data points that are not outliers in themselves but are considered anomalous within a specific	Buying heavy winter clothing during the summer

¹ <https://zhuanlan.zhihu.com/p/546002046>

² https://www.researchgate.net/figure/Collective-anomaly-corresponding-to-an-Atrial-Premature-Contraction-in-an-human_fig4_220565847

³ https://www.researchgate.net/figure/Presence-of-rare-and-collective-anomalies-together-in-a-dataset_fig2_322689930

(Conditional Anomalies)	context (Chandola et al., 2009).	season.
Collective Anomalies	A collection of related data instances that, when considered together, are anomalous in comparison to the entire dataset. These instances might not be anomalies individually, but their occurrence as a collection or sequence is anomalous (Chandola et al., 2009).	A sequence of financial transactions that might seem normal individually but are suspicious together.

Anomaly detection complexities arise from factors such as input data characteristics, label availability, and the specific output forms required from the detection process. First, anomaly detection strategies pivot significantly based on the data type. Diverse data forms, encompassing sequential time-series, spatial configurations, or networked graph data, demand tailored detection methodologies. For instance, sequence data prevalent in reviews or trajectory analyses necessitates a nuanced approach, especially when identifying collective anomalies where data instances are interrelated (Chandola et al., 2009). The analytical technique also differs for univariate (independent) and multivariate (correlated) datasets, underscoring the need for an adaptable detection mechanism. Second, the presence or absence of labeled instances (‘normal’ or ‘anomalous’) shapes the anomaly detection approach—supervised, semi-supervised, or unsupervised. Obtaining reliable labels is often a hurdle due to anomalies’ rare nature, especially in voluminous datasets, leading to a preference for methods proficient in unsupervised or semi-supervised learning with minimal reliance on labeled examples (Chandola et al., 2009). Third, anomaly detection systems produce outputs as scores or categorical labels. Anomaly scores quantify the extent of

deviation, providing a nuanced view, while labels offer a binary classification, segregating data into normal or anomalous categories. The application's demand, whether pinpointing the anomaly degree or broader classification, influences this operational choice (Chandola et al., 2009).

2.2 Manifestation and Detection of Collective Anomalies

Manifestation of Collective Anomalies

Anomaly detection, as a more widely recognized and discussed aspect in the field of anomaly research, often overshadows the equally important phase of anomaly manifestation. This initial, underlying stage is crucial as it precedes the detection process. Manifestation of an anomaly is about its occurrence or existence within a dataset, while detection pertains to the discovery and recognition of this anomaly. In essence, an anomaly must first manifest within the dataset before it can be detected (Bergmann et al., 2021; Foorthuis, 2021; Manousis et al., 2021).

The manifestation of collective anomalies is the phase where these anomalies first occur and begin to form within a dataset. Unlike detection, which is an active process of searching and identifying, manifestation is more about the natural occurrence and expression of the anomaly. This stage is typically subtle, deeply rooted in the context of the data, and not immediately apparent. Anomalies at this point exist due to underlying conditions or behaviors in the data that may not be overtly noticeable (Bergmann et al., 2021). During the manifestation phase, unusual patterns, relationships, or behaviors start to develop within a group of data points. These can range from a sudden, unexplained rise in stock prices across

multiple sectors to an unusual behavioral pattern among a social group. The focus during manifestation is on the observable characteristics and impacts of the anomaly, rather than on the anomaly itself. It deals with the effects that are visible or measurable within the system, which are often observed visually or qualitatively (Bergmann et al., 2021).

In this phase, the anomaly expresses itself and becomes evident within the group or system. For instance, a collective anomaly in social media might initially manifest as a subtle yet consistent deviation from normal spread of misinformation patterns before being identified through detailed analysis. The manifestation of an anomaly can, therefore, be seen as a trigger for its detection, such as hate speech or bot activity leading to more focused investigations (Foorhuis, 2021; Stewart et al., 2010). Another example is novelty detection, which focuses on identifying new, previously unseen patterns in data, such as emerging topics in a discussion group. Once these novel patterns are detected, they are often integrated into the standard model, distinguishing them from anomalies that remain outliers (Chandola et al., 2009). Understanding this manifestation phase is essential as it lays the groundwork for effective detection strategies. By recognizing that anomalies first have to manifest in order to be detected, analysts and researchers can develop more nuanced approaches to monitoring and analyzing data, focusing not just on finding anomalies but also on understanding their emergence and development within the dataset.

Building on this understanding of the manifestation phase, methodologies in collective anomaly manifestation are geared toward identifying these early signs and subtle indications of anomalies within a dataset, prior to their formal detection and analysis. These methodologies diverge from detection methods that rely on specific thresholds or rankings,

and instead, prioritize the initial observation and recognition of atypical patterns and behaviors. This often involves a more qualitative approach, hinging on contextual understanding and pattern recognition (Bergmann et al., 2021; Manousis et al., 2021). Analysts might monitor for unusual correlations, shifts in data trends, or deviations from historical patterns that, while not immediately alarming, could signify the beginning of an anomalous pattern. This approach necessitates a deep understanding of the normal operational range of the dataset and a keen awareness of how minor variations might herald emerging anomalies. Although manifestation methods might utilize basic statistical measures to flag potential anomalies, these are generally seen as preliminary indicators rather than conclusive evidence. The focus here is on discerning the nuances and subtleties in the data, including changes in frequency, duration, intensity, or other characteristics that collectively signal a departure from the norm. The ultimate goal is to identify these early signs, enabling a proactive approach before the anomaly becomes more conspicuous and detectable through more structured detection methodologies (Bergmann et al., 2021).

Detection of Collective Anomalies

Following the manifestation phase, the critical stage of anomaly detection begins. This stage marks a transition from the passive observation of anomaly manifestation to the active process of discovering and recognizing these manifested anomalies. Compared to manifestation, detection is a more intricate and complex process, involving detailed and sophisticated analysis to uncover patterns and relationships within the data that are not apparent at the individual data point level. It typically employs advanced analytical techniques that evaluate data in an aggregate manner, utilizing a range of methods from

statistical analysis to complex machine learning algorithms and data mining techniques (Bergmann et al., 2021; Chandola et al., 2009; Muruti et al., 2018).

The detection and flagging of collective anomalies can be achieved through three distinct methodologies: the threshold-based approach, the top-k approach, and an integrated approach that combines elements of the first two (Aggarwal, 2017; Yeh et al., 2017). In the threshold-based approach, a specific threshold is established, often derived from statistical measures such as the data's mean and standard deviation. Groups of points are flagged as anomalous when their combined values, whether in total or on average, surpass this set threshold. For instance, setting the threshold at one or more standard deviations above the mean ensures that only significantly deviant points are marked as anomalies. Conversely, the top-k approach avoids a fixed threshold. Instead, it ranks points or groups of points based on a chosen metric, such as Total Distance Traveled (TDT), and flags the highest-scoring ones as anomalies. This method is particularly useful when it is difficult to set a meaningful threshold or when the focus is on identifying the most extreme cases. Lastly, the integrated or multifaceted approach offers a more comprehensive solution by combining the threshold and top-k methods. Here, a group of points is deemed anomalous only if it satisfies both criteria: exceeding the predetermined threshold and ranking among the top-k groups. This dual-criteria approach fosters a more robust and comprehensive anomaly detection, requiring significant deviation by multiple standards (Aggarwal, 2017; Yeh et al., 2017).

Validating the results of collective anomalies manifestation and detection is a multifaceted process that encompasses both expert judgment and quantitative analysis (Aggarwal, 2017; Bergmann et al., 2021). Initially, subject matter experts play a crucial role

by reviewing the detected anomalies for practical significance and relevance. This expert validation involves assessing if the identified anomalies align with domain knowledge and real-world expectations. Additionally, correlating these anomalies with external events or known changes in the system can further support their validity. If the anomalies correspond with real-world occurrences, this adds a layer of credibility to the detection process. Beyond expert assessment, statistical validation is essential in evaluating the effectiveness of anomaly manifestation and detection methods. Implementing statistical tests, such as Multidimensional Scaling (MDS), the Wilcoxon–Mann–Whitney (WMW) U Test, t-test, or the Chi-square test, helps in assessing the significance of the detected anomalies compared to expected norms (Abdi, 2007; Castelli et al., 2017). These tests can determine if the observed anomalies are statistically significant and not just random occurrences. Furthermore, sensitivity analysis by adjusting the parameters of the detection methodology, like the threshold or top-k value, and observing the impact on the results, can provide insights into the robustness of the detection process against changes in parameters. This comprehensive approach to validation, combining expert judgment, correlation with external events, statistical analysis, and sensitivity analysis, ensures a thorough and reliable assessment of the effectiveness of collective anomalies manifestation and detection methodologies.

Table 2.2: Overview of Anomaly Manifestation vs. Detection

Aspect	Manifestation	Detection
Definition	Occurrence or existence of anomalies	Discovery and recognition of manifested anomalies
Focus	The initial, visible or measurable	Uncovering patterns and relationships not

	characteristics and impacts, often seen visually or qualitatively	evident at the individual data point level
Approach	Passive, focusing on slight variations in the data and early indications of anomalies	Active and systematic search for and identification of anomalies
Examples	Rapid spread of misinformation, hate speech or bot activity in social media	Detecting and removing harmful content
Methodologies	<ul style="list-style-type: none"> - A qualitative approach, emphasizing contextual understanding and pattern recognition - Recognition of unusual correlations, shifts in data trends, or discrepancies in data distributions - Statistical methods (MDS, WMW U Test, t-test) <p>(Abdi, 2007; Bergmann et al., 2021; Castelli et al., 2017)</p>	<ul style="list-style-type: none"> - Advanced analytical techniques, including statistical methods, machine learning algorithms, and data mining - Employs a threshold-based approach, a top-k approach, or an integrated approach combining both <p>(Aggarwal, 2017; Muruti et al., 2018)</p>
Validation	<ul style="list-style-type: none"> - Expert validation - Correlation with external events - Statistical validation (WMW U Test, t-test, Chi-square) 	<ul style="list-style-type: none"> - Expert validation - Correlation with external events - Sensitivity analysis (fine tuning)

2.3 Complexities in Collective Anomaly Identification

Collective anomalies, depending on their nature and context, can be a source of both concern and interest. For example, biases in OCR are concerning due to their impact on online sales (Öğüt & Onur Taş, 2012), new product sales (Cui et al., 2012), conversion rates (Ludwig et al., 2013), and consumer decision-making (Guo et al., 2020). The effective identification of these anomalies is beneficial as it enhances the accuracy and trustworthiness of reviews, offering deeper insights into consumer behavior and perceptions regarding bias (Eslami et al., 2017; Han & Anderson, 2020). Such insights are invaluable for businesses as they inform the development of more refined detection techniques and corrective strategies (Dellarocas & Wood, 2008), thus fostering a more transparent and trustworthy e-commerce environment. Similarly, in broader contexts, collective anomalies can signal critical issues such as fraudulent activities in financial markets or emerging public health crises in healthcare sectors. In these scenarios, the anomalies represent potential threats or problems that require immediate attention and action to prevent adverse outcomes.

Collective anomalies, while often a source of concern, can also spark considerable interest due to their potential to unveil new trends, behaviors, or previously unrecognized phenomena. The discovery of such anomalies, particularly hotspots, yields significant benefits across various domains. For instance, in public safety, climate and environmental assessments, epidemiology, and social media analytics, identifying these spatial hotspots is essential for informed decision-making and strategic planning (Hamdi et al., 2022). Their identification can be particularly transformative in fields such as transit systems, where recognizing spatial hotspots provides critical insights for transit management, urban

planning, and infrastructure development. This leads to optimized road networks, reduced congestion, and improved transportation systems, thereby enhancing travel efficiency and user experience (Castro et al., 2013). Moreover, in social media analysis and environmental monitoring, collective anomalies can reveal new viral trends or shifts in public opinion, as well as unusual environmental patterns that might signify important changes or new factors in play (Foorhuis, 2021; Stewart et al., 2010). These discoveries are invaluable for researchers, marketers, and environmental scientists, offering fresh insights and guiding future strategies. Overall, the potential of collective anomalies to reveal such crucial information underscores their dual nature as both a source of concern in some contexts and a rich source of interest and opportunity in others.

The challenges in collective anomaly manifestation stem from the subtle, context-dependent nature of early-stage anomalies in datasets. This phase requires meticulous observation and analysis to identify emerging patterns, where anomalies are often subtly embedded within the context of the data. These anomalies arise from underlying conditions or behaviors that may not be immediately obvious and are more focused on their observable characteristics and impacts (Bergmann et al., 2021; Manousis et al., 2021). For instance, in the OCR context, the manifestation challenge involves discerning the nuanced differences in review patterns that surface when the typically silent majority is prompted to provide feedback. This manifests as variations in characteristics such as opinion diversity between solicited and organic reviews. The difficulty lies in identifying these differences as actual manifestations of reporting bias, due to their understated and intricate nature, highlighting the complexity involved in recognizing collective anomalies at their initial stage.

Compounding this challenge is the difficulty in defining what constitutes normal behavior within a dataset, where the boundaries between typical and atypical patterns are often blurred (Chandola et al., 2009). This ambiguity can lead to misinterpretations, resulting in either overlooking genuine anomalies or misidentifying normal instances as anomalies. Moreover, the concept of what is considered an anomaly varies significantly across different fields (Chandola et al., 2009). For example, a pattern of reviews that seems anomalous in one product category might be typical in another, underscoring the need for a tailored approach to understand and identify manifestations of reporting bias in different contexts. This diversity in anomaly interpretation adds another layer of complexity to the manifestation phase, requiring a nuanced and context-aware approach to effectively identify and understand these early signs of anomalies.

Unlike manifestation, the detection of collective anomalies is more intricate, involving detailed analysis to uncover patterns not apparent at the individual data point level. It employs advanced analytical techniques, ranging from statistical analysis to sophisticated machine learning algorithms and data mining, to uncover non-apparent patterns and relationships within data (Chandola et al., 2009; Muruti et al., 2018). In the spatial dimension, particularly within network or graph data, where nodes and edges represent data values and their relationships respectively, specialized models are needed to handle the structural dependencies. Anomaly detection in such networks often focuses on irregularities at either the node level, such as unusual local structures, or at the edge level, where atypical connections between different communities may be present. These could be indicators of node outliers, due to their structural deviation, or relationship outliers, stemming from

unconventional linkages between communities (Akoglu et al., 2015; Noble & Cook, 2003). In the example of identifying hotspot paths within trajectory networks, the process involves evaluating multiple points and paths. This requires evaluating all possible path combinations, a process that can result in numerous combinations and lead to substantial computational complexity, potentially reaching $O(N^3)$, where N represents the number of path combinations (Rubin, 1978). Moreover, methodologies and tools effective for unweighted networks may not be suitable for weighted networks, necessitating modified approaches for successful detection of collective anomalies in such contexts. This highlights the need for tailored solutions depending on the specific nature and structure of the network under analysis.

In the temporal dimension, the intricacies of anomaly detection further intensify due to the reliance on the interrelationships between sequential data points. This complexity becomes more pronounced in scenarios requiring continuous monitoring and comparison of new data against existing sequences to spot abnormal patterns (Aggarwal, 2017; Gupta et al., 2014). Temporal data, particularly in detecting subsequence outliers, presents unique challenges that add to the complexity. One major aspect is the length of the subsequences, which consist of multiple data points. Detection methods often vary, with some tailored for fixed-length subsequences using a sliding window approach, while others handle variable-length subsequences without a fixed length parameter. The chosen length of a subsequence inversely impacts the number of subsequences analyzed, with shorter lengths leading to a higher number of subsequences (Blázquez-García et al., 2021). Another critical factor is the representation of data. Due to the comparative complexity in analyzing subsequences, many methods opt for data transformation instead of directly analyzing raw values. A common

technique in this aspect is discretization, frequently accomplished through equal-frequency binning. Lastly, the periodic nature of some subsequence outliers, which recur at intervals, is also a significant consideration. This differs from point anomalies, where periodicity is generally not a concern (Blázquez-García et al., 2021).

The complexity of collective anomaly detection escalates further when integrating multiple types of dependencies, particularly in temporal graphs that combine both spatial and temporal dimensions. In these scenarios, outliers may indicate significant changes in network communities or relational distances. This requires sophisticated models that merge network analysis with change detection techniques, often observed in large-scale networks like social, communication, or web-based platforms. Here, structural changes can be traced in community dynamics or local properties (Akoglu et al., 2015). Temporal graphs, in particular, amplify these challenges due to the vast array of potential outlier definitions. Analysts are tasked with inferring normalcy across various criteria such as node degree, connectivity structure, community dynamics, or relational metrics before they can identify deviations. In these networks, an outlier might be reflected through marked changes in node characteristics, shifts in community affiliations, or variations in inter-node distances (Noble & Cook, 2003). Thus, the delineation of outliers in temporal networks becomes a sophisticated and nuanced process, requiring a comprehensive understanding of both the spatial and temporal aspects of the network.

2.4 References

- Abdi, H. (2007). The Bonferonni and Šidák Corrections for Multiple Comparisons. *Encyclopedia of Measurement and Statistics*(3(01)).
<https://doi.org/10.4135/9781412952644>
- Aggarwal, C. C. (2017). *Outlier Analysis*. <https://doi.org/10.1016/b978-012724955-1/50180-7>
- Ahmed, M., Naser Mahmood, A., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, 60, 19-31.
<https://doi.org/10.1016/j.jnca.2015.11.016>
- Akoglu, L., Tong, H., & Koutra, D. (2015). *Graph Based Anomaly Detection and Description: A survey* (Vol. 29). <https://doi.org/10.1007/s10618-014-0365-y>
- Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., & Steger, C. (2021). The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. *International Journal of Computer Vision*, 129(4), 1038-1059.
<https://doi.org/10.1007/s11263-020-01400-4>
- Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A Review on Outlier/Anomaly Detection in Time Series Data. *ACM Computing Surveys*, 54(3).
<https://doi.org/10.1145/3444690>
- Castelli, M., Manzoni, L., Vanneschi, L., & Popovič, A. (2017). An expert system for extracting knowledge from customers' reviews: The case of Amazon.com, Inc. *Expert Systems with Applications*, 84, 117-126.
<https://doi.org/10.1016/j.eswa.2017.05.008>
- Castro, P. S., Zhang, D., Chen, C., Li, S., & Pan, G. (2013). From taxi GPS traces to social and community dynamics: A survey. *ACM Computing Surveys*, 46(2).
<https://doi.org/10.1145/2543581.2543584>
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A Survey. *Computers, Materials and Continua*, 14(1), 1-22. <https://doi.org/10.1145/1541880.1541882>
- Cui, G., Lui, H. K., & Guo, X. (2012). The effect of online consumer reviews on new product sales. *International Journal of Electronic Commerce*, 17(1), 39-58.
<https://doi.org/10.2753/JEC1086-4415170102>
- Dellarocas, C., & Wood, C. A. (2008). The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management Science*, 54(3), 460-476.
<https://doi.org/10.1287/mnsc.1070.0747>
- Eslami, M., Vaccaro, K., Karahalios, K., & Hamilton, K. (2017). "Be careful; Things can be worse than they appear" - Understanding biased algorithms and users' behavior around them in rating platforms. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*(Icws), 62-71.
- Foorthuis, R. (2021). On the nature and types of anomalies: a review of deviations in data. *Int J Data Sci Anal*, 12(4), 297-331. <https://doi.org/10.1007/s41060-021-00265-1>
- Guo, J., Wang, X., & Wu, Y. (2020). Positive emotion bias: Role of emotional content from online customer reviews in purchase decisions. *Journal of Retailing and Consumer Services*, 52(October 2018). <https://doi.org/10.1016/j.jretconser.2019.101891>

- Gupta, M., Gao, J., Aggarwal, C., Han, J., Getoor, L., Wang, W., Gehrke, J., & Grossman, R. (2014). *Outlier Detection for Temporal Data*.
- Hamdi, A., Shaban, K., Erradi, A., Mohamed, A., Rumi, S. K., & Salim, F. D. (2022). *Spatiotemporal data mining: a survey on challenges and open problems* (Vol. 55). Springer Netherlands. <https://doi.org/10.1007/s10462-021-09994-y>
- Han, S., & Anderson, C. K. (2020). Customer Motivation and Response Bias in Online Reviews. *Cornell Hospitality Quarterly*, 61(2), 142-153. <https://doi.org/10.1177/1938965520902012>
- Ludwig, S., De Ruyter, K., Friedman, M., Brüggem, E. C., Wetzels, M., & Pfann, G. (2013). More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *Journal of Marketing*, 77(1), 87-103. <https://doi.org/10.1509/jm.11.0560>
- Manousis, A., Shah, H., Milner, H., Li, Y., Zhang, H., & Sekar, V. (2021). *The shape of view* Proceedings of the 21st ACM Internet Measurement Conference,
- Muruti, G., Rahim, F. A., & Ibrahim, Z.-A. b. (2018). *A Survey on Anomalies Detection Techniques and Measurement Methods* IEEE conference on application, information and network security (AINS),
- Noble, C. C., & Cook, D. J. (2003). Graph-Based Anomaly Detection. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1, 631-636. <https://doi.org/10.1145/956750.956831>
- Öğüt, H., & Onur Taş, B. K. (2012). The influence of internet customer reviews on the online sales and prices in hotel industry. *Service Industries Journal*, 32(2), 197-214. <https://doi.org/10.1080/02642069.2010.529436>
- Rubin, F. (1978). Enumerating All Simple Paths in a Graph. *IEEE Transactions on Circuits and Systems*, 25(8), 641-642. <https://doi.org/10.1109/TCS.1978.1084515>
- Stewart, C., Shen, K., Iyengar, A., & Yin, J. (2010). *EntomoModel: Understanding and Avoiding Performance Anomaly Manifestations* 2010 IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems,
- Yeh, C.-C. M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H. A., Silva, D. F., Mueen, A., & Keogh, E. (2017). Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets. 1317-1322. <https://doi.org/10.1109/icdm.2016.0179>

CHAPTER 3: THE MANIFESTATION OF REPORTING BIAS IN ONLINE CUSTOMER REVIEWS

3.1 Introduction

Online Customer Reviews (OCR) have become integral to modern business practices, heavily influencing consumer purchasing decisions. Customers place significant trust in peer opinions, often more than marketer-initiated sources (Susan & Schuff, 2010). OCR systems provide assessments of products and services, covering aspects such as performance, price, and reliability. However, the utility of these systems is compromised when they present a biased or distorted view, failing to accurately reflect the broader user experience. Such biased reviews have substantial effects on online sales, product adoption, conversion rates, and consumer decision-making (Cui et al., 2012; Guo et al., 2020; Ludwig et al., 2013; Öğüt & Onur Taş, 2012).

Empirical studies have consistently shown systematic biases in online consumer product ratings, including the phenomenon of self-reporting bias (Bhole & Hanna, 2017; Han & Anderson, 2020). This bias results from a tendency of certain customers, especially those with extreme experiences, to disproportionately contribute to OCR, skewing the overall data.

A notable gap exists between the vocal minority who actively share their experiences and the silent majority who typically refrain from engagement (Mai et al., 2018). While the vocal minority’s views are often seen as representative of the entire customer base, they constitute only a fraction, as evidenced by the observation that top-rated products on platforms such as Google Play rarely exceed 10% of total users (refer to Table 3.1). Despite extensive research into OCR bias, the prevailing focus has been on this vocal minority, overlooking the silent majority whose participation factors are less explored.

Table 3.1: Ratio of Reviews to Total Users⁴

Google Play	Rating Count	Installs
Amazon Fire TV	235,395	10,000,000+
Subway	266,304	10,000,000+
Twitch	4,747,040	100,000,000+
Canva	6,935,138	100,000,000+

To enhance OCR visibility and volume, review solicitation has emerged as a prominent strategy. It aims to draw in a wider range of consumer voices, particularly from the silent majority, to provide a more balanced view of consumer sentiments (Litvin & Sobel, 2019). The advantages of this practice are multifold. For consumers, it expands the available information, aiding in more informed decision-making. For businesses, it leads to increased trust, greater product visibility, and potentially higher sales (Litvin & Sobel, 2019). However, while previous research presents mixed results on the impact of review solicitation on star

⁴ <https://play.google.com/store/apps>

ratings (Burtch et al., 2018; Fradkin et al., 2015), the specific influence of review solicitation on OCR bias remains an area of uncertainty.

The study addresses this gap by embarking on an in-depth exploration of the nuanced differences in OCR that emerge from the solicitation process. Our research focuses on identifying and understanding the subtle differences in review patterns that emerge when the typically silent majority is prompted to provide feedback through review solicitation. We aim to discern how this solicitation leads to divergences in review characteristics, such as opinion diversity, polarization, negative content, and review depth, between solicited and organic reviews. This involves a careful examination of these variances to effectively identify and understand these early signs of anomalies. This study integrates the "Experience Sphere" as a comprehensive theoretical framework to guide our exploration of OCR biases. This framework synergizes the herding behavior (HB) theory, spiral of silence (SOS) theory, and customer review helpfulness (CRH) model, providing a multidimensional lens to examine the nuances of review solicitation. This integrative approach allows us to effectively differentiate between the typical variations in user opinions and the specific changes that are influenced by the solicitation process. By delving into these nuanced shifts in review characteristics, our research aims to illuminate the silent majority's response to solicitation and its subsequent impact on OCR biases. This analysis is pivotal in understanding the dynamics of review solicitation, offering comprehensive insights into how it shapes the landscape of OCR and influences consumer perceptions and behavior. The following sections will discuss related work, present our methodology including our research model and hypotheses, and finally discuss our study and both research and practical implications.

3.2 Related Work

Biases in OCR

To understand the impact of OCR on businesses, researchers have investigated this relationship through a number of theories such as information economics and social comparison theories (Susan & Schuff, 2010), human communication and communication accommodation theories (Ludwig et al., 2013), and new product diffusion theory (Cui et al., 2012). OCR research also investigates the review system's reliability and bias. This includes studying bias detection methods, its effects, and corrective strategies (Dellarocas & Wood, 2008). Moreover, it delves into consumer review behavior and their perceptions of bias (Eslami et al., 2017; Han & Anderson, 2020). To further comprehend the nature of bias, researchers have utilized signaling theory (Guo et al., 2020) and benefit-cost theory (Han & Anderson, 2020). The voluntary nature of OCR often leads to a prominent reporting bias (Han & Anderson, 2020). A comprehensive summary of other types of biases, including their occurrence, detection and mitigation strategies from previous studies, and the role of solicitation in addressing these biases, is provided in Table 3.2.

Table 3.2: Types of Biases Summary

Bias Type	Occurs	Detection/Mitigating Methods	Solicitation
Reporting	When only certain customers voluntarily decide to post their reviews. Also known as underreporting, response,	Simplify and remove any barriers in the posting process (Han & Anderson, 2020). Implement a method that calculates	Addressed as Retailer-Prompted

	nonresponse, or self-selection (Han & Anderson, 2020).	how likely users are to report different outcomes, leading to fair estimates of the distribution (Dellarocas & Wood, 2008). Apply advanced model to adjust not only overall product ratings but also sub-ratings (Lim & Tucker, 2017).	
Negativity	When consumers give more weight to negative information when making evaluations and purchase decisions (Yin et al., 2014).	Not directly addressed	Not directly addressed
Positivity	When reviews are overwhelmingly positive and the distribution of reviews is positively skewed that result in a positive emotion that positively influence customers' purchase decisions (Guo et al., 2020).	Not directly addressed	Not directly addressed
Algorithm	When a review program changes a person's entered rating to a different one that is the lowest limit in the system (Eslami et al., 2017).	Some users questioned the algorithm and asked for changes to it (Eslami et al., 2017).	Not directly addressed

Self-reporting bias has become increasingly evident in the contemporary feedback ecosystem, as highlighted by the undue emphasis on achieving high ratings, illustrated in Figure 3.1. A noticeable skew arises when dissatisfied consumers opt for silence over negative feedback, leading to an overrepresentation of contented customers. This distortion is further exacerbated when businesses encourage private issue resolution, sidelining public criticism. Such skewness of reviews is not uncommon as customers with extreme experiences, either positive or negative, are more likely to review than those with moderate experience (Bhole & Hanna, 2017).



Figure 3.1: Example Cause of Self-Reporting Bias⁵

In order to mitigate bias in OCR, platforms employ various measures such as allowing users to vote on the helpfulness of reviews, verifying the authenticity of users, and enforcing posting guidelines to promote trust. Platforms are also advised to reduce the perceived burden of posting as that increases motivation to post reviews (Han & Anderson, 2020). However, these measures are considered preventive in nature. To address bias in reviews after they have been posted, platforms look for suspicious or outlier reviews and either

⁵ <https://www.pinterest.com/pin/512636370075964709/>

eliminate them or reduce their impact. TripAdvisor for instance calculates an overall rating for a hotel and repairs bias using advanced algorithms that consider the quantity, quality, and recency of reviews. Others propose a model to mitigate product rating biases by classifying reviewers into optimistic, pessimistic, realistic, or unreliable based on their rating histories and product sales rankings (Lim & Tucker, 2017). This study contributes to the development of such algorithms by examining other factors that affect bias in OCR.

Review solicitation, also known as consumer-generated media (CGM) or retailer-prompted user-generated content, not only increases the quantity of reviews but also adds value in several ways. A higher volume of reviews can improve search and ranking algorithms, attract more customers, boost sales, and enhance consumer perceptions of product quality (Litvin & Sobel, 2019). While some studies have investigated the effects of review solicitation on bias, star ratings, and voting, the results have been inconclusive. Some studies have found a positive correlation between incentivized reviews and higher star ratings (Burtch et al., 2018), while others have reported a negative correlation (Fradkin et al., 2015). Still, some have found no correlation at all at certain cases (Burtch et al., 2018). Moreover, others examined the relationship between review solicitation and users' familiarity with review platforms in addition to the associated costs and the likelihood of posting reviews (Han & Anderson, 2020). Despite these efforts, there is still a need to fully understand the impact of review solicitation on bias and the overall reliability of online review systems.

Bias Aggregation and OCR Ecology

In the lens of OCR ecology, this section summarizes common bias aggregation dynamics and processes in the literature, such as the evolution of opinion dispersion, the intensification of opinion polarization, the amplification of negative contents, and the enrichment of review depth (Fradkin et al., 2015; Sunder et al., 2019; Susan & Schuff, 2010). Furthermore, it explores the role review solicitation plays in modulating these processes.

Opinion dispersion, or diverging opinions, refers to the degree of variability or disagreement among the opinions expressed by individuals (Li, 2018). A low opinion dispersion suggest the presence of herding behavior, as consumers could be swayed by prevailing ratings and reviews, conforming to the majority viewpoint (Sunder et al., 2019). Interestingly, low dispersion can unfavorably affect sales, particularly on e-commerce platforms and for products or services that are tangible, utilitarian, newly introduced, or associated with high financial risk (Moore et al., 2019). Contrarily, high opinion dispersion could signify a broad spectrum of viewpoints and experiences concerning a product or service. However, given the individuality and varied tastes among consumers, certain products, including niche or polarizing products and categories where consumers' tastes are expected to diverge (such as experiential goods like music), tend to consistently display higher opinion dispersion than others (Moore et al., 2019). When there is a large dispersion in ratings, consumers tend to decrease their reliance on average ratings, opting instead for other quality signals such as reading individual reviews, and often perceive reviews with extreme ratings as more helpful (Sunder et al., 2019).

Opinion polarization, where group attitudes diverge toward more extreme positions over time, can be driven by biased information, groupthink, or echo chambers (Duncan et al., 2020). This tendency for reviews to lean toward extreme positivity or negativity can distort the overall product or service perception and affect purchasing decisions, further escalating reporting bias, especially between supporting and opposing groups. To further explore this issue, researchers suggest segmenting reviews based on reported experiences (Fradkin et al., 2015). Moreover, they delved into various contributing factors from product characteristics to consumer experiences, to utilizing social influences theories such as the SOS (Xing et al., 2022), which states that individuals are less likely to voice their opinions on public issues if they believe they are in the minority, for fear of social isolation or reprisal.

Negative emotions expressed in reviews have been found to be linked to the reporting bias, given that negative reviews have a higher likelihood of being shared, viewed, and persuading other consumers (Yin et al., 2014). This propensity to report negative experiences more than positive ones can skew reviews toward an overrepresentation of negative experiences and can amplify reporting bias. In addition, the anonymity of reviewers could increase the expression of negative emotions within reviews. Research indicates that reviewers who remain anonymous are more likely to express negative emotions than those required to provide their real names (Fradkin et al., 2015). Moreover, the way businesses solicit reviews can also affect the expression of negative emotions in reviews. Both financial and non-financial incentives were found to enhance the positivity of review content and the enjoyment derived from writing reviews (Woolley & Sharif, 2021). This finding implies that

businesses may have a tendency to solicit positive reviews, thereby affecting the emotional tone of the reviews they receive.

Review depth, another construct in the CRH model, is defined as the amount of detail and information contained in a review (Susan & Schuff, 2010). Review depth is found to be positively associated with the perceived helpfulness of a review as it provides comprehensive information to support consumers' decision-making processes. Moreover, other studies have demonstrated that deeper reviews are more likely to be perceived as trustworthy and credible, thereby influencing consumer behavior (Guo et al., 2020).

While numerous bias aggregation phenomena have been discussed in the literature, insufficient attention has been devoted to comprehending the impact of review solicitation on these processes. By investigating the role of review solicitation in these phenomena, we can significantly augment our understanding of OCR ecology. For example, by analyzing the relationship between review solicitation and review depth, we can explore whether solicitation strategies foster the creation of more informative and detailed reviews or inadvertently promote superficial evaluations that may perpetuate reporting bias. Moreover, by examining the association between review solicitation and negative emotional expression in reviews, we can determine if solicited reviews lead to more emotionally charged feedback, potentially introducing or amplifying bias in the overall perception of products or services.

3.3 Research Model and Hypotheses Development

In this section, we present our research model and hypotheses about the role of review solicitation in OCR. The model is centered around a customer's experience sphere,

encompassing an ideal OCR scenario, its biased counterparts, and potential corrective responses. The model is guided by the HB theory (Ali et al., 2021), the SOS theory (Askay, 2015), and the CRH model (Susan & Schuff, 2010). By examining the role solicitation plays within these interlinked theories, we aim to unravel the underlying dynamics fostering biases in OCR.

To provide a more nuanced understanding of the complex dynamics inherent in OCR, we introduce the “Experience Sphere” as a conceptual model. This sphere accommodates diverse consumer experiences, spanning the range from extreme satisfaction to extreme dissatisfaction. Though simplified, this model is rooted on the multi-layered framework for customer experience outlined in (Gretzel & Jamal, 2009) and (De Keyser et al., 2015). Within this sphere, we distinguish between two key groups: the vocal minority, who actively express their experiences and thus form the visible layer of the sphere, and the silent majority, who, although less vocal, constitute the sphere’s unobservable yet critical mass. This categorization is informed by the SOS theory, which posits that the silent majority often refrains from voicing opinions, leaving the vocal minority’s viewpoints more prominent (Askay, 2015).

In a balanced or ideal scenario, the experience sphere features a diversified distribution of consumer opinions, spanning from highly negative to highly positive experiences (as depicted in Figure 3.2: Ideal Case). Although certain reviewers might be considered extreme or outliers near one edge, the positions of other reviewers at opposite edges and the overall distribution contribute to a less biased aggregated opinion. In this case, the silent majority opt to remain so because the visible reviews, predominantly from the vocal minority, present an

aggregated opinion that aligns with their own sentiments (Duncan et al., 2020; Gearhart & Zhang, 2015).

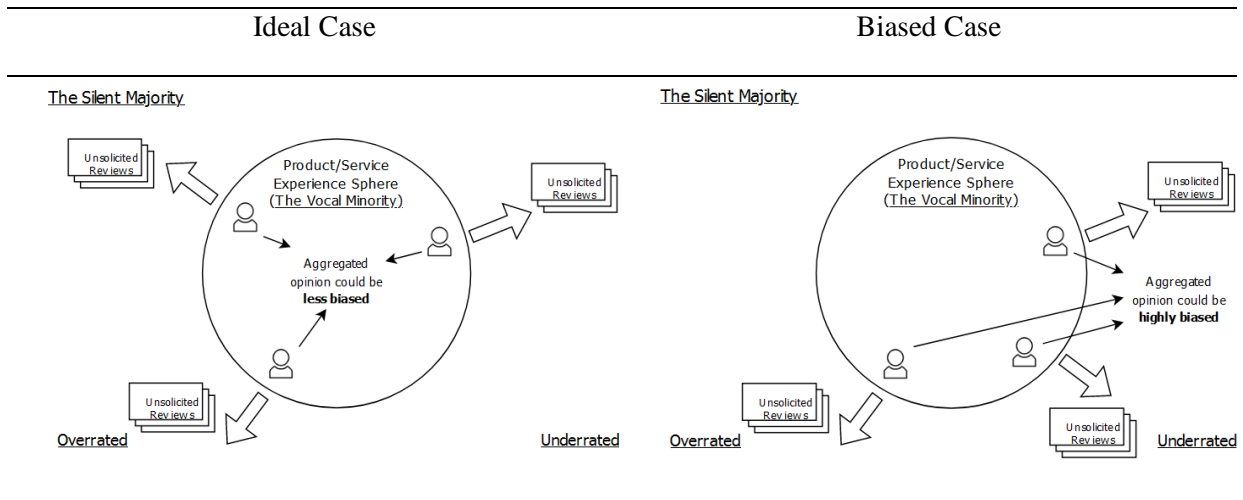


Figure 3.2: Aggregated Opinion in Ideal and Biased Cases

However, there is also a more prevalent, biased scenario, where the vocal minority’s opinions dominate and skew the observable layer of the experience sphere (depicted in Figure 3.2: Biased Case). This skewed representation is often driven by individuals with extreme experiences, either positive or negative, as they are more likely to leave reviews than those with moderate experiences (Bhole & Hanna, 2017). This phenomenon manifests in skewed rating distributions and has been substantiated by research on both negativity (Yin et al., 2014) and positivity bias (Guo et al., 2020). This imbalance may be exacerbated by herding behavior, as outlined in HB theory, which posits that individuals tend to follow dominant opinions when faced with uncertainty, thereby magnifying the vocal minority’s influence while suppressing the silent majority (Ali et al., 2021; Sunder et al., 2019).

In response to the biased OCR scenario, our primary objective is on uncovering the underlying dynamics at play, with a special attention to the role of review solicitation. While

the experience sphere embodies a broad spectrum of consumer opinions, it is crucial to investigate how external interventions, like review solicitation, can reshape its structure. Given the mixed findings in previous research regarding the impact of review solicitation on star ratings, with negative, positive, and no correlation (Burtch et al., 2018; Fradkin et al., 2015), we investigate how review solicitation may influence the herding behavior and whether it fosters or hinders the silent majority in sharing their opinions, thereby affecting the degree of opinion dispersion within the sphere.

In addition, we investigate the interplay between review solicitation and the spiral effect. While solicitation may not eradicate biases, it can push the observable layer of the experience sphere toward a balanced state by eliciting varied views from the silent majority. Thus, we aim to determine whether such initiatives mitigate or amplify opinion polarization and negative content. Moreover, as the CRH model emphasizes the significance of review depth in shaping consumer decisions (Susan & Schuff, 2010), we explore the potential of review solicitation in influencing this crucial variable, and whether such an intervention results in tangible shifts in consumer perspectives.

To this end, we explore how exactly the aforementioned mitigation effect occurs in OCR. Specifically, our objective is to uncover insights into how review solicitation influences the behavior of the silent majority in situations where a vocal minority dominates the review landscape, ultimately providing a more comprehensive understanding of the impact of review solicitation on reporting bias in OCR. To provide a deeper examination into such influence, we compare organic and solicited reviews across four dimensions: opinion dispersion, opinion polarization, negative content, and review depth.

Opinion Dispersion

According (Domingos, 2000), low opinion dispersion is often associated with high bias, while high dispersion suggests lower bias. Therefore, analyzing the dispersion of opinions in reviews can provide valuable insights into the impact of review solicitation on reporting bias. Review solicitation aims to actively encourage feedback from a diverse range of customers with varying experiences and perspectives. This diversity, in theory, culminates in a broader dispersion of opinions due to individual differences in preferences, expectations, and subjective perspectives. However, the aforementioned diversity is based on the premise that review solicitation is sampled randomly from the entire population. In reality, many factors are likely to reduce the randomness of review solicitation targets.

First, it is important to consider the potential effects of reporting bias (Han & Anderson, 2020) when soliciting reviews. There is a possibility that only a specific subset of customers will respond, such as those with extremely positive or extremely negative experiences. Particularly, the way solicitation is conducted and the incentives provided may be more likely to attract reviewers with positive opinions. For example, reviewers with negative experiences might withhold their genuine opinion when given incentives to reduce cognitive dissonance (Xi et al., 2022). Second, the timing of review solicitation can significantly influence response patterns (Brandes et al., 2022). Consumers are more likely to respond to a solicitation immediately after a positive experience, leading to an overrepresentation of positive reviews. Conversely, those with negative experiences might delay their response, contributing to non-response or delayed reporting bias. Third, solicitation might not appeal to all customers equally. Certain customers may feel

uncomfortable sharing their opinions, especially negative ones, due to fears of potential retaliation from either the company or other customers (Askay, 2015). As a result, solicitation might be less enticing to those with neutral or negative experiences, causing an overrepresentation of specific opinions in the review pool.

In light of the above factors, the random selection assumption in review solicitation becomes problematic, leading potentially to overrepresentation of specific viewpoints. Such distortion in the solicited reviews could narrow the spectrum of expressed opinions, resulting in reduced opinion dispersion, especially if those prompted to leave reviews have predominantly positive experiences. Even worse, this bias in representations can induce new reviewers to echo the prevailing opinion, succumbing to a herd mentality (Sunder et al., 2019) and becoming more susceptible to the spiral of silence effect (Duncan et al., 2020). Due to these effects, we argue that solicited reviews would exhibit less opinion dispersion compared to organic reviews. Therefore, we propose our first hypothesis:

H1: Opinion dispersion is higher in organic reviews than in solicited reviews

Opinion Polarization

Although opinion dispersion provides a broad understanding of the diversity of opinions in all reviews, research has proposed a more nuanced approach of exploring this diversity by segmenting reviews based on users' experiences (Fradkin et al., 2015). This deeper analysis is embodied by opinion polarization, which not only classifies reviews into favorable and against categories, but also investigates the extent of diversity within these segmented groups (Duncan et al., 2020). Based on synthesized data, Figure 3.4 illustrates

how high or low levels of opinion dispersion correspond to varying levels of opinion polarization.

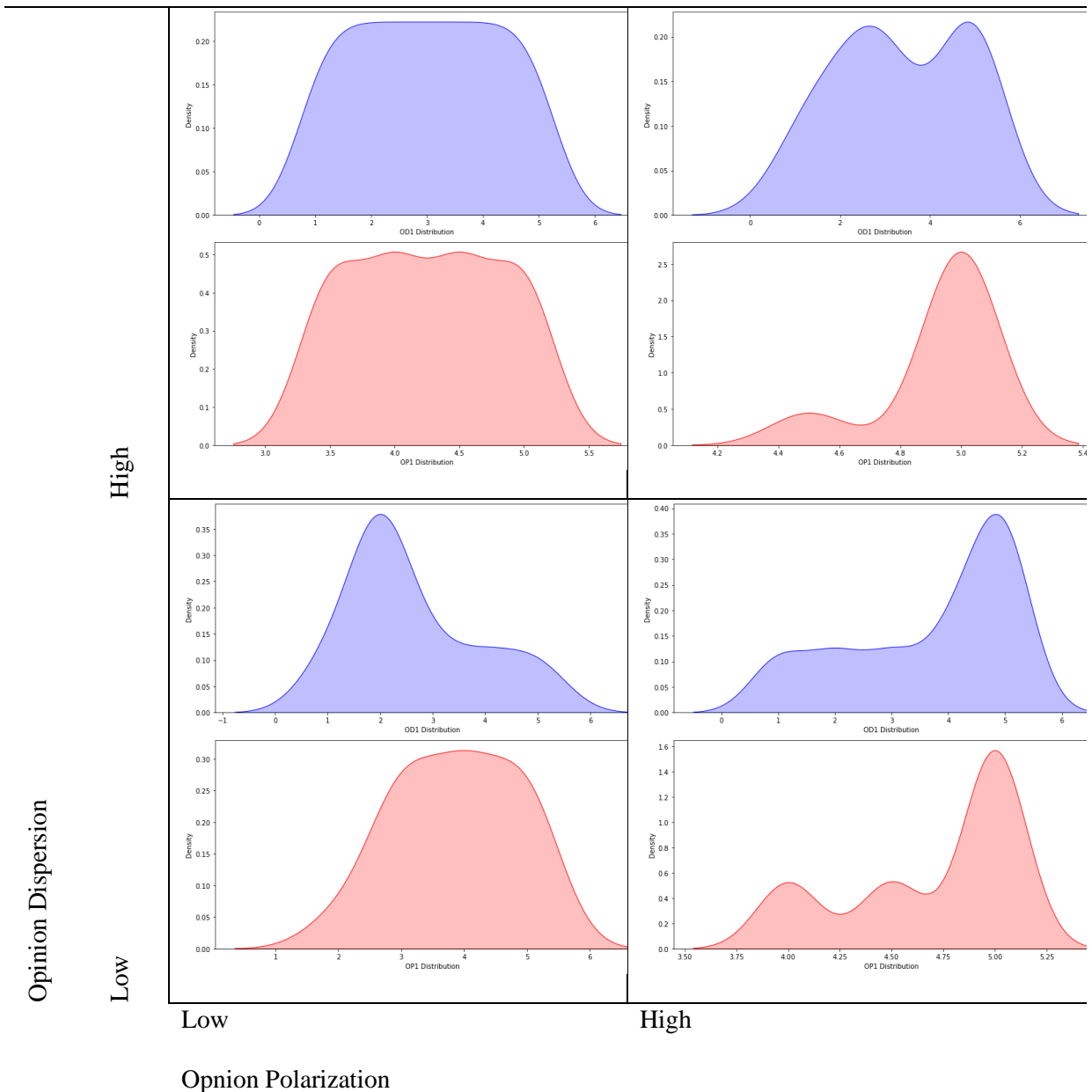


Figure 3.3: Opinion Dispersion and Polarization Matrix

Adopting the method proposed by (Han & Anderson, 2020), we categorize reviews as favorable or unfavorable by comparing a user’s rating with the aggregate platform rating.

This is further nuanced by our modification: if a user's rating exceeds or falls short of the platform's aggregate score, it signifies alignment or contradiction with the prevailing platform rating, respectively. This adjustment enables us to interpret polarization as a favorable (upvote) or unfavorable (downvote) response to the current rating. This deeper insight is crucial because a two-star rating generally suggests a near terrible experience, but when posted on a one-star average platform indicates a better-than-expected experience, suggesting an "upvote" for the existing rating and hence, polarization in its favor. Similarly, a four-star rating generally suggests a good experience, but when posted on a five-star average platform implies a less-than-expected experience, suggesting a "downvote" and hence, polarization against the existing rating.

Review solicitation, while serving as a valuable tool in gathering customer feedback, can inadvertently heighten opinion polarization in solicited reviews as compared to organic ones, for several reasons. First, the phrasing of the solicitation request itself can unknowingly guide the reviewer toward polarized opinions. The language used to request a review might evoke extreme emotions or positions, encouraging users to respond in a similar vein (Ludwig et al., 2013). For instance, requests framed in a way that accentuates positive experiences might draw out more positive reviews, pushing the distribution of ratings toward the extremes. Second, the moment at which solicitation is implemented can also influence polarization. Businesses often solicit reviews soon after purchase or interaction, a time when emotional reactions are most intense (Brandes et al., 2022). This could result in more extreme ratings, as customers who have recently had an exceptionally good or bad experience are likely to express more polarized opinions compared to those who review

organically, often at a later time when emotions have moderated. Third, the type of customers targeted for review solicitation could also play a significant role in polarization. Companies might intentionally or unintentionally target more vocal, opinionated customers or those they believe have had positive experiences (Han & Anderson, 2020), leading to more polarized feedback. On the other hand, organic reviews are typically written spontaneously by a variety of customers who might provide a wider, less polarized range of opinions. Due to these factors, we argue that solicited reviews would exhibit less opinion polarization compared to organic reviews. Therefore, we propose our second hypothesis:

H2: Opinion polarization is lower in organic reviews than in solicited reviews

Negative Content

Negative content can be present in both positive and negative reviews. When individuals receive incentives, they may feel compelled to adjust their reviews to be more positive to alleviate feelings of guilt, even if they encountered negative experiences. This tendency aligns with previous research by (Litvin & Sobel, 2019), which suggests that reviews solicited through incentives typically lean towards positivity and exhibit fewer negative sentiments compared to unsolicited ones. Furthermore, (Woolley & Sharif, 2021) found that offering both financial and non-financial incentives can heighten review positivity and enrich the reviewer's writing experience. However, these studies mainly focus on the general tone of reviews, rather than the presence of negative content. The presence of negative content in reviews, irrespective of the overall sentiment, is crucial, yet overlooked area of research. Negative reviews wield a significant influence on businesses due to their

higher likelihood of being shared, viewed, and their stronger persuasive power on other consumers (Yin et al., 2014). This impact can be extrapolated to negative content within positive reviews. Hence, we extend our analysis beyond simply categorizing reviews as positive or negative. We focus on identifying reviews that incorporate any negative elements and further analyze the intensity of these negative aspects within reviews (Askay, 2015).

While solicitation efforts primarily aim to encourage reviewers to share their positive experiences, they may inadvertently increase the appearance of negative content within both positive and negative reviews. Incentives offered during review solicitation may cause reviewers to adopt an “expert critic” mentality, feeling obliged to give a balanced perspective (Han & Anderson, 2020). This perceived obligation can make them underscore negative aspects, even in largely positive experiences, in pursuit of providing a comprehensive critique. This behavior is partly because critical thinking is often equated with negative criticism (Askay, 2015), pushing incentivized reviewers to incorporate negative content. In contrast, organic reviewers, who voluntarily share experiences without incentives, might lack this critical stance pressure, resulting in fewer criticisms and less negative content. In addition, solicited reviews typically arise from a direct business request, which could inadvertently instill a sense of obligation to provide thorough feedback (Woolley & Sharif, 2021). This sense of responsibility might magnify any negative elements of the experience, elements that might have been disregarded by organic reviewers, leading to a higher frequency of negative content within these reviews. Moreover, the process of solicitation can make customers more critical in their evaluations. Solicitation implicitly suggests that businesses are seeking feedback to improve their services (Han & Anderson, 2020). This can

cue customers to adopt a more evaluative mindset, focusing on areas of dissatisfaction that they might not have considered in an unsolicited review. Considering these aspects, we argue that solicited reviews may exhibit more negative content compared to organic reviews.

Therefore, we propose our third hypothesis:

H3: The presence of negative content is lower in organic reviews than in solicited reviews

Review Depth

Lastly, our research delves into the role of review depth in motivating the silent majority to articulate their opinions, comparing solicited reviews against organic ones. Earlier studies have noted that solicited reviews tend to be shorter and exhibit more positivity than their unsolicited counterparts (Litvin & Sobel, 2019). Furthermore, studies have underscored that review length has a tangible impact on perceived review helpfulness (Susan & Schuff, 2010). However, review depth extends beyond length, encompassing other parameters such as vocabulary richness and lexical diversity, while considering the impact of variable text lengths (Crossley et al., 2009; Ghasemaghaei et al., 2018). This consideration is especially significant given the diverse range of lengths across our review texts.

Review solicitation efforts, despite their primary aim of encouraging in-depth sharing of customer experiences, could inadvertently lead to reviews of lesser depth for several reasons. First, solicited reviews often stem from an immediate business request, prompting a quicker response (Brandes et al., 2022). Thus, reviewers might provide a condensed version of their experience, compromising the richness of their feedback. Moreover, the incentives

and rewards often associated with review solicitation might appeal to reviewers who are more driven by the lure of incentives than by the desire to share detailed experiences. This motivation could result in shorter, less comprehensive reviews. The allure of a reward may compel reviewers to write quickly or minimally, only enough to secure the incentive, thereby detracting from the depth and richness of their review. Furthermore, it is plausible that the influence of incentives could extend beyond only attracting less detailed reviews. The incentive could inadvertently foster a more transactional mindset, with reviewers focusing on the “reward” aspect of the process rather than the opportunity to share meaningful feedback. This could lead to an oversimplification of their experiences, reducing the depth and value of the insights they provide.

Second, the sentiment of the experience would generate variations in review depth. As prior studies noted that solicited reviews tends to be shorter and more positivity (Litvin & Sobel, 2019), This dynamic could affect review depth. Individuals with positive experiences may not feel the need to provide extensive details as they are satisfied with their overall experience, resulting in brief, superficial solicited reviews. On the other hand, individuals with negative experiences may be more inclined to offer a comprehensive account of the issues they encountered. Additionally, individuals with negative experiences may use online reviews as an outlet to vent their frustrations (Ghasemaghaei et al., 2018), providing comprehensive and detailed accounts, which results in deeper reviews. In light of these considerations, we argue that solicited reviews would exhibit less depth compared to organic reviews. Therefore, we propose our fourth hypothesis:

H4: Review depth is greater in organic reviews than in solicited reviews

3.4 Research Methods

We selected TripAdvisor and Trustpilot as our primary data sources due to their ability to indicate whether a review was solicited. To collect the necessary data, we developed a custom program that scraped relevant data from these two websites. From TripAdvisor, we were able to obtain approximately 2.7 million online reviews from 4,000 hotels situated in the US. The data spanned a significant period, ranging from September 9, 2002 to January 29, 2022. Similarly, we acquired roughly 0.24 million online reviews from 55 companies from Trustpilot, starting from Mar 4, 2011. Nonetheless, we carefully examined the quality of the data and restricted our analysis to a 5-year period, specifically between January 1, 2015 and December 31, 2019. We classified reviews based on their data sources, separating them into solicited and organic groups based on whether they were originally solicited.

In the operationalization of our study's constructs, we have drawn heavily from existing literature. Specifically, our constructs - opinion dispersion, negative content, and review depth - have been informed by the studies conducted by (Askay, 2015; Crossley et al., 2009; Sunder et al., 2019), while opinion polarization has been adapted from studies by (Askay, 2015; Duncan et al., 2020). A summary of our research constructs and their corresponding measurements can be found in Table 3.3. Each construct was examined using two measures as described in the table.

Table 3.3: Measurement Items and Sources

Construct	Measurement Description	Sources
Opinion Dispersion (OD)	The variance of ratings (OD1) and the variance of ratings excluding extreme ratings (OD2)	(Sunder et al., 2019)
Opinion Polarization (OP)	The rating variance of reviews aligning with the majority (OP1) and opposing the majority (OP2)	(Askay, 2015; Duncan et al., 2020)
Negative Content (NC)	The frequency (NC1) and intensity (NC2) of negative emotions in reviews	(Askay, 2015)
Review Depth (RD)	Vocabulary lexical diversity (RD1) and richness (RD2)	(Crossley et al., 2009)

To elaborate, we employed two distinct measurements for each construct. First, OD1 encompassed the entire spectrum of opinion variance, including extremes, to offer a thorough perspective on customer opinion diversity. Conversely, OD2 excluded extreme 5-star ratings, focusing on more nuanced opinion trends while keeping 1-star ratings, recognizing that consumers typically reserve 5-star reviews for exceptional experiences and use the lower scale for varying dissatisfaction levels. Second, OP1 gauged variance among ratings in line with the predominant platform rating, shedding light on the degree of consensus or diversity among those aligning with the majority. OP2, in contrast, focused on variance in ratings deviating from the main platform rating. Third, NC1 calculated the proportion of reviews containing at least one instance of negative content, defined as any explicit criticism or mention of unsatisfactory elements irrespective of the review’s overall sentiment (frequency). NC2 measured the average count of such instances of negative content per review, gauging the intensity of critical or unsatisfactory mentions within the review. Lastly,

RD1 calculated the average measure of textual lexical diversity (MTLD), while RD2 calculated the average lexical richness using Guiraud's R (Crossley et al., 2009). These comprehensive measures enabled a nuanced and thorough evaluation of the studied constructs.

Results and Analyses

Descriptive Analysis

Figure 3.4 (A) in the Appendices presents a comparative analysis of opinion dispersion (OD) in solicited and organic reviews, utilizing OD1 and OD2 measures across Trustpilot and TripAdvisor data sets. We observed a clear dominance of organic reviews in exhibiting higher opinion dispersion in the Trustpilot data set, particularly evident in both OD1 and OD2 measures. This indicated a more varied range of opinions in organic reviews. Conversely, in the TripAdvisor data set, the trend was less pronounced. Although the OD1 measure demonstrated organic reviews having slightly higher dispersion than solicited ones, this difference was not as marked as in the Trustpilot data set. However, this pattern did not replicate in the OD2 measure. Intriguingly, for OD2, both solicited and organic reviews in TripAdvisor displayed similar distributions, suggesting that opinion variability might be less pronounced in this specific industry category.

For opinion polarization (OP), figure 3.4 (B) demonstrates the contrasting distributions between solicited and organic reviews using OP1 and OP2 measures. We noted a significant increase in opinion dispersion among organic reviews, signaling a lower degree of opinion polarization compared to solicited reviews. This difference was especially pronounced in the

OP2 measure, which captured the variability of ratings from users contradicting the prevailing consensus. This trend was consistent in both Trustpilot and TripAdvisor data sets, highlighting a reduced opinion polarization in organic reviews. Interestingly, the OP1 measure, which assessed the variability of ratings from users aligning with the prevailing consensus, did not uniformly reflect this pattern. In the Trustpilot data set, the difference in opinion polarization between organic and solicited reviews was clear, echoing the pattern observed in the OP2 measure. However, in the TripAdvisor data set, the OP1 values for solicited and organic reviews were more closely aligned, indicating a less pronounced difference in opinion polarization in this context.

Figure 3.4 (C) illustrates the distribution of negative content (NC) between solicited and organic reviews. The data revealed a clear trend: solicited reviews typically had higher levels of negative content compared to organic ones. This trend was particularly evident in the NC2 measure, which calculated the average count of negative content per review, indicating a greater intensity of negativity in solicited reviews. This pattern of increased negative content in solicited reviews was consistent in both Trustpilot and TripAdvisor data sets. However, the NC1 measure, which estimated the proportion of reviews containing at least one negative aspect to reflect the frequency of negativity, presented a slightly different scenario. In the TripAdvisor data set, solicited reviews again demonstrated a higher frequency of negative content, aligning with the general trend. Yet, this correlation did not hold in the Trustpilot data set, indicating a divergence in the pattern of negative content frequency.

Finally, figure 3.4 (D) provides an illustrative comparison of review depth (RD) between solicited and organic reviews. The graphical representation distinctly illustrates that organic reviews carry a higher degree of depth in comparison to their solicited counterparts. This depth is discernible across the entire chart, suggesting a consistent trend toward more comprehensive and detailed narratives within organic reviews.

Spatial Analysis

Utilizing Multidimensional Scaling (MDS) for spatial analysis provides a nuanced perspective in comparing the organic and solicited groups. This technique excels in exploratory analysis, uncovering hidden data patterns and structures, complementing descriptive analysis. MDS effectively condenses complex relationships into more manageable dimensions, facilitating easier visualization. It achieves this by spatially arranging items, placing similar ones in proximity and disparate ones apart, thus offering a clear, intuitive understanding of group similarities and differences (Abdi, 2007; Weathers et al., 2015). Figure 3.5 in the Appendices extends the spatial analysis discussion with practical illustrations using MDS on Trustpilot and TripAdvisor data sets, encompassing both first and second measures. This analysis visually manifests the parallels and divergences between organic and solicited reviews. Notably, the overlay observed in the four figures suggests shared characteristics between the groups. Simultaneously, the distinct spread and point density, especially among outlier clusters, hint at fundamental disparities. These outlier groups potentially reveal deeper, nuanced differences between the review types, underscoring the need for additional exploration.

Wilcoxon–Mann–Whitney (WMW) U Test

To compare two sample means that come from the same population, we followed the methodology of (Borghini & Mariani, 2021; Castelli et al., 2017) by applying the WMW U test. This nonparametric statistical method is widely used in studies examining OCR. We chose the WMW U test for its effectiveness with data that does not meet the normality assumption or is ordinal. One key advantage of the WMW U test is its resilience to outliers and non-normal data distributions, making it a preferable choice over parametric tests like the t-test. Additionally, it does not require equal variances between two groups, which is beneficial when analyzing online reviews that often exhibit highly skewed distributions. The null hypothesis H_0 of the WMW U test posits that the medians or means of the two populations from which the samples are drawn are equal. The alternative hypothesis suggests that these populations have different medians or means. In our study, we applied the WMW U test to compare organic and solicited reviews in both data sets. The test results indicated whether the distribution underlying the first group (organic reviews) was stochastically less than or greater than the distribution underlying the second group (solicited reviews).

To account for errors that might generate when comparing groups across multiple data sets, we followed (Abdi, 2007) and other researchers that apply the Bonferroni correction. This correction aims to reduce the likelihood of false-positive results (type I errors) in multiple pairwise tests. It adjusts p-values to control the family-wise error rate (FWER) by dividing the critical p-value (α) by the number of comparisons made. We then calculated the statistical power of our study based on these modified p-values. This adjustment ensures greater accuracy in our results, especially when conducting multiple comparisons.

Table 3.4 presents the statistical outcomes of the WMW U test for the four dimensions examined in our research model. It details the U-Stat results and the adjusted P-value, calculated after applying the Bonferroni correction to both Trustpilot and TripAdvisor data sets. These results provide statistically significant support for our fourth hypothesis (H4) on review depth, while offering partial statistical support for hypotheses concerning opinion dispersion, opinion polarization, and negative content (H1, H2, and H3, respectively). Specifically, H4, asserting that organic reviews display greater depth than solicited reviews, received strong statistical support. However, H1, H2, and H3, which hypothesize greater opinion dispersion, lesser opinion polarization, and reduced negative content in organic reviews compared to solicited ones, only received partial statistical support.

Table 3.4: Wilcoxon–Mann–Whitney U Test Results

Measure	Hypothesis	Trustpilot		TripAdvisor		Result
		U-Stat	P-value	U-Stat	P-value	
OD1	Opinion dispersion is higher in organic reviews than in solicited reviews	29847	6.51E-	1899353	0.248076	Partially
			13***		14	Supported
OP1	Opinion polarization is lower in organic reviews than in solicited reviews	31439.5	6.21E-	1729086.	1	Partially
			18***	5		Supported
NC1	The presence of negative content is lower in organic reviews than in solicited reviews	29732.5	1	1464799.	1.31E-	Partially
				5	25***	Supported
RD1	Review depth is greater in organic reviews than in solicited reviews	30653	3.97E-15***	3417108	0	Supported

Note: P-values are the adjusted values after applying the FWER. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Robustness Test

After the initial analysis using our primary measure, we conducted a second investigation with the second measure for a more thorough robustness assessment. This approach enabled us to delve deeper into the data set and affirm the reliability of our findings. Table 3.5 presents the statistical results from the WMW U test for the four dimensions in our research model. It includes the U-Stat results and the adjusted P-value, determined after applying the Bonferroni correction, for both Trustpilot and TripAdvisor data sets. The results obtained from the second measure offer significant statistical support for our hypotheses on opinion polarization, negative content, and review depth (H2, H3, and H4, respectively). However, the hypothesis regarding opinion dispersion (H1) achieved only moderate statistical support. Specifically, H2, H3, and H4, suggesting that organic reviews have lesser opinion polarization, less negative content, and greater depth compared to solicited reviews, respectively, received full statistical support. In contrast, H1, which posits a higher degree of opinion dispersion in organic reviews than in solicited ones, gained only partial statistical support.

Table 3.5: Wilcoxon–Mann–Whitney U Test Robustness Results

Measure	Hypothesis	Trustpilot		TripAdvisor		Result
		U-Stat	P-value	U-Stat	P-value	
OD2	Opinion dispersion is higher in organic reviews than in solicited reviews	25894.5	0.000116	1849847.	1	Partially
			14***	5		Supported
OP2	Opinion polarization is lower in organic reviews than in solicited reviews	27988	1.70E-08***	2051480.	1.99E-09***	Supported

NC2	The presence of negative content is lower in organic reviews than in solicited reviews	11679	1.03E-12***	1270746.5	6.07E-58***	Supported
RD2	Review depth is greater in organic reviews than in solicited reviews	30730.5	2.37E-15***	3655565.5	0	Supported

Note: P-values are the adjusted values after applying the FWER. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

3.5 Discussion

The silent majority often refrains from expressing their opinions, leading to a dominance of perspectives from the more outspoken minority. However, when prompted by review solicitations, a portion of the silent majority may be encouraged to contribute their viewpoints. However, do these solicited reviews mirror the organic ones? To address this, we utilized two five-year data samples, delving into how solicitation impacts reporting bias in OCR. Grounded in a theory-based model that includes the concept of the experience sphere, this study explores various aspects such as opinion dispersion and polarization, the prevalence of negative content, and the depth of reviews. By examining these elements, we gained insights into the complex interplay between solicitation and the behavior of the silent majority within the experience sphere.

Our empirical findings confirm the first hypothesis: organic reviews exhibit higher opinion dispersion than solicited reviews. Although review solicitation aims to gather diverse feedback, it appears to fall short in capturing the full spectrum of user experiences. This aligns with research suggesting that solicited reviews often attract a specific segment of customers, rather than representing the entire range of experiences (Sunder et al., 2019).

Therefore, contrary to expectations, solicitation might lead to reduced opinion dispersion, introducing more bias into the OCR landscape. Factors like the timing of solicitations (Brandes et al., 2022) and fear of retaliation (Askay, 2015) could exacerbate this bias, narrowing the range of expressed opinions. Notably, our results reveal that the decreased dispersion in solicited reviews might contribute to a herding effect, as outlined in our model's experience sphere framework. Solicited reviewers could feel influenced by the dominant opinions, subconsciously aligning their feedback accordingly. This tendency suppresses nuanced opinions and individual perspectives, which are essential for countering the majority view and fostering a wider opinion dispersion. Consequently, a higher consensus emerges in solicited reviews, potentially initiating a cycle where future reviewers conform to the prevailing opinion. This perpetuates the spiral of silence effect (Duncan et al., 2020), further diminishing opinion dispersion in solicited reviews.

The partial statistical support for our hypothesis indicates varying significance of opinion dispersion across industries, as confirmed by our robustness check with a second measure. This aligns with our descriptive results, where a pattern of higher dispersion in organic reviews was notably stronger in Trustpilot than in TripAdvisor. In Trustpilot, consumer preferences are generally less diverse, focusing on common expectations like reliability and efficiency (Moore et al., 2019). Reviews from our Trustpilot sample (Table 3.6 Reviews 1 and 2) demonstrate that solicited reviews in this sector often cluster around these expectations, leading to lower opinion dispersion. This reflects the experience sphere model, where the vocal minority's similar opinions in solicited contexts overshadow broader customer experiences, resulting in less diversified opinions. Conversely, TripAdvisor's sector

is characterized by diverse consumer tastes and personal experiences (Moore et al., 2019), as evident in our TripAdvisor sample (Table 3.6 Reviews 3 and 4). Consequently, even when feedback is solicited in this sector, the resultant reviews are likely to encapsulate a vast spectrum of personal experiences and opinions, thus minimizing the significance of opinion dispersion.

Table 3.6: Sample Reviews

Review ID	Review Text
Review 1	<i>Car was available on arrival with no delays. Car was presented clean. Car gave no problems during hire period.</i>
Review 2	<i>Very easy to book. Getting the car was straightforward and the lady very helpful. Good value for money.</i>
Review 3	<i>Very quaint and charming resort offering outstanding customer service and excellent food. This was my first time and I will be back. This resort reminded me of vacations as a child. The cabins are adorable, comfortable and private yet just feet away from all the amenities.</i>
Review 4	<i>We had a fantastic time! We stayed in a 2-bedroom oceanfront cottage and absolutely loved it! We didn't want to leave. Clean, comfortable, great view. Loved the screened in porch. We'll be back for sure!</i>
Review 5	<i>We are not folks who generally return to a vacation spot - there are so many wonderful places in the world to visit. However, we will definitely return to Tween Waters. Sure, as others note, there are ways this place can be improved with updates and fine tuning - some amenities/facilities feel tired and need attention. What sets Tween Waters apart, however, is the friendly & helpful staff (we particularly enjoyed seeing Jasmine 's</i>

warm smile each morning at breakfast), the location (sandwiched between the beach & the bay/Roosevelt Channel), the setup of the resort, and the overall no-frills, no-pretense environment.

Review 6 *I like their website very much, easy to use and clear in explanations, about rules of contract, deductible in case of accidents, etc. I also never had any issues with the service provided by the actual rental companies after booking via AirportRentals. So far, great. For the first time last month however I had a phone call from them, that reminded me of some details of the reservation (good), but mostly wanted to push me to buy the additional insurance offered (not so good). I think some people might indeed need to be made aware of what the contract entails (I see many bad reviews that revolve essentially around misunderstandings), so that part is appreciated. Spending 10 minutes on the phone trying to convince me that I really need that extra insurance however was not warranted. For this reason, I give 4 * instead of 5.*

For our second hypothesis, our findings statistically support it, confirming that opinion polarization is lower in organic reviews than in solicited reviews. This insight adds depth to our previous discussion that review solicitation affects opinion diversity, demonstrating that it not only decreases overall diversity but also intensifies polarization among both supporting and opposing reviewer groups. The robustness tests and descriptive results particularly highlight a stronger polarization effect among opposing viewpoints. These findings align with prior research that attributes polarization in user-generated content to specific solicitation tactics. While (Ludwig et al., 2013) emphasized how the review solicitation language could subtly nudge users toward more extreme viewpoints, our findings support these observations, indicating that the nature and timing of solicitation can contribute to

amplified polarization. Furthermore, our results resonate with (Brandes et al., 2022) work that reviews solicited immediately after a transaction tend to capture intense emotional reactions, thus increasing polarization. Our work underscores that organic reviews, often penned after emotions have moderated, exhibit reduced polarization. In contrast, (Han & Anderson, 2020) proposition indicated that companies soliciting reviews might intentionally or unintentionally target more opinionated or satisfied customers, creating a bias toward polarized feedback. However, our study refines this argument, indicating that such strategies exacerbate polarization, particularly among those holding opposing views. This observation underscores the model's concept, where solicitation influences the vocal minority, thus skewing the observable layer of the experience sphere and leaving the silent majority's more balanced perspectives unrepresented.

Our findings indicate that the amplification effect of incentivized solicitations exceeds the self-reporting bias from emotional extremity. In the Experience Sphere model, the vocal minority, often motivated by intense experiences, dominates the sphere's visible layer (Bhole & Hanna, 2017). However, our analysis demonstrates that solicited reviews, especially when incentivized, lead to even greater opinion polarization. Incentives act as a catalyst, prompting customers to express their opinions more strongly. This dynamic distorts the ideal balanced range of opinions in the sphere, as incentives disproportionately affect the vocal minority, leading to an overrepresentation of extreme views. Table 3.6 Review 5 exemplifies this, where incentivized customers provide more intensified reviews, thus enhancing polarization and pushing the sphere toward more extreme perspectives.

Our findings support our third hypothesis that the presence of negative content is lower in organic reviews than in solicited reviews. Prior research, such as (Litvin & Sobel, 2019; Woolley & Sharif, 2021), found that solicited reviews generally contain more positive sentiments and fewer negative emotions compared to unsolicited ones. However, prior research has focused on the overall sentiment of reviews, while we have taken a more nuanced approach by examining the presence of negative content within reviews that may also contain positive feedback. For instance, one of the solicited reviews in our study stated (see Table 3.6 Review 6). Such reviews despite having positive overall sentiment, add to attention negative content that may not have been shared otherwise. Our findings suggest that solicitation efforts may inadvertently lead to increased such negative content. This can happen due to the obligation induced by explicit feedback requests enhancing negative aspect reporting (Woolley & Sharif, 2021), incentives prompting a more critical mindset, and the solicitation itself fostering a more evaluative customer approach (Han & Anderson, 2020). In other words, while solicited reviews may appear positive on the surface of the experience sphere, they often reveal underlying negative aspects.

Our finding regarding negative content in terms of negative intensity, the average count of negative content per review, was evident in all cases, as demonstrated too by our robustness tests and descriptive results. However, our findings regarding negative content in terms of negative occurrences, the proportion of reviews that incorporate at least one negative aspect, demonstrating more occurrences in TripAdvisor reviews compared to Trustpilot reviews. TripAdvisor, typically invoke a broad range of emotions, as they are closely tied to personal experiences and feelings. Consequently, customers might be more

inclined to share both positive and negative aspects of their experiences, leading to a more diverse array of expressed sentiments. When solicited for a review, a customer may feel compelled to offer a comprehensive assessment, amplifying their willingness to express negative aspects that they might have otherwise overlooked or dismissed in the absence of solicitation. In other words, the solicitation might inadvertently prompt a more meticulous evaluation of the experience, which may result in the revelation of negative sentiments that are otherwise less prevalent in organic reviews. In contrast, Trustpilot reviews are typically associated with more objective, functionality-oriented evaluations, resulting in less emotionally-charged reviews. The main concern here is whether the service performed its intended purpose effectively and efficiently. As such, unless there were significant issues affecting the service's functionality, customers might be less inclined to express negative emotions, even when solicited for a review.

Lastly, our findings provide substantial support for our fourth hypothesis stating that review depth is more profound in organic reviews compared to solicited reviews. Unlike previous studies that equated review depth with length (Litvin & Sobel, 2019; Susan & Schuff, 2010), our study enriches this understanding by considering vocabulary richness and lexical diversity. This broader approach reveals how solicitation affects the intricacy and comprehensiveness of customer feedback. Our findings suggest that solicited reviews tend to have less depth, indicating that businesses relying predominantly on solicitation may miss out on the richer, more insightful feedback typically found in organic reviews. Thus, our results highlight a potential trade-off in review solicitation strategies: while they may

increase the quantity of feedback, they might not capture the depth and richness inherent in unsolicited customer experiences.

The greater depth in organic reviews is likely as organic reviews are typically voluntary and not influenced by time constraints or incentives, enabling reviewers to provide richer and more comprehensive feedback. Moreover, individuals sharing organic reviews may be more emotionally invested in their experiences, leading to detailed accounts, especially when they've had negative encounters. This contrasts with solicited reviews, which often result from immediate business requests or incentives, and hence might lack in-depth analysis of the experience. Implications of this finding highlight the importance of organic reviews in providing detailed, valuable insights about customer experiences, which could be more useful for businesses seeking to understand and improve their products or services.

Contribution

The first key contribution of this study lies in its approach to understanding reporting bias in OCR. Previous research primarily concentrated on analyzing the perspectives of the vocal minority. In contrast, our study introduces a method for extracting and analyzing information from the typically underrepresented silent majority. By incorporating this group's perspectives, we offer a more comprehensive understanding of reporting bias in OCR, enriching the existing literature with insights from a broader range of consumer experiences and opinions.

The second main contribution of this study is the distinction and in-depth analysis of solicited versus organic reviews. Through a comprehensive examination of various dimensions, including opinion diversity, polarization, negativity content, and review depth, our research sheds light on the obscured inclinations of the silent majority. This analysis reveals the extent to which review solicitation can modulate reporting bias. By dissecting the nuanced differences between these two types of reviews, the study provides valuable insights into how the solicitation of feedback influences the representation and character of consumer opinions in the OCR landscape.

The third major contribution of this study is the introduction of our comprehensive theoretical framework, the "Experience Sphere." This framework uniquely integrates the key concepts from the HB theory, the SOS theory, and the CRH model, among other elements. By synthesizing these diverse theoretical perspectives, the Experience Sphere provides a more holistic and nuanced understanding of the dynamics in OCR. This integration allows for a deeper exploration of how individual behavior, social influence, and perceived helpfulness of reviews interact within the OCR context, offering a robust platform for analyzing the complex phenomena observed in customer review systems.

The fourth key contribution of this study is the empirical testing of our theoretical concepts using real-world review data. This practical application bridges the gap between theory and actual consumer behavior, allowing for a rigorous examination of the study's hypotheses. By analyzing actual review data from a diverse range of sources, we were able to validate and discuss the results of our theoretical framework in a real-world context. This approach not only confirms the applicability of our theoretical contributions but also provides

tangible, evidence-based insights into the dynamics of OCR. The empirical testing underscores the validity of our findings, making a significant contribution to both academic research and practical applications in the field of consumer behavior and online review systems.

Research implications:

This study significantly advances our knowledge of how the silent majority behaves in the realm of OCR. By integrating the perspectives of this typically less vocal group, the research highlights distinct patterns in review characteristics. It shows that solicited reviews, often coming from the silent majority, tend to display less diversity in opinion, higher levels of polarization, increased negativity, and less depth compared to organic reviews. These findings imply that the inclusion of the silent majority's views can lead to a different overall portrayal of consumer opinions and attitudes, which is crucial for a more accurate understanding of consumer sentiment in OCR.

The application and validation of the "Experience Sphere" theoretical framework using real-world review data represent a significant research implication. By empirically testing the intertwined concepts of HB theory, SOS theory, and the CRH model, this study provides a robust platform for analyzing complex phenomena in customer review systems. This empirical approach not only supports the theoretical model but also enhances its credibility and applicability, offering a comprehensive tool for future research to explore the dynamics of OCR and consumer behavior.

Practical implications

This study provides vital practical insights for businesses reliant on customer reviews. The insights gained from understanding the behavior of the silent majority can guide businesses in refining their review solicitation strategies. By recognizing the differences in content and tone between solicited and organic reviews, companies can tailor their approaches to encourage more genuine and representative feedback. This could involve varying the timing, wording, and medium of solicitation to capture a broader and more authentic range of consumer opinions.

In addition, the study's findings highlight the need for businesses to be aware of potential biases in solicited reviews. By understanding the tendencies of solicited reviews to display less opinion diversity and depth compared to organic ones, businesses and review platforms can develop more sophisticated tools and algorithms to detect and account for these biases. This is crucial for ensuring the reliability and credibility of the review content presented to consumers. Moreover, the research provides a framework for businesses to actively mitigate bias in their review systems. By acknowledging the different characteristics of solicited and organic reviews, companies can adjust their overall review aggregation and presentation strategies to avoid overrepresentation of certain types of feedback. This could involve weighting reviews differently based on their solicited or organic nature or developing more nuanced review analysis methods that take into account the findings of this study.

Furthermore, this research underscores for businesses that solicited reviews, especially in service industries like hospitality, may contain more negative content. This is likely

because solicitation prompts a thorough evaluation, uncovering negative sentiments. This finding is crucial for businesses aiming to fully understand and address customer dissatisfaction. Additionally, the study confirms that organic reviews typically offer greater depth, possibly due to the organic reviewer's emotional engagement and desire to share a complete experience overview. This depth is valuable for businesses, as it provides richer, more actionable feedback. Consequently, businesses should consider refining their review solicitation strategies to elicit more diverse and detailed feedback, while also being aware of the inherent biases in solicited reviews. This awareness is vital for accurately interpreting customer feedback and making informed decisions.

Limitations and Future Directions

While this study provides valuable insights into the role of review solicitation in reporting bias in OCR, it has several limitations that offer opportunities for future research. First, our study focused on data from a 5-year period from January 1, 2015, to December 31, 2019. While this provides a comprehensive view of the phenomena during this period, it does not account for changes in the online review landscape or consumer behavior that may have occurred since 2020. The study also uses a cross-sectional rather than a longitudinal design. Future research could leverage our additional data up to January 29, 2022, to perform longitudinal analysis and explore how these patterns have evolved over time.

Second, our study largely relies on quantitative data and statistical methods to examine the differences between solicited and organic reviews. Future research could benefit from employing qualitative methods, such as content analysis or interviews, to further uncover the

underlying motivations and processes that drive these observed differences. Third, while we applied our analysis on two general domains, Trustpilot and TripAdvisor, to illustrate the different dynamics in different categories, the generalizability of our findings to other industries remains unknown. Future studies could expand this research by exploring more product categories to enhance the understanding of the nuanced dynamics in each category.

Finally, we assumed that the solicited reviews were solely influenced by the solicitation itself, overlooking other potential influences such as the specific wording or presentation of the solicitation or the incentives offered. Future research could examine how these factors might further influence the content and tone of solicited reviews. Future studies should also consider investigating the impact of reviewers' demographic variables (such as age, gender, or cultural background) on the effects of review solicitation. This can further our understanding of the individual differences in review writing behaviors.

Overall, our study provides a starting point for a more nuanced understanding of the role of review solicitation in OCR. We hope our findings will stimulate further research in this area, leading to more comprehensive and detailed insights.

Conclusion

This research, delving into the role of review solicitation in OCR, makes significant contributions to the theoretical understanding of consumer behavior and offers practical insights for businesses relying on customer feedback. By employing the conceptual framework of the Experience Sphere, we have explored how solicitation reshapes the landscape of OCR, affecting opinion dispersion, polarization, negative content, and review

depth. Our findings reveal that solicited reviews often lack the depth and breadth of organic reviews. They exhibit lower opinion dispersion and more polarization, indicating a tendency toward consensus and echoing the vocal minority's viewpoints. This study highlights the potential risk of herd behavior and the spiral of silence effect in solicited reviews, where reviewers may feel pressured to align with prevailing opinions. In contrast, organic reviews, free from solicitation biases, offer a richer and more varied representation of customer experiences, providing businesses with deeper, more actionable insights. Moreover, the study underscores the inclination of solicited reviews to contain more negative content, especially in service industries like hospitality. This is a critical insight for businesses aiming to understand and address customer dissatisfaction comprehensively. In terms of review depth, the study confirms that organic reviews, driven by more emotionally engaged customers, tend to be more detailed and informative. Therefore, this research urges businesses to reconsider their review solicitation strategies, aiming for a more authentic and diverse representation of customer opinions.

3.6 References

- Abdi, H. (2007). The Bonferonni and Šidák Corrections for Multiple Comparisons. *Encyclopedia of Measurement and Statistics*(3(01)).
<https://doi.org/10.4135/9781412952644>
- Ali, M., Amir, H., & Shamsi, A. (2021). Consumer Herding Behavior in Online Buying: A Literature Review. *International Review of Management and Business Research*, 10(1), 345-360. [https://doi.org/10.30543/10-1\(2021\)-30](https://doi.org/10.30543/10-1(2021)-30)
- Askay, D. A. (2015). Silence in the crowd: The spiral of silence contributing to the positive bias of opinions in an online review system. *New Media and Society*, 17(11), 1811-1829. <https://doi.org/10.1177/1461444814535190>
- Bhole, B., & Hanna, B. (2017). The effectiveness of online reviews in the presence of self-selection bias. *Simulation Modelling Practice and Theory*.
- Borghini, M., & Mariani, M. M. (2021). Service robots in online reviews: Online robotic discourse. *Annals of Tourism Research*, 87.
<https://doi.org/10.1016/j.annals.2020.103036>

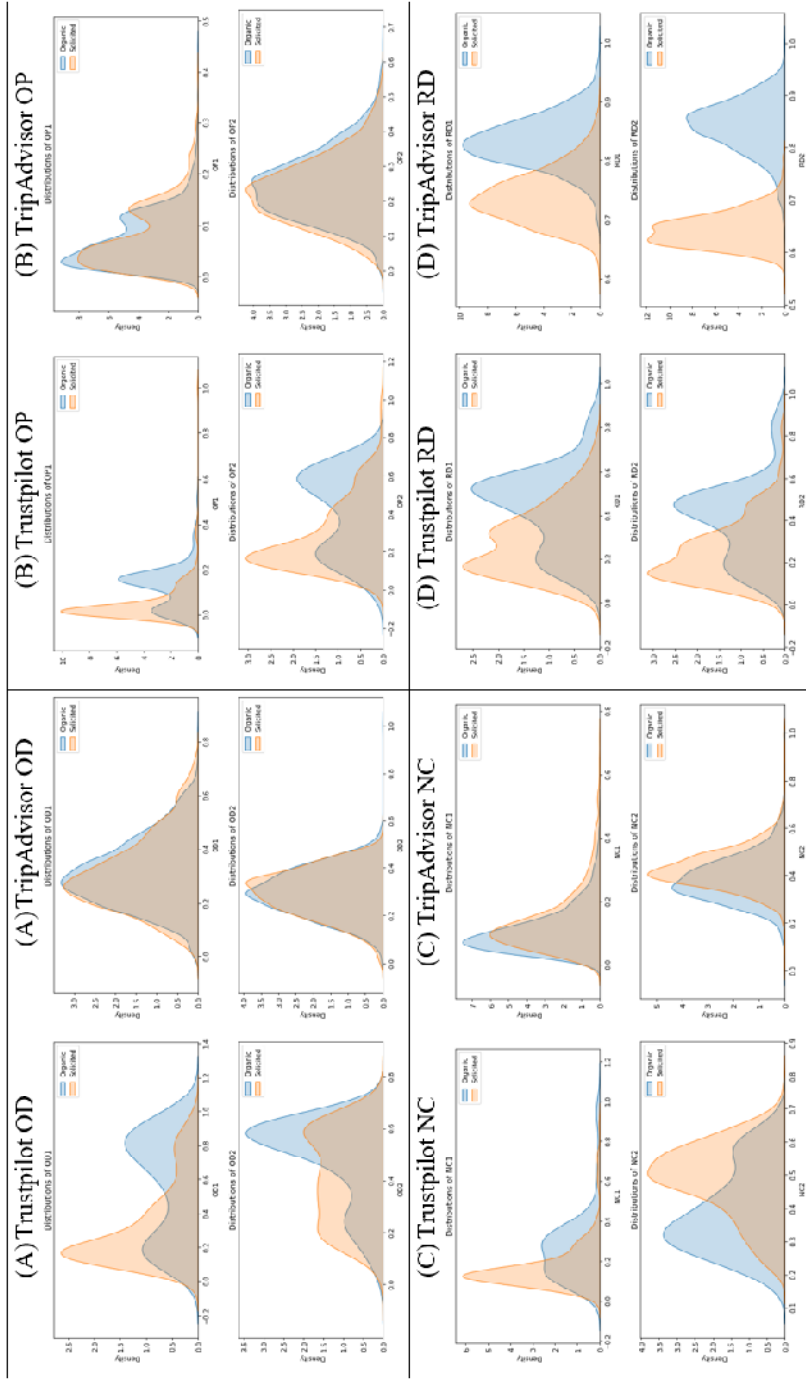
- Brandes, L., Godes, D., & Mayzlin, D. (2022). Extremity Bias in Online Reviews: The Role of Attrition. *Journal of Marketing Research*, 59(4), 675-695.
<https://doi.org/10.1177/00222437211073579>
- Burtch, G., Hong, Y., Bapna, R., & Griskevicius, V. (2018). Stimulating online reviews by combining financial incentives and social norms. *Management Science*, 64(5), 2065-2082. <https://doi.org/10.1287/mnsc.2016.2715>
- Castelli, M., Manzoni, L., Vanneschi, L., & Popovič, A. (2017). An expert system for extracting knowledge from customers' reviews: The case of Amazon.com, Inc. *Expert Systems with Applications*, 84, 117-126.
<https://doi.org/10.1016/j.eswa.2017.05.008>
- Crossley, S., Salsbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, 59(2), 307-334.
<https://doi.org/10.1111/j.1467-9922.2009.00508.x>
- Cui, G., Lui, H. K., & Guo, X. (2012). The effect of online consumer reviews on new product sales. *International Journal of Electronic Commerce*, 17(1), 39-58.
<https://doi.org/10.2753/JEC1086-4415170102>
- De Keyser, A., Lemon, K. N., Klaus, P., & Keiningham, T. L. (2015). A Framework for Understanding and Managing the CX. *Working Paper Series*, 15(121), 1-47.
- Dellarocas, C., & Wood, C. A. (2008). The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management Science*, 54(3), 460-476.
<https://doi.org/10.1287/mnsc.1070.0747>
- Domingos, P. (2000). A Unified Bias-Variance Decomposition and its Applications. *Icml*, 231-238.
- Duncan, M., Pelled, A., Wise, D., Ghosh, S., Shan, Y., Zheng, M., & McLeod, D. (2020). Staying silent and speaking out in online comment sections: The influence of spiral of silence and corrective action in reaction to news. *Computers in Human Behavior*, 102(March 2019), 192-205. <https://doi.org/10.1016/j.chb.2019.08.026>
- Eslami, M., Vaccaro, K., Karahalios, K., & Hamilton, K. (2017). "Be careful; Things can be worse than they appear" - Understanding biased algorithms and users' behavior around them in rating platforms. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*(Icwsml), 62-71.
- Fradkin, A., Grewal, E., Holtz, D., & Pearson, M. (2015). Bias and Reciprocity in Online Reviews: Evidence From Field Experiments on Airbnb. *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 1-32.
http://andreyfradkin.com/assets/reviews_paper.pdf
- Gearhart, S., & Zhang, W. (2015). "was it something i said?" "no, it was something you posted!" A study of the spiral of silence theory in social media contexts. *Cyberpsychology, Behavior, and Social Networking*, 18(4), 208-213.
<https://doi.org/10.1089/cyber.2014.0443>
- Ghasemaghaei, M., Eslami, S. P., Deal, K., & Hassanein, K. (2018). Reviews' length and sentiment as correlates of online reviews' ratings. *Internet Research*, 28(3), 544-563.
<https://doi.org/10.1108/IntR-12-2016-0394>

- Gretzel, U., & Jamal, T. (2009). Conceptualizing the creative tourist class: Technology, mobility, and tourism experiences. *Tourism Analysis*, 14(4), 471-481. <https://doi.org/10.3727/108354209X12596287114219>
- Guo, J., Wang, X., & Wu, Y. (2020). Positive emotion bias: Role of emotional content from online customer reviews in purchase decisions. *Journal of Retailing and Consumer Services*, 52(October 2018). <https://doi.org/10.1016/j.jretconser.2019.101891>
- Han, S., & Anderson, C. K. (2020). Customer Motivation and Response Bias in Online Reviews. *Cornell Hospitality Quarterly*, 61(2), 142-153. <https://doi.org/10.1177/1938965520902012>
- Li, X. (2018). Impact of average rating on social media endorsement: The moderating role of rating dispersion and discount threshold. *Information Systems Research*, 29(3), 739-754. <https://doi.org/10.1287/isre.2017.0728>
- Lim, S., & Tucker, C. S. (2017). Mitigating Online Product Rating Biases Through the Discovery of Optimistic, Pessimistic, and Realistic Reviewers. *Journal of Mechanical Design, Transactions of the ASME*, 139(11), 1-27. <https://doi.org/10.1115/1.4037612>
- Litvin, S. W., & Sobel, R. N. (2019). Organic Versus Solicited Hotel TripAdvisor Reviews: Measuring Their Respective Characteristics. *Cornell Hospitality Quarterly*, 60(4), 370-377. <https://doi.org/10.1177/1938965518811287>
- Ludwig, S., De Ruyter, K., Friedman, M., Brüggem, E. C., Wetzels, M., & Pfann, G. (2013). More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *Journal of Marketing*, 77(1), 87-103. <https://doi.org/10.1509/jm.11.0560>
- Mai, F., Shan, Z., Bai, Q., Wang, X., & Chiang, R. H. L. (2018). How Does Social Media Impact Bitcoin Value? A Test of the Silent Majority Hypothesis. *Journal of Management Information Systems*, 35(1), 19-52. <https://doi.org/10.1080/07421222.2018.1440774>
- Moore, Sarah, G., & Lafreniere, K. C. (2019). How online word-of-mouth impacts receivers. In: Society for Consumer Psychology.
- Ögüt, H., & Onur Taş, B. K. (2012). The influence of internet customer reviews on the online sales and prices in hotel industry. *Service Industries Journal*, 32(2), 197-214. <https://doi.org/10.1080/02642069.2010.529436>
- Sunder, S., Kim, K. H., & Yorkston, E. A. (2019). What Drives Herding Behavior in Online Ratings? The Role of Rater Experience, Product Portfolio, and Diverging Opinions. *Journal of Marketing*, 83(6), 93-112. <https://doi.org/10.1177/0022242919875688>
- Susan, M. M., & Schuff, D. (2010). What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. *Angewandte Chemie International Edition*, 6(11), 951-952., 34(1), 185-200.
- Weathers, D., Swain, S. D., & Grover, V. (2015). Can online product reviews be more helpful? Examining characteristics of information content by product type. *Decision Support Systems*, 79, 12-23. <https://doi.org/10.1016/j.dss.2015.07.009>
- Woolley, K., & Sharif, M. A. (2021). Incentives Increase Relative Positivity of Review Content and Enjoyment of Review Writing. *Journal of Marketing Research*, 58(3), 539-558. <https://doi.org/10.1177/00222437211010439>

- Xi, W., Baymuminova, N., Zhang, Y.-W., & Xu, S.-N. (2022). Cognitive Dissonance and Public Compliance, and Their Impact on Business Performance in Hotel Industry. *Sustainability*, 14(22). <https://doi.org/10.3390/su142214907>
- Xing, Y., Wang, X., Qiu, C., Li, Y., & He, W. (2022). Research on opinion polarization by big data analytics capabilities in online social networks. *Technology in Society*, 68(January). <https://doi.org/10.1016/j.techsoc.2022.101902>
- Yin, D., Bond, S. D., & Zhang, H. (2014). ANXIOUS OR ANGRY? EFFECTS OF DISCRETE EMOTIONS ON THE PERCEIVED HELPFULNESS OF ONLINE REVIEWS. *MIS Quarterly*, 38(2), 539-560.

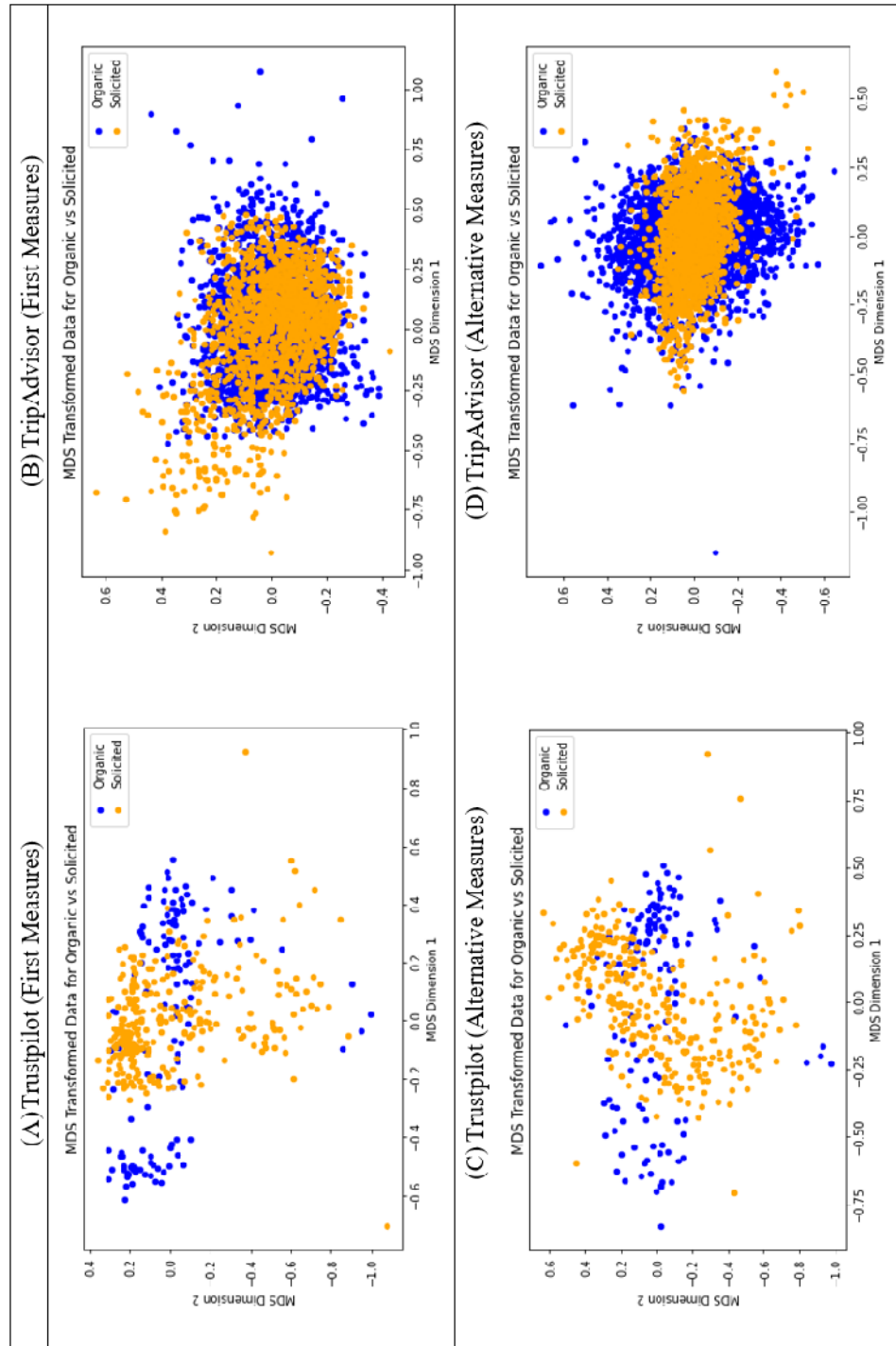
3.7 Appendices

Figure 3.4: Descriptive Analysis Results



TripAdvisor data was acquired from car rentals companies. TripAdvisor data was acquired from hotels from Florida, Massachusetts, and Nevada.

Figure 3.5: Spatial Analysis Using Multidimensional Scaling (MDS)



CHAPTER 4: DISCOVERING AHP-TDT: ANOMALOUS HOTSPOT PATHS IN TRAJECTORY NETWORKS BASED ON TOTAL DISTANCE TRAVELED

4.1 Introduction

The problem of discovering anomalous hotspot paths based on total distance traveled (AHP-TDT) aims to systematically identify and analyze pathways within a trajectory network that exhibit irregular or unexpected patterns based on the cumulative distance traveled. By focusing on the total distance traveled (TDT), this methodology seeks to shed light on collective anomalies, revealing potential areas of interest or concern. This approach extends beyond simple frequency-based metrics and delves deeper into the intricacies of movement dynamics, enabling a comprehensive understanding of spatial and temporal traffic behaviors. The ultimate objective is to extract meaningful insights from the data, which could be crucial for various applications such as urban planning, traffic management, and security surveillance, among others. Concurrently, this study is dedicated to optimizing various aspects of the travel experience.

Spatial hotspots are areas characterized by a heightened concentration of incidents or events. These concentrations manifest in diverse scenarios such as traffic bottlenecks, crime

hotbeds, disease flare-ups, and natural disasters (Qin et al., 2017). Recognizing these hotspots is invaluable across various applications, including public safety, climatic and environmental assessments, epidemiology, social media analytics, and the internet of things (Hamdi et al., 2022). In the context of transit systems, comprehending these spatial hotspots is pivotal for refining transit management, urban design, maintenance, and forward-thinking development (Castro et al., 2013). However, the traditional focus on detecting spatial hotspots has generally been confined to either spatial points or edges defined between pairs of points. This limitation leaves a critical knowledge gap, as complex travel patterns that consist of sequential edges—paths—present an unresolved challenge. This study bridges this gap by extending beyond the conventional methods of pinpointing spatial hotspots at individual points or edges. Instead, it ventures into uncharted territory, unearthing, first, a sequence of spatial hotspot edges, referred to as hotspot paths, that are woven into a network of trip trajectories, and, second, isolating those spatial hotspot paths that present anomalous travel patterns. The unveiling of these intricate paths has the potential to revolutionize our approach to spatial analysis, opening doors to richer insights and more targeted interventions.

Unlike spatial data, which usually represents a single point in a geographic coordinate system, trajectories capture the movement of objects over time. They are represented as a sequence of spatial points, arranged by timestamps. Additionally, trajectories may contain supplementary data that pertains to speed and direction (Wang et al., 2020). A combination of these features can be useful in providing valuable insights into collective human movements and can improve our understanding of social interactions defined by individuals' behavior as well as socio-dynamics characterized by group-level behavior. To determine

whether a path in the trajectory network is an anomalous hotspot, we model a city's road network as a graph $G = (V, E)$, where V represents the set of nodes (or vertices) that correspond to street crossings, and E represents the set of edges that connect these nodes and correspond to road segments. A path in this network is defined as an edge or a sequence of connected edges that does not repeat any nodes.

However, the detection of AHP-TDT from trajectory data presents significant challenges due to the multiple points and paths involved. The process requires an evaluation of all combinations of paths within the trajectory network, which can potentially result in millions of combinations. As a result, the complexity of this approach can be substantial, with a computational cost of $O(N^3)$, where N is the number of path combinations in the trajectory network (Rubin, 1978). Furthermore, considering that the trajectory network is weighted by the traffic events that occur on each edge, a solution based on extracting subgraphs using the simple connected components method is not suitable. Thus, a weighted connected component approach must be used to effectively search the trajectory network for the desired anomalous hotspots.

To address the challenge of pinpointing anomalous hotspots in trajectory data, several methods have emerged, spanning from exact to heuristic approximation techniques. These methods, as outlined by (Nogueira et al., 2018), harness computational algorithms, statistics, and machine learning to capture and classify these intricate spatial hotspots. A hybrid strategy converting a spatial hotspot problem into a tailored optimization issue has demonstrated marked efficiency, solution quality, and robustness. Notably, this strategy was adeptly used for a maximum weight independent set problem (Nogueira et al., 2018) and for

a ridesharing matching problem (Tu et al., 2019). In this research, we adopt this transformative approach, leveraging weighted connected components to unearth anomalous hotspots within the trajectory network.

In this study, we aim to advance the discovery of anomalous hotspots from trajectory data by extending the focus from single points or edges to paths based on TDT. This approach uncovers new insights into various applications, such as on-demand delivery services, shared mobility, and urban planning. The identification of anomalous hotspots based on TDT has the potential to optimize network utilization for autonomous vehicle delivery services, reduce redundancy in the delivery network, and alleviate traffic congestion. Moreover, it could play a crucial role in improving the quality of dynamic shared mobility services. Thus, this study aims to answer three fundamental questions related to AHP-TDT in trip trajectories:

R1. How does analyzing collective anomalies in paths via AHP-TDT offer a more comprehensive understanding of a network compared to traditional point- and edge-focused methods?

R2. How does defining network bounds in AHP-TDT enhance hotspot detection precision and relevance in trajectory data?

R3. What is the impact of a weighted connected component approach on identifying anomalous hotspots in trajectory networks?

The paper is organized as follows: Section 2 discusses related work, and Section 3 presents the methodology, including the question setting and recommended solution. Section 4 presents the experimental results and the discussion of findings, and Section 5 provides the conclusions and suggestions for future work.

4.2 Related Work

Spatial Analysis for Anomalous Hotspots

Spatial hotspots are a phenomenon that can be characterized by a high concentration of observations in a particular spatial location. While hotspots may exhibit some similarities with spatial anomalies, it is important to note that these terms are typically reserved for describing rare occurrences or patterns. Conversely, hotspots are often associated with a substantial number, significant number, or majority of observations (Xie et al., 2023). In some fields, certain constraints are applied to determine whether a cluster should be considered a hotspot. For example, a maximum population threshold may be used to exclude clusters that cover more than 50% of the underlying population in control data. Clusters that do not meet this criterion are generally not considered to be hotspots but, rather, represent a general phenomenon (Xie et al., 2023).

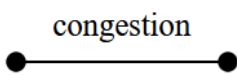

The detection of spatial hotspots in trajectory data requires careful consideration of what constitutes a hotspot and how detection will be conducted. While trajectory data contains multiple points, researchers often simplify the task by focusing solely on the starting and/or ending points of a trip. For example, (Chen et al., 2014) identified dense passenger pickup and drop-off points as candidates for future bus stops using this method, while (Li et

al., 2011) predicted pickup points with the highest passenger volume using the autoregressive integrated moving average. Other researchers have extended the search to all trajectory points and searched for the most frequently visited locations, such as (Yuan et al., 2010), who identified the most frequently traversed points by taxi drivers as landmarks for facilitating future trip planning. This approach, known as clustering, is the primary focus of most hotspot detection research and uses unsupervised machine learning methods.

In other hotspot discovery tasks, the focus shifts from trip points toward the areas between the points, such as road segments, roads, or even multiple roads. For example, (Kriegel et al., 2008) predicted areas of congestion and traffic density in a network, while (Shen et al., 2018) investigated first and last miles to and from public transportation with the objective of integrating shared autonomous vehicles. Overall, the process of hotspot detection in trajectory data requires careful consideration of the research question, available data, and appropriate methodology.

Table 4.1: Summary of Scopes in Hotspot Detection

Scope	Examples	Example Studies
Points	<ul style="list-style-type: none"> <li style="text-align: center;">● pickups <li style="text-align: center;">● drop-offs 	<p>(Chen et al., 2014) cluster dense passenger pickup and drop-off points as candidates for bus stops</p> <p>(Li et al., 2011) use the autoregressive integrated moving average to predict pickup points with the highest level of passengers</p>

Edges		(Kriegel et al., 2008) predict traffic density in a network
Paths (pre-defined)		(Shen et al., 2018) use a supply-side integration of a shared autonomous vehicle as the first and last miles to and from public transportation

The initial step in the hotspot discovery process involves selecting a representation method that can effectively model the movement of trajectory data. While GPS-based representation is commonly used, it poses significant challenges with trajectory handling, such as indexing, retrieval, and error handling. To overcome these issues, alternative representation methods have been proposed—including dimensionality reduction, binary-encoding, hashing, deep representation, and other codifications—that are suitable for large-scale data sets (Sousa et al., 2020). In this regard, several measures have been suggested for calculating the similarity of trajectories in GPS-based representation, including the longest common subsequence (LCSS), Fréchet distance, dynamic time warping (DTW), and edit distance (Toohey & Duckham, 2015). While researchers have addressed various advancements, integrations, and alternatives to these measures, the road-network-constrained trajectory approach, when correctly constructed, is more concise and precise than the GPS-based trajectory (Sousa et al., 2020).

Graph Mining for Anomalous Hotspots

The road-network-constrained trajectory is a popular representation method used for modeling the movement of trajectory data. This approach is constructed based on the GPS-

based trajectory and uses graph theory for formalization, which is a common and rigorous approach that is accessible across disciplines and is particularly suitable for routing problems and optimization (Marshall et al., 2018). The mathematical representation of a street network is defined by nodes connected by edges loaded with weights or labels, which can be directed or undirected. To measure similarities among road-network-constrained trajectories, various methods have been proposed, such as dissimilarity with length, set theory operations, similarity coefficients, dimensionality reduction, and the definition of a window parameter equal to half the size of the shortest trajectory. Another approach involves using collections of points of interest or times of interest to compute similarities, such as trajectories that pass through specific places or during peak traffic hours (Sousa et al., 2020). Each approach has its own metricity, computational complexity, and robustness to noise and local time shifts. In contrast to optimization, trajectory clustering methods require a candidate trajectory to calculate similarities and make recommendations, which can increase computational complexity. Moreover, another challenge in trajectory clustering is selecting appropriate clustering parameters, such as the number of clusters and cluster centers. In this regard, (Zhao et al., 2017) proposed an approach based on decision graphs and data fields to address this issue.

An alternative method for detecting spatial hotspots is to leverage connected components in the trajectory network. Connected components are sets of nodes in a graph that are linked to each other by paths. However, previous studies have focused mainly on simple connected components, while our trajectory networks are weighted. Additionally, the primary objectives of major prior studies, although related to transportation, do not directly

address the detection of hotspots based on total distance in the network. For example, (Kun & Vámosy, 2009) used connected components in combination with decision trees and other methods for traffic monitoring, while (Abbas et al., 2019) used connected components for image processing of traffic congestion. Therefore, this study presents an alternative approach that uses weighted connected components instead of simple connected components for identifying unique anomalous hotspots.

Another approach to identifying spatial hotspots in trajectory networks is frequent pattern mining, which consists of storing trajectories as sequences of identities and searching for the most frequent patterns in the graph. However, this method presents challenges as the data volume and spatial and temporal variations increase, leading to heavily overlapping patterns (Yang & Gidófalvi, 2018). Moreover, frequent pattern mining is suitable for detecting individual frequent points or edges, but it is not appropriate for identifying connected points or edges, such as paths. Therefore, using frequent pattern mining to detect hotspot paths in a trajectory network may not be efficient since not all frequent solutions will be connected. In contrast, the connected component approach allows for identifying clusters of connected points or edges, which is more suitable for finding hotspot paths.

Efficiently locating anomalous hotspots in a trajectory network can be a challenging task due to the large number of possible path combinations. The computational complexity of enumerating all paths within a graph can be $O(N^3)$, where N represents the number of path combinations in the trajectory network (Rubin, 1978). The problem of planning pickup and drop-off points, routes, and frequencies in transportation networks, also known as the transit route network design problem, has been demonstrated to be a nondeterministic polynomial

time (NP-hard) (Schöbel & Scholl, 2006). Furthermore, the addition of on-demand mobility services, such as those used in demand-response route systems, can transform the problem into a dial-a-ride problem, an integrated dial-a-ride problem, a vehicle routing problem with pickup and delivery, or other variations. Additionally, these can be generalized cases of the traveling salesman problem and are computationally intractable or NP-hard (Yoon et al., 2021). Locating anomalous hotspots in a trajectory network presents a similar variation of the problem and, as such, requires an innovative solution to overcome the computational challenges involved.

Table 4.2 provides a summary of themes and the associated, but not mutually exclusive, methods used in hotspot detection in trajectory data.

Table 4.2: Summary of Themes in Hotspot Detection

Theme	Methods	Example Studies
Spatial analysis	<ul style="list-style-type: none"> - Clustering starting/ending points of trips - Clustering all trajectory points - Road-network-constrained trajectory approach 	(Chen et al., 2014; Li et al., 2011; Yuan et al., 2010)
Representation methods	<ul style="list-style-type: none"> - GPS-based representation - Alternative methods (e.g., hashing, deep representation) - Similarity measures (e.g., longest common subsequence, dynamic time warping) 	(Sousa et al., 2020; Toohey & Duckham, 2015)
Graph mining	<ul style="list-style-type: none"> - Graph theory formalization - Dissimilarity with length 	(Marshall et al., 2018; Sousa et al., 2020)

	- Set theory operations, similarity coefficients	
Connected components	- Simple connected components - Weighted connected components	(Abbas et al., 2019; Kun & Vámosy, 2009)
Frequent pattern mining	- Storing trajectories as sequences of IDs - Searching for frequent patterns in the graph	(Yang & Gidófalvi, 2018)
Computational complexity	- Transit route network design problem - Vehicle routing problem with pickup and delivery	(Rubin, 1978; Schöbel & Scholl, 2006; Yoon et al., 2021)

4.3 Methodology

This section outlines our framework for discovering AHP–TDT, illustrated in Figure 4.1. Our approach begins with the standard initial steps common in spatial data analysis: defining a road network (Step A1) and extracting a trajectory network (Step A2). Following (Kriegel et al., 2008; Sousa et al., 2020), the defined road network provides the structural baseline for our data, while the extracted trajectory network captures the movement patterns within this structure. Moving beyond traditional spatial hotspot detection of nodes or edges, our methodology explores the identification of hotspot paths. This begins with a definition of the bounds of the network (Step B1), which sharpens the focus of our analysis to specific areas within the network. Then, we proceed to extract subgraphs (Step B2) using connected components. This step enriches our understanding of the data by identifying all interconnected nodes and edges, extending beyond the immediately apparent hotspots.

The subsequent step involves extracting distinct paths (Step B3), a crucial phase for ensuring accurate trajectory representation. This process helps isolate unique paths, even

when they share common nodes or edges, thereby reducing potential biases toward frequently traversed routes. The framework culminates with the setting of anomalous thresholds (Step B4), pinpointing not only relevant hotspots but also those exhibiting anomalous behavior (Step C). This final step is instrumental in highlighting areas of interest or concern within the data set. Therefore, our AHP–TDT framework stands out from previous methodologies because of its comprehensive and precise approach, marked by the incorporation of a definition of network bounds, the extraction of subgraphs, the identification of distinct paths, and the establishment of thresholds for anomaly detection.

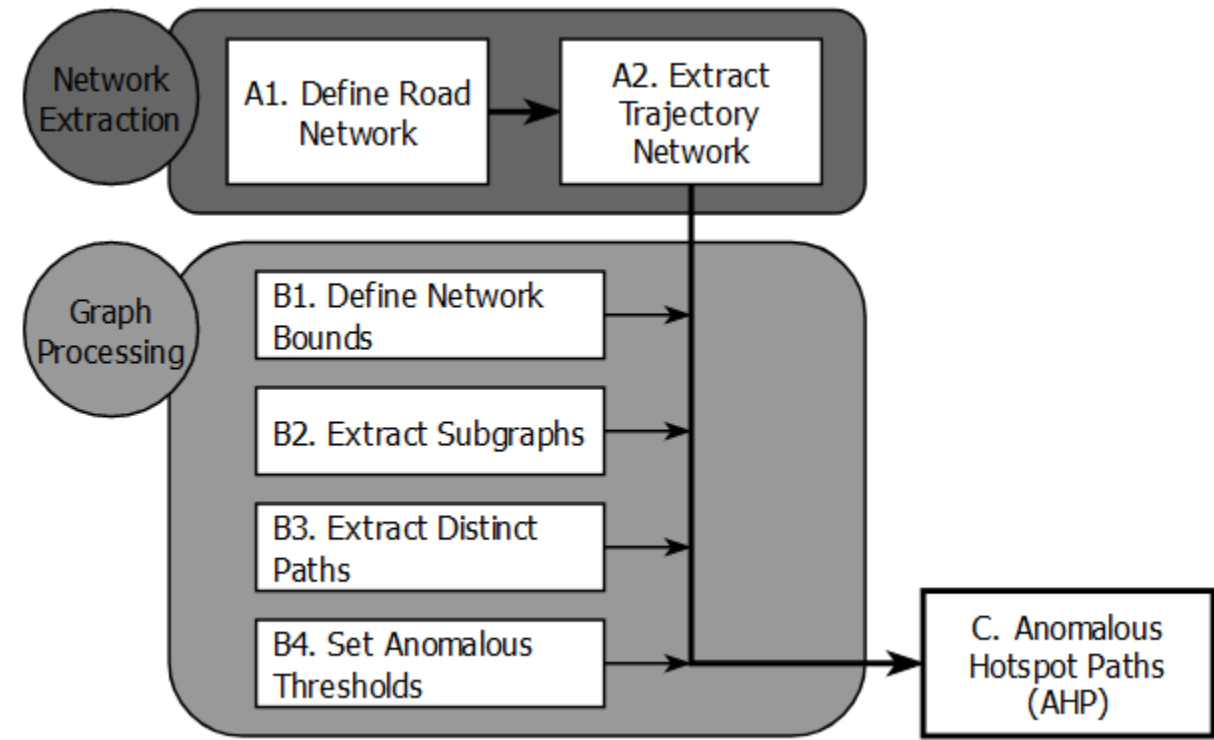


Figure 4.1: Anomalous Hotspot Paths–Total Distance Traveled Framework

Road and Trajectory Networks

The process for defining a road network (Step A1) involves creating a comprehensive representation of a geographical area, such as a city, and capturing all of its critical features, including the names of streets, intersections, and crossroads; speed restrictions; and the number of lanes, as discussed by (Marshall et al., 2018). We depict this road network as a graph $R^G = (V, E)$, where V represents a set of nodes (vertices), each corresponding to a street intersection, and E denotes a collection of edges, each of which represents a road segment linking adjacent intersections. Therefore, every pair of proximal nodes in the set V is connected by an edge from the set E . The overall graphical representation of the road network could be either directed or undirected, depending on the nature of the traffic flow in the real-world scenario it mirrors. This graph-based model aids in accurately capturing the complexity of urban mobility patterns.

The trajectory network (Step A2), which can be represented as T^G , is a derivative of the road network, reflecting only a portion of the nodes and edges used in a specific sample of on-demand trip trajectories. To put it another way, while the underlying structure of the road network remains unchanged regardless of the data sample used, the trajectory network may vary. Each trajectory represents a unique journey from a starting point to a destination, and, therefore, the trajectory network is primarily viewed as a directed network. However, where appropriate, it can be adapted into an undirected format. Each journey in the trajectory network follows a path P , comprising a sequence of edges denoted as $P = (e_1, e_2, \dots, e_n)$, where n symbolizes the total number of edges in that specific path. Each edge in this sequence shares a common node with its adjacent edge, illustrating the connectivity in the

network. In other words, $e_i = (v_l, v_k) \rightarrow e_{i+1} = (v_k, v_m)$, where v_l , v_k , and v_m belong to the set nodes.

In our analysis of the trajectory network, we concentrate on “simple paths” defined by a clear structure: a starting node, an ending node, at least two connecting edges, and, possibly, intermediate nodes. This focus on simple paths facilitates a straightforward interpretation of trajectory data, as per (Idé & Kato, 2009), untangling the complex dynamics of travel behavior. Moreover, central to our modeling approach is the adoption of the “shortest path assumption” (Idé & Kato, 2009). This assumption posits that objects within a network typically follow the most direct route from start to end, thereby streamlining our model. By applying this assumption, we effectively reduce the model’s complexity, sidestepping the need to intricately account for variables such as spatial context, temporal factors, speed, and other trip-specific data. Although these elements are vital for both a comprehensive understanding of trajectory movements and the identification of outliers, they add significant complexity to the model. In our evaluation, we considered two metrics for determining the shortest path: distance and travel time. Our empirical tests revealed negligible differences in the travel routes within our trajectory network, regardless of the metric used.

Before detecting spatial hotspots, the next stage in the process is defining what constitutes a hotspot. In the case of hotspot points, the weight is generally defined by the frequency of trips passing through each node, a metric we can represent as node frequency (NF). The process for identifying hotspot points, therefore, involves analyzing all of the nodes within the trajectory network to pinpoint those exhibiting the highest NF values or

those surpassing a predetermined NF threshold. The process can be further refined by concentrating exclusively on the starting and ending nodes of a trip, thereby facilitating the detection of significant passenger pickup and drop-off points, as exemplified in the methods employed by (Chen et al., 2014; Li et al., 2011).

Similarly, when identifying hotspot edges, the weight is typically determined by the frequency of trips traversing each edge, a metric represented as edge frequency (EF). Thus, the process for discovering hotspot edges involves inspecting all edges within the trajectory network to identify those that display the highest EF values or those that exceed a pre-established EF threshold. This method has been implemented in many applications, including detecting congestion and traffic density in a network (Kriegel et al., 2008) and identifying ideal candidate areas for integrating shared autonomous vehicles in public transportation (Shen et al., 2018).

However, in addition to EF, we can consider the distance traveled along an edge (DT) as an alternative weighting method. Given that each edge possesses a specific length (EL), the product of edge frequency and edge length yields the cumulative distance traveled for that particular edge, as denoted by $DT = EF \times EL$. Thus, by transitioning the weightage criterion from edge frequency to distance traveled, we can unveil distinct behavioral nuances within the trajectory network. For instance, an edge, despite its high frequency but small length, may weigh less than edges with longer lengths. This is because the distance traveled parameter takes into account both the frequency of use and the physical extent of the edge, offering a holistic view of overall network use. This nuanced approach forms the basis of our

methodology for discovering AHP–TDT in this study and clarifies what differentiates it from prior work.

Anomalous Hotspot Paths–Based on Total Distance Traveled Framework

Discovering hotspot paths requires enumerating all potential paths within the trajectory network. Defining key parameters such as the path length (PL) and the TDT is key in this process. PL can be computed by aggregating the lengths of all constituent edges in the path.

Formally, given n edges in a path, the path length can be calculated as $PL = \sum_{i=1}^n EL$.

Similarly, we can determine TDT within the path by accumulating the distances traveled on each edge in the path. Thus, given n edges in a path, the TDT is computed as $TDT =$

$$\sum_{i=1}^n TD = \sum_{i=1}^n (EF \times EL).$$

However, the task of enumerating all potential paths within the trajectory network requires considering all possible combinations of edges and establishing a set of all prospective paths that span all potential starting and ending points. Furthermore, the task shifts from solely evaluating edge frequencies to assessing the TDT, which is the product of edge frequencies and their lengths. For example, when seeking the path characterized by the maximum TDT value, the task can be formulated as an optimization problem, articulated as: $\text{Max TDT} = \text{Max} \sum_{i=1}^n (EF \times EL)$. This formulation is constrained by $PL, TF \geq 0$, and $\forall i = 1, \dots, n$ where n is the number of path combinations in a trajectory network. As discussed, the complexity of this enumeration challenge is $O(N^3)$, where N denotes the number of path combinations in the trajectory network (Rubin, 1978).

To address the challenge of processing numerous and often insignificant paths in our data set, we incorporate into our model a step that establishes network bounds (Step B1) by defining two key parameters: minimum path length (MinPL) and minimum path frequency (MPF). This approach, which is aligned with (Sousa et al., 2020), emphasizes relevant data representation and helps filter out less meaningful paths, thereby enhancing the accuracy of hotspot detection. Similarly, it echoes the strategy in graph mining for efficient anomaly detection introduced by (Marshall et al., 2018). For instance, defining lower MPF and/or MinPL values will yield a broader set of potential solutions, but this could affect the efficiency of the path search process. Conversely, setting high MPF and/or MinPL parameters can potentially boost search performance by focusing on more promising paths. However, this approach runs the risk of omitting suitable solutions and, possibly, returning no paths. Consequently, it is crucial to strike the appropriate balance between these parameters. This allows for strategic navigation between the comprehensive examination of potential solutions and the operational efficiency of the search process, ensuring that we identify meaningful spatial hotspot paths without burdening the system with an excessive number of paths of insignificant value.

However, a simple filtration of edges based on these lower bound parameters may not suffice. Given the intricately interconnected nature of a network, paths and edges often share significant relationships. Thus, a simplistic filtering approach could inadvertently eliminate essential information from neighboring edges, disrupting the overall understanding of the network. To address this challenge, our model integrates an additional step that is focused specifically on the extraction of all subgraphs using connected components within the

subnetwork (Step B2). This process acknowledges the interconnected relationships between the network elements, ensuring that crucial inter-edge information is preserved and used effectively. Consequently, this allows for a robust and insightful exploration of the network, significantly augmenting our capacity to identify meaningful hotspot paths.

Though this step, extracting all subgraphs, acknowledges the complex interplay within network elements, it primarily manifests subgraphs rather than simple paths. To optimize the use of these subgraphs, we need to transform these derived subgraphs into all possible simple paths. Our model's main benefit becomes evident at this stage. Instead of indiscriminately searching the entire network, the incorporation of lower network-bound parameters and subgraphs along with the regeneration of simple paths focuses the search effort on the most significant areas in the network and their surrounding clusters. As a result, this approach streamlines the search process and significantly enhances the efficiency and effectiveness of hotspot detection. This focus on the areas of utmost importance not only mitigates the computational strain but also sharpens our insights into the trajectory network's pivotal aspects.

The integration of an upper network bound parameter—specifically, the maximum path length (MaxPL)—into our model is another step toward enhancing its effectiveness. However, the timing of this incorporation is key. The introduction of this parameter prior to the extraction of the subgraphs may unintentionally lead to the removal of vital subgraphs from our consideration, thus potentially compromising the quality of our detection. Instead, we strategically incorporate this upper bound parameter after the extraction of connected components and the transformation of subgraphs into simplified paths. At this stage, applying

the MaxPL provides a focused lens to sift through the multitude of paths, selectively retaining those that fall within the defined length constraint. This approach further streamlines our search process, effectively reducing computational demands while simultaneously ensuring a robust and comprehensive examination.

The next step in our approach is the extraction of distinct paths within the candidate solutions (Step B3) by examining how paths overlap among each other. An overlap instance within a network occurs when two paths share common nodes or edges. If unaddressed, these overlaps could distort our understanding and interpretation of the network, leading to potential issues. An overlooked overlap could cause certain paths or edges that are part of numerous paths to be disproportionately represented. This overrepresentation could misleadingly elevate their perceived importance or frequency, thereby skewing the outcome of the discovery. Conversely, the discovery could fail to highlight unique, albeit less frequent, paths that possess equivalent significance if excessive attention is accorded to overlapping sections. Moreover, computational efficiency could be compromised by overlaps. Repetitive processing of common nodes or edges across different paths could lead to unnecessary computational expenditure.

To tackle these challenges, we employ the acceptable similarity (AS) parameter. This parameter uses Jaccard similarity to gauge the degree of similarity between two paths. It achieves this by comparing the size of the intersection of the two paths to the size of their union, as per (Sousa et al., 2020). By strategically fine-tuning this parameter, we can tailor the experimental direction to suit our intended objectives. Lower AS values will yield a larger pool of potential solutions but may sustain fewer distinct paths, resulting in fewer

unique paths. Conversely, higher AS values will generate a greater number of distinct paths but may inadvertently exclude key paths. Therefore, the systematic extraction of distinct paths ensures that our model is both robust and insightful, capable of reflecting the network's intricacies without losing sight of the underlying objectives.

The final critical step (B4) in our AHP–TDT methodology involves the establishment of anomalous thresholds. This step is pivotal in identifying paths that exhibit significant deviations from normal traffic patterns and highlighting areas that warrant special attention. To define these thresholds, we employ two primary methods: specifying top-k anomalies and applying a specific threshold value (Aggarwal, 2017; Yeh et al., 2017). In the top-k anomalies method, paths with the highest TDT values are flagged as anomalous. These heavily traversed paths could reveal notable patterns or points of interest, necessitating further examination. Alternatively, the threshold method identifies paths where the TDT equals or exceeds the network-wide average TDT. This technique effectively highlights paths that stand out from typical traffic flow, thus identifying possible hotspots. For enhanced statistical accuracy, our approach deems paths as anomalous if their TDT is significantly higher than the mean, specifically, one or more standard deviations above. This ensures that the anomalies we identify are truly remarkable, not simply marginally above the average. While the threshold method provides a binary, clear-cut classification of anomalies, it may lead to inaccuracies if it is not carefully calibrated. Conversely, the top-k method, offering a ranked analysis of anomalies, yields a more detailed perspective, which is particularly beneficial in scenarios in which a fixed threshold is not feasible. However, it is important to note that in a predominantly regular time series, the most anomalous paths identified by the

top-k method might not be exceptionally anomalous in an absolute sense. To address these considerations, our framework integrates these methods, considering a path as anomalous if it is flagged by more than one of these approaches (Aggarwal, 2017; Yeh et al., 2017). This multifaceted strategy ensures a thorough and robust identification of potential hotspots, thus significantly enhancing our capacity to accurately map and interpret the complexities within trajectory networks.

Table 4.3: Notations of all Variables

Notation	Comment
NF	Node frequency
EF	Edge frequency
EL	Edge length
PL	Path length
DT	Distance traveled
TDT	Total distance traveled
MPF	Minimum path frequency
MinPL	Minimum path length
MaxPL	Maximum path length

4.4 Research Methods

To validate our model, we used real-world trajectory data collected from an on-demand transportation agency operating in Porto, Portugal.⁶ This data set includes an array of approximately 1.7 million individual trips, chronicled over a full annual cycle from July 1,

⁶ <https://www.kaggle.com/crailtap/taxi-trajectory>

2013, to June 30, 2014. Every trip is meticulously represented by consecutive GPS coordinates that are recorded at regular 15-second intervals. Such recordings create polylines that allow for a detailed analysis of movement and traffic patterns within the city. An example of this raw data can be found in Table 4.4. The selected data set not only aligns with the overarching goals of this research but also has been leveraged in similar studies, underscoring its relevance and validity for our investigation. While our approach would have been enriched by examining an array of similar trajectory data sets, recent regulatory shifts and increasing privacy concerns, including those articulated in geolocation privacy legislation in countries such as the United States, have placed significant constraints on the accessibility of other similar data sets.

Table 4.4: A Sample of the Raw Data

TRIP_ID	CAL L_TY PE	ORIGI N_CAL L	ORIGI N_STA ND	TAX I_ID	TIMESTA MP	DAY _TYP E	MISSIN G_DAT A	POLYLINE
1379415 6366200 00000	A	2002		2000 0653	9/17/2013 07:00	A	FALSE	[[[-8.625798,41.157342], [-8.625789,41.15736], [- 8.625744,41.157369], ...]
1379415 6146200 00000	B		6	2000 0657	9/17/2013 07:00	A	FALSE	[[[-8.582598,41.180202], [-8.582346,41.180211], [- 8.582265,41.180769], ...]
1379415 7416200 00000	B		32	2000 0011	9/17/2013 07:02	A	FALSE	[[[-8.627589,41.157684], [-8.627607,41.157702], [- 8.627913,41.157909], ...]

To define the road network, we used two powerful Python libraries. First, we employed NetworkX to model and manage the intricacies of the road network as mirrored in the trajectory data.⁷ NetworkX is renowned for its ability to effectively construct, manipulate, and study the structure and dynamics of complex networks. Subsequently, we used OSMnx for accessing and harnessing spatial data such as street networks, which are fundamental to our discovery task.⁸ This tool is particularly useful due to its ability to extract, model, analyze, and visualize a wide range of spatial objects using data from OpenStreetMap (Boeing, 2017 and visualizing complex street networks). Our experiment focused on the city of Porto, Portugal, with specific geographical coordinates (41.155, -8.63) serving as our central point of interest. To ensure a manageable and relevant area of study, we limited our scope to a radius of 2,500 meters around this focal point. Following this, we engaged in data pre-processing, which included dismissing trips of less than one-minute duration as well as extracting key data points such as pickup and drop-off locations, which played a crucial role in subsequent stages of our experiment. The result of these procedures was a meticulously defined road network, which accurately reflected the urban layout of Porto, comprising a total of 6,159 edges and 2,993 nodes. This road network formed the base upon which our further discovery was conducted; it is graphically represented in Part A1 of Figure 4.2. This approach employed in defining the road network ensures that it is an accurate, comprehensive, and analytically valuable representation of the real-world urban environment.

⁷ <https://networkx.org/>

⁸ <https://github.com/gboeing/osmnx>

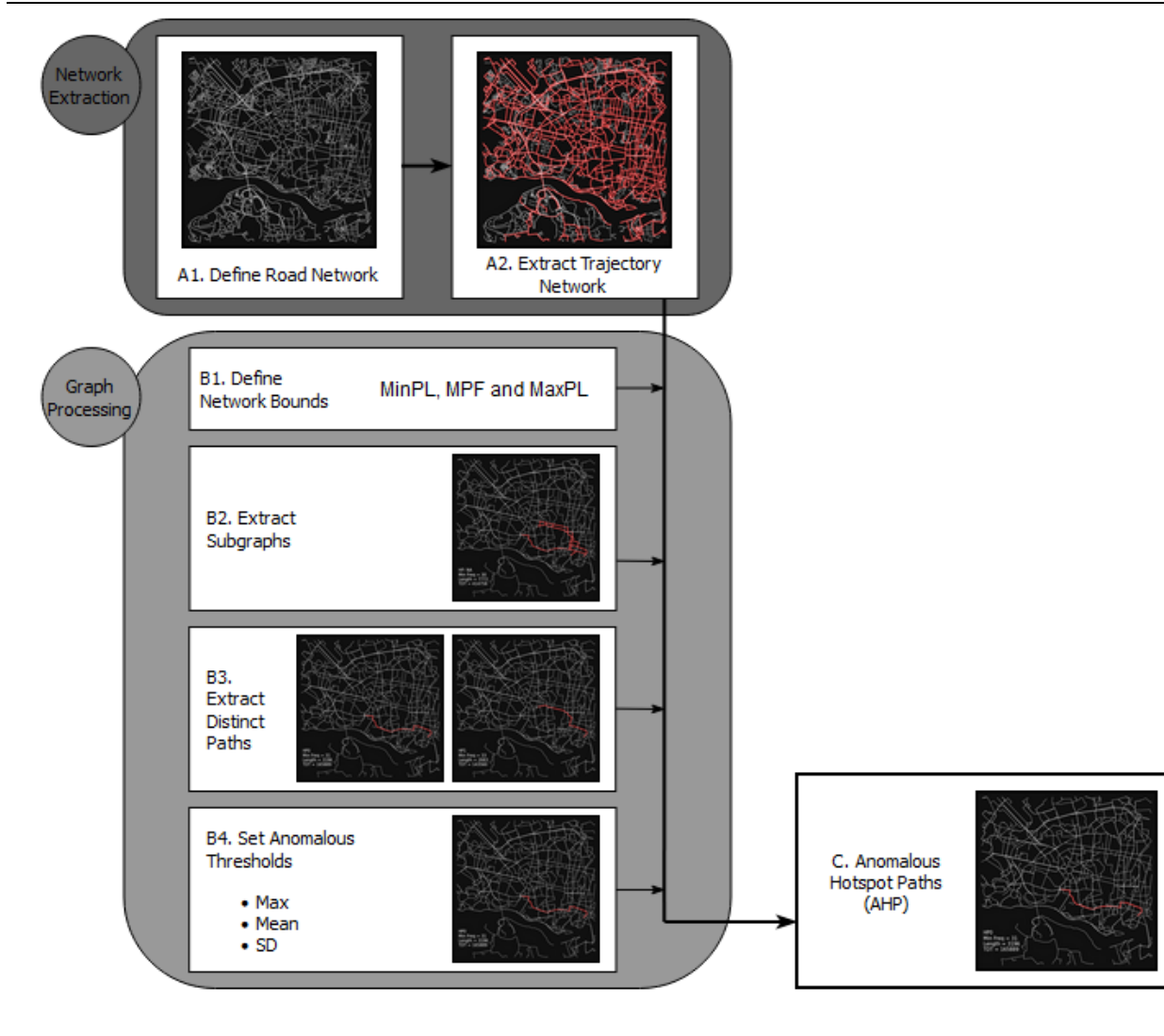


Figure 4.2: Results of Applying the Anomalous Hotspot Paths Framework

To distill the trajectory network from the overarching road network, we strategically aligned nodes and edges drawn from trip trajectories to their corresponding elements within the road network. This alignment was driven primarily by the specific pickup and drop-off locations, which are linked to the nearest nodes and edges within the road network. This procedure not only facilitated a seamless synchronization of the road and trajectory network

structures but also effectively eliminated any nodes and edges from the road network that remained unused within the scope of our trajectory data. While the structural integrity of the road network remained consistent across different data samples, the trajectory network displayed a certain degree of variability. In other words, if we were to choose a different subset of trajectory data, the road network would retain its form, whereas the trajectory network would adapt accordingly.

In tackling the challenge of analyzing trajectory data sets, such as ours that encompassed approximately 1.7 million trips, our objective was to extract hotspot points, edges, and paths using various thresholds and parameters, such as MPF, MinPL, and AS. Drawing inspiration from the methodologies of (Djenouri et al., 2019; Zhang et al., 2021), we employed a time-window sampling strategy. This approach effectively segmented the vast data set into smaller, more manageable units, enhancing our analytical precision and efficiency. Moreover, it prepared our model for future extensions, wherein multiple samples from varying time windows could be used to highlight spatio-temporal anomalies, thereby providing insights into the influence of time on travel patterns. For our analysis, we selected a representative sample that encapsulated a critical aspect of urban mobility: the morning rush hour. Specifically, we focused on data from 7 a.m. to noon on Tuesday, September 17. This period is characterized by heightened network activity, providing a fertile ground for studying congestion and travel dynamics. By concentrating on this time window, we aimed to derive insights into traffic patterns during peak use, offering a window into the most intense operational challenges faced by the network. After a meticulous process of data cleaning and preprocessing, we refined our sample to 1,085 trips. Guided by the principle of

the shortest distance for the shortest path analysis, this subset yielded a trajectory network comprising 1,936 nodes and 2,833 edges. This network, visualized in Step A2 of Figure 4.2, served as the foundational framework for our hotspot discovery process. Our approach, grounded in specific data handling and strategic sampling, positioned us to extract detailed and meaningful insights into the patterns of trajectory movement, highlighting our contribution to the understanding of urban mobility dynamics.

Results and Analyses

Hotspot Points and Edges

To enhance our understanding and interpretation of the entire AHP–TDT framework, we first explored conventional methodologies that are typically employed to extract spatial hotspots in the form of both points and edges. This allowed us to create a broader context to examine and compare various methodologies and their resultant findings. When detecting hotspot points, we focused primarily on nodes in our trajectory network that exhibited high NF values. These nodes represented points that are frequently visited or traversed during trips, thereby contributing to a higher NF score. On a similar note, hotspot edges were identified based on their EF values, and edges with higher EF values were considered to be spatial hotspots due to their frequent use in trips. Moreover, in our quest for a more comprehensive view, we extended our scope beyond the standard EF measure for edge hotspots, and we considered the distance traveled along an edge as an alternate weighting metric. This subtle change in perspective offered a rich understanding by taking into account

both the frequency and the length of the edge, thereby unveiling additional dimensions of our trajectory network. This nuanced lens of interpretation is illustrated in Figure 4.3.

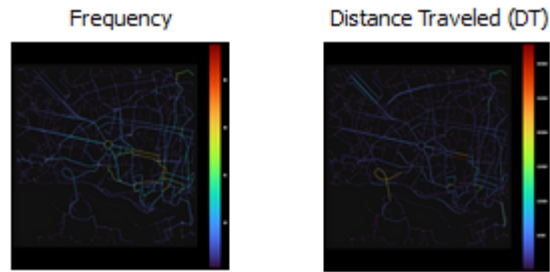


Figure 4.3: Changing Weight from Frequency to Total Distance Traveled

To gain a clearer and more tangible insight into these hotspots, we collated our findings in Tables 4.5 and 4.6. These tables present the nodes and edges, respectively, identified as hotspots within our network. Our findings in these tables indicated that the hotspot points were also detected as participating nodes in the hotspot edges. To provide a more spatial perspective, Figure 4.4 serves as a graphical representation of detected hotspot points and edges, providing a vivid confirmation of their respective positions within our network.

Table 4.5: The Top Hotspot Points

Node ID	Frequency
n324	110
n346	102
n11	100

Table 4.6: The Top Hotspot Edges

From Node ID	To Node ID	Frequency
n1758	n346	95
n324	n12	93
n11	n10	77

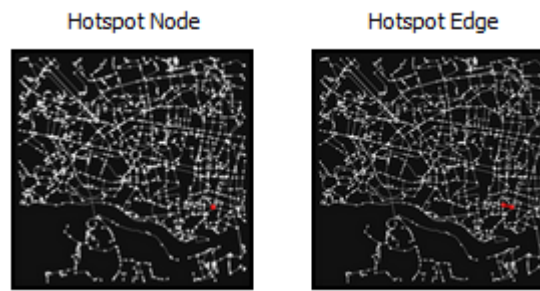


Figure 4.4: Spatial Hotspot Nodes and Edges

Anomalous Hotspot Paths Based on Total Distance Traveled

As introduced, our AHP–TDT approach can be differentiated from previous work because it introduces a more granular and comprehensive process, incorporating network bounds setting, extracting subgraphs, extracting distinct paths, and specifically targeting AHP, ensuring precision and depth in our discovery.

To define the network bounds in our AHP–TDT approach, we conducted experiments to calibrate the MinPL and MPF parameters. In the first experiment, we set the MPF values at [20, 30, 40, 50, 60] and the MinPL at [1,000, 1,500, 2,000] meters. For the second experiment, we broadened the MPF range to [20, 25, 30, 35, 40, 45, 50, 55, 60], while

keeping the MinPL unchanged. These parameters were vital in filtering out insignificant paths, aiming to capture only the most distinct patterns. This approach was designed to act as a sentinel, guarding against overlooking hidden anomalies and enriching our understanding of the data landscape. This pivotal stage in our framework is visually encapsulated in Step B1 of Figure 4.2.

The subsequent step in our AHP–TDT approach centered on extracting all subgraphs using connected components within our trajectory network. This phase transitioned from basic filtration to a nuanced mapping of intricate network relationships. Recognizing the complex interplay within the network elements, this step transcended the limitations of simple filtration and edge isolation, extending beyond the simple identification of individual paths and unfolding the intricate relationships shared across different paths and edges. Step B2 of Figure 4.2 vividly captures this pivotal procedure, offering a snapshot of a typical subgraph within our trajectory landscape.

As we advanced, we set MaxPL to 3,500 meters. Implemented after the extraction of subgraphs, this upper network bound parameter served as a discerning filter, refining our search to consider only paths within this length constraint. This streamlined incorporation of MaxPL ensured optimal performance while preserving the richness of our analytical depth. Following that, our approach delved into transforming subgraphs into simple paths, ensuring a clearer representation. Subsequently, the AS parameter was calibrated at 0.3. This calibration allowed us to systematically extract distinct paths, striking a balance between capturing unique paths and averting the pitfalls of overrepresentation. This step further

optimized computational efficiency and honed the model's precision. Step B3 of Figure 4.2 graphically delineates this crucial juncture in our methodology.

Based on the above specifics, our first experiment resulted in 14 different models that yielded 13 distinctive hotspot paths, illustrated in Table 4.7 in the Appendices. Of particular note, the path labeled "HP0" showcased a remarkable TDT of 166.144 km, traversing a path length of 3.487 km, and resonating at a base frequency of 20. In another example, the path labeled "HP7" charted a TDT of 23.319 km, measuring 1.014 km, echoing at a frequency of 23. For our second experiment that broadened the MPF intervals, the results were 24 models that yielded 19 unique hotspot paths, illustrated in Table 4.8 in the Appendices. Interestingly, the anchors of paths with the minimum and maximum TDTs mirrored those from our initial experiment.

In the culminating phase of our methodology (Step B4), we focused on establishing anomalous thresholds to effectively pinpoint AHP. Our strategy was to identify paths exhibiting significant deviations from standard traffic patterns. To ensure statistical robustness, we adopted a multi-method approach, labeling a path as anomalous only if it was identified as such by more than one analytical method. This included examining paths with the highest TDT values, those exceeding the average TDT, and paths significantly above the mean TDT, measured by one or more standard deviations. The results of this approach are depicted in Step C of Figure 4.2. This visualization provides a clear representation of an exemplary AHP identified in our trajectory network. Comprehensive outcomes from our dual experiments can be found in Tables 4.7 and 4.8 in the Appendices, offering a deep dive into

both overt and intricate findings, leading us to identify potentially significant hotspots that beckon further scrutiny.

4.5 Discussion

This section presents our key findings regarding the discovery of AHP in trajectory data. Initially, our experimentation focused on identifying spatial hotspot nodes and edges, highlighted by a high concentration of events such as traffic density, as shown in Figures 4.1 and 4.2. These spatial hotspots, however, are confined to discrete points or edges and are insufficient for identifying continuous hotspot paths. To overcome this challenge, we introduce the AHP-TDT framework. This approach effectively identifies larger hotspots that extend beyond individual nodes and edges, capturing more expansive paths. Our findings, illustrated in Figure 4.3, showcase the AHP-TDT framework's ability in pinpointing spatial hotspots and offering in-depth analysis of movement patterns in trajectory networks.

Specifically, while spatial hotspot nodes and edges are useful for identifying ideal pickup and drop-off points for passengers and for locating traffic congestion areas, AHP-TDT provides a sophisticated approach for trajectory collective movement analysis that can be useful for detailed trajectory planning and optimization. By considering the TDT and trip frequency, AHP-TDT enables detailed trajectory planning and can be particularly useful for applications that require a high level of control over the travel experience. For example, the use of AHP-TDT can enhance data-driven mobility by providing granular information about the frequency of trips and the TDT by passengers. This approach can help in optimizing

mobility routes to reduce travel time, improve passenger experience, and lessen the environmental impact.

Our experiments also reveal that the anomalous hotspot with the highest TDT does not necessarily coincide with the spatial hotspot nodes or edges. For instance, the third anomalous hotspot in the first AHP–TDT scenario has a higher weight in terms of trip frequency than the first and second anomalous hotspots, but it ranks third in terms of TDT when factoring in hotspot path length. Furthermore, we find that discovering AHP–TDT at higher lengths requires lowering the required trip frequency, as not all trips pass through the same path and some deviate at various locations. These observations suggest that the presence of both measures is critical for applications that require precise control over hotspot path distance length and trip frequency, without which the results may be ineffective or inefficient.

From a methodological perspective, discovering all types of spatial hotspots requires defining a representation that can model the movement of trajectory data such as the road-network-constrained approach. Our findings indicate that the methods of discovering spatial hotspot nodes and edges involve graph searching, which commonly falls within clustering and unsupervised machine learning. However, discovering AHP–TDT cannot use the same methods since searching for all of these added measures in a network is computationally complex. To overcome this issue, we transform the search into an optimization problem and leverage weighted connected components to discover AHP–TDT with significant control over other aspects, such as hotspot path distance length and trip frequency, as Figure 4.3 illustrates.

Contributions

This study makes several key contributions. First, it pioneers the definition and exploration of the AHP–TDT problem. This novel approach emphasizes the importance of analyzing cumulative distances in trajectory networks to identify irregular movement patterns, offering a fresh perspective in the field of spatial data analysis. Second, it advances a comprehensive methodology for discovering anomalous hotspot paths. This framework is distinct in its thorough approach, integrating steps such as defining network bounds, extracting subgraphs, identifying distinct paths, and establishing thresholds for anomaly detection. This framework marks a significant leap from traditional spatial hotspot detection methods, focusing on the intricacies of movement dynamics in a trajectory network. Third, this study tests its framework using a real-world data set from an on-demand transportation agency in Porto, Portugal. This empirical validation, involving about 1.7 million individual trips, not only demonstrates the practical applicability of the AHP–TDT framework but also underscores its relevance and effectiveness in analyzing complex traffic and movement patterns in urban settings.

Implications

Theoretical Implications

This research enhances our comprehension of spatial data analysis by transitioning the focus from conventional hotspots at points and edges within urban mobility networks to encompass entire paths. This shift offers a more holistic perspective of urban dynamics, thereby broadening the scope and depth of spatial analysis. The study’s methodology extends

beyond conventional analysis by emphasizing the collective nature of spatial data, particularly in trajectory paths within urban environments. This perspective facilitates a deeper and more nuanced understanding of spatial phenomena, recognizing the importance of broader patterns and the complex interplay of sequential and collective spatial events in interpreting the significance of data. Consequently, it contributes to refining the frameworks of spatial data analysis, network analysis, and graph theory. The introduction of a weighted connected component approach provides a nuanced understanding of complex urban trajectory networks, thereby augmenting traditional theoretical models in these fields.

Furthermore, the focus on trajectory paths as sequences of spatial events represents a significant theoretical advancement in trajectory data analysis. This approach is pivotal in revealing the intricacies of movement patterns and their anomalies, contributing to a holistic understanding of spatial dynamics. The methodology employed in the AHP-TDT framework augments the fields of data mining and machine learning, with a particular emphasis on enhancing anomaly detection algorithms. This nuanced form of analysis necessitates rethinking existing models and algorithms. Lastly, the concept of collective anomalies in trajectory data has cross-disciplinary theoretical implications. It extends its relevance to fields such as urban planning, environmental studies, and social network analysis, demonstrating the broad utility of this research in a data-centric global context.

Practical Implications

From a practical standpoint, the findings of this study have substantial implications for urban planning and traffic management. By identifying and analyzing AHP, urban planners

and traffic managers are equipped with insights that can lead to efficient road network optimization, congestion reduction, and overall transportation system enhancement. In the realm of public safety and emergency response, the ability to detect and analyze hotspot paths is invaluable. This research aids in developing preemptive strategies to manage areas susceptible to high traffic incidents or other anomalies, thereby enhancing public safety. The implications of this research extend to smart city initiatives, offering a data-driven framework that supports intelligent urban infrastructure management. This is particularly relevant in the context of integrating advanced technologies such as the internet of things, which are pivotal in the evolution of smart cities. Additionally, the insights garnered from this study are instrumental for optimizing on-demand mobility services. They provide a foundation for improving route planning in ridesharing services and streamlining delivery logistics, thereby enhancing the efficiency and effectiveness of these services.

Conclusions and Future Work

As cities grow larger and more crowded, traffic congestion increases infrastructure costs, accidents, and pollution, reducing the overall quality of life and economic productivity (Litman, 2015). In the United States, transportation consumes approximately 17% of the average household income, and in some areas, this figure exceeds 50% (Ridewithvia, 2021). Consequently, improving mobility and travel experiences is critical, and discovering AHP–TDT is one way to achieve this. In this research, we sought to answer critical questions that could pave the way for better urban mobility solutions. First, we examined how analyzing collective anomalies in paths via AHP–TDT could offer a more comprehensive understanding of network dynamics compared to traditional point- and edge-focused

methods. Second, we explored the role of defining network bounds in enhancing the precision and relevance of hotspot detection in trajectory data. Third, we investigated the impact of employing a weighted connected component approach in identifying anomalous hotspots within trajectory networks.

By identifying AHP–TDT in areas with high individual on-demand trips, we can more effectively optimize travel services in areas with high demand for individual on-demand trips. This could involve tailoring transportation services for efficient connectivity between an airport and a specific neighborhood at peak arrival times, or between a stadium and another area post a major sporting event. Similarly, understanding the flow from residential areas to business districts during rush hours, or from shopping malls to other regions during holiday seasons, can significantly enhance service planning and resource allocation. Prior research on route optimization, such as matching and dispatching, has been limited to managing the demand side only, while AHP–TDT can generate additional supply. Thus, by learning from previous trips, AHP–TDT can identify when and where abnormal demand occurs and offer improved mobility services in these temporal and spatial areas.

Future research can investigate the use of other network bounds and constraints in an AHP–TDT model to further optimize the detection of anomalous hotspots. The AHP–TDT approach can be extended to detect spatial hotspots in other types of trajectory data, such as public transit or animal tracking. Additionally, the approach can be used to study the spatiotemporal dynamics of hotspots and their changes over time. The impact of the use of AHP–TDT on other transportation modes, such as cycling and walking, can also be explored.

Lastly, the AHP–TDT approach can be incorporated into an intelligent transportation system to improve the management of traffic and enhance the travel experience.

4.6 References

- Abbas, Z., Sigurdsson, T. T., Al-Shishtawy, A., & Vlassov, V. (2019). Evaluation of the use of streaming graph processing algorithms for road congestion detection. *Proceedings - 16th IEEE International Symposium on Parallel and Distributed Processing with Applications, 17th IEEE International Conference on Ubiquitous Computing and Communications, 8th IEEE International Conference on Big Data and Cloud Computing, 11t*, 1017-1025. <https://doi.org/10.1109/BDCloud.2018.00148>
- Aggarwal, C. C. (2017). *Outlier Analysis*. <https://doi.org/10.1016/b978-012724955-1/50180-7>
- Boeing, G. (2017). OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65, 126-139. <https://doi.org/10.1016/j.compenvurbsys.2017.05.004>
- Castro, P. S., Zhang, D., Chen, C., Li, S., & Pan, G. (2013). From taxi GPS traces to social and community dynamics: A survey. *ACM Computing Surveys*, 46(2). <https://doi.org/10.1145/2543581.2543584>
- Chen, C., Zhang, D., Li, N., & Zhou, Z. H. (2014). B-planner: Planning bidirectional night bus routes using large-scale taxi GPS traces. *IEEE Transactions on Intelligent Transportation Systems*, 15(4), 1451-1465. <https://doi.org/10.1109/TITS.2014.2298892>
- Djenouri, Y., Belhadi, A., Lin, J. C. W., Djenouri, D., & Cano, A. (2019). A Survey on Urban Traffic Anomalies Detection Algorithms. *IEEE Access*, 7, 12192-12205. <https://doi.org/10.1109/ACCESS.2019.2893124>
- Hamdi, A., Shaban, K., Erradi, A., Mohamed, A., Rumi, S. K., & Salim, F. D. (2022). *Spatiotemporal data mining: a survey on challenges and open problems* (Vol. 55). Springer Netherlands. <https://doi.org/10.1007/s10462-021-09994-y>
- Idé, T., & Kato, S. (2009). Travel-time prediction using gaussian process regression: A trajectory-based approach. *Society for Industrial and Applied Mathematics - 9th SIAM International Conference on Data Mining 2009, Proceedings in Applied Mathematics*, 3, 1177-1188. <https://doi.org/10.1137/1.9781611972795.101>
- Kriegel, H. P., Renz, M., Schubert, M., & Zuefle, A. (2008). Statistical density prediction in traffic networks. *Society for Industrial and Applied Mathematics - 8th SIAM International Conference on Data Mining 2008, Proceedings in Applied Mathematics* 130, 2, 692-703. <https://doi.org/10.1137/1.9781611972788.63>
- Kun, A. J., & Vámosy, Z. (2009). Traffic monitoring with computer vision. *SAMI 2009 - 7th International Symposium on Applied Machine Intelligence and Informatics, Proceedings*, 131-134. <https://doi.org/10.1109/SAMI.2009.4956624>
- Li, X., Pan, G., Qi, G., & Li, S. (2011). Predicting Urban Human Mobility Using Large-Scale Taxi Traces Prediction of Human Mobility for Hotspots.

- Litman, T. (2015). www.vtppi.org Evaluating Complete Streets. *Victoria Transport Policy Institute*(September).
- Marshall, S., Gil, J., Kropf, K., Tomko, M., & Figueiredo, L. (2018). Street Network Studies: from Networks to Models and their Representations. *Networks and Spatial Economics*, 18(3), 735-749. <https://doi.org/10.1007/s11067-018-9427-9>
- Nogueira, B., Pinheiro, R. G. S., & Subramanian, A. (2018). A hybrid iterated local search heuristic for the maximum weight independent set problem. *Optimization Letters*, 12(3), 567-583. <https://doi.org/10.1007/s11590-017-1128-7>
- Qin, K., Zhou, Q., Wu, T., & Xu, Y. Q. (2017). Hotspots detection from trajectory data based on spatiotemporal data field clustering. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 42(2W7), 1319-1325. <https://doi.org/10.5194/isprs-archives-XLII-2-W7-1319-2017>
- Ridewithvia. (2021). Here are counties where residents spend at least 50% of their income on transportation - Via Transportation. In.
- Rubin, F. (1978). Enumerating All Simple Paths in a Graph. *IEEE Transactions on Circuits and Systems*, 25(8), 641-642. <https://doi.org/10.1109/TCS.1978.1084515>
- Schöbel, A., & Scholl, S. (2006). Line planning with minimal traveling time. *OpenAccess Series in Informatics*, 2.
- Shen, Y., Zhang, H., & Zhao, J. (2018). Integrating shared autonomous vehicle in public transportation system: A supply-side simulation of the first-mile service in Singapore. *Transportation Research Part A: Policy and Practice*, 113(June 2017), 125-136. <https://doi.org/10.1016/j.tra.2018.04.004>
- Sousa, R. S. D., Boukerche, A., & Loureiro, A. A. F. (2020). Vehicle Trajectory Similarity: Models, Methods, and Applications. *ACM Computing Surveys*, 53(5). <https://doi.org/10.1145/3406096>
- Toohy, K., & Duckham, M. (2015). Trajectory similarity measures. *SIGSPATIAL Special*, 7(1), 43-50. <https://doi.org/10.1145/2782759.2782767>
- Tu, M., Li, Y., Li, W., Tu, M., Orfila, O., & Gruyer, D. (2019). Improving ridesplitting services using optimization procedures on a shareability network: A case study of Chengdu. *Technological Forecasting and Social Change*, 149(March). <https://doi.org/10.1016/j.techfore.2019.119733>
- Wang, D., Miwa, T., & Morikawa, T. (2020). *Big Trajectory Data Mining: A Survey of Methods, Applications and Services*.
- Xie, Y., Shekhar, S., & Li, Y. (2023). Statistically-Robust Clustering Techniques for Mapping Spatial Hotspots: A Survey. *ACM Computing Surveys*, 55(2). <https://doi.org/10.1145/3487893>
- Yang, C., & Gidófalvi, G. (2018). Mining and visual exploration of closed contiguous sequential patterns in trajectories. *International Journal of Geographical Information Science*, 32(7), 1282-1304. <https://doi.org/10.1080/13658816.2017.1393542>
- Yeh, C.-C. M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H. A., Silva, D. F., Mueen, A., & Keogh, E. (2017). Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets. 1317-1322. <https://doi.org/10.1109/icdm.2016.0179>

- Yoon, G., Chow, J. Y. J., & Rath, S. (2021). A simulation sandbox to compare fixed-route , flexible-route transit , and on-demand microtransit system designs. *arxiv*(arXiv:2109.14138), 1-27.
- Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Sun, G., & Huang, Y. (2010). T-Drive: Driving Directions Based on Taxi Trajectories Jing. 99-99. <https://doi.org/10.1145/1869790.1869807>
- Zhang, Z., Qi, G., Ceder, A., Guan, W., Guo, R., & Wei, Z. (2021). Grid-Based Anomaly Detection of Freight Vehicle Trajectory considering Local Temporal Window. *Journal of Advanced Transportation*, 2021. <https://doi.org/10.1155/2021/8103333>
- Zhao, P., Qin, K., Ye, X., Wang, Y., & Chen, Y. (2017). A trajectory clustering approach based on decision graph and data field for detecting hotspots. *International Journal of Geographical Information Science*, 31(6), 1101-1127. <https://doi.org/10.1080/13658816.2016.1213845>

4.7 Appendices

Table 4.7: Results of Experiment 1

Model	MinPL (m)	MPF	Path ID	TDT (m)	Length (m)	Min Freq	Nodes	Edges
<i>1</i>	<i>1000</i>	<i>20</i>	<i>HP0*</i>	<i>166144*</i>	<i>3487</i>	<i>20</i>	['n15', 'n1638', 'n1636', 'n1996', ...]	[('n15', 'n1638'), ('n1638', 'n1636'), ...]
<i>1</i>	<i>1000</i>	<i>20</i>	<i>HP1</i>	<i>152561</i>	<i>3376</i>	<i>20</i>	['n488', 'n167', 'n683', 'n571', ...]	[('n488', 'n167'), ('n167', 'n683'), ...]
<i>1</i>	<i>1000</i>	<i>20</i>	<i>HP2</i>	<i>129811</i>	<i>3207</i>	<i>20</i>	['n671', 'n675', 'n678', 'n673', ...]	[('n671', 'n675'), ('n675', 'n678'), ...]
1	1000	20	HP3	77053	2075	29	['n2785', 'n205', 'n211', 'n512', ...]	[('n2785', 'n211'), ('n2785', 'n2786'), ...]
1	1000	20	HP4	33583	1320	21	['n380', 'n856', 'n1653', 'n224', ...]	[('n380', 'n853'), ('n380', 'n379'), ...]
1	1000	20	HP5	26990	1155	21	['n1199', 'n1213', 'n1198', 'n1979', ...]	[('n1199', 'n1198'), ('n1199', 'n1213'), ...]
1	1000	20	HP6	25822	1097	20	['n1588', 'n115', 'n784', 'n2350', ...]	[('n1588', 'n1544'), ('n1588', 'n1974'), ...]
1	1000	20	HP7	23319	1014	23	['n68', 'n1685', 'n2318']	[('n68', 'n1685'), ('n1685', 'n2318')]
<i>2</i>	<i>1000</i>	<i>30</i>	<i>HP8*</i>	<i>165889*</i>	<i>3196</i>	<i>31</i>	['n287', 'n2828', 'n291', 'n1112', ...]	[('n287', 'n2828'), ('n2828', 'n291'), ...]
2	1000	30	HP9	143039	2622	31	['n278', 'n1649', 'n1648', 'n680', ...]	[('n278', 'n1649'), ('n1649', 'n1648'), ...]

2	1000	30	HP10	76334	2050	31	['n2785', 'n205', 'n211', 'n512', ...]	[('n2785', 'n211'), ('n2785', 'n2786'), ...]
3	1000	40	HP11 *	124061*	2201	43	['n247', 'n346', 'n2180', 'n1622', ...]	[('n247', 'n249'), ('n247', 'n2038'), ...]
3	1000	40	HP12	109540	1702	52	['n10', 'n2172', 'n324', 'n11', ...]	[('n10', 'n11'), ('n10', 'n1637'), ...]
3	1000	40	HP13	81864	1407	48	['n279', 'n280', 'n1173', 'n44', ...]	[('n279', 'n280'), ('n280', 'n2001'), ...]
4	1000	50	HP12 *	109540*	1702	52	['n10', 'n2172', 'n324', 'n11', ...]	[('n10', 'n11'), ('n10', 'n1637'), ...]
5	1000	60	NA	NA	NA	NA	NA	NA
6	1500	20	HP0*	166144*	3487	20	['n15', 'n1638', 'n1636', 'n1996', ...]	[('n15', 'n1638'), ('n1638', 'n1636'), ...]
6	1500	20	HP1	152561	3376	20	['n488', 'n167', 'n683', 'n571', ...]	[('n488', 'n167'), ('n167', 'n683'), ...]
6	1500	20	HP2	129811	3207	20	['n671', 'n675', 'n678', 'n673', ...]	[('n671', 'n675'), ('n675', 'n678'), ...]
6	1500	20	HP3	77053	2075	29	['n2785', 'n205', 'n211', 'n512', ...]	[('n2785', 'n211'), ('n2785', 'n2786'), ...]
7	1500	30	HP8*	165889*	3196	31	['n287', 'n2828', 'n291', 'n1112', ...]	[('n287', 'n2828'), ('n2828', 'n291'), ...]
7	1500	30	HP9	143039	2622	31	['n278', 'n1649', 'n1648', 'n680', ...]	[('n278', 'n1649'), ('n1649', 'n1648'), ...]
7	1500	30	HP10	76334	2050	31	['n2785', 'n205',	[('n2785', 'n211'),

							'n211', 'n512', ...]	('n2785', 'n2786'), ...]
8	1500	40	HP11 *	124061*	2201	43	['n247', 'n346', 'n2180', 'n1622', ...]	[('n247', 'n249'), ('n247', 'n2038'), ...]
8	1500	40	HP12	109540	1702	52	['n10', 'n2172', 'n324', 'n11', ...]	[('n10', 'n11'), ('n10', 'n1637'), ...]
9	1500	50	HP12 *	109540*	1702	52	['n10', 'n2172', 'n324', 'n11', ...]	[('n10', 'n11'), ('n10', 'n1637'), ...]
10	1500	60	NA	NA	NA	NA	NA	NA
11	2000	20	HP0*	166144*	3487	20	['n15', 'n1638', 'n1636', 'n1996', ...]	[('n15', 'n1638'), ('n1638', 'n1636'), ...]
11	2000	20	HP1	152561	3376	20	['n488', 'n167', 'n683', 'n571', ...]	[('n488', 'n167'), ('n167', 'n683'), ...]
11	2000	20	HP2	129811	3207	20	['n671', 'n675', 'n678', 'n673', ...]	[('n671', 'n675'), ('n675', 'n678'), ...]
11	2000	20	HP3	77053	2075	29	['n2785', 'n205', 'n211', 'n512', ...]	[('n2785', 'n211'), ('n2785', 'n2786'), ...]
12	2000	30	HP8*	165889*	3196	31	['n287', 'n2828', 'n291', 'n1112', ...]	[('n287', 'n2828'), ('n2828', 'n291'), ...]
12	2000	30	HP9	143039	2622	31	['n278', 'n1649', 'n1648', 'n680', ...]	[('n278', 'n1649'), ('n1649', 'n1648'), ...]
12	2000	30	HP10	76334	2050	31	['n2785', 'n205', 'n211', 'n512', ...]	[('n2785', 'n211'), ('n2785', 'n2786'), ...]
13	2000	40	HP11 *	124061*	2201	43	['n247', 'n346', 'n2180', 'n1622', ...]	[('n247', 'n249'), ('n247', 'n2038'), ...]

							...]	
14	2000	50	NA	NA	NA	NA	NA	NA

Bold: TDT \geq average, *Italics:* TDT \geq standard deviation, *: MaxTDT

Table 4.8: Results of Experiment 2

Model	MinPL (m)	MPF	Path ID	TDT (m)	Length (m)	Min Freq	Nodes	Edges
1	1000	20	HP0*	166144*	3487	20	['n15', 'n1638', 'n1636', 'n1996', ...]	[('n15', 'n1638'), ('n1638', 'n1636'), ...]
1	1000	20	HP1	152561	3376	20	['n488', 'n167', 'n683', 'n571', ...]	[('n488', 'n167'), ('n167', 'n683'), ...]
1	1000	20	HP2	129811	3207	20	['n671', 'n675', 'n678', 'n673', ...]	[('n671', 'n675'), ('n675', 'n678'), ...]
1	1000	20	HP3	77053	2075	29	[n2785', 'n205', 'n211', 'n512', ...]	[('n2785', 'n211'), ('n2785', 'n2786'), ...]
1	1000	20	HP4	33583	1320	21	[n380', 'n856', 'n1653', 'n224', ...]	[('n380', 'n853'), ('n380', 'n379'), ...]
1	1000	20	HP5	26990	1155	21	[n1199', 'n1213', 'n1198', 'n1979', ...]	[('n1199', 'n1198'), ('n1199', 'n1213'), ...]
1	1000	20	HP6	25822	1097	20	[n1588', 'n115', 'n784', 'n2350', ...]	[('n1588', 'n1544'), ('n1588', 'n1974'), ...]
1	1000	20	HP7	23319	1014	23	[n68', 'n1685', 'n2318', ...]	[('n68', 'n1685'), ('n1685', 'n2318'), ...]
2	1000	25	HP8*	165889*	3196	31	['n287', 'n2828', 'n291', 'n1112', ...]	[('n287', 'n2828'), ('n2828', 'n291'), ...]
2	1000	25	HP9	158549	3197	25	['n287', 'n2828', 'n291', 'n1112', ...]	[('n287', 'n2828'), ('n2828', 'n291'), ...]
2	1000	25	HP3	77053	2075	29	[n2785', 'n205', 'n211', 'n512', ...]	[('n2785', 'n211'), ('n2785', 'n2786'), ...]

2	1000	25	HP10	29987	1149	25	['n380', 'n856', 'n1653', 'n224', ...]	[('n380', 'n853'), ('n380', 'n379'), ...]
3	1000	30	HP8*	165889*	3196	31	['n287', 'n2828', 'n291', 'n1112', ...]	[('n287', 'n2828'), ('n2828', 'n291'), ...]
3	1000	30	HP11	143039	2622	31	['n278', 'n1649', 'n1648', 'n680', ...]	[('n278', 'n1649'), ('n1649', 'n1648'), ...]
3	1000	30	HP12	76334	2050	31	['n2785', 'n205', 'n211', 'n512', ...]	[('n2785', 'n211'), ('n2785', 'n2786'), ...]
4	1000	35	HP13 *	195778*	3473	35	['n680', 'n279', 'n280', 'n2001', ...]	[('n680', 'n279'), ('n279', 'n280'), ...]
4	1000	35	HP14	152488	2607	36	['n10', 'n1654', 'n11', 'n1637', ...]	[('n10', 'n11'), ('n10', 'n1637'), ...]
5	1000	40	HP15 *	124061*	2201	43	['n247', 'n346', 'n2180', 'n1622', ...]	[('n247', 'n249'), ('n247', 'n2038'), ...]
5	1000	40	HP16	109540	1702	52	['n10', 'n2172', 'n324', 'n11', ...]	[('n10', 'n11'), ('n10', 'n1637'), ...]
5	1000	40	HP17	81864	1407	48	['n279', 'n280', 'n1173', 'n44', ...]	[('n279', 'n280'), ('n280', 'n2001'), ...]
6	1000	45	HP16 *	109540*	1702	52	['n10', 'n2172', 'n324', 'n11', ...]	[('n10', 'n11'), ('n10', 'n1637'), ...]
6	1000	45	HP17	81864	1407	48	['n279', 'n280', 'n1173', 'n44', ...]	[('n279', 'n280'), ('n280', 'n2001'), ...]
6	1000	45	HP18	62722	1047	49	['n1658', 'n27', 'n763', 'n295', ...]	[('n1658', 'n294'), ('n1658', 'n295'), ...]
7	1000	50	HP16	109540*	1702	52	['n10', 'n2172',	[('n10', 'n11'),

			*				'n324', 'n11', ...]	('n10', 'n1637'), ...]
8	1000	55	HP19 *	95498*	1435	56	['n10', 'n324', 'n11', 'n1637', ...]	[('n10', 'n11'), ('n10', 'n1637'), ...]
9	1000	60	NA	NA	NA	NA	NA	NA
10	1500	20	HP0*	166144*	3487	20	['n15', 'n1638', 'n1636', 'n1996', ...]	[('n15', 'n1638'), ('n1638', 'n1636'), ...]
10	1500	20	HP1	152561	3376	20	['n488', 'n167', 'n683', 'n571', ...]	[('n488', 'n167'), ('n167', 'n683'), ...]
10	1500	20	HP2	129811	3207	20	['n671', 'n675', 'n678', 'n673', ...]	[('n671', 'n675'), ('n675', 'n678'), ...]
10	1500	20	HP3	77053	2075	29	['n2785', 'n205', 'n211', 'n512', ...]	[('n2785', 'n211'), ('n2785', 'n2786'), ...]
11	1500	25	HP8*	165889*	3196	31	['n287', 'n2828', 'n291', 'n1112', ...]	[('n287', 'n2828'), ('n2828', 'n291'), ...]
11	1500	25	HP9	158549	3197	25	['n287', 'n2828', 'n291', 'n1112', ...]	[('n287', 'n2828'), ('n2828', 'n291'), ...]
11	1500	25	HP3	77053	2075	29	['n2785', 'n205', 'n211', 'n512', ...]	[('n2785', 'n211'), ('n2785', 'n2786'), ...]
12	1500	30	HP8*	165889*	3196	31	['n287', 'n2828', 'n291', 'n1112', ...]	[('n287', 'n2828'), ('n2828', 'n291'), ...]
12	1500	30	HP11	143039	2622	31	['n278', 'n1649', 'n1648', 'n680', ...]	[('n278', 'n1649'), ('n1649', 'n1648'), ...]
12	1500	30	HP12	76334	2050	31	['n2785', 'n205', 'n211', 'n512', ...]	[('n2785', 'n211'), ('n2785', 'n2786'), ...]
13	1500	35	HP13	195778*	3473	35	['n680', 'n279',	[('n680', 'n279'),

			*				'n280', 'n2001', ...]	('n279', 'n280'), ...]
13	1500	35	HP14	152488	2607	36	['n10', 'n1654', 'n11', 'n1637', ...]	[('n10', 'n11'), ('n10', 'n1637'), ...]
14	1500	40	HP15 *	124061*	2201	43	['n247', 'n346', 'n2180', 'n1622', ...]	[('n247', 'n249'), ('n247', 'n2038'), ...]
14	1500	40	HP16	109540	1702	52	['n10', 'n2172', 'n324', 'n11', ...]	[('n10', 'n11'), ('n10', 'n1637'), ...]
15	1500	45	HP16 *	109540*	1702	52	['n10', 'n2172', 'n324', 'n11', ...]	[('n10', 'n11'), ('n10', 'n1637'), ...]
16	1500	50	HP16 *	109540*	1702	52	['n10', 'n2172', 'n324', 'n11', ...]	[('n10', 'n11'), ('n10', 'n1637'), ...]
17	NA	55	NA	NA	NA	NA	NA	NA
18	NA	60	NA	NA	NA	NA	NA	NA
19	2000	20	HP0*	166144*	3487	20	['n15', 'n1638', 'n1636', 'n1996', ...]	[('n15', 'n1638'), ('n1638', 'n1636'), ...]
19	2000	20	HP1	152561	3376	20	['n488', 'n167', 'n683', 'n571', ...]	[('n488', 'n167'), ('n167', 'n683'), ...]
19	2000	20	HP2	129811	3207	20	['n671', 'n675', 'n678', 'n673', ...]	[('n671', 'n675'), ('n675', 'n678'), ...]
19	2000	20	HP3	77053	2075	29	['n2785', 'n205', 'n211', 'n512', ...]	[('n2785', 'n211'), ('n2785', 'n2786'), ...]
20	2000	25	HP8*	165889*	3196	31	['n287', 'n2828', 'n291', 'n1112', ...]	[('n287', 'n2828'), ('n2828', 'n291'), ...]
20	2000	25	HP9	158549	3197	25	['n287', 'n2828',	[('n287', 'n2828'),

							'n291', 'n1112', ...]	('n2828', 'n291'), ...]
20	2000	25	HP3	77053	2075	29	[<i>'n2785', 'n205', 'n211', 'n512', ...]</i>	[<i>('n2785', 'n211'), ('n2785', 'n2786'), ...]</i>
21	2000	30	<i>HP8*</i>	<i>165889*</i>	<i>3196</i>	<i>31</i>	<i>[<i>'n287', 'n2828', 'n291', 'n1112', ...]</i></i>	<i>[<i>('n287', 'n2828'), ('n2828', 'n291'), ...]</i></i>
21	2000	30	HP11	143039	2622	31	[<i>'n278', 'n1649', 'n1648', 'n680', ...]</i>	[<i>('n278', 'n1649'), ('n1649', 'n1648'), ...]</i>
21	2000	30	HP12	76334	2050	31	[<i>'n2785', 'n205', 'n211', 'n512', ...]</i>	[<i>('n2785', 'n211'), ('n2785', 'n2786'), ...]</i>
22	2000	35	<i>HP13*</i>	<i>195778*</i>	<i>3473</i>	<i>35</i>	<i>[<i>'n680', 'n279', 'n280', 'n2001', ...]</i></i>	<i>[<i>('n680', 'n279'), ('n279', 'n280'), ...]</i></i>
22	2000	35	HP14	152488	2607	36	[<i>'n10', 'n1654', 'n11', 'n1637', ...]</i>	[<i>('n10', 'n11'), ('n10', 'n1637'), ...]</i>
23	2000	40	<i>HP15*</i>	<i>124061*</i>	<i>2201</i>	<i>43</i>	<i>[<i>'n247', 'n346', 'n2180', 'n1622', ...]</i></i>	<i>[<i>('n247', 'n249'), ('n247', 'n2038'), ...]</i></i>
24	2000	45	NA	NA	NA	NA	NA	NA

Bold: TDT ≥ average, Italics: TDT ≥ standard deviation, *: MaxTDT

CHAPTER 5: TACKLING MICROTRANSIT BOTTLENECKS: A SPATIO-TEMPORAL COLLECTIVE ANOMALY DISCOVERY FRAMEWORK

5.1 Introduction

The anomalous hotspot paths based on total distance traveled (AHP–TDT) framework offers a valuable approach for analyzing and pinpointing anomalous hotspot paths within trajectory networks. It stands out by shifting the focus from traditional point- and edge-centric techniques to the total distance traveled (TDT). This shift enables the AHP–TDT framework to reveal complex movement patterns, thus uncovering collective anomalies. Such insights hold significant potential for enhancing urban mobility by providing a deeper understanding of traffic flows and behaviors. However, the AHP–TDT model, in its current form, has limitations due to its lack of integration with the temporal dimension. Only focusing on spatial aspects — the paths and distances — restricts the model’s ability to provide a comprehensive analysis of movement patterns. This is particularly crucial in urban mobility contexts where time plays a pivotal role. By incorporating a spatio-temporal approach, the AHP–TDT framework can gain a more nuanced understanding of movement dynamics.

Integrating the temporal dimension would enable the model to account for variations in traffic flow and mobility patterns at different times. For instance, the same route may exhibit varying characteristics during peak hours versus off-peak hours, weekdays versus weekends, or during special events versus normal days. This temporal analysis is vital for understanding the full spectrum of urban mobility patterns. Ultimately, a spatio-temporal approach in the AHP-TDT framework not only enriches the analysis of urban mobility but also enables the development of more targeted and efficient transportation solutions, adapting to the dynamic nature of city life.

However, implementing a spatio-temporal approach to extract anomalous hotspot paths in trajectory networks presents significant challenges, primarily due to the complexity of identifying collective anomalies across both temporal and spatial dimensions. Conventional urban mobility studies have primarily focused on point and edge anomalies (Chen et al., 2014; Kriegel et al., 2008; Li et al., 2011), often overlooking the intricate collective anomalies in trajectory networks. This shift to analyzing collective anomalies requires sophisticated methodologies ranging from heuristic approaches to advanced graph theory algorithms (Nogueira et al., 2018; Tu et al., 2019; Zhou et al., 2020). Each of these methodologies contributes to a more refined understanding of traffic dynamics and congestion management, yet they also introduce new complexities in data analysis and interpretation.

Furthermore, applying a spatio-temporal approach to the extraction of anomalous hotspot paths within the rapidly evolving domain of urban mobility services adds another layer of complexity. The urban mobility landscape, which extends beyond traditional

vehicular transit, is increasingly shifting towards flexible, demand-responsive models like microtransit (Litman, 2015; Yoon et al., 2021). This evolution is largely driven by digital platforms of transport network companies (TNCs), revolutionizing access to services such as ride-hailing and delivery (Hou et al., 2020). The dynamic nature of these services, combined with the complexity of analyzing traffic flow and congestion within a spatio-temporal frame (Asadi & Regan, 2019), makes extracting meaningful patterns a formidable challenge. The rapidly changing nature of urban mobility, fueled by technological advancements and evolving consumer behaviors, requires a sophisticated, adaptable approach that can effectively integrate and analyze both spatial and temporal data to provide actionable insights for urban planning and traffic management.

This study introduces the spatio-temporal collective anomaly discovery (STCAD) framework. This approach targets microtransit bottlenecks, enhancing urban mobility by integrating spatial insights from anomalous hotspot paths based on total distance traveled (AHP-TDT) and the temporal nuances from the matrix profile (MP), thereby extending the AHP-TDT spatial framework into a comprehensive spatio-temporal discovery. Therefore, this paper is structured around two pivotal research questions:

R1. How does the integration of spatial and temporal data in the STCAD framework enhance the discovery and characterization of AHP?

R2. How does STCAD improve insights into urban mobility over conventional methods?

The rest of the paper is organized as follows: Section 2 discusses related work, and Section 3 presents the methodology. Section 4 presents the research method and results, and Section 5 discusses the findings. Lastly, Section 6 presents the conclusions and future work.

5.2 Related Work

Mobility and Microtransit

Mobility, a term sometimes used interchangeably with transportation, encompasses a wider range of movement than only vehicular transit (Litman, 2015). The evolution of transport systems is moving toward more flexible models, manifesting in services such as ride-hailing and on-demand microtransit (Yoon et al., 2021). These are often grouped under the term TNC, which leverage digital platforms to connect riders with various services, including delivery services (Hou et al., 2020). Specifically, microtransit denotes small-group transit. Yet, to truly harness its potential, a comprehensive grasp of its spatio-temporal dimensions is paramount.

The AHP–TDT methodology presents a forward leap in refining travel experiences and streamlining fleet operations. Its relevance in microtransit is reinforced by (Hou et al., 2020), which highlight the criticality of specific spatial factors such as pickup and drop-off points. Echoing this sentiment, (Stiglic et al., 2015) stress on integrating “meeting points” in ride-sharing systems to bolster efficiency and reduce expenses, drawing parallels with the AHP in our context. Moreover, strategies like the short-turn method proposed by (Chen et al., 2018) advocate for targeted service regions in microtransit to drive down costs. However, despite AHP–TDT’s spatial innovations, it has yet to incorporate the temporal facet.

Research such as (Liu et al., 2020) tackles route optimization and pooling challenges by pairing riders with appropriate vehicles and optimizing dispatch based on spatial and temporal factors, aiming for cost reduction and enhanced efficiency. However, these methods are often constrained to operational or tactical planning and necessitate pre-existing demand or orders for activation.

Collective anomaly detection identifies groups or patterns of anomalies that, when aggregated, suggest irregular behaviors, rather than isolating individual anomalies (Chandola et al., 2009). In microtransit systems, while single anomalies may appear insignificant, their collective presence can indicate major bottlenecks or systemic challenges. The basic brute force approach means comparing each subsequence in a time series with others, leading to intensive quadratic pairwise comparisons.

A paradigm shift in this arena was brought about by the MP. Proposed by (Yeh et al., 2017), the MP offers a transformative perspective on time series analysis. By providing a z-normalized Euclidean distance metric, it encapsulates the similarity between different subsequences in a time series. Such a representation streamlines the identification of motifs (repeated patterns) and discords (anomalies). Given our goal of pinpointing collective anomalies in spatio-temporal data for microtransit bottlenecks, integrating the MP's capabilities becomes indispensable. Its ability to swiftly capture temporal nuances can significantly aid in enhancing urban mobility strategies.

In time series analysis, the decomposition of data into distinct components—trend, seasonality, and residuals—serves as a pivotal methodology for nuanced understanding and

anomaly detection. Each component embodies specific characteristics of the data, and discerning this can significantly influence the interpretation of anomalies and the strategic responses they mandate (Zhang et al., 2022). The trend reflects a systematic, linear or nonlinear trajectory that the series follows over a period, indicating a general upward or downward movement. Anomalies aligned with this component often signify more than simple incidental fluctuations; they hint at systematic or sustained shifts, potentially categorizable as trend anomalies. These irregularities suggest widespread disturbances, like a long-term industrial slowdown or a persistent shift in resource usage, which are intrinsic to the entire data set or system rather than isolated events. Seasonality accounts for periodic fluctuations that recur over regular intervals, be it daily, weekly, monthly, or annually. These oscillations are predictable and can be extrapolated into the future. While anomalies in this domain are less frequent, their occurrence could indicate disruptions in the usual cyclical patterns, termed seasonal anomalies, such as an unexpected dip in sales during the peak season or an uncharacteristic spike in energy usage during off-peak hours. Residuals encompass the irregularities and noise within the data, essentially capturing the randomness or unforeseen events not explained by the trend or seasonal components. Anomalies within this component, often abrupt and ephemeral, typify point anomalies. These could stem from spontaneous, unforeseeable events, like sudden equipment malfunctions or flash crowd incidents, punctuating the data with brief spikes or dips (Zhang et al., 2022).

In their study, (Zhang et al., 2022) delineate the inadequacy of non-sequential anomaly definitions in sequential data sets and propose a nuanced classification. They categorize sequential anomalies into point anomalies, encompassing global and context point anomalies,

and pattern anomalies, which include shapelet, seasonal, and trend anomalies. This taxonomy aligns with the understanding that anomalies in the trend and seasonality components represent broader, more systemic issues (pattern anomalies), while those in the residuals are more transient and situational (point anomalies) (Zhang et al., 2022). Recognizing these distinctions is paramount in crafting informed, effective responses or interventions. It necessitates a comprehensive understanding of the underlying data's behavior and the contextual implications of these anomalies, ensuring that measures taken are responsive to the specific type of anomaly, be it a momentary aberration or an indication of a more deep-seated, systemic issue.

Spatio-Temporal Anomaly Discovery in Trajectory Network

Analyzing and extracting meaningful patterns from traffic flow and congestion remains an intricate task, especially within a spatio-temporal frame, given the dynamic behavior of transportation networks. This dynamism, coupled with diverse dependencies and recurrent events, accentuates the complexity (Asadi & Regan, 2019). Specifically, challenges such as the dial-a-ride problem and the ride-sharing conundrum are earmarked as nondeterministic polynomial time (NP-hard) issues (Tu et al., 2019).

To address these NP-hard challenges, researchers typically lean on exact algorithms such as dynamic programming and Lagrangian relaxation. However, their limited scalability has led to the adoption of heuristic approaches, including genetic algorithms and tabu searches. A notable mention in this domain is the hybrid approach ILS-VND, lauded for its efficiency and robustness, as highlighted by (Nogueira et al., 2018; Tu et al., 2019).

The modern wave of research employs graph theory algorithms for dissecting these challenges, with studies focusing on collective taxi movements to make predictions about travel-time (Idé & Kato, 2009), understand traffic density (Kriegel et al., 2008), and detect anomalies for enhanced passenger experiences (Atluri et al., 2018). An expansive perspective on congestion detection and management is articulated by (Aashtiani & Magnanti, 1981; Grundy & Radenkovic, 2010). Similarly, (Wang et al., 2019) delve deep into aggregation effects using large-scale car trajectory data.

In our quest to refine microtransit bottlenecks, we recognize the potential of travel time as a metric over simple trip frequency. By juxtaposing expected versus actual travel times, potential congestion zones can be identified. Nevertheless, it is crucial to adjust for variables such as unforeseen stops or meteorological changes to glean accurate insights. While (Salamanis et al., 2017) have ventured into abnormality detection in traffic, their perspective diverges slightly from ours, emphasizing the diverse events like accidents or weather shifts. Given our microtransit lens, our approach remains more focused.

Lastly, with the rise of intricate systems and methodologies, deep learning emerges as a frontrunner for spatio-temporal anomaly discovery. Graph neural networks (GNN), and especially the spatial-temporal GNN (ST-GNN), epitomize this transition, balancing both spatial and temporal aspects. This paradigm shift, along with a more detailed overview of ST-GNN applications, is elaborated upon in studies by (Bui et al., 2022; Zhou et al., 2020). The latter offers rich insights on traffic forecasting via ST-GNN frameworks, aligning seamlessly with our objective to tackle microtransit bottlenecks.

Table 5.1: Spatial and Temporal Thematic Methods Overview

Theme	Methods	Example Studies
Spatial	AHP–TDT Framework	(Chen et al., 2018)
	Short-turn method for targeted service regions	
Temporal	MP	(Yeh et al., 2017)
	Time series decomposition	(Zhang et al., 2022)
	Classification of anomalies into point, shapelet, seasonal, and trend anomalies	
Spatio-Temporal	Dynamic programming	(Nogueira et al., 2018)
Temporal	Heuristic approaches	(Tu et al., 2019)
	Hybrid approach (such as ILS-VND)	(Idé & Kato, 2009)
	Graph theory	(Zhou et al., 2020)
	Deep learning (including GNN and ST-GNN)	(Bui et al., 2022)

5.3 Methodology

This section details the STCAD framework. The framework represents a fusion of the AHP–TDT approach and the MP method for temporal anomaly detection. Illustrated in Figure 5.1, the STCAD methodology unfolds through a series of structured steps, meticulously blending spatial and temporal data analyses to yield insightful results. It initiates in the spatial dimension, beginning with the delineation of a road network. This crucial first step, define road network (A1), lays the groundwork by establishing the spatial framework within which the trajectory data will be analyzed. Following this, the extract trajectory network (A2) step is undertaken, where trajectory data is meticulously extracted,

mapping movement patterns onto the pre-established road network and setting the stage for deeper analysis.

Transitioning to the temporal dimension, the framework advances to extract temporal TDT (B1), a step crucial for understanding the extent of movement within the network over time. This is closely followed by adjust appropriate window size (B2), a critical process that calibrates the temporal analysis for optimal anomaly detection. The subsequent step, validate with time series decomposition (B3), ensures the robustness of the temporal findings. Finally, collective temporal anomalies (B4) are identified, highlighting significant temporal patterns that might indicate systemic issues or recurrent trends in trajectory network.

In its final phase, the STCAD framework shifts to a spatio-temporal dimension. This begins with define network bounds (C1), which delineates the extent of the analysis within the spatial-temporal matrix. Extract subgraphs (C2) is the next step, where specific segments of the network are isolated for detailed examination. The extract distinct paths (C3) phase involves identifying unique pathways within these subgraphs, which are crucial for understanding movement trends. The process culminates in set anomalous thresholds (C4), where criteria are established to flag significant anomalies within these paths. The objective of this comprehensive methodology is succinctly captured in the final step, spatio-temporal anomalous hotspot paths (ST-AHP) (D). Here, the framework consolidates its spatial and temporal analyses to pinpoint and characterize hotspot paths that are anomalous, providing vital insights into the dynamics of trajectory network and urban mobility.

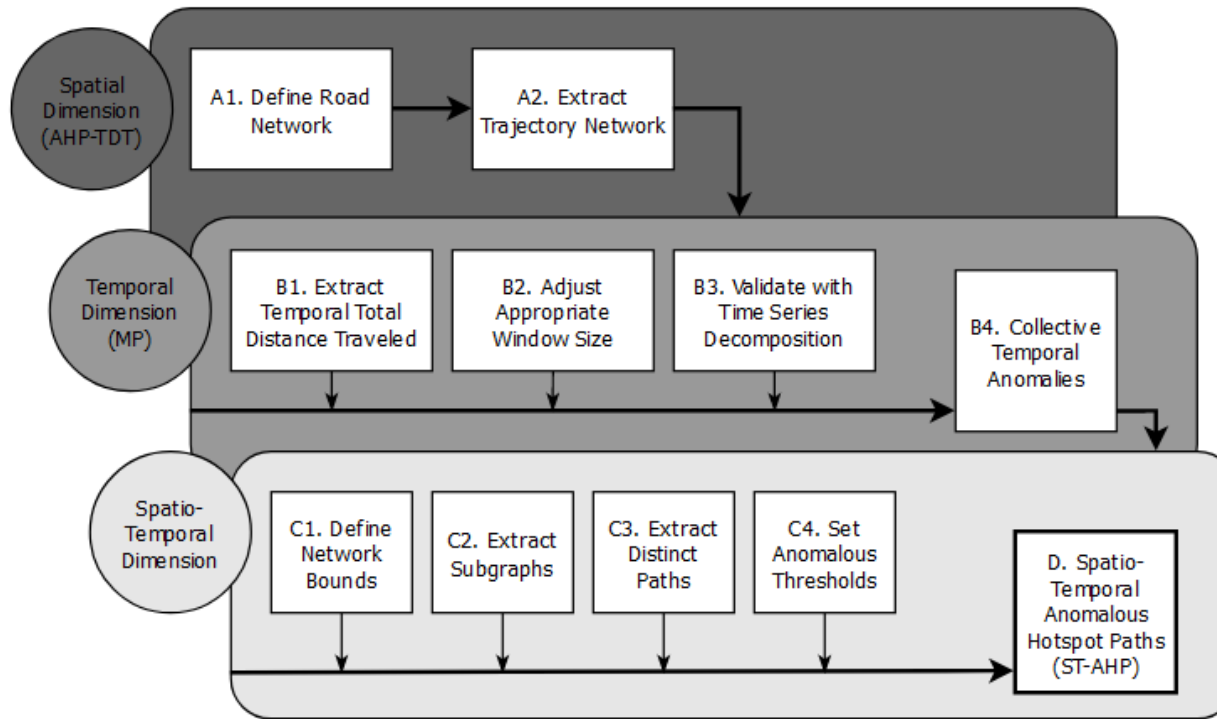


Figure 5.1: Spatio-Temporal Collective Anomaly Discovery Framework

As the STCAD framework transitions into its temporal dimension, its primary aim shifts from pinpointing specific points to unveiling collective temporal anomalies within the trajectory network. This critical objective, encapsulated in step B4 of the framework, is initiated with the extraction of temporal TDT in step B1. This initial phase is pivotal for grasping the movement’s scope within the network over time, setting the stage for a refined temporal analysis. Key to this temporal analysis is the deployment of the MP, a tool instrumental in uncovering and distinguishing time-specific irregularities, patterns, and nuances that may be concealed in a purely spatial or point-based analysis. The MP’s implementation within the STCAD framework transcends conventional temporal analysis, delving deeper to highlight not only the immediate anomalies but also those patterns that

unfold over time, revealing time-sensitive constraints and recurrent themes in trajectory patterns (Yeh et al., 2017).

Following the extraction of TDT, the STCAD framework navigates to step B2 adjust appropriate window size. This step is essential in tailoring the temporal analysis for optimal detection of anomalies. The selection of the appropriate window size within the MP framework is a nuanced process, significantly affecting the analysis's resolution, sensitivity to anomalies, computational demands, and the accuracy of the data interpretation. Larger window sizes, while offering a broader view of the trends, increase computational requirements, whereas smaller windows, though computationally efficient, may miss overarching patterns due to their narrow focus (Djenouri et al., 2019). Thus, the decision-making process requires a careful balance between detailed insight extraction and maintaining a comprehensive perspective on the trends.

Deciding on an optimal window size is not a standardized process but one necessitating specific data and contextual considerations, especially in urban mobility frameworks. This decision significantly affects the system's precision in recognizing and reporting relevant incidents. Established literature seldom prescribes fixed window sizes for traffic-related anomalies, given their inherent reliance on the unique aspects of each data set. In this realm, expertise in urban dynamics and traffic flow becomes invaluable, guiding the selection process toward window sizes that reflect common durations of urban transit events (Djenouri et al., 2019). Alternatively, the optimal window size is often arrived at through empirical fine-tuning. This iterative process involves experimenting with multiple window sizes until the most effective one is found, often measured by anomaly detection accuracy (Zhang et al.,

2021). Some researchers adopt a data-driven methodology, adjusting the window size in response to data traits such as event frequency or typical durations of interest-specific occurrences or anomalies. For instance, a window might mirror the average length of traffic jams if congestion is a focal concern (Zhang et al., 2021). This study embraces a data-oriented strategy, corroborated by evidence suggesting a prevalent window size range between 4 and 10 in comparable research scenarios (Djenouri et al., 2019; Zhang et al., 2021).

Following the strategic determination of the optimal window size, our framework proceeds to validate the anomaly discovery efficacy of the MP through a meticulous time series decomposition of the TDT, as illustrated in step B3. Moreover, this step helps validating the MP findings as it involves deconstructing the TDT into its constitutive elements: trend, seasonality, and residuals (Zhang et al., 2021). By isolating these components, we could cross-verify the anomalies pinpointed by the MP. Specifically, we scrutinized the residuals for abrupt, unaccounted fluctuations that characterize point anomalies, contrasting these against the systemic pattern anomalies evident within trend and seasonal deviations. This comparative inspection emphasizes the precision of the MP in identifying significant anomaly candidates and enhances our confidence in its role as a robust tool for anomaly discovery within the intricate dynamics of our network.

As the framework transitions from its detailed temporal analysis, it advances into the spatio-temporal dimension. This integration begins with Step C1, define network bounds, which establishes the structural parameters of the network, ensuring a clearly delineated space for further analysis. This definition of boundaries is vital for contextualizing the

subsequent steps within a well-defined spatial framework. Following this, in Step C2, extract subgraphs, the framework uses the connected components method to isolate specific segments of the network. This extraction is instrumental in segmenting the network into manageable subgraphs for focused analysis. In Step C3, extract distinct paths, the framework meticulously identifies and delineates distinct movement paths within these subgraphs. This step is essential in clarifying and understanding the various trajectories and patterns of movement that occur within the network, providing a clear and unambiguous representation of urban mobility flows. The culmination of this spatio-temporal integration is achieved in Step C4, set anomalous thresholds, where the framework establishes criteria to identify significant spatio-temporal anomalies. This step is critical in pinpointing AHP, which are the focus of the final objective, ST-AHP in Step D. This culmination phase is where the STCAD framework identifies and characterizes those paths that exhibit unusual spatio-temporal characteristics, thereby revealing insights into the complex dynamics of trajectory data and urban movement.

5.4 Research Methods

To validate our model, we used real-world trajectory data collected from an on-demand transportation agency operating in Porto, Portugal.⁹ This data set includes an impressive array of approximately 1.7 million individual trips, chronicled over a full annual cycle from July 1, 2013 to June 30, 2014. Every trip is meticulously represented by consecutive GPS coordinates that are recorded at regular 15-second intervals. Such recordings create polylines that allow for a detailed analysis of movement and traffic patterns within the city. An

⁹ <https://www.kaggle.com/craiptap/taxi-trajectory>

example of this raw data can be found in Table 5.2. The selected data set not only aligns with the overarching goals of this research but also has been leveraged in similar studies, underscoring its relevance and validity for our investigation. While our approach would have been enriched by examining an array of similar trajectory data sets, recent regulatory shifts and increasing privacy concerns, including those articulated in geolocation privacy legislation in countries such as the United States, have placed significant constraints on the accessibility of other similar data sets.

Table 5.2: A Sample of the Raw Data

TRIP_ID	CAL L_TY PE	ORIGI N_CAL L	ORIGI N_STA ND	TAX I_ID	TIMESTA MP	DAY _TYP E	MISSIN G_DAT A	POLYLINE
1379415 6366200 00000	A	2002		2000 0653	9/17/2013 07:00	A	FALSE	[[[-8.625798,41.157342], [-8.625789,41.15736], [- 8.625744,41.157369], ...]]
1379415 6146200 00000	B		6	2000 0657	9/17/2013 07:00	A	FALSE	[[[-8.582598,41.180202], [-8.582346,41.180211], [- 8.582265,41.180769], ...]]
1379415 7416200 00000	B		32	2000 0011	9/17/2013 07:02	A	FALSE	[[[-8.627589,41.157684], [-8.627607,41.157702], [- 8.627913,41.157909], ...]]

To define the road network (Step A1), we used two powerful Python libraries. First, we employed NetworkX to model and manage the intricacies of the road network as mirrored in

the trajectory data.¹⁰ NetworkX is renowned for its ability to effectively construct, manipulate, and study the structure and dynamics of complex networks, thereby serving as an apt tool for our task. Subsequently, we used OSMnx for accessing and harnessing spatial data such as street networks, which were fundamental to our discovery task.¹¹ This tool was particularly useful due to its ability to extract, model, analyze, and visualize a wide range of spatial objects using data from OpenStreetMap (Boeing, 2017 and visualizing complex street networks). Our experiment focused on the city of Porto, Portugal, with specific geographical coordinates (41.155, -8.63) serving as our central point of interest. To ensure a manageable and relevant area of study, we limited our scope to a radius of 2,500 meters around this focal point. Following this, we engaged in data pre-processing, which included dismissing trips of less than one-minute duration as well as extracting key data points such as pickup and drop-off locations, which played a crucial role in subsequent stages of our experiment. The result of these procedures was a meticulously defined road network, which accurately reflected the urban layout of Porto, comprising a total of 6,159 edges and 2,993 nodes. This road network formed the base upon which our further discovery will be conducted. This approach employed in defining the road network ensures that it is an accurate, comprehensive, and analytically valuable representation of the real-world urban environment.

To derive the trajectory network (Step A2) from the established road network, we methodically matched nodes and edges from trip trajectories to their analogous components in the road network. Central to this alignment were the specific pick-up and drop-off points,

¹⁰ <https://networkx.org/>

¹¹ <https://github.com/gboeing/osmnx>

which we associated with the closest nodes and edges in the road network. This method not only ensured a cohesive integration of both road and trajectory networks but also removed any unused nodes and edges from the road network based on our trajectory data. Although the road network's structure remained unaltered across data subsets, the trajectory network exhibited variability. Meaning, for a different trajectory data subset, the road network remained unchanged, but the trajectory network evolved. To compute the daily total traveled distance (Step B1), we generated a distinct trajectory network for each time segment and its associated spatial intricacies. This approach, though time-intensive, provided a nuanced understanding and precise representation, as it considered the actual distances of trips rather than only their count.

Figure 5.2 showcases the raw TDT data, offering a glimpse into the inherent patterns which, though perceptible, required transformation for meaningful interpretation. The MP emerged as an instrumental transformation tool. Primarily, it was efficient in reducing spatial complexity to $O(n)$ by retaining only the smallest non-trivial distances from each distance profile. These distances, maintained in Euclidean space, signified sub-sequence similarities or disparities within the time series. Specifically, a distance nearing 0 indicated high similarity to another sub-sequence, whereas a distance considerably diverging from 0, such as 100, denoted distinctiveness. Therefore, extracting the largest distances identified the discords (anomalies) (Yeh et al., 2017).

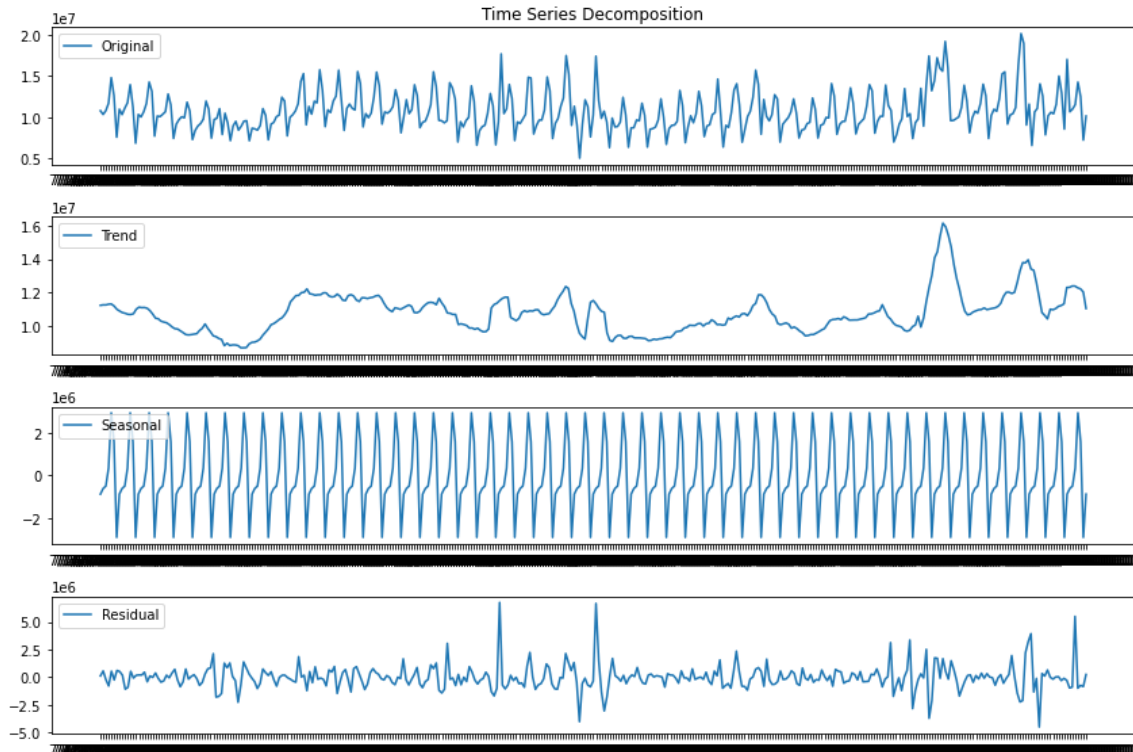


Figure 5.2: Raw and Temporal Decomposition of Total Distance Traveled

In the following phase (Step B2), we embarked on a rigorous experimental to refine the parameters influencing MP performance, specifically focusing on the critical aspect of window size selection. Following comparable research scenarios (Djenouri et al., 2019; Zhang et al., 2021), our exploration covered a range from 4 to 10, a process meticulously detailed in Figure 5.3. This phase was instrumental in shedding light on the dynamics of identified top anomalies, revealing their sensitivity to alterations in window size. A pivotal discovery was the emergence of two major collective anomalies, whose presence and breadth notably expanded as the window size increased. In pursuit of analytical precision and operational efficiency for the subsequent phases of our study, we converged on a window

size of 4. This selection unveiled a substantial collective anomaly, unfolding over the period from June 21, 2014, to June 24, 2014.



Figure 5.3: Matrix Profile Results

To ascertain the validity of the anomalies pinpointed by MP, we subsequently engaged in time series decomposition of TDT (Step B3), as thoroughly illustrated in Figure 5.2. This analytical approach enhanced the robustness of our findings, corroborating the MP's competence in discerning significant disturbances within the data landscape. Furthermore, the exercise bolstered our conviction in the appropriateness of the chosen window size, as the anomalies detected concurred with pronounced inflections in the trend component, thereby attesting to their substantive nature. This congruence between the MP's insights and the trend disruptions revealed through time series decomposition underscores the reliability of the selected window size and the discovered anomalies.

Following our temporal mining, the experiment shifted to a spatio-temporal dimension in the STCAD framework. Post meticulous data cleansing and preprocessing efforts, our finalized data set encompassed 20,959 trips. Adopting the shortest distance approach (Idé & Kato, 2009) for measuring the shortest path in in a trip, these trips evolved into a trajectory network, structured with 2,993 nodes and 6,159 edges. This intricate network, illustrated in Figure 5.4, served as the foundational layer for our ensuing experiment.



Figure 5.4: Trajectory Network

To establish the lower bounds within our STCAD framework (Step C1), experiments were initiated to fine-tune the minimum path length (MinPL) and minimum path frequency (MPF) parameters. Given that our data set surpasses the scale of the AHP–TDT study, our first experiment engaged with an expanded MPF value range of [400, 800, 1,200, 1,600] paired with MinPL values spread across [1,000, 1,500, 2,000] meters. This was subsequently augmented by broadening the MPF range to [400, 600, 800, 1,000, 1,200, 1,600], while keeping the MinPL consistent. This parameterization was essential for excluding inconsequential paths, thereby spotlighting the traffic patterns of utmost relevance. Such network bounds setting served as a mechanism to ensure no anomalous activity, however subtle, evaded detection, consequently providing a richer, more nuanced comprehension of the urban mobility landscape. In the ensuing phase of our collective anomaly discovery, the experimentation incorporated comprehensive steps within our STCAD framework, commencing with the extraction of subgraphs from the trajectory network using connected components (Step C2). Subsequently, we established a maximum path length (MaxPL) threshold at 3,500 meters, facilitating the transformation of all subgraphs into unambiguous, simple paths. The subsequent stage (Step C3) introduced an acceptable similarity parameter set at 0.3, a critical determinant allowing for the extraction of distinct paths by identifying unique paths and preventing excessive overlap.

Within this structured environment, our preliminary experiment generated 12 models, unraveling 9 prominent hotspot paths, illustrated in Table 5.3 in the Appendices. Significantly, the path denoted as “HP0” registered an exceptional TDT of 3,056.116 km, covering an area of 3.198 km, with 413 trip instances. At another example, “HP1” revealed

TDT of 2,967.668 km, covering a path of 3.484 km, with 408 trip instances. The subsequent experiment, characterized by an expanded MPF range, produced 18 models, subsequently identifying 13 distinct hotspot paths, illustrated in Table 5.4 in the Appendices. Notably, the topographical characteristics of the primary and secondary hotspot paths were reflective of those discerned in the former experiment, underscoring the consistency and reliability of our analytical methodology.

In the culmination of the STCAD framework (Step C4), the focus was on establishing precise thresholds to effectively identify anomalous spatial hotspot paths. This critical process incorporated two principal methodologies as per (Aggarwal, 2017; Yeh et al., 2017): the top-k anomalies method and a threshold-based approach. The top-k anomalies method prioritized paths with the highest TDT values, marking them as anomalous. This selection was based on the premise that the most traversed paths were likely to exhibit unusual patterns or highlight areas of concern, warranting closer scrutiny. Conversely, the threshold method flagged paths whose TDT met or surpassed the average TDT of all paths, singling out routes that significantly deviated from the normative traffic flow of the network. For added statistical rigor, paths were deemed anomalous if their TDT was notably higher than the mean, specifically one or more standard deviations above it. This ensured that the anomalies identified were not only above average but also distinctly significant. Furthermore, to bolster the robustness of our findings, the framework adopted a multi-method approach. A path was classified as anomalous only if it was recognized as such by multiple analytical methods (Aggarwal, 2017; Yeh et al., 2017). This included paths with the highest TDT, those exceeding the average TDT, and those substantially above the mean TDT. The integration of

these methods ensures a comprehensive and reliable identification of potential spatial hotspots.

The result of this multifaceted approach to anomaly detection is illustrated in the final objective of the framework, Step D, where ST-AHP are identified. Figure 5.5 provides a lucid representation of the most pronounced ST-AHP identified, specifically "HP0" and "HP1", offering insights into typical anomalous patterns within our trajectory network. A comprehensive breakdown of outcomes from both experiments can be found in Tables 5.3 and 5.4 in the Appendices, capturing both overt and subtle observations. Through these results, our research underscores the importance of recognizing key spatio-temporal hotspots that merit further discovery, aiming to alleviate microtransit bottlenecks and enhance urban mobility.

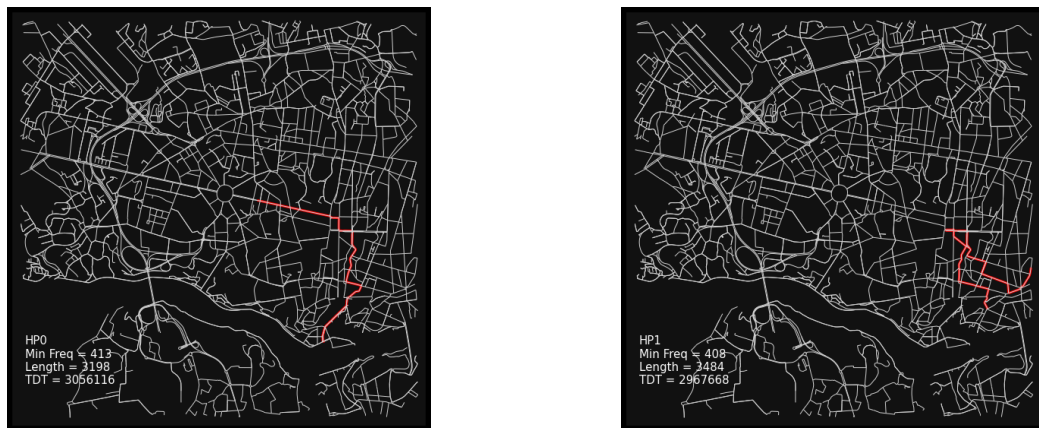


Figure 5.5: Anomalous Hotspot Paths Within our Trajectory Network

5.5 Discussion

In exploring the capabilities of the STCAD framework, our study has delved into the intricacies of identifying collective temporal anomalies within trajectory networks, offering a

nuanced perspective that significantly advances our understanding in the field of urban mobility. This exploration is particularly critical when considering the distinct nature and implications of collective temporal anomalies as opposed to point temporal anomalies, each presenting unique challenges and requiring different detection methodologies. Collective temporal anomalies, unlike their point-in-time counterparts, represent extended irregular patterns occurring across multiple data points or trajectories. Their detection is essential for understanding broader trends and systemic issues within trajectory data, such as recurring traffic congestions or consistent deviations in travel patterns. These anomalies provide insights crucial for long-term planning and strategic decision-making, offering a comprehensive view that point anomalies, typically isolated incidents, cannot.

However, the detection of collective anomalies presents its own set of challenges, primarily due to the complexity of the data and the need to analyze patterns over extended periods. Traditional methods that focus on point anomalies may lack the depth and scope required to capture these broader patterns. The STCAD framework addresses these challenges through its integration of specific spatial and temporal data analysis devices. By employing the MP method, STCAD effectively mines detailed temporal information from trajectory networks, uncovering patterns and irregularities that might be overlooked in a solely spatial study. For example, a critical aspect of STCAD's methodology is the determination of an appropriate time window size within the MP framework. This decision directly influences the resolution of detected patterns and the framework's sensitivity to anomalies, demonstrating a nuanced departure from conventional trajectory hotspots

methodologies. This data-driven approach ensures that the temporal analysis is attuned to the specific nuances of the trajectory network context.

Our pivotal experiment within the framework revealed a significant collective anomaly from June 21 to June 24, 2014. During this period, traffic patterns showed notable deviations from the norm, indicating an extended phase of unusual urban mobility. This finding highlights the framework’s ability to capture not just AHP but also ST–AHP, offering deeper insights into urban mobility dynamics. Upon validating the data, we confirmed that the São João festival, one of Porto’s largest events, occurred on June 23. This festival, renowned for its lively street parties, music, dancing, and traditional activities, attracts both locals and tourists. It culminates in fireworks display over the Douro River at night, with the riverfront being the main viewing area (Penedo, 2024), precisely where our path starts, as illustrated in Figure 5.6.



Figure 5.6: Alignment of ST Path Initiation with São João Festival on June 23rd¹²

When comparing STCAD to existing methodologies, including AHP–TDT, and spatial-only analyses, it becomes clear that STCAD provides a more comprehensive analysis

¹² <https://portuguesediner.com/tiamaria/dia-de-sao-joao-traditions/>

of urban mobility. The spatio-temporal path extends to the riverfront where the fireworks occurred, whereas paths from spatial-only analyses were further into the city. These findings validate our results, demonstrating that the framework offers a holistic view of urban mobility. While previous studies have focused on immediate, isolated incidents, STCAD's approach allows for understanding larger, more complex patterns, which are crucial for effective urban planning and management. This comprehensive perspective not only contributes significantly to academic research but also has practical implications for developing efficient and responsive urban transit systems.

The identification of ST paths "HP0" and "HP1", which are characterized by their significant TDT rather than frequency, unveils patterns that extend beyond fleeting or isolated incidents. These ST-AHP represent consistent, recurring deviations from typical trajectory patterns, highlighting deeper systemic trends in movement and pinpointing critical routes essential to improving a city's overall flow and dynamics. Therefore, the STCAD framework's approach to identifying ST-AHP offers a nuanced and in-depth understanding of trajectory movement and urban mobility, and reveals significant pathways that are central to a city's transportation dynamics, providing insights into long-term trends and systemic issues that are crucial for effective urban planning and the development of sustainable transportation strategies.

Contributions

This study makes several key contributions. First, it introduces the STCAD framework, a pioneering approach that enhances urban mobility analysis. This framework integrates

spatial insights from AHP–TDT with temporal nuances from the MP, thereby transforming the AHP–TDT spatial framework into a comprehensive tool for spatio-temporal anomaly discovery. Second, this study goes beyond theoretical propositions by empirically testing the STCAD framework using real-world trajectory data. This practical application demonstrates the framework’s effectiveness in identifying and analyzing collective anomalies in urban mobility. The empirical testing not only validates the framework but also showcases its potential in real-life urban mobility scenarios, enhancing its credibility and applicability. Third, this study provides a thorough discussion on the actionable insights that can be derived from implementing the STCAD framework in urban mobility contexts. It delves into how this framework can practically contribute to identifying and alleviating microtransit bottlenecks. This aspect of the study bridges the gap between theoretical research and practical application, offering valuable strategies for improving urban transit operations and overall mobility in urban settings.

Implications

Theoretical Implications

The STCAD framework represents a significant advancement in the theoretical landscape of trajectory networks and spatial data analysis by integrating both spatial and temporal dimensions. This integration marks a substantial shift from traditional spatial-only analyses to a more holistic perspective in trajectory network analysis. By focusing on collective anomalies that manifest across both spatial and temporal dimensions, STCAD enriches the domain of spatial analysis. It offers profound insights into broader, systemic

trends and patterns in trajectory networks. This approach aligns with the evolving focus in spatial data analysis on understanding the complex interplay of sequential and collective events, thereby contributing to a deeper and more nuanced understanding of trajectory networks.

The STCAD framework contributes significantly to the improvement of urban mobility services research. By identifying significant pathways based on TDT in a spatio-temporal context, it goes beyond conventional analysis of points and edges, uncovering intricate movement patterns and their anomalies within urban mobility networks. This theoretical advancement aids in developing new methodologies for data mining and machine learning, particularly in anomaly detection algorithms. Furthermore, the framework's ability to recognize recurring or consistent deviations from typical traffic patterns has critical implications for urban planning, environmental studies, and social network analysis. This interdisciplinary approach highlights the importance of a holistic perspective in spatial data analysis and paves the way for further theoretical developments across various domains, emphasizing the framework's pivotal role in enhancing the efficacy and sustainability of urban mobility services.

Practical Implications

The STCAD framework enables the discovery of systemic and seasonal trends in urban mobility, which are crucial for reaching the highest operational opportunities. By analyzing anomalous paths in conjunction with time, this framework aids in optimizing travel services in areas with high demand for individual on-demand trips. For example, it can facilitate the

tailoring of transportation services for efficient connectivity between airports and specific neighborhoods during peak times, or between stadiums and other areas post-major events. This understanding also extends to managing traffic flow from residential areas to business districts during rush hours, or from shopping malls to other regions during holiday seasons, significantly enhancing service planning and resource allocation.

The STCAD framework advances the management of various urban mobility resources by providing insights into traffic patterns and anomalies. This tool is invaluable for urban planners and traffic managers, as it enables the development of strategies to reduce congestion, optimize road networks, and enhance the overall transportation system. In public safety and emergency response, detecting and analyzing hotspot paths can lead to preemptive measures for managing high-incident areas, thus improving public safety. The framework also supports smart city initiatives by providing a data-driven approach to managing urban infrastructure. This is particularly relevant with the integration of technologies like IoT in smart cities. Additionally, the insights from STCAD can optimize on-demand mobility services, improving route planning for ridesharing and streamlining delivery logistics, thereby enhancing the efficiency and effectiveness of these services.

Conclusions and Future Work

In addressing the complexities of urban mobility and microtransit bottlenecks, our study has harnessed the STCAD framework, yielding pivotal insights and advancing the field of urban mobility analysis. Reflecting on our research, we revisited two fundamental questions that guided our exploration: 1) How does the integration of spatial and temporal

data in the STCAD framework enhance the discovery and characterization of AHP? This investigation demonstrated the enhanced capacity of STCAD in uncovering intricate patterns within urban mobility networks. By marrying spatial and temporal dimensions, the framework offered a nuanced and holistic perspective, enabling a deeper understanding of both the location and timing of microtransit bottlenecks. 2) How does STCAD improve insights into urban mobility over conventional methods? Compared to traditional approaches, STCAD's integration of spatio-temporal data provided a comprehensive view of urban mobility trends. This methodological advancement allowed for the identification of significant spatio-temporal anomalies, providing crucial insights for urban planning and microtransit optimization.

This research underscores the critical importance of adopting a spatio-temporal lens in urban mobility studies. The STCAD framework, with its ability to detect significant collective anomalies in trajectory networks, paves the way for informed and strategic urban planning. Future work should focus on refining this framework further, potentially integrating advanced machine learning techniques and expanding its applicability to other urban contexts. The journey toward smarter, more efficient urban mobility systems continues, with STCAD marking a significant step forward in this endeavor.

5.6 References

- Aashtiani, H. Z., & Magnanti, T. L. (1981). Equilibria on a Congested Transportation Network. *SIAM Journal on Algebraic Discrete Methods*, 2(3), 213-226. <https://doi.org/10.1137/0602024>
- Aggarwal, C. C. (2017). *Outlier Analysis*. <https://doi.org/10.1016/b978-012724955-1/50180-7>
- Asadi, R., & Regan, A. (2019). A Spatial-Temporal Decomposition Based Deep Neural Network for Time Series Forecasting. 1-17.

- Atluri, G., Karpatne, A., & Kumar, V. (2018). Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys*, 51(4), 1-37. <https://doi.org/10.1145/3161602>
- Boeing, G. (2017). OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65, 126-139. <https://doi.org/10.1016/j.compenvurbsys.2017.05.004>
- Bui, K. H. N., Cho, J., & Yi, H. (2022). Spatial-temporal graph neural network for traffic forecasting: An overview and open research issues. *Applied Intelligence*, 52(3), 2763-2774. <https://doi.org/10.1007/s10489-021-02587-w>
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A Survey. *Computers, Materials and Continua*, 14(1), 1-22. <https://doi.org/10.1145/1541880.1541882>
- Chen, C., Zhang, D., Li, N., & Zhou, Z. H. (2014). B-planner: Planning bidirectional night bus routes using large-scale taxi GPS traces. *IEEE Transactions on Intelligent Transportation Systems*, 15(4), 1451-1465. <https://doi.org/10.1109/TITS.2014.2298892>
- Chen, J., Liu, Z., Wang, S., & Chen, X. (2018). Continuum approximation modeling of transit network design considering local route service and short-turn strategy. *Transportation Research Part E: Logistics and Transportation Review*, 119(October), 165-188. <https://doi.org/10.1016/j.tre.2018.10.001>
- Djenouri, Y., Belhadi, A., Lin, J. C. W., Djenouri, D., & Cano, A. (2019). A Survey on Urban Traffic Anomalies Detection Algorithms. *IEEE Access*, 7, 12192-12205. <https://doi.org/10.1109/ACCESS.2019.2893124>
- Grundy, A., & Radenkovic, M. (2010). Promoting congestion control in opportunistic networks. *2010 IEEE 6th International Conference on Wireless and Mobile Computing, Networking and Communications, WiMob'2010*, 324-330. <https://doi.org/10.1109/WIMOB.2010.5645048>
- Hou, Y., Garikapati, V., Weigl, D., Henao, A., Moniot, M., & Sperling, J. (2020). Factors Influencing Willingness to Pool in Ride-Hailing Trips. *Transportation Research Record*, 2674(5), 419-429. <https://doi.org/10.1177/0361198120915886>
- Idé, T., & Kato, S. (2009). Travel-time prediction using gaussian process regression: A trajectory-based approach. *Society for Industrial and Applied Mathematics - 9th SIAM International Conference on Data Mining 2009, Proceedings in Applied Mathematics*, 3, 1177-1188. <https://doi.org/10.1137/1.9781611972795.101>
- Kriegel, H. P., Renz, M., Schubert, M., & Zuefle, A. (2008). Statistical density prediction in traffic networks. *Society for Industrial and Applied Mathematics - 8th SIAM International Conference on Data Mining 2008, Proceedings in Applied Mathematics* 130, 2, 692-703. <https://doi.org/10.1137/1.9781611972788.63>
- Li, X., Pan, G., Qi, G., & Li, S. (2011). Predicting Urban Human Mobility Using Large-Scale Taxi Traces Prediction of Human Mobility for Hotspots.
- Litman, T. (2015). www.vtppi.org Evaluating Complete Streets. *Victoria Transport Policy Institute*(September).

- Liu, C., Sun, J., Jin, H., Ai, M., Li, Q., Zhang, C., Sheng, K., Wu, G., Qie, X., & Wang, X. (2020). Spatio-Temporal Hierarchical Adaptive Dispatching for Ridesharing Systems. *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, 227-238. <https://doi.org/10.1145/3397536.3422212>
- Nogueira, B., Pinheiro, R. G. S., & Subramanian, A. (2018). A hybrid iterated local search heuristic for the maximum weight independent set problem. *Optimization Letters*, 12(3), 567-583. <https://doi.org/10.1007/s11590-017-1128-7>
- Penedo, Z. E. (2024). 5 Interesting São João Festival Traditions in Portugal. <https://www.portugal.com/history-and-culture/5-interesting-traditions-of-sao-joao-in-portugal/>
- Salamanis, A., Margaritis, G., Kehagias, D. D., Matzoulas, G., & Tzovaras, D. (2017). Identifying patterns under both normal and abnormal traffic conditions for short-term traffic prediction. *Transportation Research Procedia*, 22, 665-674. <https://doi.org/10.1016/j.trpro.2017.03.063>
- Stiglic, M., Agatz, N., & Savelsbergh, M. (2015). The Benefits of Meeting Points in Ridesharing Systems. *Elsevier*, 1-42.
- Tu, M., Li, Y., Li, W., Tu, M., Orfila, O., & Gruyer, D. (2019). Improving ridesplitting services using optimization procedures on a shareability network: A case study of Chengdu. *Technological Forecasting and Social Change*, 149(March). <https://doi.org/10.1016/j.techfore.2019.119733>
- Wang, D., Fan, J., Xiao, Z., Jiang, H., Chen, H., Zeng, F., & Li, K. (2019). Stop-and-Wait: Discover Aggregation Effect Based on Private Car Trajectory Data. *IEEE Transactions on Intelligent Transportation Systems*, 20(10), 3623-3633. <https://doi.org/10.1109/TITS.2018.2878253>
- Yeh, C.-C. M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H. A., Silva, D. F., Mueen, A., & Keogh, E. (2017). Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets. 1317-1322. <https://doi.org/10.1109/icdm.2016.0179>
- Yoon, G., Chow, J. Y. J., & Rath, S. (2021). A simulation sandbox to compare fixed-route , flexible-route transit , and on-demand microtransit system designs. *arxiv(arXiv:2109.14138)*, 1-27.
- Zhang, C., Zhou, T., Wen, Q., & Sun, L. (2022). TFAD: A Decomposition Time Series Anomaly Detection Architecture with Time-Frequency Analysis. *International Conference on Information and Knowledge Management, Proceedings*, 2497-2507. <https://doi.org/10.1145/3511808.3557470>
- Zhang, Z., Qi, G., Ceder, A., Guan, W., Guo, R., & Wei, Z. (2021). Grid-Based Anomaly Detection of Freight Vehicle Trajectory considering Local Temporal Window. *Journal of Advanced Transportation*, 2021. <https://doi.org/10.1155/2021/8103333>
- Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, 1, 57-81. <https://doi.org/10.1016/j.aiopen.2021.01.001>

5.7 Appendices

Table 5.3: Results of Experiment 1

Model	MinPL (m)	MPF	Path ID	TDT (m)	Length (m)	Min Freq	Nodes	Edges
1	1000	400	HP0*	3056116*	3198	413	['n278', 'n1649', 'n1648', 'n680', ...]	[('n278', 'n1649'), ('n1649', 'n1648'), ...]
1	1000	400	HP1	2967668	3484	408	['n184', 'n252', 'n2442', 'n2026', ...]	[('n184', 'n252'), ('n252', 'n2442'), ...]
1	1000	400	HP2	1318089	2471	435	['n272', 'n68', 'n1685', 'n2318', ...]	[('n272', 'n271'), ('n68', 'n1685'), ...]
1	1000	400	HP3	728218	1292	410	['n1098', 'n1101', 'n305', 'n60', ...]	[('n1098', 'n1101'), ('n1101', 'n305'), ...]
1	1000	400	HP4	640751	1141	416	['n134', 'n133', 'n441', 'n2760', ...]	[('n134', 'n133'), ('n133', 'n441'), ...]
1	1000	400	HP5	640745	1519	404	['n138', 'n839', 'n92', 'n93', ...]	[('n138', 'n349'), ('n138', 'n93'), ...]
2	1000	800	HP6*	3192158*	2607	824	['n10', 'n1654', 'n11', 'n1637', ...]	[('n10', 'n11'), ('n10', 'n1637'), ...]
2	1000	800	HP7	1888958	2050	860	['n2785', 'n205', 'n211', 'n512', ...]	[('n2785', 'n211'), ('n2785', 'n2786'), ...]
2	1000	800	HP8	1357666	1151	881	['n671', 'n675', 'n678', 'n673', ...]	[('n671', 'n675'), ('n675', 'n678'), ...]
3	1000	1200	HP9*	1979531*	1435	1203	['n10', 'n324',	[('n10', 'n11'),

							'n11', 'n1637', ...]	('n10', 'n1637'), ...]
4	1000	1600	NA	NA	NA	NA	NA	NA
5	1500	400	HP0*	3056116*	3198	413	['n278', 'n1649', 'n1648', 'n680', ...]	[('n278', 'n1649'), ('n1649', 'n1648'), ...]
5	1500	400	HP1	2967668	3484	408	['n184', 'n252', 'n2442', 'n2026', ...]	[('n184', 'n252'), ('n252', 'n2442'), ...]
5	1500	400	HP2	1318089	2471	435	['n272', 'n68', 'n1685', 'n2318', ...]	[('n272', 'n271'), ('n68', 'n1685'), ...]
5	1500	400	HP5	640745	1519	404	['n138', 'n839', 'n92', 'n93', ...]	[('n138', 'n349'), ('n138', 'n93'), ...]
6	1500	800	HP6*	3192158*	2607	824	['n10', 'n1654', 'n11', 'n1637', ...]	[('n10', 'n11'), ('n10', 'n1637'), ...]
6	1500	800	HP7	1888958	2050	860	['n2785', 'n205', 'n211', 'n512', ...]	[('n2785', 'n211'), ('n2785', 'n2786'), ...]
7	1500	1200	NA	NA	NA	NA	NA	NA
8	1500	1600	NA	NA	NA	NA	NA	NA
9	2000	400	HP0*	3056116*	3198	413	['n278', 'n1649', 'n1648', 'n680', ...]	[('n278', 'n1649'), ('n1649', 'n1648'), ...]
9	2000	400	HP1	2967668	3484	408	['n184', 'n252', 'n2442', 'n2026', ...]	[('n184', 'n252'), ('n252', 'n2442'), ...]
9	2000	400	HP2	1318089	2471	435	['n272', 'n68', 'n1685', 'n2318', ...]	[('n272', 'n271'), ('n68', 'n1685'), ...]
10	2000	800	HP6*	3192158*	2607	824	['n10', 'n1654',	[('n10', 'n11'),

							'n11', 'n1637', ...]	('n10', 'n1637'), ...]
10	2000	800	HP7	1888958	2050	860	['n2785', 'n205', 'n211', 'n512', ...]	[(<i>'n2785', 'n211'</i>), (<i>'n2785', 'n2786'</i>), ...]
11	2000	1200	NA	NA	NA	NA	NA	NA
12	2000	1600	NA	NA	NA	NA	NA	NA

Bold: TDT \geq average, *Italics:* TDT \geq standard deviation, *: MaxTDT

Table 5.4: Results of Experiment 2

Model	MinPL (m)	MPF	Path ID	TDT (m)	Length (m)	Min Freq	Nodes	Edges
1	1000	400	HP0*	3056116*	3198	413	['n278', 'n1649', 'n1648', 'n680', ...]	[('n278', 'n1649'), ('n1649', 'n1648'), ...]
1	1000	400	HP1	2967668	3484	408	['n184', 'n252', 'n2442', 'n2026', ...]	[('n184', 'n252'), ('n252', 'n2442'), ...]
1	1000	400	HP2	1318089	2471	435	['n272', 'n68', 'n1685', 'n2318', ...]	[('n272', 'n271'), ('n68', 'n1685'), ...]
1	1000	400	HP3	728218	1292	410	['n1098', 'n1101', 'n305', 'n60', ...]	[('n1098', 'n1101'), ('n1101', 'n305'), ...]
1	1000	400	HP4	640751	1141	416	['n134', 'n133', 'n441', 'n2760', ...]	[('n134', 'n133'), ('n133', 'n441'), ...]
1	1000	400	HP5	640745	1519	404	['n138', 'n839', 'n92', 'n93', ...]	[('n138', 'n349'), ('n138', 'n93'), ...]
2	1000	600	HP6*	3425352*	3161	606	['n287', 'n2828', 'n291', 'n1112', ...]	[('n287', 'n2828'), ('n2828', 'n291'), ...]
2	1000	600	HP7	786386	1149	634	['n380', 'n856', 'n1653', 'n224', ...]	[('n380', 'n853'), ('n380', 'n379'), ...]
2	1000	600	HP8	615060	1014	604	['n68', 'n1685', 'n2318', ...]	[('n68', 'n1685'), ('n1685', 'n2318'), ...]
3	1000	800	HP9*	3192158*	2607	824	['n10', 'n1654', 'n11', 'n1637', ...]	[('n10', 'n11'), ('n10', 'n1637'), ...]
3	1000	800	HP10	1888958	2050	860	['n2785', 'n205', 'n211', 'n512', ...]	[('n2785', 'n211'), ('n2785', 'n2786'), ...]

3	1000	800	HP11	1357666	1151	881	['n671', 'n675', 'n678', 'n673', ...]	[('n671', 'n675'), ('n675', 'n678'), ...]
4	1000	1000	HP12 *	2185669*	1640	1005	['n10', 'n324', 'n11', 'n1637', ...]	[('n10', 'n11'), ('n10', 'n1637'), ...]
5	1000	1200	HP13 *	1979531*	1435	1203	['n10', 'n324', 'n11', 'n1637', ...]	[('n10', 'n11'), ('n10', 'n1637'), ...]
6	1000	1400	NA	NA	NA	NA	NA	NA
7	1500	400	HP0*	3056116*	3198	413	['n278', 'n1649', 'n1648', 'n680', ...]	[('n278', 'n1649'), ('n1649', 'n1648'), ...]
7	1500	400	HP1	2967668	3484	408	['n184', 'n252', 'n2442', 'n2026', ...]	[('n184', 'n252'), ('n252', 'n2442'), ...]
7	1500	400	HP2	1318089	2471	435	['n272', 'n68', 'n1685', 'n2318', ...]	[('n272', 'n271'), ('n68', 'n1685'), ...]
7	1500	400	HP5	640745	1519	404	['n138', 'n839', 'n92', 'n93', ...]	[('n138', 'n349'), ('n138', 'n93'), ...]
8	1500	600	HP6*	3425352*	3161	606	['n287', 'n2828', 'n291', 'n1112', ...]	[('n287', 'n2828'), ('n2828', 'n291'), ...]
9	1500	800	HP9*	3192158*	2607	824	['n10', 'n1654', 'n11', 'n1637', ...]	[('n10', 'n11'), ('n10', 'n1637'), ...]
9	1500	800	HP10 *	1888958*	2050	860	['n2785', 'n205', 'n211', 'n512', ...]	[('n2785', 'n211'), ('n2785', 'n2786'), ...]
10	1500	1000	HP12	2185669	1640	1005	['n10', 'n324', 'n11', 'n1637', ...]	[('n10', 'n11'), ('n10', 'n1637'), ...]
11	1500	1200	NA	NA	NA	NA	NA	NA
12	1500	1400	NA	NA	NA	NA	NA	NA

13	2000	400	HP0*	3056116*	3198	413	[<i>'n278', 'n1649', 'n1648', 'n680', ...</i>]	[<i>('n278', 'n1649'), ('n1649', 'n1648'), ...</i>]
13	2000	400	HP1	2967668	3484	408	[<i>'n184', 'n252', 'n2442', 'n2026', ...</i>]	[<i>('n184', 'n252'), ('n252', 'n2442'), ...</i>]
13	2000	400	HP2	1318089	2471	435	[<i>'n272', 'n68', 'n1685', 'n2318', ...</i>]	[<i>('n272', 'n271'), ('n68', 'n1685'), ...</i>]
14	2000	600	HP6*	3425352*	3161	606	[<i>'n287', 'n2828', 'n291', 'n1112', ...</i>]	[<i>('n287', 'n2828'), ('n2828', 'n291'), ...</i>]
15	2000	800	HP9*	3192158*	2607	824	[<i>'n10', 'n1654', 'n11', 'n1637', ...</i>]	[<i>('n10', 'n11'), ('n10', 'n1637'), ...</i>]
15	2000	800	HP10	1888958	2050	860	[<i>'n2785', 'n205', 'n211', 'n512', ...</i>]	[<i>('n2785', 'n211'), ('n2785', 'n2786'), ...</i>]
16	2000	1000	NA	NA	NA	NA	NA	NA
17	2000	1200	NA	NA	NA	NA	NA	NA
18	2000	1400	NA	NA	NA	NA	NA	NA

Bold: TDT \geq average, *Italics:* TDT \geq standard deviation, *****: MaxTDT

CHAPTER 6: CONCLUSION

This dissertation embarks on an explorative journey into the phenomenon of collective anomalies, particularly focusing on their manifestation and detection within data-driven systems. It addresses the subtle, context-dependent nature of these anomalies as they emerge and are identified within datasets. The dissertation is structured into three interconnected papers, each delving into different aspects of collective anomalies. The first paper examines the manifestation phase in Online Customer Reviews (OCR), analyzing shifts in review patterns due to review solicitation. The second paper shifts focus to detecting anomalies in trajectory networks, contributing to urban planning and traffic management. The third paper further explores detection through a spatio-temporal collective anomaly discovery framework, enhancing urban mobility by identifying microtransit bottlenecks. Each chapter builds upon the previous, cumulatively providing a comprehensive understanding of collective anomalies, from their subtle emergence in consumer feedback systems to their more evident presence in urban mobility networks.

The findings from the three studies within this dissertation collectively offer a holistic view of collective anomaly manifestation and detection phases. The first study makes

significant strides in understanding reporting bias in Online OCR. It shifts the focus from the traditionally studied vocal minority to the often-neglected silent majority, revealing a broader spectrum of OCR biases. The study uncovers that solicited reviews, primarily coming from the silent majority, generally show less diversity in opinion, more polarization, increased negativity, and shallower content compared to organic reviews. This finding indicates that solicited reviews can significantly alter the overall portrayal of consumer opinions. The study introduces and empirically validates the “Experience Sphere” framework, which integrates various theories like herding behavior and spiral of silence, offering a comprehensive perspective on OCR dynamics. This approach not only strengthens the theoretical understanding of OCR but also provides practical insights. It highlights the need for businesses to refine their review solicitation strategies to elicit more genuine feedback and to be mindful of potential biases in solicited reviews. The study also reveals the tendency of solicited reviews in service industries to contain more negative content, a crucial insight for businesses aiming to address customer dissatisfaction effectively. Additionally, the richer, more actionable feedback from organic reviews is underscored, emphasizing its value for comprehensive customer insights. Overall, the research suggests that businesses should adapt their review aggregation and analysis strategies to accurately interpret customer feedback, thereby making informed decisions. The study’s nuanced exploration of the silent majority’s feedback patterns and the impact of review solicitation offers a critical perspective on early anomaly indicators, underscoring the importance of recognizing these signs for effective anomaly detection.

The second study shifts the focus towards the detection phase of collective anomalies. This chapter introduces a pioneering approach to analyzing trajectory networks, focusing on the AHP–TDT problem. This approach emphasizes the significance of cumulative distances in identifying irregular movement patterns, offering a fresh perspective in spatial data analysis. The study presents a comprehensive methodology for discovering anomalous hotspot paths, significantly advancing beyond traditional hotspot detection methods. This methodology, integrating various analytical steps, marks a major leap in understanding the intricacies of movement dynamics within trajectory networks. The empirical application of this framework, using data from an on-demand transportation agency in Porto, Portugal, demonstrates its practical applicability and underscores its effectiveness in revealing complex urban movement patterns. The research enriches spatial data analysis by shifting focus from points and edges to entire paths in urban mobility networks, providing a more holistic view of urban dynamics. This broader perspective refines frameworks in spatial analysis, network analysis, and graph theory, introducing advanced approaches like the weighted connected component for a sophisticated interpretation of urban trajectory networks. Practically, this research has profound implications for urban planning and traffic management, providing crucial insights for road network optimization and congestion reduction. It supports public safety and emergency response by enabling the identification of high-incident areas. Additionally, the findings contribute to smart city initiatives and the optimization of on-demand mobility services. The research's interdisciplinary implications extend its relevance to urban planning, environmental studies, and social network analysis, highlighting its broad applicability and importance in shaping data-driven urban environments.

The third study takes a deeper dive into the detection phase of collective anomalies, introducing the spatio-temporal collective anomaly discovery (STCAD) framework. This approach significantly enhances urban mobility analysis by integrating spatial insights from AHP-TDT with temporal nuances, offering a comprehensive tool for spatio-temporal anomaly discovery. The STCAD framework's practical application, through empirical testing with real-world trajectory data, demonstrates its effectiveness in identifying and analyzing collective anomalies in urban mobility. The STCAD framework represents a theoretical milestone in spatial data analysis, uniquely integrating spatial and temporal dimensions to examine collective anomalies. This holistic approach unlocks deeper insights into systemic patterns in trajectory networks, influencing new data mining and machine learning techniques, particularly in anomaly detection. Practically, STCAD provides essential strategies for enhancing urban mobility and transit operations, aiding traffic management and optimizing urban infrastructures. Crucial for urban planners and smart city developments, this framework drives intelligent urban management and boosts the efficacy of on-demand mobility services, highlighting its pivotal role in advancing data-driven urban landscapes. The study demonstrates how temporal nuances, coupled with spatial insights from the second study, can enhance our ability to detect and understand collective anomalies in urban mobility, offering a more complete picture of these complex phenomena.

Together, the studies collectively reveal the complex and dynamic nature of collective anomalies, highlighting their varied implications across different contexts. The initial study delves into the subtle manifestation of reporting biases in OCR, shedding light on the dual aspects of these biases as sources of concern for business decision-making and as

opportunities for gaining richer insights into consumer behavior. The latter studies shift focus to the detection of anomalies in urban mobility networks, uncovering critical issues such as traffic bottlenecks, while also unveiling opportunities for advanced urban planning and smart city initiatives. These investigations into spatial and spatio-temporal trajectories demonstrate the potential of collective anomalies to inform and enhance urban mobility solutions. Collectively, these investigations offer a holistic perspective on collective anomalies, emphasizing their role in fostering innovative enhancement applications across diverse data-driven domains.

BIOGRAPHICAL SKETCH OF THE AUTHOR

Mohammad K. Bakhsh

Before joining the Information Systems for Data Science and Management Ph.D. program at the University of Massachusetts Boston, Mohammad K. Bakhsh earned a Bachelor of Science degree in Marine Operations from Northumbria University and an MBA with a focus on Decision Sciences from San Francisco State University. His professional career began when he was selected by Saudi Aramco right after high school to join its High-Talent Program, where he played a key role in the company's marine fleet of oil tankers. He also took on leadership roles, serving as the head of the management department and lecturer at the University of Prince Mugrin, and leading product development at a startup company. After completing his doctoral program, he will join a consulting firm as a Data Science Consultant. Mohammad's research focuses on uncovering anomalies, particularly within text, networks, and time series data.