New York University

Graduate School of Arts and Science

Computational Frameworks for Analyzing Large-Scale Urban Mobility Patterns

by

Marcus J. Chen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Department of Computer Science

May 2024

Abstract

The accelerating pace of urbanization has placed extraordinary demands on metropolitan transportation infrastructure, yet the computational tools available to urban planners remain fragmented and limited in scale. This dissertation introduces a unified computational framework for analyzing large-scale urban mobility patterns by integrating heterogeneous data streams — including transit card records, road sensor networks, and mobile location signals — into a coherent analytical pipeline. The framework leverages graph-theoretic representations of multimodal transit networks, Bayesian inference for trip-purpose imputation, and distributed optimization techniques to process city-scale datasets with sub-hour latency.

The proposed methodology is validated through an extensive case study conducted in partnership with the Meridian Urban Data Initiative, encompassing over 2.1 billion transit records collected across sixteen metropolitan regions. Results demonstrate a 34% improvement in passenger flow prediction accuracy relative to prevailing gravity-model baselines, while reducing computational cost by an order of magnitude through principled subsampling and incremental recomputation strategies. The framework also introduces a novel congestion cascading model that identifies vulnerable nodes in a transit graph up to forty-five minutes before service degradation becomes apparent to riders.

Beyond methodological contributions, this work provides open-source reference implementations of all core algorithms as part of the Nimbus Transit Simulation Platform, a toolkit designed to lower barriers to entry for municipal transportation agencies. The dissertation concludes with a discussion of policy implications, ethical considerations surrounding mobility data privacy, and directions for extending the framework to incorporate real-time adaptive routing and equitable access optimization.

Acknowledgments

I am deeply grateful to my advisor, Dr. Elena Vasquez, whose intellectual generosity and unwavering belief in this project carried it from a nascent idea to a completed dissertation. Her insistence on rigor balanced with pragmatism shaped not only this work but my entire approach to research.

I thank the members of my dissertation committee — Dr. James Whitfield, Dr. Priya Nair, and Dr. Robert Okonkwo — for their incisive feedback and patience through numerous revisions. The faculty and staff of the Department of Computer Science at New York University provided an environment where interdisciplinary inquiry could flourish.

The Meridian Urban Data Initiative generously provided the datasets that anchor the empirical contributions of this dissertation. I am particularly indebted to the engineering team at Nexa Transit Analytics for their collaboration on the Nimbus platform. This research was supported in part by a Graduate Research Fellowship from the Apex Foundation for Computational Science.

To my family — my parents, Li and David Chen, and my sister, Naomi — thank you for a lifetime of encouragement and for never asking when I would finish. And to Sofia Reyes, for everything: this exists because you believed it could.

Table of Contents

Contents

Chapter 1

Introduction

1.1 Motivation and Context

The twenty-first century has witnessed an unprecedented concentration of human populations within urban centers. By current estimates, metropolitan regions now house over fifty-six percent of the global population and generate more than eighty percent of global economic output [1]. This demographic transformation places extraordinary strain on transportation infrastructure originally designed for far smaller populations and fundamentally different mobility patterns.

Traditional approaches to transportation planning rely heavily on periodic household travel surveys, manual traffic counts, and static four-step models that assume equilibrium conditions rarely observed in practice. While these methods served adequately during an era of slower urban growth, they are increasingly mismatched to the demands of cities where mobility patterns shift on timescales of hours rather than decades. The emergence of ubiquitous sensing infrastructure — from automated fare collection systems to inductive loop detectors to the inertial measurement units embedded in consumer mobile devices — presents both an opportunity and a challenge: an opportunity to observe urban mobility at unprecedented resolution, and a challenge to develop computational methods capable of extracting actionable insight from petabytes of heterogeneous, noisy, and privacy-sensitive data.

1.2 Problem Statement

This dissertation addresses a central question in computational urban science: how can we construct, calibrate, and continuously update high-fidelity models of metropolitan-scale human mobility that are simultaneously accurate enough to inform policy decisions, efficient enough to run on commodity hardware available to municipal agencies, and principled enough to provide guarantees about privacy preservation and uncertainty quantification?

Existing approaches fall short along at least one of these dimensions. Gravity models and their radiation-model descendants [2] offer computational simplicity but systematically mispredict flows along low-probability routes that nonetheless account for significant aggregate congestion. Agent-based microsimulation platforms achieve high fidelity but require calibration to dozens of latent parameters and scale poorly beyond midsize metropolitan areas [3].

Deep learning methods applied to mobility prediction have demonstrated impressive accuracy on benchmark datasets but typically function as black boxes, offering no interpretable mechanism by which planners can assess counterfactual scenarios such as the introduction of a new transit line or a congestion pricing zone.

1.3 Contributions

This dissertation makes four principal contributions to the field of computational urban mobility analysis.

First, it introduces the *Hierarchical Multimodal Transit Graph* (HMTG), a unified representation that encodes road, rail, bus, pedestrian, and micromobility networks within a single layered graph structure. The HMTG formulation enables existing graph algorithms for shortest paths, centrality, and community detection to operate seamlessly across modalities without requiring ad hoc impedance functions or mode-choice nesting structures.

Second, it develops a *Bayesian Trip-Purpose Imputation* (BTPI) procedure that assigns latent purposes — commute, errand, leisure, social, and medical — to individual transit transactions using a hierarchical Dirichlet process prior over origin-destination-time triplets. The BTPI model is trained on a small annotated subset and generalizes to millions of unlabeled records with calibrated uncertainty estimates for each imputation.

Third, it presents a *Distributed Incremental Flow Estimation* (DIFE) algorithm that partitions a metropolitan transit graph into loosely coupled subgraphs, computes passenger flows within each subgraph using parallel convex optimization, and reconciles boundary flows through an alternating direction method of multipliers (ADMM) scheme. DIFE achieves a 34% reduction in root-mean-square error relative to the industry-standard gravity model while reducing wall-clock computation time from approximately fourteen hours to under forty-five minutes on a sixteen-node commodity cluster for a city the size of the Meridian metropolitan area.

Fourth, it proposes a *Congestion Cascading Model* (CCM) that treats transit network congestion as a percolation process on the HMTG. The CCM identifies nodes whose failure probability exceeds a threshold under current load conditions and predicts cascading degradation paths up to forty-five minutes before service quality impacts become measurable to passengers. This predictive capability enables proactive intervention — rerouting buses, adjusting signal timing, or issuing real-time rider alerts — rather than reactive response to already-materialized congestion.

1.4 Validation and Case Study

All components of the proposed framework were validated using data from the Meridian Urban Data Initiative, a multi-agency consortium that curates and anonymizes transit records from sixteen metropolitan regions in North America and Europe. The primary case study focuses on the greater Meridian metropolitan area (population approximately 4.8 million), for which the dataset includes 2.1 billion individual transit transactions spanning a thirty-six-month period from January 2021 through December 2023.

Ground-truth passenger flows were established through a combination of manual onboard surveys conducted by the Meridian Regional Transit Authority and automated passenger counting systems deployed on a stratified sample of bus and rail routes. The BTPI procedure was validated against a held-out set of 47,000 trips manually labeled by trained annotators, achieving a weighted F1 score of 0.82 across the five trip-purpose categories.

1.5 Thesis Organization

The remainder of this dissertation is organized as follows. Chapter 2 surveys related work in urban mobility modeling, graph-based transportation analysis, and Bayesian inference for spatiotemporal data. Chapter 3 formally defines the Hierarchical Multimodal Transit Graph and derives its mathematical properties. Chapter 4 presents the Bayesian Trip-Purpose Imputation procedure, including the hierarchical Dirichlet process formulation and variational inference algorithm. Chapter 5 describes the Distributed Incremental Flow Estimation algorithm with convergence guarantees and empirical benchmarks. Chapter 6 introduces the Congestion Cascading Model and evaluates its predictive performance on held-out congestion events. Chapter 7 discusses policy implications, privacy considerations, and limitations of the current framework. Chapter 8 concludes with a summary of contributions and directions for future research.

Bibliography

[1] L. Chen and M. Okazaki, *Urban Mobility in the Twenty-First Century: Data, Models, and Policy*. Meridian University Press, 2023.

[2] F. Simini, M.C̃. González, A. Maritan, and A.-L. Barabási, "A universal model for mobility and migration patterns," *Nature*, vol. 484, pp. 96–100, 2012.

[3] M. Balmer, K. Meister, R.Ã. Waraich, M. Rieser, and K. Nagel, "Agent-based simulation of travel demand: Structure and computational performance of MATSim-T," presented at the Innovations in Travel Modeling Conference, Austin, TX, 2007.

[4] E. Vasquez and P. Thornton, "Graph convolutional networks for multimodal transit assignment," *Journal of Transport Geography*, vol. 93, art. 103076, 2021.

[5] P. Nair, J. Whitfield, and S. Okada, "Hierarchical Bayesian models for latent trip-purpose inference from smart-card data," *Transportation Research Part C*, vol. 138, art. 103624, 2022.

[6] Meridian Urban Data Initiative, *Annual Data Report: Transit Flows and Network Topology, 2021–2022*. Meridian Regional Transit Authority, Technical Report MUI-TR-2022-04, 2022.