Unveiling Roadway Network Safety: Application of Statistical Physics to Crowdsourced Velocity Data

By

Meshkat Botshekan

Master of Science, University of Massachusetts, Dartmouth (2019)
Bachelor of Science, Isfahan University of Technology (2016)

Submitted to the Department of Civil and Environmental Engineering in partial fulfillment of the requirements for the degree(s) of
Doctor of Philosophy in Civil and Environmental Engineering
at the

Massachusetts Institute of Technology

February 2024

Authored by: Meshkat Botshekan
Civil and Environmental Engineering
Sep. 12, 2023

Certified by: Franz-Josef Ulm
Professor of Civil and Environmental Engineering
Faulty Director, Concrete Sustainability Hub
Thesis supervisor

Accepted by: Heidi Nepf
Professor of Civil and Environmental Engineering
Chair, Graduate Program Committee

Unveiling Roadway Network Safety: Application of Statistical Physics to Crowdsourced Velocity Data

by

Meshkat Botshekan

Submitted to the Department of Civil and Environmental Engineering
on Sep 12, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Civil and Environmental Engineering

Abstract

In an increasingly mobile world, traffic safety poses stark realities. In 2022, roadway incidents in the U.S. claimed over 38,824 lives, surpassing the fatality rate of the COVID-19 pandemic within the country. Despite these alarming statistics, understanding the overarching patterns of traffic safety presents a complex challenge due to myriad influencing factors such as driver behavior, roadway network geometry, weather conditions, and vehicle design.

This study revisits traffic safety from the perspective of statistical physics, positing universal temporal and memory effects to delve into the internal structure of traffic by exploring higher-order statistics. The examination of the internal structure enables the uncovering of near-miss incident risks in congested traffic flow—risks positively correlated with collision risks derived from historic accident records. By integrating the complex dynamics of traffic flow, the near-miss risk is ascertained from the crowdsourced velocity measurements of vehicles, thereby offering a computationally efficient framework with potential for real-time implementation.

We apply this framework to extensive velocity datasets collected anonymously across multiple states in the U.S., enabling the derivation of the spatial distribution of expected near-miss risk on a large scale. Moreover, we assess and compare the reliability and robustness of these networks, merging graph theory with our physics-inspired near-miss risk approach. Our findings consistently reveal patterns across different states, facilitating the identification of the most and least reliable/robust networks. This framework lays the foundation for a real-time, proactive maintenance of roadway networks, a major stride towards creating a safer transportation infrastructure.

Thesis Supervisor: Franz-Josef Ulm
Title: Professor of Civil and Environmental Engineering
Faculty Director, Concrete Sustainability Hub

Acknowledgments

As I write this acknowledgment section, one of the last pages of my PhD journey, I'm overwhelmed with a bittersweet feeling. Though this section stands apart from the scientific nature of my other writings, it bears a familiar challenge: articulating my thoughts.

Embarking upon my PhD has been a roller coaster of experiences and emotions. A defining chapter of this adventure was tragically penned with the outbreak of the COVID-19 pandemic, which led to a prolonged national lockdown that began in March 2020. We were thrust into a new reality of online classes, quarantine, and separation from family, each of which presented unique challenges. Despite these obstacles, my advisor, role model, and friend, Professor Franz-Josef Ulm, made the journey more than worthwhile. His boundless scientific curiosity, critical thinking, and stimulating conversations not only fueled my motivation but also opened new scientific horizons after every meeting. I hold dear the conversations we had during our walks in the Killian Court, often accompanied by coffee. I am profoundly grateful to have been under your guidance, Franz. Your resolute support, inspiration, and encouragement have indeed been instrumental throughout this journey.

I would also like to express my gratitude to my committee members, Professors Ali Jadbabaie and Patrick Jaillet, for taking the time to meet with me and providing invaluable insights through their comments.

As I reflect on my years as a graduate student, the strength I gained from the constant presence of my parents and sister defies all words and cannot be adequately expressed. Being thousands of miles apart from family was undeniably challenging, yet every video call and every message bridged the distances and made me feel as though you were right beside me. During moments when my research seemed to falter or when spirits were low, your faith in me, your unconditional love, and your unwavering support rekindled my drive. From the bottom of my heart, I thank you. It's because of you that I have been able to achieve this milestone.

Contents

List of Figures

Chapter 1

Introduction

1.1 Research Context

In today's globalized world, road transportation has become a vital component of our societal infrastructure, generating numerous job opportunities and serving as the main mode of transport to reach destinations, access resources, and engage with markets [7–11]. The burgeoning of economies, coupled with rapid urbanization and increasing demands for mobility, has led to a significant expansion in the global road transportation network [12]. However, this widespread growth has brought with it an escalating concern of paramount importance - traffic safety. With its vast societal, and economic implications, traffic safety forms a substantial challenge in our quest for a safe and efficient transport system.

1.1.1 Societal Impact

From a societal perspective, the ramifications of traffic safety manifest themselves in diverse ways that are profound, pervasive, and underscore the urgency of this issue. The World Health Organization, a leading authority in global health issues, provides a stark snapshot of the situation [13]. As per their data, road traffic injuries have ascended the list to become the eighth leading cause of death globally [Fig. 1-1], and, shockingly, they represent the primary cause of mortality for individuals aged

between 5 and 29 years. This age group represents a significant portion of society's active participants, highlighting the tragic loss of potential contributions to societal progress. Furthermore, approximately 1.3 million people find their lives prematurely cut short due to car accidents every year across the globe, translating to an average of 3,287 deaths per day [2, 14]. When viewed through the lens of comparative analysis, these figures are not far removed from the cumulative Covid-19 deaths reported at the peak of the pandemic in 2020 [15], highlighting the extent of the human toll brought about by traffic accidents.

Shifting focus to a more localized setting, the United States presents its own set of alarming statistics. Approximately, 4.1% of the 300 million registered vehicles find themselves involved in one of the 5 million police-reported car accidents annually [16]. This leads to a fatality rate of 1.2×10^{-4} per registered vehicle [17], or, when taking into account the overall mileage, the rate of 1.3×10^{-6} accidents per vehicle miles driven. At first glance, 1.2×10^{-4} may not elicit alarm. However, when this figure is projected onto the canvas of the vast vehicular landscape and the extensive roadway network in the United States, the implications rapidly scale up to concerning levels of significance. These startling statistics serve as a somber backdrop to the urgent need for effective strategies to enhance traffic safety, a need that, given the scale of the problem, translates into a significant opportunity for impact on both individual lives and broader societal health.

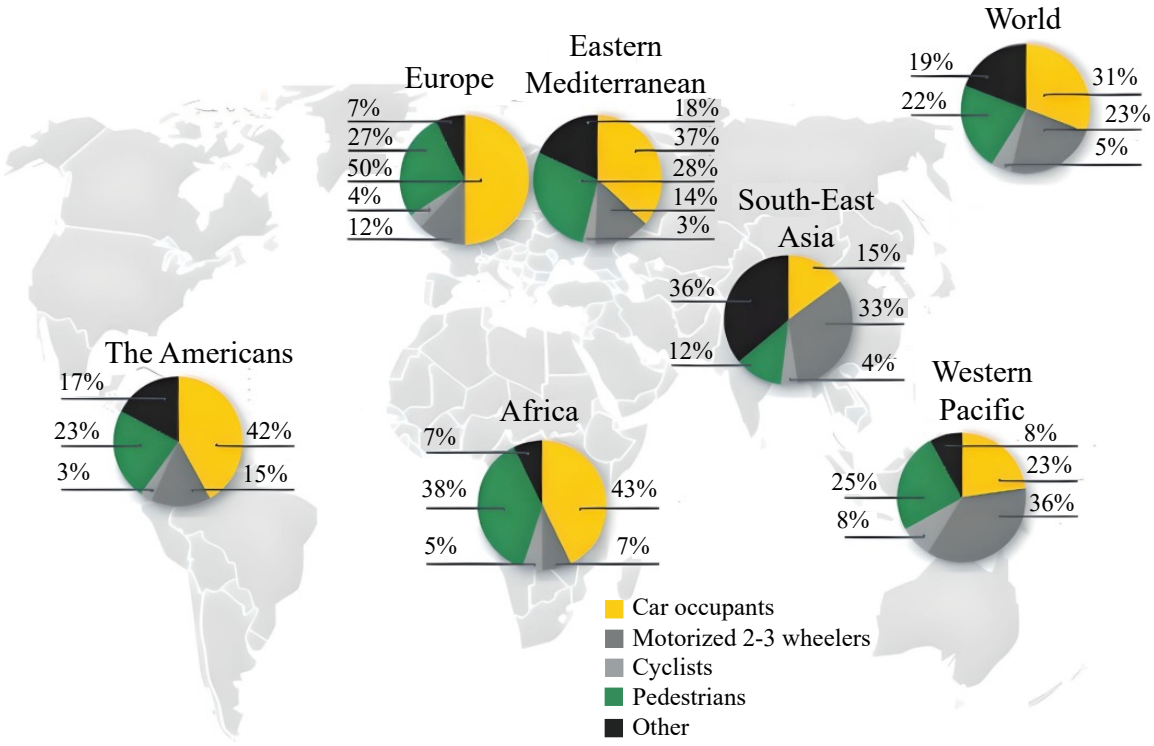

Figure 1-1: Road fatalities across the globe. In various continents, motorized accidents account for over half of the road fatalities, with car collisions and crashes involving motorized two- to three-wheelers being the primary causes. Figure is adapted from Ref. [1].

1.1.2 Economic Impact

The significance of the economic implications associated with traffic safety is as profound as it is complex, encapsulating multiple facets that intricately weave together to create a tangible impact on society at large. This impact parallels, and indeed amplifies, with the continuous growth and expansion of our globally interconnected road transportation systems [18], further heightening the exigency to understand and mitigate road safety effectively. At the heart of the matter lie traffic accidents, notorious for their capability to impose an astronomical economic burden on nations worldwide. Literature on the impact of road safety and road collisions on the economy includes studies on the effectiveness of interventions in reducing road injuries [19] and the cost-effectiveness of these interventions [20], as well as studies that attempt to quantify the economic strain of road injuries [21–24]. Broadly, the total cost associated with traffic safety can be bifurcated into direct and indirect costs. Direct economic costs of car accidents are relatively straightforward to quantify, encompassing expenses associated with healthcare provisions to treat victims of accidents, as well as the intricate web of legal proceedings that often ensue in the aftermath of such incidents. Indirect costs, on the other hand, are less tangible but equally consequential. These costs primarily stem from the lost productivity that society incurs when an individual is rendered incapacitated due to injury or tragically loses their life prematurely, thereby diminishing their potential contributions to the economic engine. Most methodologies estimate the economic toll by aggregating the direct and indirect costs of traffic accidents (cost of illness approach) or by determining the willingness of individuals to pay to avoid risks (value of statistical life approach). The annual worldwide economic cost of road accidents is estimated at about $518 billion [25], which can cost most countries an estimated 1-3% of their Gross Domestic Product (GDP) [2]. Given its extensive and heavily utilized national roadway network, the United States, along with other highly industrialized nations, experiences a significant macroeconomic impact due to poor road safety [Fig. 1-2]. The sheer weight of this economic loss throws into sharp relief the exigent necessity for instituting effective traffic safety measures.

Looking ahead, the picture becomes even more worrisome. Road transportation networks are expected to grow, driven by the demands of burgeoning industries and an increasingly mobile population. In the absence of successful interventions, this implies an escalation of the economic encumbrance, thereby emphasizing the pivotal role of traffic safety within the broad spectrum of our infrastructure.

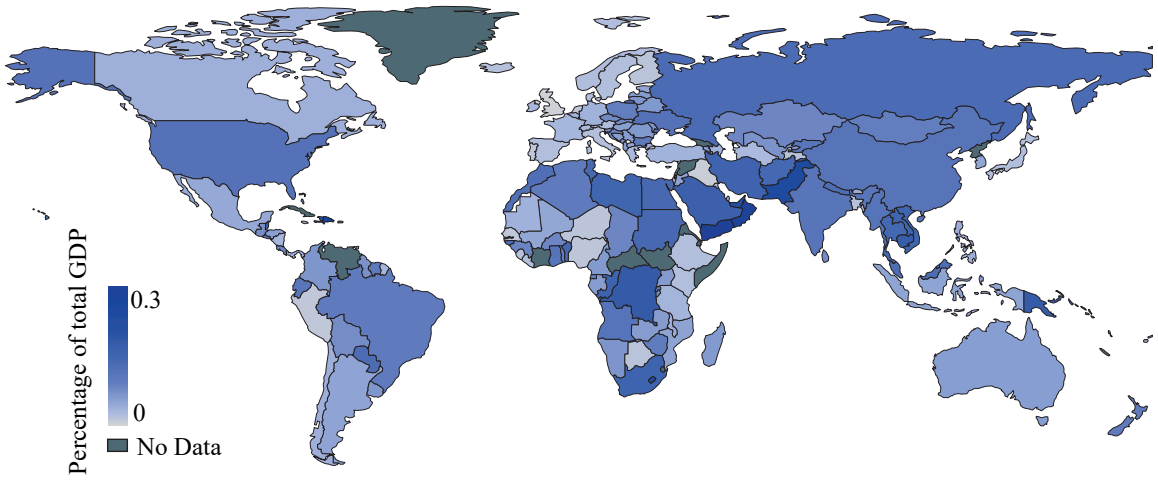

Figure 1-2: Macroeconomic impact due to road injuries as a percentage of total GDP ver the period 2015–2030. It highlights that the United States is among the nations where the economic burden of road transportation is substantial, largely owing to the extensive size of its transportation system. Figure is adapted from Ref. [2]

1.2 Research Background

The substantial impact of road transportation has motivated scientific investigation into the inherent patterns within traffic. The dynamics of traffic and the possibility to quantify risk of accident can be viewed from two distinct perspectives: *(i)* the physics of traffic flow, and *(ii)* data-driven models. While the physics of traffic involves theories that attempt to understand interactions at multiple scales, from vehicle-to-vehicle interaction to macroscopic properties, data-driven models use mathematical models and machine learning techniques to identify the key predictors of accidents using large data sets and records of accidents.

1.2.1 Physics of Traffic Flow

The physics community aims to propose models that elucidate the fundamental dynamics of traffic at multiple scale, from an individual driver to macroscopic properties. At their core, these models typically begin by treating vehicles as particles that interact with each other. This assumption gives rise to complex behavior of many-body systems, which offers insight into the intricate interactions seen at various scales in traffic, including the occurrence of rare events, shock waves, and phase transition.

Physics-based models of traffic flow often draw parallels with particle physics, providing a theoretical foundation for the interpretation of traffic dynamics as the collective behavior of individual particles. Specifically, traffic resembles a many-body system of interacting particles exhibiting characteristic internal structures at different levels of density. The most distinctive characteristic of traffic flow is the phase transition. At low density, in what is known as the *free flow* state, particles can move freely. However, as density increases, vehicles begin to interact with each other, leading to the formation of traffic jams, thereby transitioning into the *congested regime* [26–28]. This transition is akin to the gas-liquid transition, where the free flow and congested regime can be compared to gas and liquid states, respectively [29]. Traffic jams often exhibit nonlinear wave behaviors similar to solitons, triangular shocks, and kinks, which are commonly described by equations such as the Korteweg-de-Vries (KdV),

Modified KdV (MKdV), and Burgers equations [30–33]. From a physics perspective, accurately modeling the internal structure of traffic flow and reproducing its characteristics at different time and length scales require a thorough understanding of the interplay among individual vehicles, traffic density, boundary conditions, and all the other external forces that could influence not only individual driver behavior, but also the collective dynamics of the vehicles. In pursuit of this understanding, various models have been proposed subsequent to the pioneering works of Biham et al. [34], Nagel and Schreckenberg [35], and Kerner and Kohnhauser [36] aiming at explaining the wide array of physical phenomena observed in different traffic settings - from urban zones to single and multi-lane traffic, freeways, and bottlenecks. Broadly, these microscopic traffic models can be classified into one of the three categories which approach physics of traffic from different perspectives: car-following models, gas-kinetic models, and cellular automata models.

Car-following Models

The car-following model represents one of the earliest categories of traffic models that are continuous in both time and space. In these models, the behavior of a vehicle is determined by the car directly in front of it [37], the so-called "follow the leader" approach [30]. The foundational concept of these models is that a driver adjusts their speed in response to the positioning of nearby vehicles. In particular, the velocity of any given vehicle is determined by an optimal velocity function, describing the ideal speed a vehicle should maintain, taking into account its headway (the distance to the car in front) [37–39]. However, early optimal velocity models often overlooked the driver's response to the relative velocity of the vehicle ahead. As a result, these models predicted an increased occurrence of collisions with increased delay times, which are the periods required for drivers to perceive, process, and respond to changes in the preceding vehicle's movement. Further modifications have been introduced to optimal velocity models to boost their performance and more accurately replicate real-world phenomena [40, 41]. For instance, one of the most commonly-used models is the Intelligent Driver Model (IDM), which takes into account relative velocity

and reproduces realistic patterns under controlled conditions [42–44]. Generalized models have also been proposed to take into account multiple types of vehicles [45], incorporate the behavior of the vehicle behind in the optimal velocity function [46], or to include the impact of the next-nearest-neighbor interaction [47]. While these models provide a good description of the microscopic behavior of traffic, they often fail to give a proper macroscopic description of the system such as the fundamental diagram [48]. Additionally, the complexity of these models, with their extensive parameters, requires data-intensive calibration procedures [49–51].

Gas-kinetic Models

In this class of models, vehicles are considered as interacting gas particles [52, 53]. The foundation of this approach was laid by Prigogine and Herman, who applied the Boltzmann equation to traffic flow [54]. Relying on the same theoretical foundation, a generalized gas-kinetic model was later proposed to incorporate different personalities of drivers [55]. Essentially, these models are generally characterized by variables such as passing probability, desired velocity distribution function, relaxation time, and inter-particle interaction. They have been instrumental in investigating complex traffic phenomena, including the hysteretic phase transition [52], power-law platoon formation akin to aggregation phenomena [56–59], and the dynamic growth and evolution of traffic congestion [60]. Moreover, gas-kinetic models have been extended to encompass two-dimensional flow for urban and multi-lane traffic flow [61–63]. Despite their utility, gas-kinetic traffic models encounter challenges due to their simplified assumptions, computational complexity, calibration and validation difficulties, sensitivity to input parameters, and a perceived lack of realistic features.

Cellular Automata Models

Cellular automata (CA) models employ a discretized approach for traffic representation, partitioning space into *cells* and time into *time steps*. These models typically use heuristic rules for particle movements, eschewing the use of differential equations common in gas-kinetic and car-following models. The CA models, despite having

fewer parameters, successfully emulate traffic attributes at both micro and macro levels [64–67]. The Nagel-Schreckenberg (NaSch) model is a one-lane, collision-free CA traffic model with periodic boundary conditions [35,68–71], which serves as the backbone of many other refined traffic CA models. The NaSch model's simplicity belies its capability to simulate critical traffic features like backward-moving shock waves, the fundamental diagram of traffic, and the internal structure of traffic flows [72–74]. Given its computational efficiency and promising results, the NaSch model has seen numerous modifications, allowing for a greater range of applications. Extensions include adaptations for multi-lane roads [75,76], urban traffic scenarios [77,78], highways with junctions [79], at-grade roundabout crossings [80], and interactions between on-ramps and main roads [81]. These enhancements reflect the model's flexibility and potential in addressing diverse traffic conditions.

1.2.2 Data-driven Models for Traffic Safety

Given the recent advancements in data analytics and data collection methodologies, data-centric approaches have been proposed to understand the patterns in vehicular accidents [82–85]. Data-driven models are contingent upon the examination of multidimensional collision datasets. These evaluations employ statistical methods and machine learning techniques to decipher the parameters that demonstrate the strongest correlation with collision incidences. Following the identification of these critical parameters, statistical models are then formulated with the objective of predicting areas or conditions that are associated with a heightened risk of collisions.

Identification of Risk Factors

Early studies on driver fatigue have identified key factors such as irregular shifts, long load wait times, starting the workweek tired, difficulty in finding rest spots, and scheduling practices to be linked to fatigue and risk of accidents [86,87]. Conditions like obstructive sleep apnea, sleep debt, and excessive daytime sleepiness have also been associated with increased accidents, whereas breaks and naps have been found

to decrease them [88]. Fatigue risk have shown to be positively correlated with events such as hard braking via automatic data collection [89]. Furthermore, the significant impact of fatigue, highlighted by controlled experiments alongside observational studies, has led to developments like the FAST driver scheduling optimization algorithm [90] designed to reduce fatigue. In addition to fatigue, distracted driving, especially due to mobile phone use, has been recognized as a significant contributor to accidents. With an 11% increase in cell phone usage leading to quadrupling crash chances, distractions account for a significant portion of accidents [91]. Distraction-related fatalities rose by 28% between 2005 and 2008 [92], and a substantial percentage of crashes and near-crashes involved drivers performing non-driving tasks [93,94]. Consequently, countries have implemented laws and companies have adopted policies against distracted driving, with technologies being developed to enforce safe driving behavior.

Several external factors, such as weather conditions, traffic flow, and road geometry, significantly influence the probability and severity of crashes [95–97]. The crash risk during snowy seasons was found to be twice as high as in dry seasons, potentially influenced by the interaction between icy road surface conditions and steep road segments [98]. Adverse weather conditions coupled with critical roadway conditions were also found to increase crash probability significantly, with common predictors like precipitation and average speed in both single-vehicle and multiple-vehicle crash models [99]. Among various factors, the percentage of home-based work production emerged as a significant predictor for accident risk [100]. Depending on crash severity levels, the effect of environmental variables was found to vary; while adverse weather conditions generally increase crash risk, they potentially reduce the chance of crashes resulting in injuries and fatalities [97]. These findings underline the intricate interplay of these external factors in shaping crash likelihood and severity.

Statistical and Predictive Modeling

Traffic safety research predominantly employs retrospective case-control studies, where data analysis leverages logistic regression or alternative classification models [95, 96,

101–105]. Within these studies, "crashes" are defined as cases, and "non-crashes" as controls. With the latter considerably outnumbering the former, a system of aligning cases with multiple controls is required [106–109]. Although methodologies may vary, matching ratios can escalate up to 10:1 in the alignment of non-crashes to crashes [110–113].

Critical contributing factors to accident likelihood have been identified across various research initiatives. Safety has been revealed as a nonlinear function of traffic flow, with parameters such as speed limits and precipitation impacting accident frequency significantly [95]. Factors such as speed variation in the vicinity of the crash site, speed difference, average traffic volume, and average speed have been spotlighted as contributors to accident odds [96]. Detailed analysis of rear-end crashes have identified peak hour, high volume upstream, low speed downstream, and high congestion index downstream as important influencers [101]. Classification and regression trees (CART) have been instrumental in distinguishing high-risk and low-risk situations, achieving a success rate of 75% for high-risk scenarios [102]. Subsequent research leveraging a multilevel perception neural network identified occupancy downstream and average speed upstream as key factors [106]. Logistic regression and penalized maximum likelihood approaches found average speed as a significant determinant of accident risk [107]. A proposal for a frequent pattern (FP) tree for variable selection suggested 10-minute intervals as more efficient than 5-minute intervals [108]. Implementing a Markov model highlighted a non-linear relationship between speed and risk [109]. Exploration of weaving impact indicated factors such as speed at the beginning of the weaving segment, difference in speed between the start and end, and the logarithm of traffic volume as noteworthy variables [103, 104]. A combined Poisson regression and logistic regression model surpassed standalone models by offering a higher receiver operating characteristic (ROC) curve [104].

1.3 Research Motivation: Bridging the Gap

A recent article published by the New York Times, entitled “The Exceptionally American Problem of Rising Roadway Deaths", draws attention to a critical and counter-intuitive trend in road safety [114]. The article reports an increase in fatal car accidents in countries like the United States, Switzerland, and Ireland during the Covid pandemic period, a time when roads were ostensibly emptier. This rise in road fatalities, even with fewer vehicles on the road, underscores the complexity of traffic safety and raises questions about our current understanding of accident causality. It points out that the causes of car accidents extend beyond first-order descriptors of traffic such as the mean traffic density, and require a higher-order analysis of the internal structure of traffic.

In this light, our research motivation is shaped by the need to develop an enhanced understanding of the internal structure of traffic; that is a more granular insight into the complex interactions that govern traffic flow and occurrence of rare events and accidents. The aim is to bridge the gap that exists between physics-based traffic modelling and data-driven predictive modelling of accident collisions. While powerful in simulating complex traffic dynamics, physics-based modelling often overlooks safety considerations, including the prediction of collision risks. On the other hand, data-driven predictive modelling, while exhaustive in its consideration of accident factors, falls short in integrating the internal structure and interactions modelled in physics-based traffic studies. Our research intends to provide the handshake between these two disparate approaches, relating the internal traffic structure to accident precursors, comparing it with historical collision data, and applying this enriched framework to large-scale roadway networks. This approach is expected to provide a more comprehensive understanding of road transportation safety, thereby facilitating the development of effective measures to prevent road fatalities.

1.4 Research Significance

The findings from this research play a pivotal role in enhancing traffic safety, serving as a bridge connecting physics-based traffic modeling with data-driven predictive methods. By examining the intricate structure of traffic via statistical physics and high-order statistics, the study provides a thorough comprehension of the complex interplay that controls traffic movement and the onset of extreme incidents. The proposed framework is capable of predicting the risk of near-miss events in congested traffic flow, a factor that is positively correlated with collision risks obtained from the historic accident records. This proactive approach to safety, facilitated by the integration of real-time velocity data, has the potential to prevent accidents before they happen, improving traffic safety on a large scale. Furthermore, by assessing the reliability and robustness of roadway networks, the study offers beneficial insights for forward-looking infrastructure maintenance, aiding transportation authorities and urban planners in making informed decisions. The outcomes of this research have far-reaching implications for a wide range of stakeholders, including transportation authorities, urban planners, traffic safety researchers, the automotive industry, insurance companies, and the everyday public, ultimately leading to safer roads and enhanced driving experiences.

1.5 Thesis Outline

To achieve our research goals, we initiate our investigation into the internal structure of traffic in Chapter 2. We start by determining the conditions that allow us to consider traffic flow as a stationary ergodic random process over properly chosen time and length scales. We demonstrate that the two-point correlation function of vehicle occupancy provides access to spatial memory effects, such as headway, and the velocity autocovariance function reveals temporal memory effects, such as traffic relaxation time and traffic compressibility. We elucidate that this internal structure not only offers insight into the long-range correlations in the system but also contains critical

information relevant to the spatial and temporal mapping of traffic density, mean velocity, and driver behavior from individual driver velocity recordings. Particularly, under the provision of ergodicity and stationarity, the stochasticity-density plots enable a direct determination of traffic properties from crowdsourced measurements of vehicle velocities.

In Chapter 3, we utilize traffic's internal structure to introduce accident precursor metrics. We show that the statistics of accident precursors can be mapped onto a phase diagram that elevates observables, such as traffic density and velocity fluctuations, to a predictive metric of accident risk. To this end, we derive the near-miss collision risk from the velocity state transition matrix, considering the excessive deceleration of particles in steady-state traffic flow. We show that there exists an intrinsic near-miss collision risk when we apply the developed methodology to model-generated and crowdsourced velocity signals. It is confined to congested flow, predicts the highest risk of actual collisions, and decreases with increasing randomness in driver behavior, a feature it has in common with other many-body systems with long-range correlations triggered by randomness.

In Chapter 4, we extend the methodologies and outcomes from Chapters 2 and 3 to large-scale crowdsourced velocity data across various states in the United States. By visualizing the roadway network as a graph, we provide deeper insights into near-miss distribution and statistics, moving beyond the fundamental representations. In this context, we study the occurrence patterns of near-miss events along paths of varying lengths, integrating both the spatial distribution of near-miss risks along the network and the network's topology. Specifically, we examine the reliability and resiliency of roadway networks at state scale, focusing on their ability to sustain functionality and connectivity amid potential disruptions.

Finally, Chapter 5 serves as the culmination of our preceding efforts, as we summarize the main findings, limitations, and future directions of this work.

1.6 Summary

In conclusion, the issue of traffic safety is more relevant today than ever before. As the reach of our road transportation networks widens, the challenge of ensuring safety on the roads scales up correspondingly. concerted research efforts and innovative solutions. Despite significant progress in recent decades, traffic-related fatalities and injuries remain unacceptably high, accentuating the urgent need for addressing this issue with renewed vigour and a more profound understanding of its complexities. The multifaceted problem of traffic safety, deeply rooted in the fabrics of our society and economy calls for development of a physics-inspired framework which sheds light into the patterns of traffic safety from the internal structure of traffic.

Chapter 2

Internal Structure of Traffic: Memory Effects for Crowdsourced Traffic Property Determination

It has long been recognized that traffic exhibits complex dynamics of a many-body system from free flow to jamming, which defies simple averaging rules [35, 64, 69, 115, 116]. Analogous to various interacting many-body systems - including ideal gases, galaxy clusters, flock of birds, and nuclear systems composed of protons and neutrons - traffic too manifests characteristic structural behaviors at micro, meso, and macro scales. This includes short- and long-range interactions, statistics of specific events in traffic flow, and first-oder descriptors which are pivotal in the spatial and temporal representation of traffic density, average velocity, and driver behavior. These features are central mobility indicators amid the backdrop of an ever-increasing demand for reliable navigation systems designed to mitigate intrinsic trade-offs between mobility, safety and sustainability of our road networks.

In this chapter, we hypothesize that the internal structure of traffic holds critical information about the "mood" of the road. We explore this hypothesis by considering the NaSch-model as a stochastic process with intrinsic homogeneity which stems from the analogous collective random behavior of drivers and the periodic boundary condition of the system. We demonstrate that, owing to the unique internal structure of

the NaSch-model, memory in a vehicle's velocity time history provides the handshake between velocity fluctuations and traffic properties. Furthermore, we elaborate on the universal structural patterns aid in deriving macroscopic variables such as traffic density, mean velocity, and driver behavior from individual driver velocity recordings. This approach is complementary to prevalent navigation systems that approximate traffic density using user count and mean real-time vehicle velocity measurements, often sourced spatially via devices like smartphones to generate traffic flux estimates, the so-called floating car data [117]. These techniques, however, are data-intensive and often associated with high costs. Conversely, our proposed method capitalizes on the intrinsic traffic patterns, presenting a more cost-effective, and real-time solution using crowdsourced measurements.

2.1 Spatial and Temporal Memory Effects in the NaSch Model

Among the many traffic flow models ranging from continuum to agent-driven approaches [118], the Nagel-Schreckenberg (NaSch) cellular automaton model [35] has emerged as a powerful tool not only to reproduce critical features of traffic flow [68], such as backward moving shock waves [72], and the fundamental diagram of traffic [119], but foremost to track the internal "model" traffic structure by means of simulations [70, 120–123]. Of particular advantage is the apparent simplicity of the NaSch-model heuristics as a discrete lattice-gas-like model — acceleration $v_j = \min(v_j + 1, v_{max})$; deceleration $v_j = \min(d_j, v_j)$ (with d_j the headway); and random deceleration $v_j = \max(v_j - 1, 0)$ with probability p — which permit an update of the j^{th} vehicle position $x_j \to x_j + v_j$ in units of cells (or integer velocities) in the cellular automaton (for a more comprehensive review of the model, and its implementation details, see Appendix A). In addition to v_{max}, driver behavior in the NaSch-model is condensed into the stochasticity parameter p, bounding the maximum speed of vehicles in free flow. Indicative of the intrinsic unpredictability in

driver behavior, the stochasticity parameter drives fluctuations of vehicles' velocity, flux and occupancy, intrinsic to the internal model traffic structure captured by the NaSch-model. The random velocity fluctuations inherent in the NaSch model permit us to interpret it as a stochastic process. For a fundamental review of stochastic processes, readers are encouraged to refer to Appendix B.

Focusing on the NaSch mode, we start investigating the internal structure by ascertaining spatial and temporal memory effects in form of unique scaling relations through the application of the two-point correlation function to cell occupancy, and the velocity autocovariance function, respectively.

2.1.1 Two-Point Correlation Function and Spatial Memory Effects

The two-point correlation function is a statistical tool used to describe the spatial structure/texture of a material or a medium [124, 125]. For a random medium, such as a textured material or a turbulent fluid, the two-point correlation function can provide information about the degree of disorder and the characteristic length scales of the medium's internal structure. Herein, we apply the two-point correlation to cell occupation numbers to gain insight into the second-order spatial structure of the model [Fig. 2-1.(a)]. Formally,

$$S_2(r_1, r_2) = \mathbb{E}[I_\eta(r_1, t) I_\eta(r_2, t)] \tag{2.1}$$

where $\mathbb{E}[.]$ and η respectively denote the ensemble average operator and jamming random variable, and the binary cell occupation number $I_\eta(r, t) = 1$ if cell r is occupied at time t, otherwise $I_\eta(r, t) = 0$. In the context of random media theory, the occupation number serves as a broad indicator of the medium's phase such as solid, fluid or void. Given stationarity *and* ergodicity of the system, the first-order moment of the occupation number remains invariant over time, and is equal to the traffic density ρ, which corresponds to the porosity in the context of porous materials, i.e., $\rho = \mathbb{E}[I_\eta(r, t)]$. Furthermore, the two-point correlation function degenerates to

$S_2(r_1, r_2) = S_2(\delta_r = |r_2 - r_1|) = \mathbb{E}[I_\eta(r, t_0) I_\eta(r + \delta_r, t_0)]$ which represents the autocorrelation of a snapshot of the occupation numbers at a given time t_0. Specifically, the stationarity of the process implies that the two-point correlation function does not depend on the absolute values of r_1 and r_2, but is rather a function of the lag between the two, i.e., $\delta_r = |r_2 - r_1|$; that is, $S_2(r_1, r_2) = S_2(r_3, r_4)$ as long as $|r_2 - r_1| = |r_4 - r_3|$. The ergodicity, on the other hand, implies that a snapshot of the system, referred to as a realization of the ensemble, can statistically represent the entire ensemble if the system is sufficiently large. The bounds of the two-point correlation function $S_2(\delta_r)$ are obtained by examining the asymptotic behavior of the function as [126],

$$S_2(\delta_r = 0) = \rho; \quad S_2(\delta_r \to \infty) = \rho^2 \tag{2.2}$$

Specifically, the function's behavior at the origin $r = 0$ is derived from the binary nature of the occupation number, allowing for the replacement of the expected value operator with probability operator. Conversely, the function's behavior at $r \to \infty$ is obtained from the statistical independence between occupation numbers at large spatial lags, i.e., $\lim_{\delta_r \to \infty} \mathbb{E}[I_\eta(r, t_0) I_\eta(r + \delta_r, t_0)] = \mathbb{E}[I_\eta(r)] \mathbb{E}[I_\eta(r + \delta_r)]$. A critical information about the internal traffic structure is provided by the slope of S_2 at the origin in terms of the chord length ℓ_c [127],

$$\frac{dS_2}{d\delta_r}|_{\delta_r = 0} = -\frac{S_2(0)}{\ell_c} \tag{2.3}$$

The chord length serves as a measure of the intrinsic structure of a random process. For instance, in texture characterization of random porous materials, the chord length represents the mean linear distance between the pores [126, 128]. Applied to traffic, the chord length defines the probability of the immediate neighboring cell ($\delta_r = 1$) of an occupied cell to be occupied or not:

$$\mathbb{P}[I_\eta(r + 1, t_0) \mid I_\eta(r, t_0) = 1] = \frac{S_2(1)}{\rho} = 1 - \frac{1}{\ell_c} \tag{2.4}$$

where $\mathbb{P}[.]$ stands for the probability operator. In free flow $\ell_c = 1$, and hence $\mathbb{P}[I_\eta(r+1, t_0) \mid I_\eta(r, t_0) = 1] = 0$, whereas in complete jamming $S_2(1) = \rho$, and thus $1/\ell_c \to 0$. In between these asymptotes, jamming transition occurs at a critical occupancy ρ_c^p. More specifically, when plotted against the probability of the subset $\eta \in N$ to be part of a cluster of jammed vehicles, that is, $\overline{\eta} = \mathbb{P}[\eta \cap \rho > \rho_c^p] \approx (\rho - \rho_c^p)/(1 - \rho_c^p) \in [0, 1]$, we find from simulations that the occupation probability of the next cell neighbor scales as [Fig. 2-1.(b)],

$$S_2(1) = \overline{\eta}^\alpha; \quad \alpha \in [1, 2] \tag{2.5}$$

The exponent α depends on v_{max} [inset of Fig. 2-1.(b)] for all v_{max} and p, when ρ_c^p is estimated from $\rho_c^p = v_J(v_J + \overline{v}_F)^{-1}$ where $\overline{v}_F = v_{max} - p$ and $v_J = 1 - p$ stand for, respectively, the average free flow velocity and the analytic upper bound of the jammed dissolution velocity of the NaSch-model [129]. Moreover, in the congested regime, we find that $S_2(\delta_r)$ converges rapidly for $\delta_r > 1$ to the asymptotic value $S_2(\delta_r \to \infty) = \rho^2$ [Fig. 2-1.(a)]. This is indicative of statistical independence of occupancy of non-adjacent cells, much akin to a Markov process, that is:

$$\mathbb{E}[I_\eta(r, t_0) I_\eta(r + \delta_r, t_0)] \approx \mathbb{E}[I_\eta(r, t_0)]\mathbb{E}[I_\eta(r + \delta_r, t_0)] = \rho^2 \tag{2.6}$$

for $\delta_r > 1$ and $\overline{\eta} > 0$. This shows that the underlying spatial ergodicity and stationarity of the NaSch-model give rise to traceable *spatial* memory effects, which will turn out to be critical for linking internal structure to traffic flow properties. For instance, taking into account the Markovian property of occupancy I_η, the average headway of a vehicle can be estimated from the two-point correlation function (for derivation, see Appendix C):

$$\mathbb{E}[d] = \overline{d} = \frac{1}{S_2(0)} - 1 \tag{2.7}$$

with $S_2(0) = \rho$.

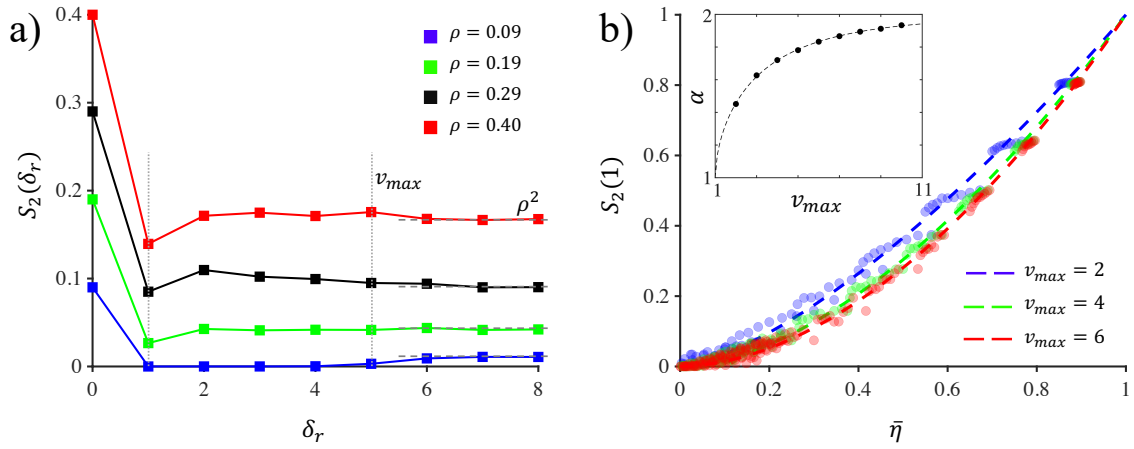

Figure 2-1: Two-point correlation function of the NaSch-model representing probability of occupancy: (a) $S_2(\delta_r)$ with asymptotes $S_2(0) = \rho$ and $S_2(\delta_r \to \infty) = \rho^2$ for $p = 0.5$ and $v_{max} = 5$; $\rho = 0.09 < \rho_c^p$ is the free flow regime and shows no correlation with the occupation number of the neighboring cells for $\delta_r \leq v_{max}$. (b) $S_2(1)$ for the congested flow as a function of $\overline{\eta} = (\rho - \rho_c^p)/(1 - \rho_c^p)$ and $p \in \{0, 0.1, ..., .9\}$; $S_2(1) \sim$ scales as $S_2(1) \sim \overline{\eta}^{\alpha}$ with $1 \leq \alpha \approx 1 + \tanh[0.5(v_{max} - 1)^{0.6}] \leq 2$.

2.1.2 Velocity Autocovariance Function and Temporal Memory Effects

A second quantity of interest to ascertain *temporal* memory effects in the internal structure is the velocity autocovariance function. This function is formulated as the expected value, taken over an ensemble, of the product of the velocity's deviation from its mean value at two distinct time instances separated by a time delay, δ_τ. In essence, it provides a statistical measure of the temporal correlation in the velocity fluctuations, thereby indicating the persistence of such fluctuations over time. For stationary processes, the canonical definition of this function thus reads,

$$C_{vv}(\delta_\tau = |t_2 - t_1|) = \mathbb{E}[v(t_1, \zeta)v(t_2, \zeta)] - \overline{v}^2 \tag{2.8}$$

where $v(t, \zeta)$ represents velocity (in NaSch-units) as a stochastic process, while ζ denotes a random event with $\mathbb{P}[\zeta]$ probability of occurrence and $\overline{v} = \mathbb{E}[v(t, \zeta)]$ its mean value. The autocovariance function quantifies the correlation between velocity and itself at different points in time, providing insight into the temporal structure. From the definition [Eq. (2.8)], one can realize that *(i)* the autocovariance function is symmetric around the origin ($C_{vv}(\delta_\tau) = C_{vv}(-\delta_\tau)$), and *(ii)* the maximum value occurs at lag zero $\delta_\tau = 0$, which is the variance of the process σ^2_{vv}. A rapid decay of the autocovariance function to zero indicates that the process exhibits short-term dependencies, implying a low memory process. Conversely, a slow decay suggests the presence of long-range correlations, indicative of a process with strong memory.

In free flow, the velocity variation of a single vehicle resembles a Bernoulli process of velocity trials v_{max} and $v_{max} - 1$ with probabilities $1 - p$ and p, respectively. Thus, the free-flow velocity variance reads $\sigma^2_{vv,F} = C_{vv,F}(\delta_\tau = 0) = p(1 - p)$ and $C_{vv,F}(\delta_\tau \neq 0) = 0$ (due to the independence of trials in a Bernoulli process). In contrast, upon jamming [Fig. 2-2.(a)], a distinct time memory effect builds up similar to Ornstein-Uhlenbeck stochastic processes [130–132], which fades exponentially from $C_{vv}(\delta_\tau = 0) = \sigma^2_{vv}$ to $C_{vv}(\delta_\tau \to \infty) = 0$. Analogous to the chord length in the two-point correlation function [Eq. (2.3)], we capture the time memory effect in form of

a characteristic decay time, τ_c, from the slope of the autocovariance function,

$$\frac{dC_{vv}}{d\delta_\tau}|_{\delta_\tau \to 0} = -\frac{C_{vv}(0)}{\tau_c} \tag{2.9}$$

where $C_{vv}(0) = \sigma_{vv}^2$. The velocity is expected to have the strongest memory upon transitioning from free flow to jammed flow, and to reduce to zero (the so-called memoryless process) for fully jammed traffic. In between these asymptotes, we find from simulations that the decay time, τ_c, scales with the jamming probability, $\overline{\eta} = P(\eta \cap \rho > \rho_c^p)$ [Fig. 2-2.(b)] as $\tau_c \sim \overline{\eta}^{-\gamma_v}$ and can be approximated by,

$$\tau_c \approx \hat{\tau}_c(\overline{\eta}^{-\gamma_v} - 1) \tag{2.10}$$

where $\hat{\tau}_c = 1.88$ and $\gamma_v = 0.56$ are fitting parameters. The τ_c isotherm, which captures all (p, ρ)-pairs with same τ_c, reads as $p = (1-\rho(1+v_{max})+v_{max}\theta_v)/(1-2\rho+\theta_v)$ with $0 \leq \theta_v = (\tau_c/\hat{\tau}_c + 1)^{-1/\gamma_v} \leq 1$ and $\theta_v \leq \rho \leq (1 + v_{max}\theta_v)/(1 + v_{max})$. The universal decay pattern of memory suggests that the correlation between velocities at different time lags is invariant with respect to the system parameter v_{max}, and, instead, is exclusively a function of the system's congestion level. Additionally, memory exhibits a behavior indicative of first-order phase transitions during the transition from free flow to congested regime, a characteristic bearing remarkable similarity to transitions observed in liquid crystals [133]. Furthermore, akin to the kinetic theory of gases, we define a kinetic compressibility of vehicle speed from the expected value ratio of jammed-to-free flow velocity fluctuations [Fig. 2-2.(c)]:

$$\kappa_v = \frac{\mathbb{E}[v^2]}{\mathbb{E}[v_F^2]}; \quad \mathbb{E}[v_F^2] = p(1-p) + \overline{v}_F^2 \tag{2.11}$$

The kinetic compressibility, κ_v, represents the adaptability of a medium's density under changing conditions. When the system is unimpeded or in free flow ($\rho \leq \rho_c^p$), the kinetic compressibility is at its maximum ($\kappa_v = 1$), signifying full compressibility. However, as the jamming probability $\overline{\eta}$ increases, κ_v diminishes according to a power-law relationship as $\kappa_v \approx (1 - \overline{\eta})^\beta$, eventually reaching zero at maximum density

($\rho = 1$), a state indicating incompressibility. This power-law decay is controlled by the exponent $\beta = \beta(p)$, which depends on stochasticity parameter [inset of Fig. 2-2.(c)], linking the system's randomness to its compressibility behavior.

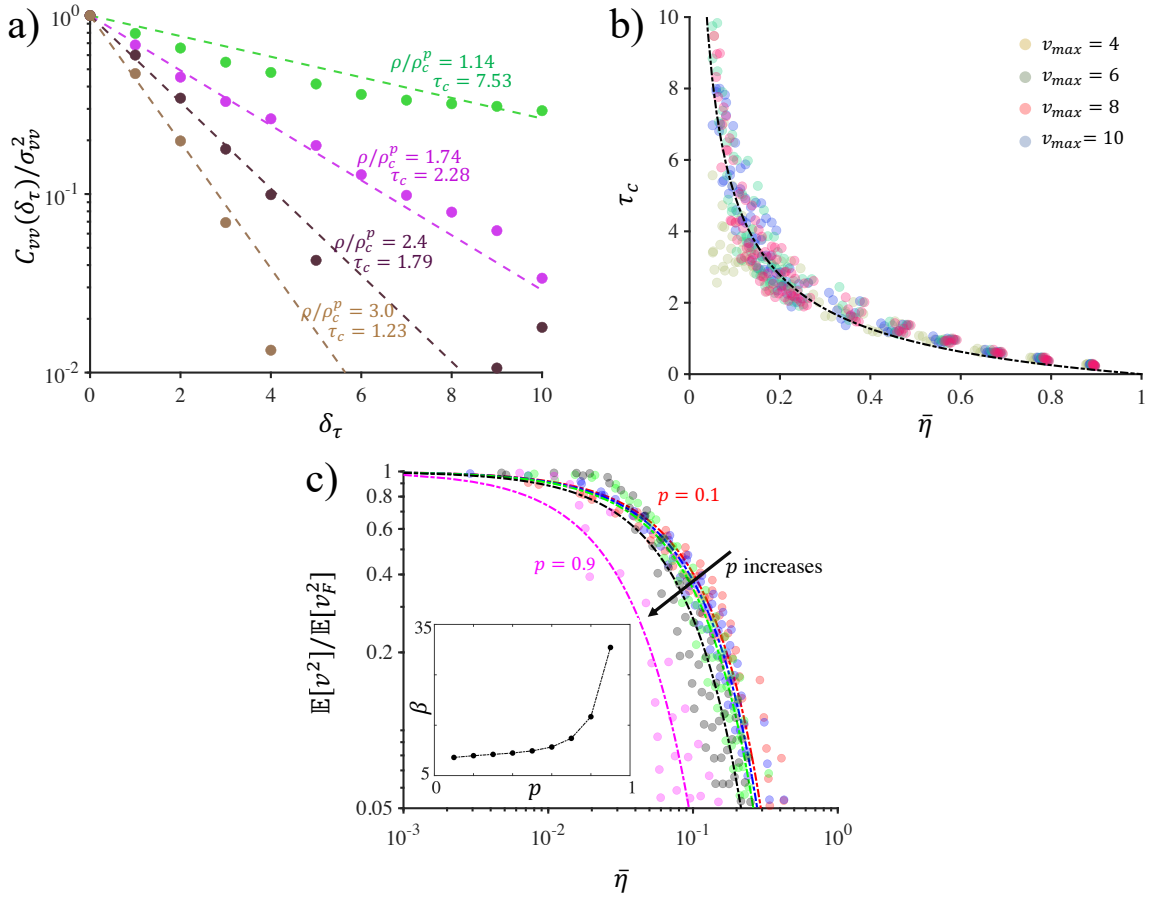

Figure 2-2: (a) Velocity autocovariance in NaSch-model decays exponentially, $C_{vv}(\delta_\tau) \approx \sigma_{vv}^2 \exp(-\delta_\tau/\tau_c)$, resembling an Ornstein-Uhlenbeck process ($v_{max} = 5, p = 0$). (b) Decay time scale τ_c as a function of $\overline{\eta}$ in the interaction-dominated regime for $p \in \{0.1, 0.2, ..., 0.9\}$; Velocity approaches memorylessness as traffic density increases with $\tau_c \approx 1.88\,(\overline{\eta}^{-0.56} - 1)$. (c) Jammed-to-free flow velocity fluctuations κ_v for $v_{max} \in \{4, 6, 8, 10\}$ decays monotonically from 1 (kinetically compressible vehicle speed) at $\overline{\eta} = 0$ to 0 (kinetically incompressible vehicle speed) upon jamming in a power form $\kappa_v \approx (1 - \overline{\eta})^\beta$

2.1.3 Local Density Autocovariance Function and Temporal Memory Effects

We proceed by exploiting memory effects in local density $\tilde{\rho}$ defined as the traffic density on a randomly selected subset of adjacent cells of the system $l_{\tilde{\rho}}^{\zeta}$ with a prescribed size of $|l_{\tilde{\rho}}|$; that is, $\tilde{\rho}(t,\zeta) = \sum_{j \in l_{\tilde{\rho}}^{\zeta}} I_{\eta}(j,t)/|l_{\tilde{\rho}}|$ where $|l_{\tilde{\rho}}|$ is significantly smaller than the system size to be representative of local properties. Conservation of system-wide (global) density ρ results from the system's periodic nature as the number of particles is a conserved quantity. This, however, doesn't apply to the localized density, which behaves as a random process. Localized density variation originates from the unequal rates of vehicle influx and efflux in the subset of cells, signifying the imbalance between vehicles entering and departing the subset. Nonetheless, the local density maintains an expected value of ρ, since, on average, the total number of vehicles within that subset is proportional to the total number of vehicles in the system. Nonetheless, the stationarity inherent in the overall process induces stationarity in local density behavior for sufficiently large local road subsets.

To understand the temporal structure , we study the second-order moment of the local density as a stochastic process. To this end, the autocovariance function of local density, $C_{\tilde{\rho}\tilde{\rho}}(|t_1 - t_2| = \delta_\tau) = \mathbb{E}[\tilde{\rho}(t_1,\zeta)\tilde{\rho}(t_2,\zeta)] - \rho^2$, adopts a characteristic linear behavior illustrated in Fig. 2-3.(a). The two-point correlation function implies that for high v_{max} and in the congested regime, we have $S_2(r, r+\delta_r) \approx \rho^2$ [Sec. 2.1.1] which is indicative of statistical independence between $I_{\eta}(r,t)$ and $I_{\eta}(r+\delta_r,t)$ for $\delta_r > 0$. Therefore, using the Bernoulli assumption of $I_{\eta}(r,t)$, the variance of local density can be approximated as $\sigma^2_{\tilde{\rho}\tilde{\rho}} \approx \rho(1-\rho)/|l_{\tilde{\rho}}|$ [inset of Fig. 2-3.(a)]. The autocovariance function thus permits the following simplification:

$$C_{\tilde{\rho}\tilde{\rho}}(\delta_\tau) \approx \sigma^2_{\tilde{\rho}\tilde{\rho}} \left(1 - \frac{1}{\nu}\frac{\delta_\tau}{\delta_\tau^c}\right) \geq 0 \tag{2.12}$$

where $\delta_\tau^c = |l_{\tilde{\rho}}|/\bar{v}$ denotes the average residence time of a vehicle travelling on $l_{\tilde{\rho}}$, and $\nu\delta_\tau^c$ stipulates the temporal memory of the local density process. In free flow, local density at two time steps are correlated only if their time lag is less than δ_τ^c; and thus

$\nu = 1$. Akin to temporal memory of velocity in the congested flow, memory can be parametrized as a power function of $\bar{\eta}$ [Fig. 2-3.(b)],

$$\nu \approx v_{max}\hat{\nu}(\overline{\eta}^{-\gamma_\nu} - 1) \tag{2.13}$$

where $\hat{\nu} = 0.5$ and $\gamma_\nu = 5 \times 10^{-4}|l_{\tilde{\rho}}| + 0.31$ [inset of Fig. 2-3.(b)] are fitting parameters. Local density shows strong memory upon transitioning to congested flow and approaches memorylessness as $\bar{\eta} \to 1$. Analogous to temporal memory [Eq. (2.10)], the congestion level of the system plays a pivotal role in ascertaining the memory of local density. This relationship demonstrates a linear trend with respect to v_{max}. Moreover, the consistency of these outcomes suggests a more comprehensive understanding: provided that the statistical assumptions of stationarity and ergodicity remain valid over appropriately selected temporal and spatial scales, one can potentially characterize the universal behavior of the system, which can be achieved by introducing normalized parameters, in this case ν/v_{max}.

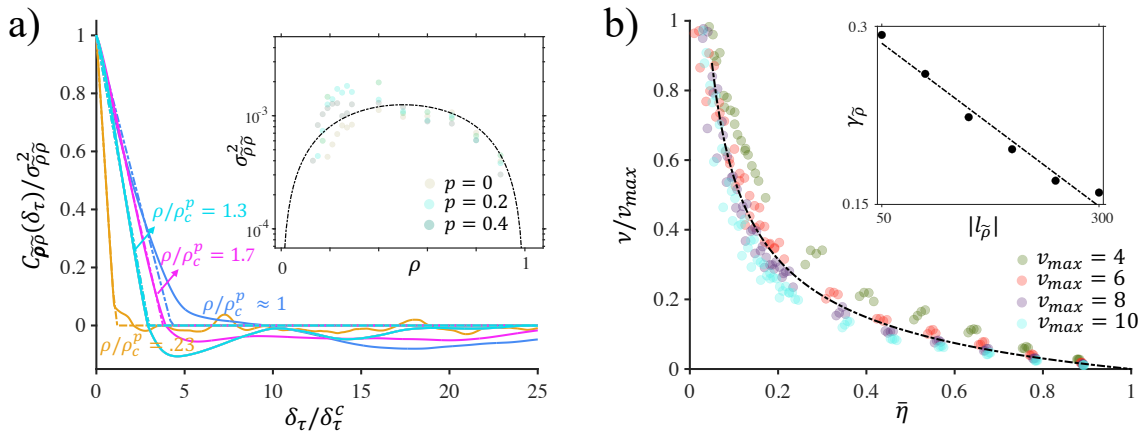

Figure 2-3: (a) Normalized autocovariance function of local density exhibiting a distinct linear decay considering $v_{max} = 5$, $p = 0.3$ and $|l_{\tilde{\rho}}| = 100$ for free flow $\rho/\rho^p_c \leq 1$, transitioning to congested flow $\rho/\rho^p_c \approx 1$ and congested flow $\rho/\rho^p_c > 1$; the inset shows the variance of local density $\sigma^2_{\tilde{\rho}\tilde{\rho}}$ and the black curve is the variance for Bernoulli process $\sigma^2_{\tilde{\rho}\tilde{\rho}} \approx \rho(1-\rho)/|l_{\tilde{\rho}}|$. (b) Universal pattern of ν/v_{max} as a function of jamming probability approximated by $\nu/v_{max} \approx 0.5(\bar{\eta}^{-\gamma_\nu} - 1)$ where $\gamma_\nu = 5 \times 10^{-4}|l_{\tilde{\rho}}| + 0.31$ as shown in the inset.

2.2 Link between Internal Structure and Traffic Density and Stochasticity

In this section, we use the scaling relations developed in Section 2.1 to obtain the isotherms of first- and second-order velocity statistics for traffic density and driver behavior determination derived. To this end, we proceed by matching the found internal traffic structure of spatial and temporal memory effects with traffic density ρ and stochasticity parameter p.

2.2.1 Fundamental Diagram

The traffic fundamental diagram operates as a critical framework analogous to a phase-space plot, encapsulating the dynamic interplay of speed, flow, and density in a traffic stream. Similar to phase transitions in physical systems, this diagram demarcates two main regimes, free-flow and congested traffic, governed by the traffic density and the stochasticity parameter, illustrating the intricate balance between spatial and temporal variables using principles of statistical mechanics. Therefore, we proceed by constructing the fundamental diagram using the results of the two-point correlation function, employing the traffic flux definition $\mathbb{E}[I_\eta]\mathbb{E}[v] = \rho\overline{v}$. In the free-flow regime, the independence of cell occupation and vehicle speed provides a linear relation between flux and the average free flow velocity $\overline{v} = \overline{v}_F = v_{max} - p$:

$$J = \rho\overline{v} = v_F\rho, \quad \rho < \rho_c^p \tag{2.14}$$

An equally linear relationship between flux and density approximates flux in jammed flow, when considering (i) a first-order estimate of the average velocity –in NaSch-units of number of cells or integer velocity– from the average headway, $\overline{d}$, in the form, $\overline{v} = (1-p)\overline{d}$; and (ii) an estimate of the average headway from the two-point correlation function $\overline{d} = 1/S_2(0) - 1$ [Eq. (2.7) and Appendix C], while (iii) assuming the statistical independence of I_η and v. That is, letting $S_2(0) = \rho$, the traffic flux–

density relation for congested flow is obtained:

$$J = \rho\overline{v} \approx \rho(1-p)\overline{d} = (1-p)(1-\rho), \quad \rho > \rho_c^p \tag{2.15}$$

Finally, a combination of Eqs. (2.14) and (2.15) leads to the well-known bi-linear approximation of the fundamental diagram [134] (cited by [118]):

$$J = \rho\overline{v} = v_F(\rho - \overline{\eta}) \tag{2.16}$$

where $\overline{\eta} = |\rho - \rho_c^p|/(1-\rho_c^p)$. It should be noted that measured flux-density relations typically exhibit large scatter that can be attributed to non-equilibrium traffic conditions. The regularity of the bi-linear form we here derive with the help of the two-point correlation function of the NaSch-model relates to its underlying stationarity and ergodicity. In return, given stationarity and ergodicity, the fundamental diagram provides a further relation between a measurable mean velocity $\overline{v}$, density ρ and stochasticity p.

2.2.2 Stochasticity-Density Plot

We are now ready to match spatial or temporary memory effects for the determination of traffic density, ρ, and stochasticity parameter, p. From the two-point correlation function, we retain the bilinear flux–density relation (the fundamental diagram), to construct mean velocity isotherms,

$$p = 1 + \frac{\rho\overline{v}}{\rho - 1} \tag{2.17}$$

with

$$0 \leq \rho \leq \frac{1}{1+\overline{v}} \tag{2.18}$$

In the velocity stochasticity-density (p, ρ) plot [Fig. 2-5.(a)], we overlay these mean velocity isotherms with isotherms of the decay time τ_c as a function of (p, ρ). Akin

to a phase diagram, the curve

$$0 \leq p = \frac{1 - \rho(1 + v_{max})}{1 - 2\rho} \leq 1 \tag{2.19}$$

separates free flow from congested flow. In the congested flow, the mean velocity $\overline{v}$ and temporal memory τ_c isotherms intersect at a unique point in the (p, ρ)-plane provided that $\Omega(\overline{v}, \theta_v) = \overline{v}/v_{max} + (1 + \overline{v})\theta_v \leq 1$, where $\theta_v = (\tau_c/\hat{\tau}_c + 1)^{-1/\gamma_v}$, with fitting parameters $\hat{\tau}_c = 1.88$ and $\gamma_v = 0.56$ [Fig. 2-2.(b)]. That is, $\Omega(\overline{v}, \theta_v) \leq 1$ is the velocity sample space of the NaSch-model indicating the region of all possible $\overline{v}$ and θ_v outcomes [Fig. 2-4]. In a similar fashion, we construct the stochasticity-density plot for the local density [Fig. 2-5.(b)], which allows for relating the initial slope of the normalized autocovariance function, $(\sigma^2_{\tilde{\rho}\tilde{\rho}})^{-1} dC_{\tilde{\rho}\tilde{\rho}}/d\delta_\tau$, and the expected value of local density, $\mathbb{E}[\tilde{\rho}] = \bar{\tilde{\rho}}$, to the stochasticity parameter and traffic density. The ν-isotherm reads

$$p = \frac{1 + \theta_\nu v_{max} - \rho(1 + v_{max})}{1 + \theta_\nu - 2\rho} \tag{2.20}$$

with $\theta_\nu = (\nu/(\hat{\nu} v_{max}) + 1)^{-1/\gamma_\nu}$ which intersects the vertical $\bar{\tilde{\rho}}$ isotherms at a unique point if $(\rho(v_{max} + 1) - 1)/v_{max} \leq \theta_\nu \leq \rho$.

In summary, given stationarity and ergodicity, first- and second-order ensemble statistics of velocity or local density, namely mean and initial slope of normalized autocovariance function, provide a means to determine stochasticity parameter and traffic density. This introduces a novel framework for deducing the macroscopic attributes of traffic based on vehicle velocity fluctuations and their correlations, offering computational efficiency when contrasted with traditional methodologies like the floating-car method.

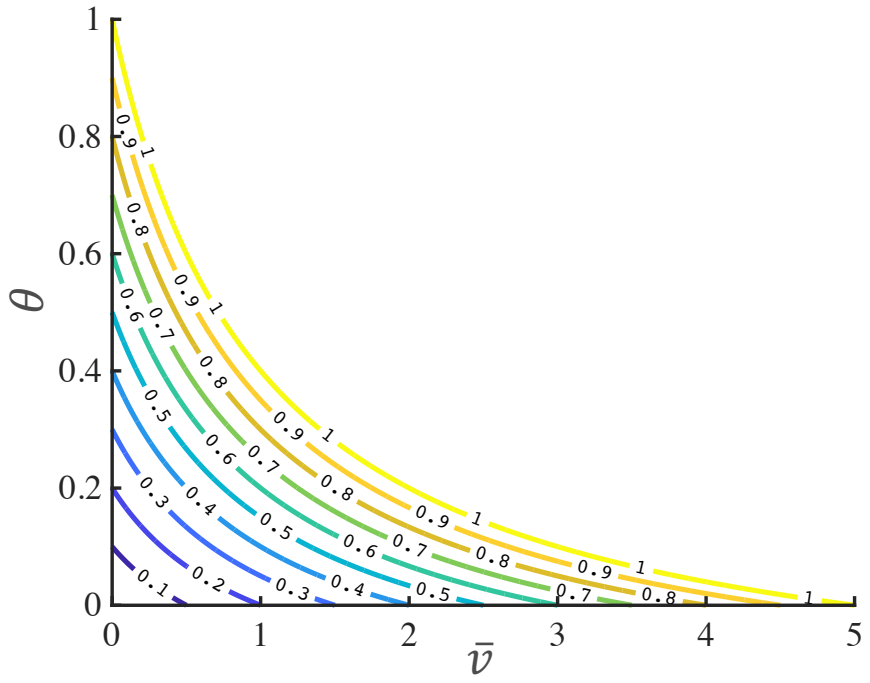

Figure 2-4: The event space of the NaSch model: the isotherms of expected velocity and memory intersect at a unique point only if $\Omega = \bar{v}/v_{max} + (1 + \bar{v})\theta \leq 1$ where $\theta = (\tau/\tau_c + 1)^{-1/\gamma_v}$ is the normalized memory and $v_{max} = 5$. The contours represent the value of the control parameter Ω. The nonlinear domain of the event space in the NaSch model enables the evaluation of a velocity signal's alignment with the NaSch model heuristics, up to second-order statistics.

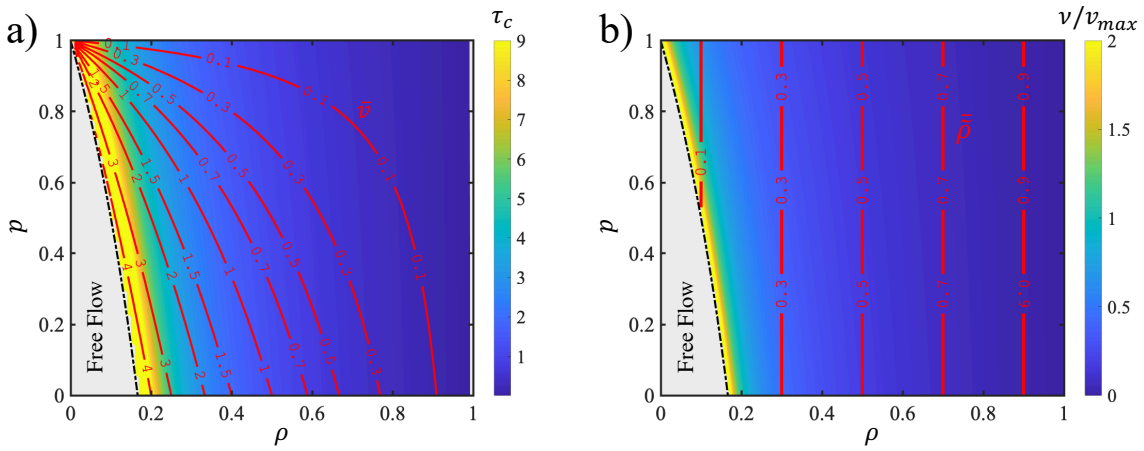

Figure 2-5: (a) Velocity and (b) local density stochasticity-density plot showing the interplay between the first-, second-order moments, occupancy ρ and stochasticity parameter p for $v_{max} = 5$. Stochasticity-occupancy $(p - \rho)$ interaction provides the means to estimate macroscopic traffic properties, p and ρ, from two statistical observable, which are $\overline{v}$ and τ_c [Eq. (2.10)] for velocity, and $\overline{\tilde{\rho}}$ and ν [Eq. (2.13)]for local density.

2.3 Application

We now investigate the predictive prowess of our approach for estimating traffic properties from local density measurements and crowdsourced velocity data of individual vehicles. Since our approach relies on ergodicity, we first investigate the conditions under which a realization of the process can statistically represent the ensemble moments. We then employ this condition with empirical local density measurements and crowdsourced velocity data of vehicles.

2.3.1 Ergodicity and Entropy

Invoking the ergodic theorem of statistical mechanics [135], a vehicle eventually explores the entire phase space in a uniform sense over long (enough) time scales, resulting in an overall ergodic behavior of vehicles in the NaSch-model. That is, one realization of the process, say ζ_0, is statistically rich enough to approximate the ensemble averages from its temporal moments. We apply the ergodic principle to velocity $v(t, \zeta_0)$ of a randomly selected vehicle ζ_0. That is, recalling the approximate equality of time and phase averages in ergodic mechanical systems, the mean and autocovariance are estimated from the following time averages:

$$\overline{v} \approx \frac{1}{T_r(\zeta_0)} \sum_{t=1}^{T_r(\zeta_0)} v(t, \zeta_0) \tag{2.21}$$

and,

$$C_{vv}(\delta_\tau) \approx \frac{1}{T_r(\zeta_0)} \sum_{t=1}^{T_r(\zeta_0)} v(t, \zeta_0) v(t + \delta_\tau, \zeta_0) - \overline{v}^2 \tag{2.22}$$

where $T_r(\zeta_0)$ denotes the representative time scale of realization ζ_0. This time scale is the shortest time interval over which random event ζ is statistically representative of the ensemble. Focusing on the probability distribution of velocity, such time scale is controlled by the entropy (also known as the expected information content) of velocity [136]:

$$H(p, \rho) = -\sum_{v_i=0}^{v_{max}} \mathbb{P}[v_i] \ln(\mathbb{P}[v_i]) \tag{2.23}$$

Entropy, measured in 'nats' (1 nat ≈ 0.6931 bits), serves as a quantitative tool to assess the level of uncertainty or randomness in a dataset. High entropy signifies a large degree of uncertainty or randomness, while low entropy indicates that the data is predictable. The lower and upper bounds of entropy are 0 and $\ln(v_{max}+1)$, respectively. The lower bound is reached in deterministic scenarios, such as when the system is in free flow with a stochasticity parameter of zero ($p = 0$), or when the system is jammed and all vehicles are at a standstill. In a continuous limit, the entropy reaches its maximum when the probability density function adopts the form of a Dirac delta function. Conversely, maximum entropy is attained when the velocity distribution is uniform, creating a state of maximum uncertainty in the system.

The behavior of representative time scale $T_r(\zeta_0)$ with respect to entropy [Fig. 2-6] is similar to the relaxation time τ_c: it diverges at the transition density [122, 123], and increases exponentially as,

$$\mathbb{E}[T_r] = T_r^0 \exp(\Gamma \tilde{H}(p,\rho)) \tag{2.24}$$

where $\tilde{H}(p,\rho) = H(p,\rho)/\ln(v_{max}+1)$ is the normalized entropy with $\ln(v_{max}+1)$ corresponding to the entropy of a uniform distribution of velocity. While fitting parameter T_r^0 increases linearly as a function of v_{max} [with $T_r^0 \approx 10(v_{max}+1)$ for $2 \leq v_{max} \leq 10$], the prefactor $\Gamma \approx 2.65$ is independent of v_{max}. An ensemble with entropy $H(p,\rho)$ is, therefore, expected to be statistically identical to one of its realizations of minimum length $T_r^0 \exp\{\Gamma \tilde{H}(p,\rho)\}$. Eq. (2.24) provides an expected lower bound for achieving ergodicity of velocity in the NaSch-model. Similar to velocity, ergodicity in NaSch model allows us to approximate the ensemble statistics of local density from a subset of space when observed over long time interval.

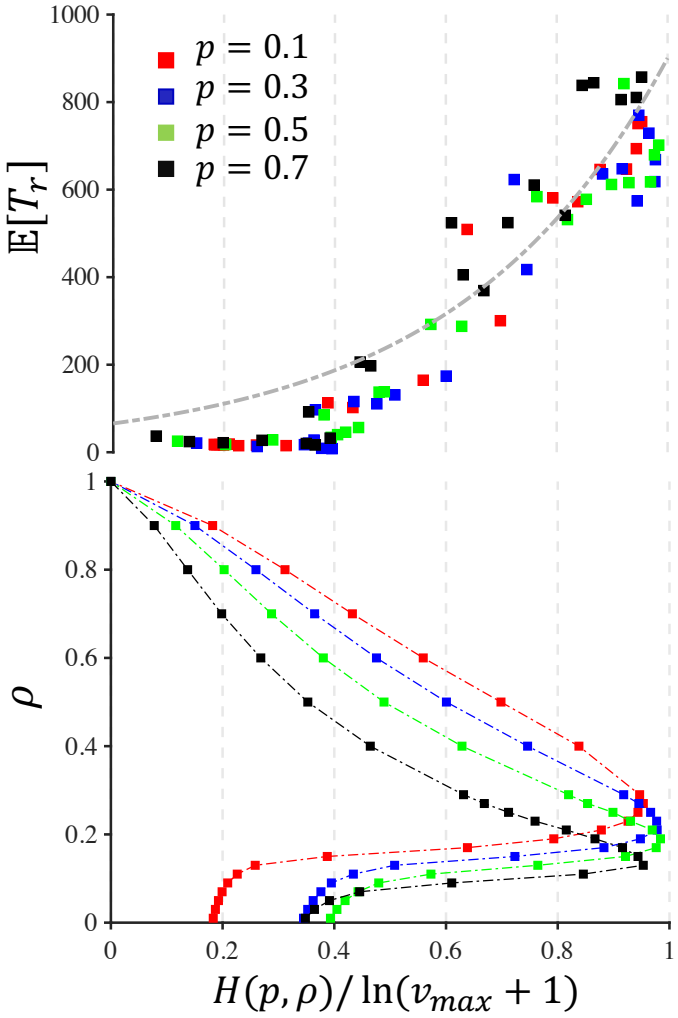

Figure 2-6: Variation pattern of entropy $H(p, \rho)$ and its impact on expected representative time scale $\mathbb{E}[T_r]$ for $v_{max} = 5$. Representative time scale is controlled by entropy and increases as the system becomes more uncertain, corresponding to higher levels of entropy with more possible configuration. Entropy degenerates to $H(p, \rho < \rho_c^p) = -\ln p^p(1-p)^{1-p}$ for free flow regime, and reaches its upper bound at $\rho \sim \rho_c^p$ where velocity has an almost uniform distribution with $H(p, \rho \sim \rho_c^p) \approx \ln(v_{max} + 1)$.

2.3.2 Application to Local Density Data

The first application considers classical traffic density measurements achieved by a fixed sensor (e.g., cameras) along a road. Such measurements provide a means to estimate local density $\tilde{\rho}$ over sufficiently large time interval. The data reported in Ref. [3] were obtained from measurements carried out on the German freeway A1 near an intersection with German freeway A59 in June 1996 [Fig. 2-7.(a)]. Translated in NaSch-units, we consider the road capacity to be N_c (in number of vehicles per hour), which we attribute to the deterministic limit density of the NaSch-model, $\rho_c^{p=0} = (v_{max} + 1)^{-1}$. Denoting the cell length by L_c, a first-order conversion between NaSch-units and real (time-length) units is provided by:

$$\frac{V_{max}/L_c}{(1 + v_{max})N_c} = \text{const.} \tag{2.25}$$

For a single-lane road capacity $N_c = 1900$, and for $v_{max} = 5$, each second corresponds to approximately 1 NaSch time unit considering $L_c \approx 7.5$ meters (as suggested in [35]). The local density shows an expected value of $\rho = \overline{\tilde{\rho}} = 0.22$ and its normalized autocovariance function exhibits a characteristic linear behavior [Fig. 2-7.(b)] with a fitted slope of $-(\nu\delta_\tau^c)^{-1} = -0.0155$. From the variance of local density $\sigma^2_{\tilde{\rho}\tilde{\rho}} = 2.5 \times 10^{-4}$, we readily recognize from Eq. (2.12) that $|l_{\tilde{\rho}}| \approx 60$, which implies that the local density values are averaged over an approximately 450-meter spatial window. From Eqs. (2.13) and (2.15) we obtain the stochasticity parameter $p = 0.3$. Next, from the fundamental diagram [Eq. (2.15)], we estimate the average velocity, $\bar{v}/v_{max} = 0.49$. This value which we obtain from the statistical moments of local density, is in remarkable agreement with the average velocity obtained from the recorded measurements, i.e., $\mathbb{E}[V]/V_{max} = 0.47 \approx \overline{v}/v_{max} = 0.49$. This shows that the internal structure provides an independent means to estimate not only average traffic speed, but as well an estimate of the stochasticity parameter reminiscent of driver behavior.

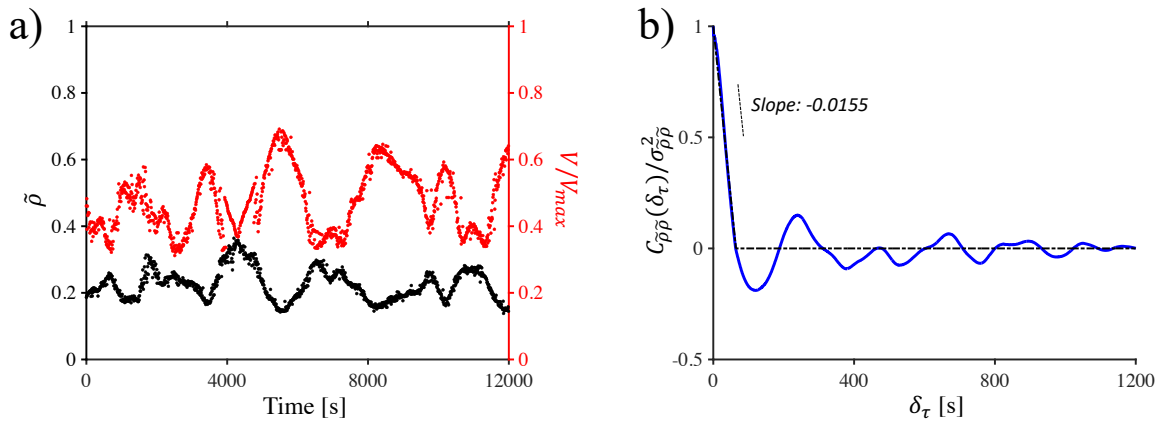

Figure 2-7: (a) Summary of measurements adopted from Ref. [3]; $\mathbb{E}[V]/V_{max} = 0.47$ with $V_{max} = 120$ km/h and $\mathbb{E}[\tilde{\rho}] = 0.22$. (b) The normalized autocovariance function of the local density displays a characteristic linear decay. The statistical properties of NaSch model allows us to estimate the expected value of velocity $\mathbb{E}[V]$ from the first- and second-order moments of the local density.

2.3.3 Application to Crowdsourced Vehicle Velocity Data

The second application we consider illustrates the use of our approach for crowdsourced velocity data. Here, in contrast to the fixed camera application, sensors are installed in moving vehicles. The vehicle speed data were collected by anonymous users on main roads in the Commonwealth of Massachusetts, USA, through the Carbin Educational App [4,5] which records the velocity from GPS position at a 1 Hz frequency. In order to avoid interference with road regulations such as traffic lights, school zones, speed bumps, and etc., we focus on the data recorded on roads with speed limits $V_{max} \geq 45$ mph (≥ 72.4 km/h). As drivers are observed to drive 10% to 20% faster than the posted speed limit in free flow, this speed limit can be considered as a real-life lower bound of speed in free-flow regime. The conversion of velocity measurements into NaSch-units is thus performed via $v = \lfloor v_{max}(V/V_{max}) + 0.5 \rfloor \leq v_{max}$, where $\lfloor . \rfloor$ denotes the floor operator. We consider this conversion in our analysis of more than 31,000 miles of speed measurements acquired over a time span of one year by anonymous users covering almost the entire main road network of Massachusetts. The length of the time-window, T_m, was checked a priori to satisfy the ergodicity condition [Eq. (2.24)] which is an underlying assumption for our analysis; that is, $T_m - T_r^0 \exp(\Gamma \tilde{H}) \leq 0$ with $\tilde{H}$ denoting the normalized entropy of velocity distribution over time window of T_m. Furthermore, the results of the analysis were checked a posteriori to satisfy the NaSch event space condition, i.e., $\Omega(\overline{v}, \theta_v) \leq 1$, and, moreover, $\kappa_v \approx (1 - \overline{\eta})^{\beta}$. By way of example, Fig. 2-8 displays the analysis of (a) a sample velocity measurement $V(t)/V_{max}$ of a 20-minute trip on interstate highway I-95 together with (b) its conversion into NaSch (velocity) units v; (c) evolution of traffic density ρ and transition density ρ_c^p predicted from the velocity profile; and (d) examples of autocovariance functions at densities higher and close to the transition density, showing the intimate interplay of decay time with traffic density.

The so-obtained results were partitioned into four time intervals: (i) 6:00 to 10:00; (ii) 10:00 to 15:00; (iii) 15:00 to 19:00; and (iv) 19:00 to 00:00; where in each time interval there are at least 10^3 analysis results. By taking into account the probability of

observing memoryless velocity profiles given $v_{max} - \overline{v} \leq 1$, i.e., $\mathbb{P}[\tau_c \approx 0 \mid v_{max} - \overline{v} \leq 1]$, we find that the free-flow probability for time intervals (i) and (iii) is around 10%, whereas for time intervals (ii) and (iv) it increases to almost 15%. This is in agreement with the average weekday daily traffic data reported by the Boston Metropolitan Planning Organization [137]. Fig. 2-9 depicts the geospatial distribution of traffic density and stochasticity parameter for time interval (iii). It is found that the expected traffic density and stochasticity parameter around the urban area of Boston (inset in Fig. 2-9) is respectively 1.3 and 1.15 times the one of rural area (the region outside the red box), implying higher average velocity in rural areas. The network-level expected traffic densities are $\mathbb{E}[\rho] = [0.12, 0.11, 0.13, 0.10]$, and expected stochasticity parameters are $\mathbb{E}[p] = [0.7, 0.63, 0.67, 0.62]$ for time intervals (i) to (iv). The traffic density and stochasticity parameter in time intervals (i) and (iii) are higher than in time intervals (ii) and (iv). This suggests that drivers show a more erroneous driving behavior during rush hours. Furthermore, the inferred traffic parameters imply that, in an average sense, traffic is predominantly in the congested flow regime.

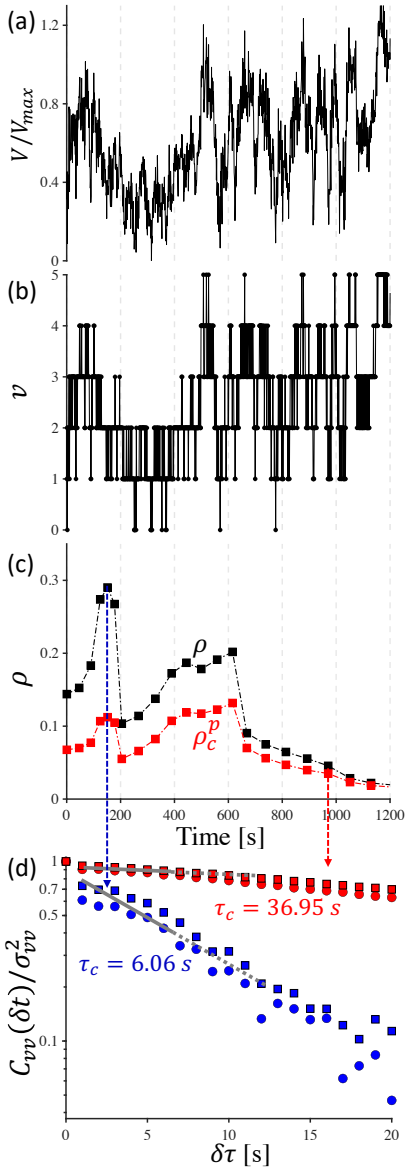

Figure 2-8: Sample analysis: (a) Velocity time history of a vehicle driving on I-95, a north–south interstate highway in MA, USA, (b) NaSch representation of velocity profile, (c) evolution of inferred traffic density ρ and transition density ρ_c^p, and (d) autocovariance functions at two times (squares and circles represent the autocovariance of velocity measurement and its NaSch representation, respectively). The initial slope of the autocovariance function is inversely proportional to τ_c [Eq. (2.9)]; memory of velocity signal is shorter at the higher density.

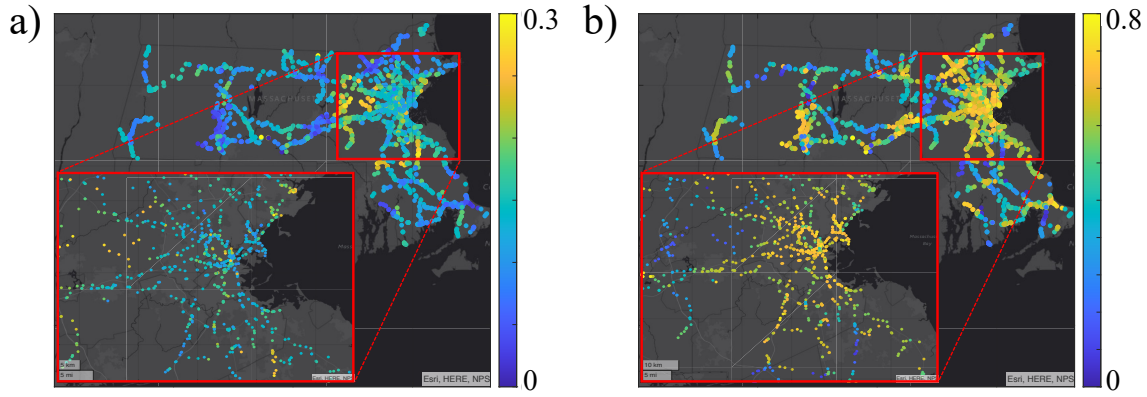

Figure 2-9: Geospatial distribution of (a) traffic density ρ and (b) stochasticity parameter p for the time interval 15:00 to 19:00 pm, on main roads in Massachusetts, USA, determined from crowdsourced 1 Hz vehicle velocity recordings (data collected over a 12-month period with Carbin Educational App [4, 5]).

2.4 Summary

In this chapter, we have demonstrated quantifiable nature of the internal structure of traffic by investigating the system's higher-order statistical properties. We have leveraged the ergodic theorem, a fundamental concept from statistical physics, demonstrating that the fluctuations in speed profile experienced in congested flow can effectively characterize the statistical properties of the collective behavior of the traffic flow as an ensemble. Specifically, we have highlighted that the memory effects in both space and time, expressed by second-order moments of occupancy and velocity hold critical information relevant for the spatial and temporal mapping of traffic density and driver behavior, which can be assessed from individual driver velocity recordings provided ergodicity and stationary. More precisely, the two-point correlation function of occupancy enables us to understand spatial memory effects, such as headway; while the velocity autocovariance function sheds light on temporal memory effects, in terms of decay time and traffic compressibility. In particular, the two-point correlation function of occupancy provides access to spatial memory effects, such as headway; whereas the velocity autocovariance function provides access to temporal memory effects in form of the decay time and traffic compressibility. Taken together, the isotherms provide a means to access traffic density and stochasticity from density-stochasticity plots. The fact that these higher-order statistical moments are directly accessible by crowdsourced velocity measurements provides a powerful alternative to classical traffic property estimates from spatially distributed user counts. Ultimately, these moments quantify the internal structure of traffic and present a wealth of information which can characterize intricate interactions within the system, thereby allowing for the quantification of various statistical properties of the system.

Chapter 3

Phase Diagram of Near-Miss Collision Risk in Congested Traffic Flow

Within the complex interaction of factors governing the dynamics of traffic, first-order descriptors such as average traffic density, velocity, flux, as well as various geometric attributes of the roadway network, struggle to accurately quantify the accident risk inherent to the internal structure of traffic. Traditionally, in fact, studies have attributed the occurrence of accidents to extraneous variables, such as global fluctuations in traffic density, weather conditions [138,139], geometric designs of roads and crossings design parameters [140,141], or erroneous driver behavior [142,143]. Instead of investigating the occurrence of accidents as an externality of traffic flow, we investigate the risk of near-miss events from the internal structure of traffic. Specifically, we propose an approach which is driven by the consideration that if such an accident risk intrinsic to the structure of traffic exists and is significant, it should be predictable and preventable, much akin to the probability of molecular collisions in gas physics from pressure and kinetic temperature.

In this chapter, we investigate the near-miss event statistics in vehicular traffic, i.e., the probability of accident precursors that have the potential to cause collision, but actually do not result in human injury, vehicle damage and interruption of operation (for a generic definition of near-miss events, see [144]). We propose a probability definition for this near-miss collision risk using the velocity state transition matrix

as the sole input. We corroborate our proposition by utilizing both model-based traffic flow simulations and extensive crowdsourced measurements of vehicle velocities collected via smartphones. Furthermore, using the network-level dataset of accident records, we illustrate the positive correlations between this definition of near-miss events with the actual collision risk.

3.1 Statistics of Near-miss Events in the Reduced-unit Representation

Our starting point is the common-sense consideration that the statistics of near-miss events, as collision-free accident precursors, are characterized by decelerations in excess of normal operation braking events - from maximum speed to zero. To simplify the analysis, we employ a reduced-unit representation of traffic in NaSch units [Fig. 3-1.(a)], after the Nagel-Schreckenberg cellular automaton model [35]. In such representation, time and space are both discretized quantities and the units of speed and position coincide. The state of a particle is represented by its velocity which is a non-negative integer not greater than v_{max}, i.e., $v_k(t) \in \{0, 1, ..., v_{max}\}$ with $v_k(t)$ representing the velocity state of k^{th} vehicle at time t. The likelihood of observing an excessive deceleration, $a_k(t) = -v_{max}$ with acceleration $a_k(t) = v_k(t+1) - v_k(t)$, for the k^{th} vehicle transitioning from a current velocity state $v_k(t)$ to a future velocity state $v_k(t+1)$ is derived from the law of total probability [Appendix D], in the form:

$$\mathbb{P}[a_k(t) = \theta | j = -\theta = v_{max}] = \pi_j \, q_{j \to j+\theta} \tag{3.1}$$

Herein, $\pi_j(t) := \mathbb{P}[v_k(t) = j]$ stands for the probability of observing a vehicle at velocity state j, whereas tranisition probability $q_{j \to j+\theta} := \mathbb{P}[v_k(t+1) = j+\theta \mid v_k(t) = j]$ represents the probability of finding a particle transitioning from state j to state $j+\theta$. It is important to emphasize that Eq. (3.1) broadly remains valid under various statistical conditions and remains applicable irrespective of the specific traffic model selected, provided that the model employs the reduced-unit representation. In what

follows, we explain that, given the Markovian nature of traffic, the probability of observing maximum deceleration can be quantified using the statistics derived from velocity signals. This will turn out useful in determining the structural patterns of traffic in a broader context using crowdsourced velocity measurements.

3.1.1 Markovianity of Velocity State

We resort ourselves to steady-state homogeneous traffic flow on simple road networks. Under these circumstances, velocity state exhibits stationarity; that is, in strict sense, the joint probability of the process remains insensitive with respect to time lag [Appendix B]. Furthermore, velocity state in such traffic flow is essentially of Markovian nature, hence the *future* state of the system at time $t+1$ only depends on the *current* state at time t. As a result, the state space of the system is entirely defined by the velocity state transition matrix, $\underline{q} \equiv q_{i \to j}$, where row i and column j represent the current and future states of the particle, respectively. Otherwise said, the state transition matrix $\underline{q}$ contains all the information required to evaluate the statistics of near-miss events from Eq. (3.1). To this end, we are particularly interested in expressing the moments of velocity as a function of the transition matrix. In what follows, we illustrate the derivation of the expected value and autocovariance from $\underline{q}$. While it goes beyond the scope of this work, higher-order moments can be obtained is a similar fashion.

We proceed by relating the probability distribution function of velocity to the transition matrix, which proves beneficial in deriving the mean and variance of velocity. First, using the Markov property, the probability distribution of velocity at time $t+\tau$ can be related to that of time t via,

$$\vec{\pi}(t+\tau) = (\underline{q}^T)^\tau \cdot \vec{\pi}(t) \tag{3.2}$$

with,

$$\vec{\pi}(t) = (\pi_0(t), \pi_1(t), ..., \pi_{v_{max}}(t))^T \tag{3.3}$$

where $\vec{\pi}(t)$ and $\vec{\pi}(t+\tau)$ are, respectively, the discrete velocity probability distribution at time t and $t+\tau$, linked by the transition matrix $\underline{q}$.

We recall statistical stationarity, induced by homogeneous particle randomness and a closed-system behavior, which leads to a constant velocity probability distribution at all times, i.e., $\vec{\pi}(t+\tau) = \vec{\pi}(t) = \vec{\pi}$. By applying the Perron–Frobenius theorem, the velocity probability distribution $\vec{\pi}$ is the eigenvector of the transposed transition matrix $\underline{q}^T$ associated with the unit eigenvalue; that is,

$$(\underline{q}^T - \underline{I}) \cdot \vec{\pi} = \vec{0} \tag{3.4}$$

where $\underline{I}$ is the identity matrix [Appendix D]. Thus, the velocity distribution corresponds to the unit eigenvalue of the transposed state transition matrix, the so-called stationary distribution of $\underline{q}^T$ [145, 146]. Having the velocity distribution function, we can easily compute the mean and variance of the process as $\bar{v} = \vec{s}^T \cdot \vec{\pi}$, and $\sigma_v^2 = \vec{s}^T \cdot \mathrm{diag}(\vec{s}) \cdot \vec{\pi} - \bar{v}^2$ where $\vec{s} = (0, 1, ..., v_{max})^T$ stands for the state vector [Fig. 3-1.(a)].

We now shift our focus to deriving the second-order moment of velocity, expressed as a function of the transition matrix. The Markovian property of velocity allows us to derive the autocovariance function, which in its canonical definition reads $C_{vv}(\tau) = \mathbb{E}[(v_k(t_1) - \bar{v})(v_k(t_2) - \bar{v})]$. Essentially, Markovianity quantifies the correlation between successive steps directly. Expanding upon this concept, treating time $t+1$ as an intermediate state with certain state probabilities allows for determining the correlation between velocities at times t and $t+2$. By iterating this process, one can analytically obtain the correlation between for any time lag τ, and thus derive a dual definition of the velocity autocovariance function (for detailed derivation, refer to Appendix E),

$$C_{vv}(\tau) = \vec{s}^T \cdot (\underline{q}^\tau)^T \cdot \mathrm{diag}(\vec{s}) \cdot \vec{\pi} - \bar{v}^2 \tag{3.5}$$

which establishes a link between the transition matrix and the second-order central moment. In a similar fashion, higher-order statistics can be obtained from the wealth of infromation that the transition matrix provides for a Markovian process.

3.1.2 From Velocity Signal to Near-Miss Risk

The reduced-unit model of traffic flow, combined with the Markovian property, therefore enables us to describe the statistics of acceleration and, in particular, near-miss events using the state transition matrix. Specifically, among all possible transition probabilities [see Fig. 3-1.(a)], the transition matrix includes the extremal deceleration transition probability $q_{v_{max}\to 0}$. Moreover, the probability of observing a vehicle at the highest velocity state, $\pi_{v_{max}}$, is ascertained from the velocity probability distribution, $\vec{\pi}(t) = (\pi_0, \pi_1, ..., \pi_{v_{max}})^T$ [Fig. 3-1.(b)], which is the eigenstate associated with the unit eigenvalue of the transposed transition matrix. Finally, we remind ourselves that all statistical measures of velocity (here, we only focus up to the second order) can be expressed in terms of the state transition matrix, i.e., mean ($\bar{v}$), variance (σ_v^2), and velocity autcovariance function (C_{vv}), as follows,

$$\begin{cases} \bar{v} & = \mathbb{E}[v_k(t)] = \vec{s}^T \cdot \vec{\pi} \\ \sigma_v^2 & = \mathbb{E}[(v_k(t) - \bar{v})^2] = \vec{s}^T \cdot \mathrm{diag}(\vec{s}) \cdot \vec{\pi} - \bar{v}^2 \\ C_{vv}(\tau) & = \mathbb{E}[v_k(t_1)v_k(t_2)] - \bar{v}^2 \\ & = \vec{s}^T \cdot (\underline{q}^\tau)^T \cdot \mathrm{diag}(\vec{s}) \cdot \vec{\pi} - \bar{v}^2 \end{cases} \tag{3.6}$$

where $\tau = |t_1 - t_2|$ is the time lag between two velocity states, $v_k(t_1)$ and $v_k(t_2)$.

By leveraging the ergodicity of velocity as a stochastic process, it becomes evident that all statistical moments, derived in terms of the state vector and transition matrix, can actually be acquired from the velocity time series over a sufficiently long time interval. Specifically, by matching the expected velocity, fluctuations about the mean, and correlation levels of velocity for various time lags, the state transition matrix can be determined. This, in turn, enables the complete characterization of the process as a Markovian process, and further allows us to calculate the near-miss risk by focusing on the maximum deceleration, i.e., $q_{v_{max}\to 0}$ and $\pi_{v_{max}}$.

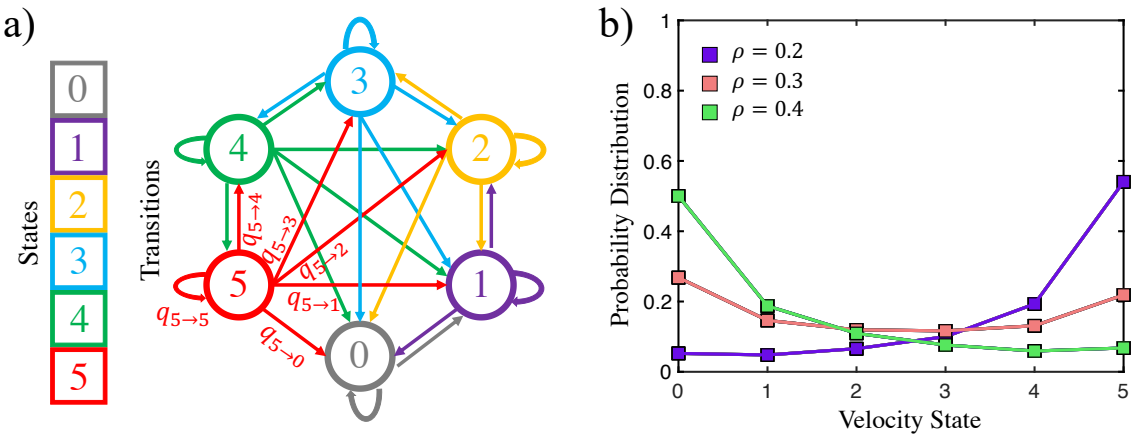

Figure 3-1: (a) Graphical representation of velocity state vector $\vec{s}^T = (0, 1, ..., v_{max})$ (in NaSch-units), and transition matrix components $q_{i\to j}$ for $v_{max} = 5$. (b) Velocity distribution $\vec{\pi} = \vec{\pi}(\vec{s})$ obtained from the eigenstate analysis of the transition matrix, $\underline{q} \equiv q_{i\to j}$ for congested flow below, at and above the critical jamming probability, $\bar{\eta}_{crit}$, at which the deceleration probability $\mathbb{P}[a_k(t) = v_{max}]$ exhibits a maximum.

3.2 Application to the NaSch Model

We analyze general aspects of near-miss collision risk using velocity data generated by the NaSch cellular automaton model due to its capacity to replicate key traffic flow features. These features include backward-moving shock waves, the first-order fundamental diagram, phase transitions from free to congested flow, formation and dissolution of traffic jams similar to dynamic arrest, interactions between velocities of consecutive vehicles, and temporal and spatial memory effects [Appendix A]. Here, we analyze the velocity state transition matrix from the ensemble statistics of particle velocities generated by the (Markovian) heuristics of the NaSch model, which we remind ourselves: (1) acceleration, $v_j^{(1)} = \max(v_k(t), v_{max})$; (2) deceleration, $v_j^{(2)} = \min(v_j^{(1)}, d_k(t))$ with $d_k(t)$ representing the headway; (3) randomized braking, $v_k(t+1) = v_j^{(2)}$ with probability $1-p$ and $v_k(t+1) = \max(v_j^{(2)} - 1, 0)$ with probability p; (4) an update of the k^{th} vehicle position, $x_k(t+1) = x_k(t) + v_k(t+1)$, where the stochasticity parameter p takes into account the drivers' individual freedom to regulate their speed.

As a first point of reference, we recognize that neither the free flow regime nor the jammed regime exhibit extreme decelerations (see Appendix F), since:

$$\lim_{\rho/\rho_c^p \to 0} q_{v_{max}\to 0} = 0; \;\; \lim_{\rho \to 1} \pi_{v_{max}} = 0 \tag{3.7}$$

with ρ_c^p the phase transition density. Essentially, in the free flow, a vehicle travels at $v_{max} > 1$ with a probability of $(1-p)$, and at a velocity of $v_{max} - 1$ with a probability of p. Therefore, there are no vehicle transitioning from v_{max} to zero velocity. Conversely, as the system approaches a completely jammed state, nearly all vehicles are at standstill state, and the probability of observing a vehicle traveling at v_{max} diminishes to zero. In both these limiting scenarios, the probability of observing maximum deceleration is zero.

The intrinsic probability of a near-miss collision is hence confined to the congested flow regime, wherein any individual particle has an equal likelihood of being part of a traffic jam. We map the near-miss collision risk, considering that traffic

flow in the congested regime exhibit temporal memory effects akin to the Ornstein-Uhlenbeck random process for which the autocovariance funtion decays exponentially as $dC_{vv}/d\tau|_{\tau=0} = -\sigma_v^2/\tau_c$ [130, 131] where the characteristic decay time τ_c is the so-called temporal memory [Chapter 2]. The velocity state transition matrix, $\underline{q}$, can thus be determined from the measured velocity variance σ_v^2, and the decay time τ_c:

$$\min_{\underline{q}} \sum_{\tau=0}^{\infty} \left[\left(\vec{s}^T (\underline{q}^\tau)^T \mathrm{diag}(\vec{s}) \vec{\pi} - \bar{v}^2 \right) - \sigma_v^2 \exp(-\tau/\tau_c) \right]^2 \tag{3.8}$$

subject to $(\underline{q}^T - \underline{I}) \cdot \vec{\pi} = \vec{0}$ and $\underline{q} \cdot \vec{1} = \vec{1}$.

3.2.1 Deterministic Model: $p = 0$

For the deterministic model ($p = 0$), we use the analytical expressions for the binary-state NaSch model obtained from 2-cluster solution [147], to derive generalized solutions for the multi-state model ($v_{max} > 1$). The transition probability $q_{v_{max}\to 0}$ exhibits a universal pattern, analogous to the behavior of the two-state NaSch model, as a function jamming probability $\bar{\eta}$ [Appendix F.3.1],

$$q_{v_{max}\to 0} = \frac{2\bar{\eta}}{1+\bar{\eta}}, \qquad (p = 0) \tag{3.9}$$

Similarly, the behavior of likelihood of observing maximum velocity for multi-state model can also be derived through power generalization and satisfying the boundary conditions of probability at $\bar{\eta} = 0$ and $\bar{\eta} = 1$ as,

$$\mathbb{P}[v_k(t) = v_{max}] = 1 - \beta \left((1 + v_{max}\bar{\eta})^{1-\alpha} - 1 \right), \qquad (p = 0) \tag{3.10}$$

with $\beta = (1 + v_{max})^\alpha / (1 + v_{max} - (1 + v_{max})^\alpha)$ and the fitting parameter $\alpha \approx 2 + \tanh\sqrt{(v_{max} - 1)/2}$ [Fig. 3-2]. The probability of observing near-miss event for the determinsitc model can readily be obtained at any level of congestion as the multiplication of Eqs. (3.9) and (3.10) since $\mathbb{P}[a_k(t) = -v_{max}] = q_{v_{max}\to 0}\pi_{v_{max}}$.

The asymptotic behavior of near-miss risk can be comprehended as follows: $q_{v_{max}\to 0}$

suggests that as $\bar{\eta} \to 0$, the system's congestion diminishes, nudging the transition probability towards that of the free-flow regime, i.e., $q_{v_{max}\to 0} \to 0$. In contrast, as the system leans towards a fully jammed regime with $\bar{\eta} \to 1$, the transition probability edges closer to one. Conversely, $\mathbb{P}[v_k(t) = v_{max}]$ displays the inverse asymptotic behavior; as the system inclines towards a free-flow regime, the probability of observing v_{max} approaches 1, whereas in a fully jammed system, $\mathbb{P}[v_k(t) = v_{max}]$ tends towards zero. These findings imply that the occurnece likelihood of a near-miss event is primarily influenced by the transition probability as $\bar{\eta} \to 0$ and by the frequency of maximum velocity as $\bar{\eta} \to 1$. In a more formal context, as $\bar{\eta}$ approaches 0, the transition density $q_{v_{max}}$ becomes very small, $q_{v_{max}} = \epsilon_0$. At the same time, observing v_{max} becomes very probable, with $\pi_{v_{max}} = 1 - \epsilon_1$. Here, $\epsilon_{0,1} << 1$ represent very small positive numbers. This relationship allows us to approximate the probability of maximum deceleration as $\mathbb{P}[a_k(t) = -v_{max}] \approx q_{v_{max}}$. A similar rationale can be used to demonstrate that $\mathbb{P}[a_k(t) = -v_{max}] \approx \mathbb{P}[v_k(t) = v_{max}]$ as $\bar{\eta} \to 1$.

3.2.2 Stochastic Model: $p \neq 0$

The asymptotic behavior of the likelihood of observing maximum velocity at boundaries for stochastic multi-state NaSch model reads [Appendix F.3.2],

$$\lim_{\bar{\eta}\to 0} \mathbb{P}[v_k(t) = v_{max}] = 1 - \sqrt{p}; \quad \lim_{\bar{\eta}\to 1} \mathbb{P}[v_k(t) = v_{max}] = 0 \tag{3.11}$$

Decoupling the impact of randomness while satisfying the asymptotic behavior behavior, one can approximate the likelihood of observing maximum velocity in the congested regime as,

$$\mathbb{P}[v_k(t) = v_{max}|p \neq 0] \approx (1 - \sqrt{p}) \left(\mathbb{P}[v_k(t) = v_{max}|p = 0]\right)^{\xi} \tag{3.12}$$

with exponent $\xi = \xi_0 + \xi_1 \tanh(v_{max} - 1)$ where $\xi_1/\xi_0 \approx .85$ and $\xi_0 = 0.77$. The likelihood of observing a stationary vehicle and the transition probability are estimated

using power generalization of the 0-1 model, given as:

$$\mathbb{P}[v_k(t) = 0] \approx 1 - \mathbb{P}[v_k(t) = v_{max}]^{1-\nu_v} \tag{3.13}$$

$$q_{v_{max}\to 0} \approx 1 - \mathbb{P}[v_k(t) = v_{max}]^{1-\nu_q} \tag{3.14}$$

where $\nu_v = \hat{\nu}_v \tanh(v_{max} - 1)$ and $\nu_q = \hat{\nu}_q \tanh(v_{max} - 1)$ as depicted in Fig. 3-3.(b).

We proceed by exploring the intricate interaction between the transition density $q_{v_{max}\to 0}$, and probability of observing a vehicle at velocity states 0 (standing) and v_{max}. In particular, the probabilities of observing a standing vehicle and the transition probability are approximated via power generalization as,

$$\mathbb{P}[v_k(t) = 0] \approx 1 - \mathbb{P}[v_k(t) = v_{max}]^{1-\nu_v} \tag{3.15}$$

$$q_{v_{max}\to 0} \approx 1 - \mathbb{P}[v_k(t) = v_{max}]^{1-\nu_q} \tag{3.16}$$

with $\nu_v = \hat{\nu}_v \tanh(v_{max} - 1)$ and $\nu_q = \hat{\nu}_q \tanh(v_{max} - 1)$ [Fig. 3-3.(b)].

Equations (3.15) and (3.16) capture the variation in near-miss risk as a function of both density and stochasticity parameter. Similar to the deterministic system, the transition density is leading term of the near-miss risk as $\bar{\eta} \to 0$. Conversely, the probability of v_{max} emerges as the primary factor when $\bar{\eta} \to 1$, satisfying the limiting behaviors of near-miss event [Eq. (3.7)].

3.2.3 Phase Diagram of the NaSch Model

We proceed by focusing on the congested phase in between the limiting cases of free flow and jammed regime. By sweeping the NaSch phase space (ρ, p), we map the near-miss collision risk of the NaSch model onto a phase diagram [Fig. 3-4.(a)]. Two observations deserve particular attention. First, we observe a ridge along which the near-miss collision risk exhibits a maximum. This ridge is indicative of a critical behavior of near-miss collisions close to a critical jamming probability $\overline{\eta} = (\rho - \rho_c^p)/(1 - \rho_c^p) = \overline{\eta}_{crit}$, where $\partial \mathbb{P}[a_k(t) = -v_{max}]/\partial \bar{\eta} = 0$ (see Fig. 3-4.(b), and detailed discussion

in Appendix F). The critical jamming probability indicates that highest likelihood of maximum deceleration lies within the congested regime. Second, we find a counter-intuitive response to heightened levels of stochasticity in driver behavior. Instead of an increase of the near-miss collision risk, we observe a decrease in the maximum near-miss collision risk, $\max_{\bar{\eta}}(\mathbb{P}[a_k(t) = -v_{max}])$, with increase in stochasticity p [Fig. 3-4.(c)]. In fact, increasing stochasticity leads to larger traffic jams, and thus decreases the number of clusters, and promotes closer alignment of vehicle velocities [148]. Specifically, we attribute this diminishing risk to the long-range correlations triggered by random velocity fluctuations. These long-range correlations increase the formation of traffic jams, lower the average ensemble velocity, $\bar{v} \sim 1 - p$, and entail a lower probability of both $\pi_{v_{max}}$ and $q_{v_{max}\to 0}$. As a consequence, the near-miss collision risk (3.1) decreases from its maximum deterministic limit at $p = 0$ [Fig. 3-4.(b)], for which clustering effects are negligible, to zero in the limit case of $p \to 1$ [Fig. 3-4.(c)], for which traffic comes to a halt due to erroneous braking of all particles in the system.

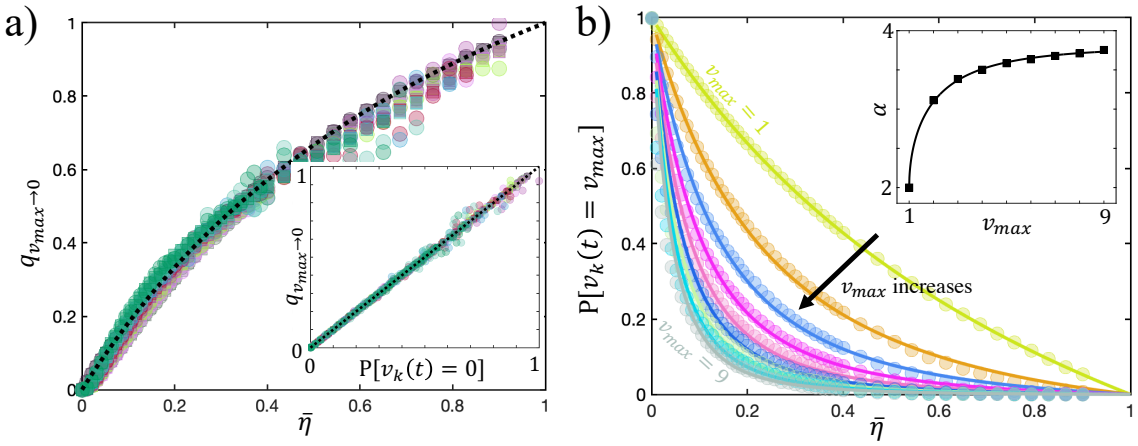

Figure 3-2: Behavior of the deterministic NaSch model: (a) the universal increasing pattern of the transition probability $q_{v_{max}\to 0}$ as a function of jamming probability [Eq. (F.11)] with different colors representing different $v_{max} \in \{1, 2, ..., 9\}$ (circles: simulations, squares: far-field behavior); the inset shows the duality between the transition probability $q_{v_{max}\to 0}$ and the probability of observing standing vehicle: $q_{v_{max}\to 0} = \mathbb{P}[v_k(t) = 0]$. (b) Decreasing pattern of probability of observing a vehicle driving at v_{max} as jamming probability increases (circles: simulation, lines: semi-analytical solution [Eq. (F.13)]); the inset shows the power scaling of maximal velocity probability $\alpha \approx 2 + \tanh\sqrt{(v_{max}-1)/2}$.

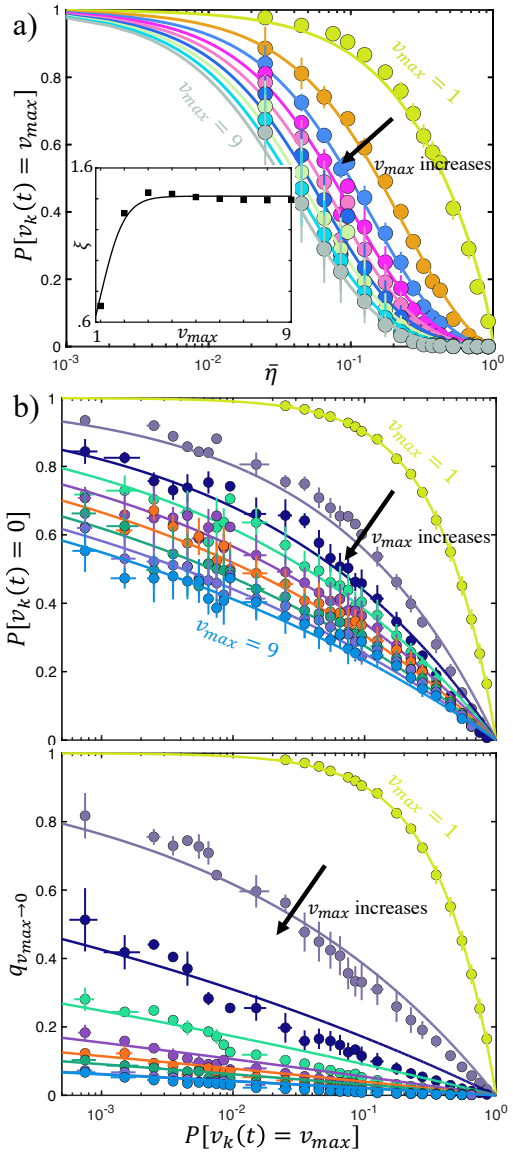

Figure 3-3: Approximate behavior of the stochastic NaSch model for different levels of v_{max}: (a) approximation of the probability of observing maximum velocity through decoupling the impact of stochasticty parameter and jamming probability (the inset shows the behavior of the exponent of the approximation $\zeta = \zeta_0 + \zeta_1 \tanh(v_{max} - 1)$ with $\zeta_0 = .77$ and $\zeta_1 = .65$), and (b) the universal behavior of probability of observing standing vehicle and transition density as a function of probability of observing maximum velocity; the bars represent standard deviation and the lines are semi-analytical power approximations: $\mathbb{P}[v_k(t) = 0] \approx 1 - \mathbb{P}[v_k(t) = v_{max}]^{1-\nu_v}$ and $q_{v_{max}\to 0} \approx 1 - \mathbb{P}[v_k(t) = v_{max}]^{1-\nu_q}$ with $\nu_v = .83 \tanh(v_{max} - 1)$ and $\nu_q = \tanh(v_{max} - 1)$.

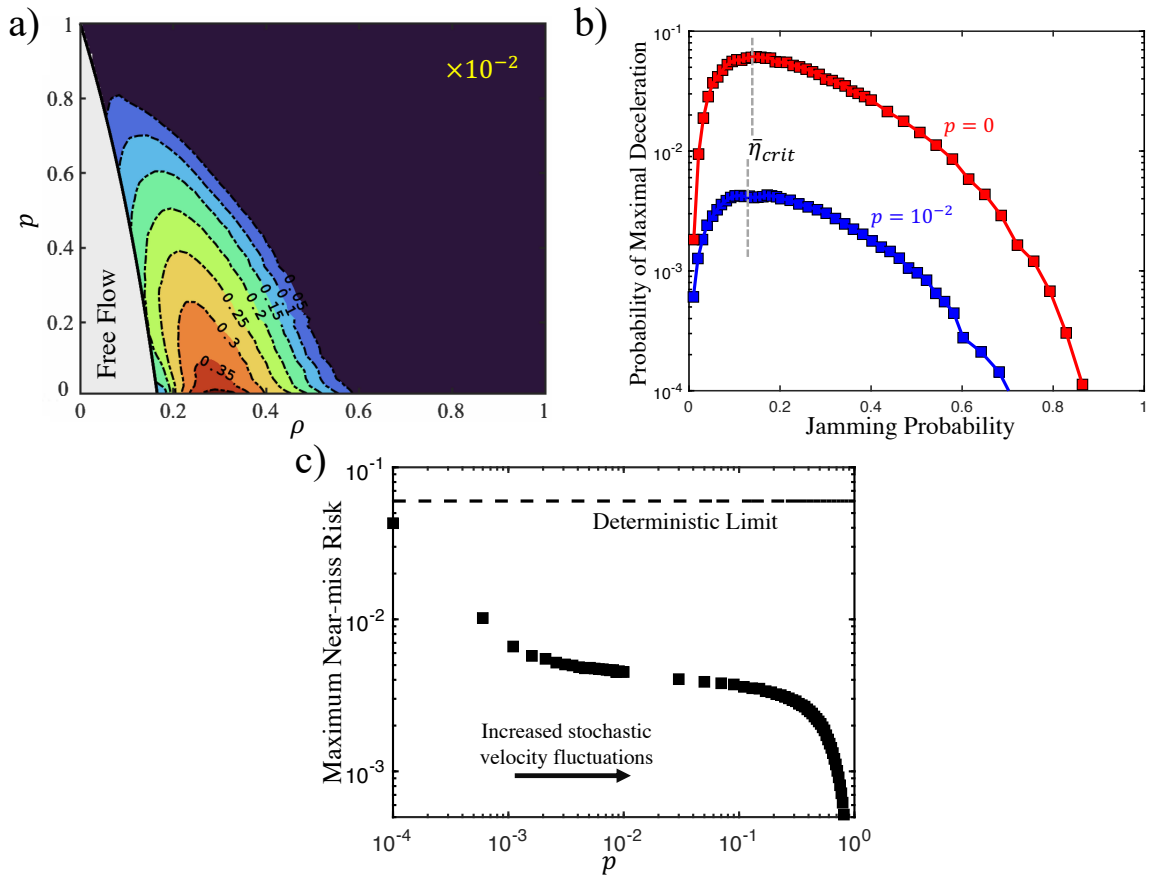

Figure 3-4: Model-based phase diagram of near-miss collision risk $\mathbb{P}[a_k(t) = -v_{max}]$ in (a) the phase space of the NaSch model (traffic density ρ, stochasticity p); and (b) in function of the jamming probability $\overline{\eta} = (\rho - \rho_c^p)/(1 - \rho_c^p) \in]0, 1[$. (c) Maximum near-miss collision risk, $\max_{\overline{\eta}}(\mathbb{P}[a_k(t) = -v_{max}])$, vs. stochasticity parameter p showing the decrease in near-miss collison risk due to long-range correlations induced by randomness (Results for $v_{max} = 5$. For other v_{max} values, see Appendix F.)

3.3 Application to Crowdsourced Vehicle Velocity Data

We now turn to investigate these model-based features using crowdsourced vehicle velocity data recorded between 2019 and 2021 at 1 Hz frequency by anonymous users via the Carbin educational app [149]. After data processing to ensure statistical stationarity and ergodicity (see Appendix D), we considered a data set of approximately 8×10^5 trips of 400 seconds covering approximately 60% of the roadway network of the Commonwealth of Massachusetts, USA. For each trip, the near-miss collision risk was determined from the state transition matrix using as sole input the velocity autocovariance function [Eq. (3.6)] independent of any model attributes (for algorithmic implementation, see Appendix H). In addition, the near-miss risk was compared with the expected collision risk derived from the vehicular accident database of the Massachusetts Department of Transportation [6], which has detailed information of all accidents between 2019 and 2021, incl. accident location, road properties (number of lanes and road type), Annual Average Daily Traffic (AADT), number of cars involved, and severity of the accident.

3.3.1 Direct Comparison: Near-miss vs. Collision

Our data analysis centers around a graph representation of the primary roadway network in the state of Massachusetts, USA [150]. The graph representation consists of 195,138 nodes and 196,906 edges, with each edge and node representing a road segment and intersection (or endpoint of a road segment), respectively. By analyzing the Carbin dataset for the state of Massachusetts and utilizing crowdsourcing, we assign an expected near-miss risk to each edge [Fig. 3-5.(a)]. Additionally, we accessed the crash data set provided by the Massachusetts Department of Transportation (MassDOT) [6], which includes detailed information about all the accidents that occurred between 2019 and 2021, such as the accident's location, number of vehicles involved, traffic volume (e.g., AADT), and road properties. For each edge, we calculated the

collision risk defined as the number of accidents on an edge divided by the total volume of traffic per length [Fig. 3-5.(b)]. A primary insight is offered by the positive correlation between near-miss and actual collision risk from a segment-by-segment comparison [Fig. 3-6.(a)]. This positive correlation suggests an association between the proposed near-miss risk and the actual collision risk, positioning the former as a potential accident precursor. Specifically, upon comparing the two parameters, i.e., near-miss and collision risks, we observed a positive correlation between them, with the strongest correlation evident for roads that have an expected traffic density of $\rho \approx .2$ [Fig. 3-5.(c)]. This observation has two important implications. First, it highlights the potential for near-miss events to escalate into actual collisions. Second, the identification of a "critical regime" regime, where the likelihood of accidents transcends the realm of erroneous driving behavior and encompasses the internal structure of traffic, is of crucial importance. This highlights the need for a deeper understanding of the complex interplay between traffic dynamics and driver behavior in order to develop more comprehensive and effective strategies for accident prevention.

3.3.2 Clustering Analysis

To gain a more detailed understanding of the data, we implement clustering analysis—a method widely employed within the physics community for deciphering phase transitions and critical phenomena allowing for the quantification of "clusters" or groups of similar elements within a many-body system, with applications in various areas such as percolation theory, cosmology, and structure identification. A refined picture emerges from a clustering analysis of the data considering traffic density, near-miss and crash probabilities [Figs. 3-6.(b)]. Particularly, the Gaussian mixture approach permits identifying two distinct clusters (phases) separating the data into a low traffic density phase ("Cluster I" in Fig. 3-6.(b)) and a high density phase ("Cluster II" in Fig. 3-6.(b)), which affirms the existence of the fundamental diagram of traffic at network scale [72]. Moreover, it is at the interface between the two phases that both the highest near-miss collision risk and the highest actual collision risk occurs [Fig. 3-6.(c)]. This finding is significant for two reasons. First, the overlap of low and

high density regime can be viewed as a critical regime, in which accident precursors predicted by the intrinsic near-miss collision risk have the highest likelihood to turn into actual accidents. Second, the fact that this high probability occurs at a traffic density of $\rho \approx 0.2$ [Figs. 3-6.(b-c)], is not a coincidence, but a clear evidence that (near-miss and actual) collisions are predominant in the congested flow regime at a critical jamming probability, in full agreement with the model-based phase diagram [Fig. 3-4]. The results thus confirm the existence of an intrinsic near-miss collision risk which is significant in its strong correlations with the actual crash probability.

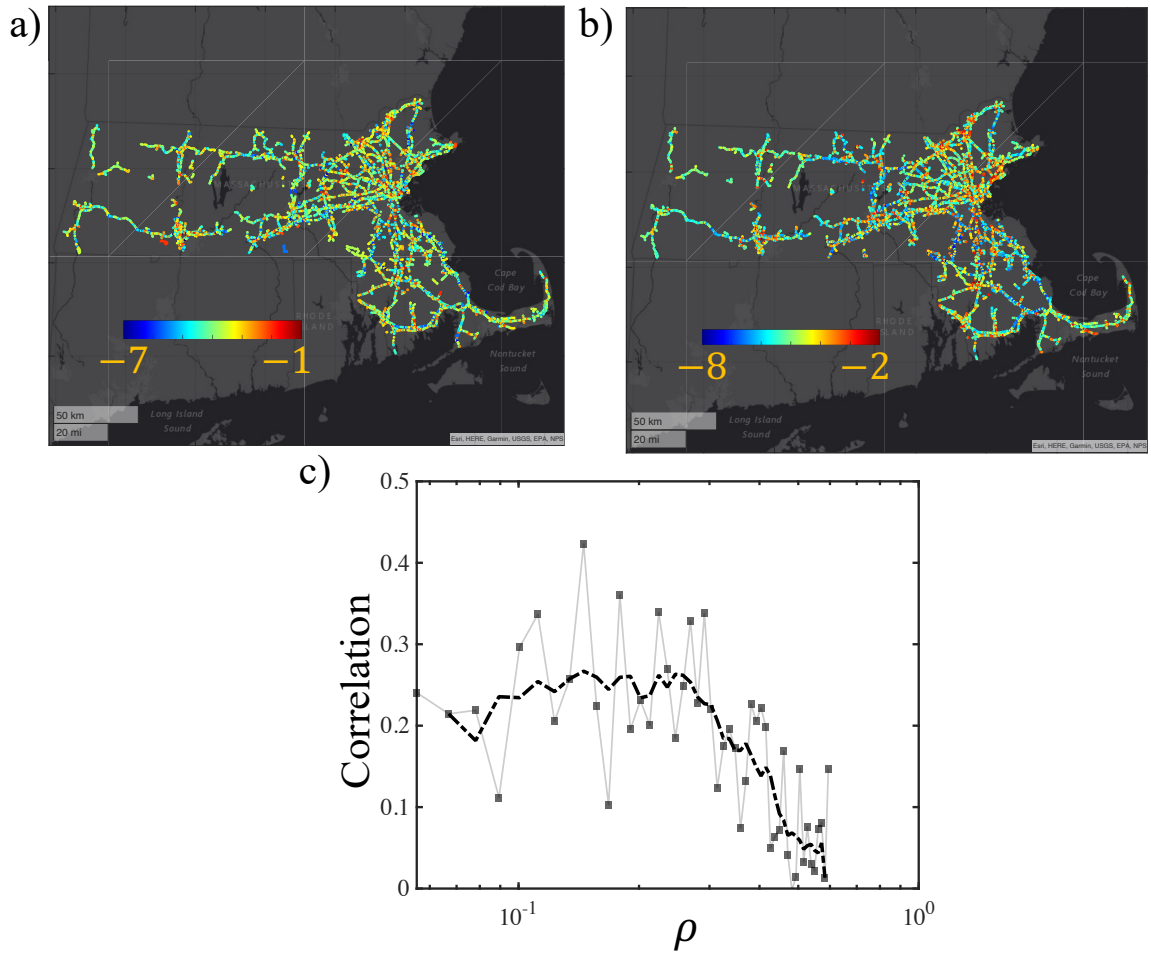

Figure 3-5: The geo-spatial distribution of the logarithm of the expected (a) near-miss risk, and (b) collision risk. (c) The correlation between the near-miss and collision risk as a function of traffic density ρ. The road segments with a density of approximately $\rho = 0.2$ exhibit the highest correlation levels, although the correlation is consistently positive.

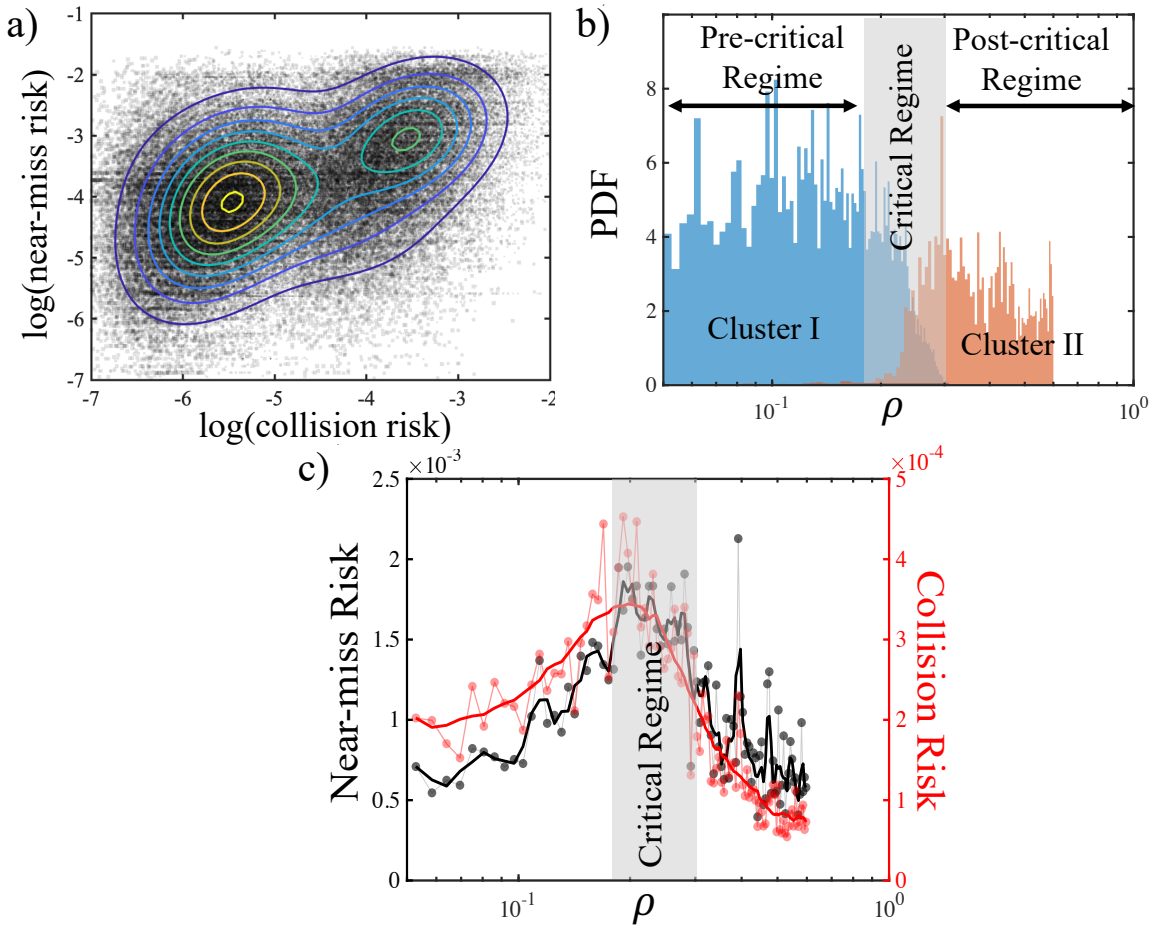

Figure 3-6: Phase diagram of near-miss and actual collision risk: (a) Correlation plot of near-miss risk derived from crowdsourced velocity measurements and actual collision risks derived from 2019-21 accident data for the Commonwealth of Massachusetts [6] (size of squares is proportional to traffic density). (b) Gaussian clustering of traffic density, near-miss risk and collision probability showing the existence of two dominant clusters associated with traffic density interfaced by a critical regime. (c) Phase diagram of near-miss and actual collision risk highlighting the critical regime in which accident precursors predicted by the intrinsic near-miss collision risk have the highest likelihood to turn into actual accidents.

3.4 Summary

In summary, considering a long-standing question in traffic safety, curiously never explicitly addressed before, we have uncovered a near-miss collision risk in congested traffic flow that is an integral part of the internal structure of traffic. This intrinsic nature makes the risk predictable using vehicle velocity data as only input to determine the velocity state transition matrix and to subsequently map model-based generated or crowdsourced vehicle velocity data onto a phase diagram of accident precursors. The phase diagram highlights that the intrinsic near-miss collision risk is confined to the congested regime, where accident precursors have the highest likelihood to turn into actual accidents. In return, jamming or an increase in randomness in driver behavior entail long-range interactions between particles that tend to diminish the intrinsic near-miss collision risk. The intrinsic near-miss collision risk in traffic flow thus appears to have similar features as many other many-body systems with long range correlations triggered by randomness, ranging from fracture risk in random porous materials [151], and the collective behavior of coupled pendulums with random connections [152], to the control of extreme events in complex networks through randomness [153, 154] and the moderation of wild price fluctuations in financial markets via random trading agents [155]. Of course, other externality-driven vehicle accidents which are not part of our near-miss event statistics, need to be considered in the development of a comprehensive accident risk mitigation strategy. We believe however, that the availability of a predictable metric of the intrinsic near-miss collision risk is a first step to curbing the incessant rise in vehicle-caused fatality rates. True to this goal, we expect that the scalability of our approach for crowdsourced vehicle velocity will enable development and implementation of active collision risk mitigation strategies in both human-driven and autonomous driving networks.

Chapter 4

Network-Level Safety Analysis: Application of Near-Miss Risk Analysis to Crowdsoruced Data

Advancements in data analytics and collection methodologies have catalyzed a paradigm shift towards data-centric approaches in the analysis of vehicular accident patterns [82–85]. These approaches utilize statistical methods and machine learning techniques to scrutinize multidimensional collision datasets, unraveling critical parameters that strongly correlate with collision incidents, e.g., such as driver fatigue, weather conditions, traffic flow, and road geometry [86–88, 95–97]. The predictive modeling which can identify high-risk zones, however, is mainly anchored in retrospective case-control studies [95, 96, 101–105]. While helpful in pinpointing high-risk areas, the existing data-centric studies largely neglect the dynamic internal structure of traffic and fail to quantify network-level near-miss risk systematically. Furthermore, their heavy reliance on historical datasets limits their utility as real-time predictive tools. Hence, current models experience the limitation of lacking comprehensiveness, as their application is primarily driven by the constraints of "training" dataset.

Addressing these critical gaps, we center our discussion on two pivotal elements of roadway networks: reliability and robustness. Reliability assesses the probability of maintaining a functional network, while robustness measures the large-scale effects

of road segment disturbances. To quantify the reliability and robustness of roadway network, we proceed by representing a roadway network as an undirected connected graph, denoted by $G = (V, E)$, where each node $v \in V$ represents an intersection or a terminal point of a road segment, and each edge $e = (u, v) \in E$ represents a bi-directional path between nodes u and v. This abstraction facilitates the analysis of road networks at a large scale, allowing us to conveniently incorporate the topology and connectivity of the network in our analysis.

This chapter introduces an approach that incorporates graph theory with near-miss risk analysis in the congested traffic flow to assess the safety of roadway networks at large scales. We illustrate the significance of crowdsourced velocity data collected via smartphones, utilizing it as the input to the frameworks developed in the preceding chapters. In particular, although the suggested near-miss framework doesn't factor in the "surprise" aspect of accidents, it does measure the complex risk associated with accident precursors inherent to the internal structure. Hence, we provide a comprehensive discussion asserting that the near-miss framework not only enables us to identify high-risk zones, but it also gives insight into the reliability and robustness of the roadway network on a broader scale. This approach is expected to set the stage for a comprehensive network-level safety assessment, offering transformative potential for our understanding of road safety and fueling the development of safer, more efficient transportation systems.

4.1 Network Reliability

Network reliability can be broadly defined as the statistical probability that a network retains its interconnectivity given a specified set of potential failure scenarios [156]. Reliability of roadway networks can therefore be interpreted as the probability that the road network can successfully facilitate transportation between various points without failure.

The reliability of a path between two nodes is determined by the absence of near-miss events along that path, and is thus defined as the corresponding probability.

Consider p_e as the likelihood of near-miss event occurring on edge e. The reliability of a path connecting nodes u and v, denoted as $\theta(u, v)$, depends on the near-miss probabilities of all the edges on the path. The path's reliability is therefore a function of the probabilities of each edge along the path, i.e., $\theta(u, v) = f(p_e)$, for all $e \in \mathcal{P}_{(u,v)}$, where $\mathcal{P}_{(u,v)}$ is the path from node u to v. Assuming that the occurrence of near-miss events on any specific edge is independent of the events on other edges, the reliability of the path $\mathcal{P}_{(u,v)}$ is given by,

$$\theta(u, v) = \prod_{e \in \mathcal{P}_{(u,v)}} (1 - p_e) \tag{4.1}$$

The network reliability, $q(d)$, is hence defined as the expected path reliability for all node pairs (u, v), given that the shortest path connecting them is of length d, i.e., $\mathcal{L}_{(u,v)} = d$ with path length $\mathcal{L}_{(u,v)}$ defined as $\mathcal{L}_{(u,v)} := \sum_{e \in \mathcal{P}_{(u,v)}} l_e$, where l_e represents the length of edge e. In other words,

$$q(d) = \mathbb{E}[\theta(u, v) | \mathcal{L}_{(u,v)} = d] = \frac{\sum_{\{(u,v) \in V \times V | \mathcal{L}_{(u,v)} = d\}} \theta(u, v)}{\#\{(u, v) \in V \times V : \mathcal{L}_{(u,v)} = d\}} \tag{4.2}$$

Thus, the network reliability $q(d)$ constitutes a coarse-grained metric that captures the network's macroscopic behavior by averaging the microscopic path reliabilities. Analogous to the emergence of macroscopic properties in statistical physics through microscopic state averaging, $q(d)$ embodies the network's reliability state at a particular "scale" denoted by the path length d. This measure provides an aggregated view across all node pairs (u, v), given a path length of d, thereby distilling the network's reliability while obfuscating individual path or node specifics. Furthermore, theoretically, ergodicity proposes that observing a single vehicle over an extended period offers a reliable mirror to the broader fleet behavior. Specifically, an absence of near-miss events for a given vehicle implies a similar non-occurrence across the entire vehicle network, given an assumed homogeneity of risk. This encapsulates the fundamental principle of ergodicity in stochastic systems, where long-term observations of an individual instance can faithfully represent the global system's statistical behavior.

4.2 Network Robustness

In network analysis, robustness signifies the ability of the network to sustain its key operations despite disturbances. This trait, in the context of roadway networks, corresponds to the preservation of connectivity and functionality amidst near-miss incidents. In this section, we initially establish the concept of disturbance and its impact on the scale of a road segment or edge. We then suggest network functionality metrics that quantify alterations in the network's interconnectedness at various length scales post-disturbance. Lastly, we discuss the Monte-Carlo approach as a method to simulate disturbances, explore various potential outcomes, and ascertain the network's expected robustness.

4.2.1 Disturbance Mechanism

Our analysis commences with the modeling of edge disruptions. Consider an edge e with N_e lanes. A near-miss event, equivalent to a disturbance, can occur on this edge with a probability p_e. Once edge e is disturbed, its corresponding travel time will be extended by ΔT_e. If one lane in a N_e lane road fails, the additional travel time for the remaining lanes is computed as $\Delta T_e = T_e/(N_e - 1)$, with T_e as the pre-failure travel time of edge e. This estimate presumes uniform traffic distribution and lane capacity. However, instead of time, our analysis focuses on the edge length. A disruption increases the perceived length of the edge from l_e to $l'_e = l_e(1 + 1/(N_e - 1))$. Unlike traditional network reliability studies that perceive failure as edge removal ($l'_e \to \infty$), our approach aids network connectivity by increasing the length of the disrupted edge. This method of representing failure loss complies with boundary conditions. More specifically, for a single lane road, a failure halts traffic, thus $l'_e \to \infty$. Conversely, for a large number of lanes ($N_e \to \infty$), a single lane failure has minimal impact, thus $l'_e \approx l_e$.

4.2.2 Network Functionality

The robustness of a network is often assessed via the interconnectedness index, which is typically embodied by the network's diameter - the average shortest path length between all node pairs. The diameter is inversely proportional to the communication efficiency between nodes - a smaller diameter indicates a shorter expected path between any two nodes. Following a similar approach, we incorporate a dynamic factor by considering network disturbances at varying scales. Specifically, we examine the change in the shortest path length, $\mathcal{L}_{(u,v)}$, between nodes u and v before and after disturbance. The path length increases to $\mathcal{L}'_{(u,v)}$ post-disturbance. Thus, the change in path length, $\Delta\mathcal{L}_{(u,v)}$, is:

$$\Delta\mathcal{L}(u,v) = \mathcal{L}'_{(u,v)} - \mathcal{L}_{(u,v)} \tag{4.3}$$

Network robustness metric is represented by the average change in shortest path length across all node pairs that have a pre-disturbance shortest path of d, defined as:

$$\bar{\Delta}\mathcal{L}(d) = \mathbb{E}[\Delta\mathcal{L}_{(u,v)}|\mathcal{L}_{(u,v)} = d] = \frac{\sum_{\{(u,v)\in V\times V|\mathcal{L}_{(u,v)}=d\}} \Delta\mathcal{L}_{(u,v)}}{\#\{(u,v) \in V \times V : \mathcal{L}_{(u,v)} = d\}} \tag{4.4}$$

This robustness measure provides a coarse-grained view of the expected change in the shortest path at different scales controlled by d. It offers an aggregated perspective across all node pairs (u, v) with a shortest path length of d, thereby encapsulating the network's robustness without focusing on specific paths or nodes.

4.2.3 Disturbance Simulation

We employ Monte Carlo simulations to systematically explore the potential outcomes of various disturbances, thereby modeling edge failures in the roadway network. Each edge e is associated with a failure probability p_e, representing the likelihood of a near-miss event. Our model assumes that edge failures are independent, implying that a near-miss event's occurrence on one edge does not influence the probability of

a similar event on another edge. This assumption aligns with the operational reality of roadway networks, where incidents across different road segments generally occur independently.

4.3 Data Processing

In this section, we explain the application of the near-miss analysis framework to large-scale graph representations of roadway networks. Our primary focus is on the roadway networks of Massachusetts, California, Ohio, Maine, and Oregon due to the abundance of crowdsourced velocity data we have amassed from the Carbin educational app [149]. Such data set allows us to perform comprehensive crowdsourced and statistical analysis at the state scale. To this end, we first explain the process of obtaining these networks' graph representations, and further outline the simplification of these graphs as well as edge weight assignments. We then implement the near-miss analysis framework to the velocity data, which allows us to derive the expected near-miss risk for the entirety of the network.

4.3.1 Graph Represenation

To construct the graph representation of roadway networks, we utilized the open-access shapefiles of primary and secondary roads for different states, provided by the US Census Bureau [157]. These shapefiles were transformed into undirected graphs, reflecting the interconnectivity of the roadway network. We utilized the 2018 Highway Performance Monitoring System (HPMS) Public Release dataset [158], to assign each edge with corresponding attributes: speed limit, number of lanes, and traffic density values.

The original graphs extracted from the shapefiles are typically large mostly due to the high number of nodes with degree 2, which meticulously capture the geometric intricacies of the roadway system such as U-turns and roundabouts. To streamline our analysis, we implemented a process of strategic reduction, judiciously eliminating nodes of degree two and subsequently replacing the associated edges, provided each

was less than 2 km in length. This simplification served a dual purpose. Primarily, it facilitated a substantial reduction in graph size by almost more than 95% [Table. 4.1], increasing the efficiency of subsequent analyses such as shortest path and inbetweenness indices. Secondly, and perhaps more critically, it allows us to consider the occurrence of near-miss events on each edge as an independent process. The assumption of independence is grounded on the notion that longer edges are less likely to share influencing factors for near-miss events. Thus, by transforming into a simplified network, we enhance our ability to analyze these near-miss incidents as independent occurrences, undistorted by shared or common characteristics.

4.3.2 Phase Diagram and Spatial Correlation

We perform our analysis of near-miss risk on a state-by-state basis, applying the near-miss risk framework to all recorded trips within each of the five state [Table 4.1]. Each trip had an average duration of approximately three minutes, sufficient to ensure both stationarity and ergodicity. We ensured a uniform distribution of trip frequency spanning the period from 2019 to 2021. The near-miss risk, corresponding to each trip, was computed. Utilizing the trip coordinates, this risk was allocated to the respective edge within the network. The risk-weighted centrality of each edge was subsequently calculated as $p_e B_e$, where p_e and B_e denote the near-miss risk and the edge betweenness centrality, respectively [Fig. 4-1]. These risk-weighted centrality indices not only offer an exhaustive understanding of how near-miss risks are distributed throughout the network, but they also take into account the network's topology. This primary analysis lays the groundwork for identifying high-risk zones.

Phase Diagram

The first significant observation is the distinctive phase diagram behavior shared by all the states under examination [Fig. 4-2.(a)], mirroring the phase diagram we derived for Massachusetts in the prior chapter [Fig. 3-6]. This pattern carries significant weight as it indicates that the risk of near-miss events peaks at a certain density within

the congested traffic regime. This finding also bears resemblance to the behavior predicted by the NaSch model, where the highest likelihood of a near-miss event emerges within the confines of the congested regime [Fig. 3-4]. As evidenced through a clustering analysis conducted in the previous chapter, this critical regime of peak risk is expected to be situated at the intersection of low-density and high-density congested regimes. In particular, when this was compared with actual collision data from the state of Massachusetts, the strongest correlation was observed in the critical regime [Fig. 3-6.(a)]. This implies that traffic density is a vital determinant in shaping the internal traffic structure. Of course, this is not to undermine the roles played by other influential factors such as the geometric configuration of the road network and the behaviors of drivers. These parameters, to a certain extent, are implicitly encompassed in the state transition matrix as it quantifies the internal structure of traffic.

Spatial Correlation

As a secondary perspective, the computation of spatial covariance between each edge's near-miss risk was carried out. This analysis aimed to glean insights into the near-miss risk distribution across the network. More specifically, we calculated the normalized spatial covariance between the near-miss risks associated with all pairs of edges (e_1, e_2) located at a liner distance r from each other via,

$$\tilde{C}_{pp}(r) = \frac{1}{\sigma_p^2}\left(\mathbb{E}[p_{e_1} p_{e_2}] - \bar{p}^2\right) \tag{4.5}$$

where $\bar{p}$ and σ_p^2 denote the network-wide first- and second-order central moments of near-miss risks, respectively. The observed covariance demonstrates a rapidly decaying exponential pattern [Fig. 4-2.(b)], suggesting that the spatial correlation between the near-miss risks across different edges is minimal. Such low correlation levels, coupled with the large edge lengths, indicate that occurrence of near-miss events on each edge are likely to be triggered by unique factors. Consequently, these events can be approximated as nearly independent processes. In other words, the failure of

one edge does not significantly depend on the failure of other edges in the network. It is, however, important to note that selecting small edge lengths would inherently amplify correlation levels on smaller scales. Consequently, the cause of edge failure would exhibit shared characteristics, which in turn requires the consideration of correlated failures. A common approach to take into account correlated failures is the application of Markov Chains, which provides an understanding of an edge's failure probability based on its previous state [159].

	Before Simplification		After Simplification		
	# Nodes	# Edges	# Nodes	# Edges	# Trips
Massachusetts	195138	196906	3902	5320	1.6×10^6
California	630616	632418	8298	9751	6×10^5
Ohio	613672	617366	10058	13389	5.2×10^5
Oregon	252516	253040	2931	3345	3.4×10^5
Maine	238410	239289	3092	3842	3.3×10^5

Table 4.1: Summary of each state's graph characteristics and total trip count. Significant reduction in the total number of edges and nodes is achieved by strategically removing degree-two nodes which linking two short edges and their subsequent consolidation into a singular, extended edge.

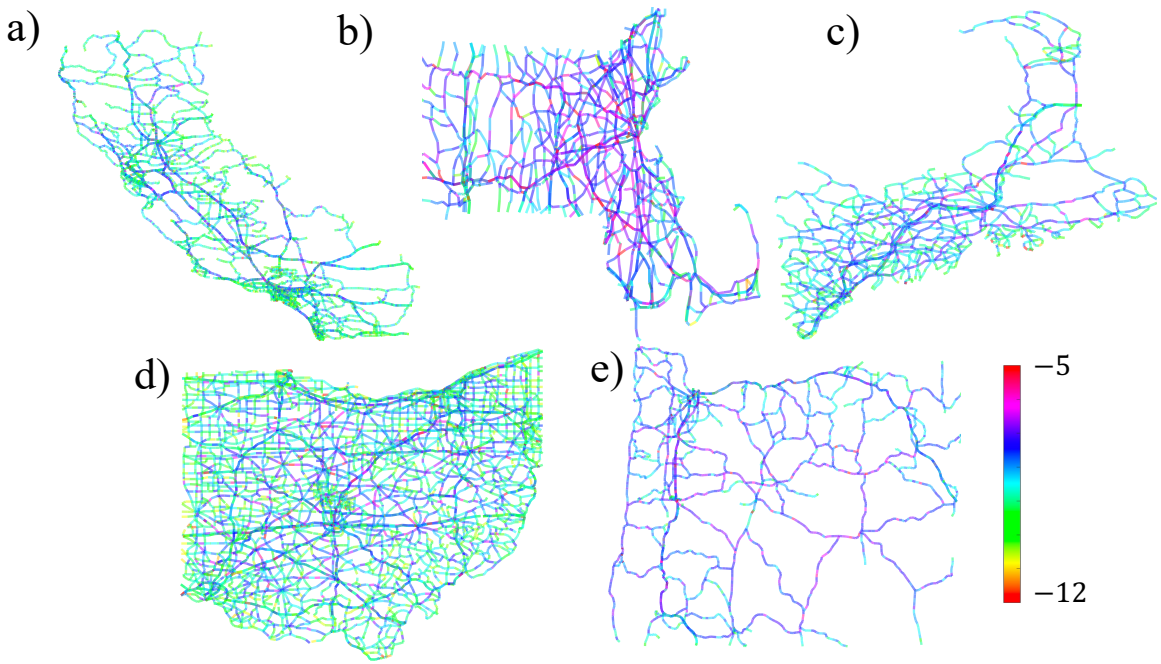

Figure 4-1: Risk-weighted centrality for each edge in the network. The plot depicts the normalized value of $p_e B_e / \sum_e B_e$, where B_e indicates the edge betweenness index computed using Ulrik Brandes' algorithm, and p_e is the near-miss risk. The expected network-scale average of risk-weighted centrality is calculated as $\sum_e p_e B_e / \sum_e B_e$ for each state. The calculated averages are: a) California: $10^{-3.40}$, b) Massachusetts: $10^{-2.27}$, c) Maine: $10^{-2.75}$, d) Ohio: $10^{-2.86}$, and e) Oregon: $10^{-3.10}$.

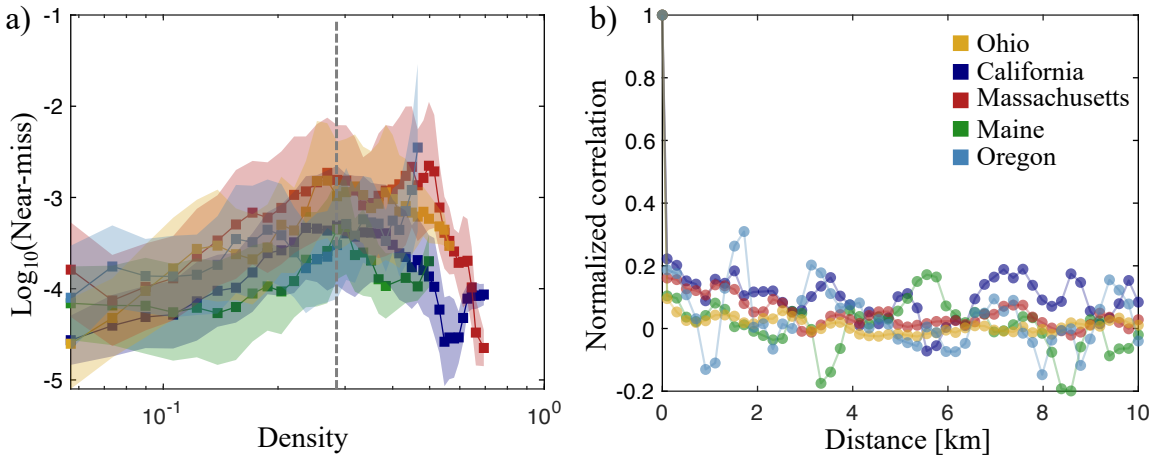

Figure 4-2: (a) The phase diagram of near-miss risk for different states. A universal pattern emerges, revealing that the probability of a near-miss risk peaks at a density of approximately 0.2. It then approaches zero as traffic veers towards either the jammed or free-flow regime. As indicated in the preceding chapter, this density is anticipated to correlate most strongly with the probability of collisions. (b) The normalized spatial correlation between near-miss risks across edges. The correlation function quickly decays and approaches zero (indicating a lack of correlation) even over short distances. This suggests that the granularity of network discretization offers enough distance for edge failures to be treated as independent events. This independence allows us to assume that the causes of near-misses are not identical for adjacent edges. The shaded area represents the mean value plus and minus one standard deviation.

4.4 Results and Discussion

We now have all the necessary components to conduct state-scale reliability and robustness analyses for the states of Massachusetts, Ohio, Maine, California, and Oregon. These components consist of the graphs of roadway networks, with large enough edges to assume the independence of the failure mechanism. The subsequent analysis elevates the network's topology and the spatial distribution of near-miss risks to derive the network's expected behavior at various length scales, based on the shortest path length.

4.4.1 Network Reliability

We advance by examining the reliability of the roadway networks to quantify the probability of encountering no disturbances. The reliability of the network decreases exponentially as the trip distance increases, which can be expressed as:

$$q(d) \approx e^{-\frac{\mathcal{L}}{\alpha_c L_c}} \tag{4.6}$$

where $\mathcal{L}$ represents the length of the shortest path to the destination, while L_c is the representative length of the state, calculated as $L_c = \sqrt{A}$, where A stands for the area of the state. We coin α_c as the State Failure Index, demonstrating the rate at which the network's reliability diminishes as the trip length augments. A higher failure index indicates that near-miss event is observed over a shorter distance, thereby reducing the network's reliability. The choice of the exponential function in Equation (4.6) is apt as it satisfies the probability boundary conditions. Namely, if the length of the trip converges to zero, the likelihood of observing a disturbance also approaches zero, implying a fully reliable network. Conversely, if the trip length tends towards infinity, the occurrence of a near-miss event becomes a certainty; that is,

$$\lim_{d \to 0} q(d) = 1, \quad \lim_{d \to \infty} q(d) = 0 \tag{4.7}$$

Considering their respective sizes, Massachusetts has the least reliable network with a state failure index (α_c) of 67.09, while California boasts the most reliable network, with an α_c value of 43.45.

4.4.2 Network Robustness

The robustness of the networks is examined through Monte Carlo simulations, as outlined in Section 4.2.3. For each trial of the Monte Carlo simulations, a random number within the range [0,1] is assigned to every edge. Edge failure occurs when the generated random number falls below p_e, resulting in an increase of its length from l_e to l'_e, as detailed in Section 4.2.1. We conducted a total of 10^4 Monte Carlo trials for each state to assess and compare the network's functionality pre- and post-disturbance.

Our initial metric of reference is the change in network diameter, a macroscopic measure of the network's interconnectedness. By definition, the network diameter represents the mean length of the shortest paths between any two nodes in the network [160]. Prior to any disturbances, the network diameters for the states of California, Oregon, Ohio, Maine, and Massachusetts stood at 525.25, 368.97, 207.39, 179.97, and 108.32 km, respectively. Following disturbances, the expected increases in network diameter for these states were estimated to be 0.022, 0.067, 0.08, 0.11, and 0.2 km respectively; this suggests that the network interconnectedness for Massachusetts and California are at most and least risk, respectively.

We further investigated the network functionality by examining the expected change in shortest path $\bar{\Delta}\mathcal{L}(d)$ at various scales, determined by the length of the shortest path d. For relatively short journeys of less than 100 km, all of the road networks in question displayed a linear pattern in the expected length increase of the shortest path, with the function controlled by the pre-disturbance shortest path. The slope of this linear pattern, $\beta_c = d\bar{\Delta}\mathcal{L}/d\mathcal{L}$, is inversely proportional to the robustness of the network. When evaluating the robustness of the networks relative to the characteristic length of each state ($\beta_c L_c$), we observe a distinct hierarchy. California stands at the forefront with a robustness factor of $\beta_c L_c = 0.04$. Following California, Oregon

has a robustness of 0.11, while Ohio has a slightly higher factor at 0.13. Maine's network exhibits a robustness of 0.24, with Massachusetts exhibiting the least robust network with a factor of 0.33. These measurements further underscore the relative resilience of each state's network to potential disturbances. These results suggest that the network in Massachusetts is relatively more vulnerable to disturbances, while the network in California exhibits greater robustness under the same conditions.

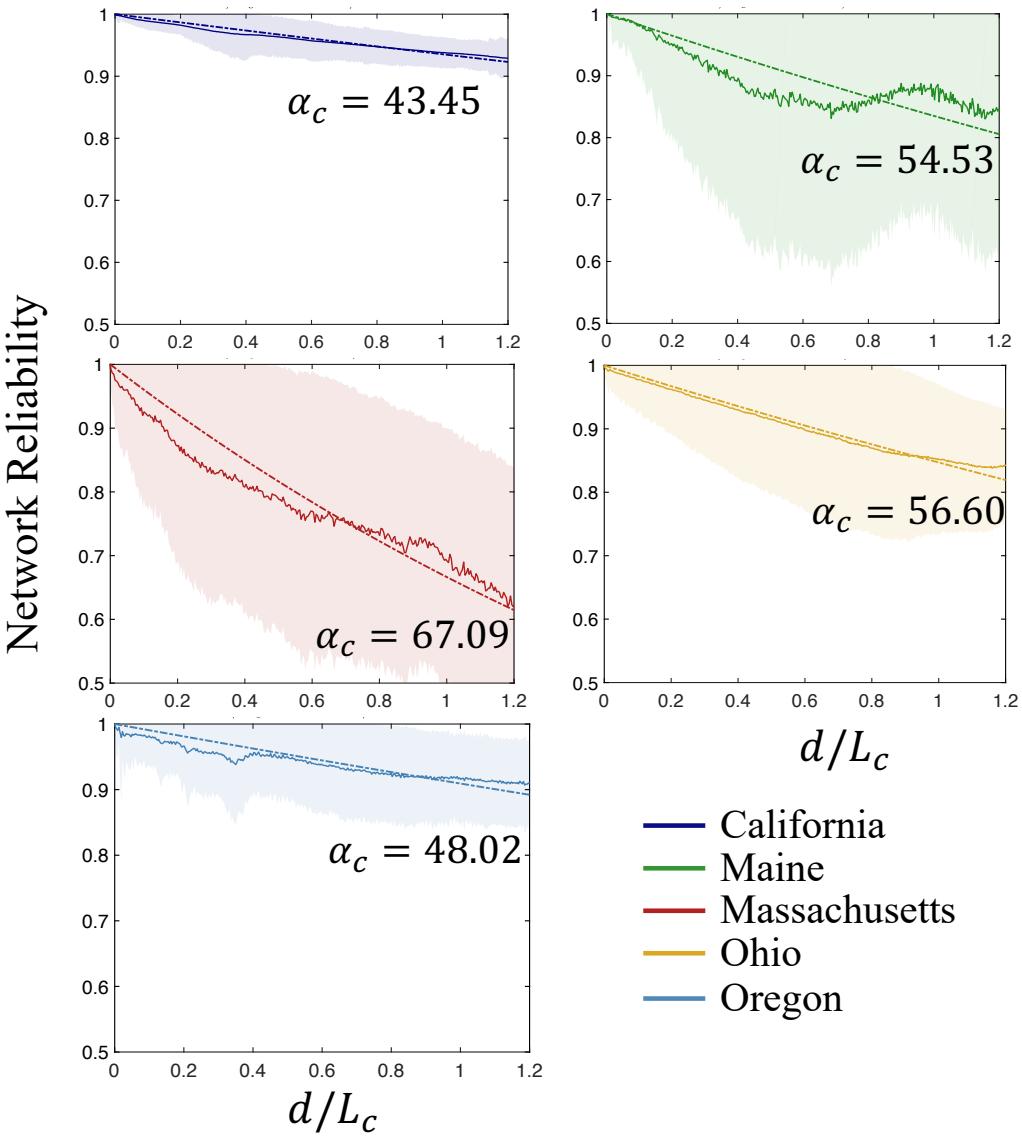

Figure 4-3: Network reliability for different states as a function of the normalized trip length $\mathcal{L}/L_c$, where L_c is the characteristic length of each state, computed as the square root of its area ($L_c = \sqrt{A}$). Network reliability exhibits an exponential decrease as trip length increases. Among the five states under consideration, Massachusetts demonstrates the quickest decay, implying it has the least reliable network. Conversely, California exhibits the slowest decay, suggesting it possesses the most reliable network. The shaded area represents the mean value plus and minus one standard deviation.

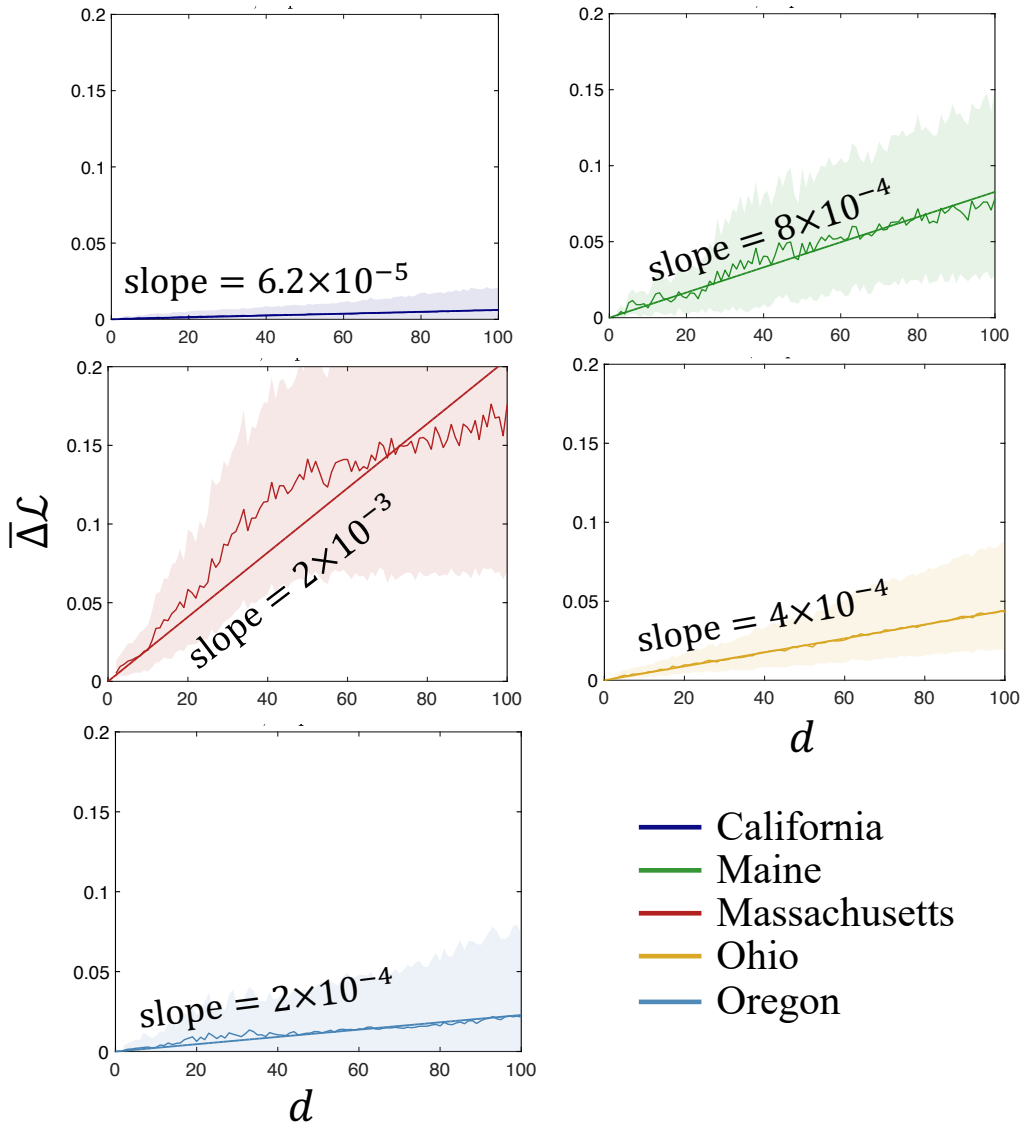

Figure 4-4: The expected excess length of trips $\bar{\Delta}\mathcal{L}$ after network failure as a function of original trip length $\mathcal{L}$ prior to the failure. The additional trip length due to network-level failures demonstrates a linear relationship with the original trip length for short trips (less than 100km). The slope of this line serves as an indicator of network robustness. Among the analyzed states, the roadway networks of Massachusetts and California have the steepest and gentlest slopes, respectively. This implies that the network in California is the most robust, while the one in Massachusetts is the least. The shaded area in the plot represents the 5% and 95% percentiles of the data.

4.5 Summary

In this chapter, we elaborate on the large-scale application of the near-miss risk framework by introducing robustness and reliability metrics, which serve the dual purpose of identifying high-risk zones and providing insights into safety patterns across at larger context. Specifically, we employed the near-miss framework across five states - Massachusetts, California, Ohio, Maine, and Oregon - using graph representation of networks. We demonstrated through each state's phase diagram that the risk of accident precursor remains minimal at both low and high limiting densities, but it adopts to its maximum at the critical density within congested flow. In our pursuit of understanding safety patterns, we introduced two metrics applicable at the network-scale, specifically, reliability and robustness, integrating not just the distribution and statistical moments of near-miss risks but also the network's topology. The network reliability, $q(d)$, offers a macroscopic insight into the network's behavior at different length scales controlled by d, denoting the probability of not observing any near-miss event as a function of trip length d. We found that the reliability of all networks decreased exponentially, with California and Massachusetts showing the most and least reliable networks, respectively. We defined network robustness as the expected increase in trip length for a trip of length d in an undisturbed network, offering a broad view of the expected change in the shortest path at varying scales. We demonstrated that for relatively short trips, the trip length is expected to increase linearly with respect to d. Similar to reliability, Massachusetts and California were the extremes, presenting the least and most robust networks among the five states studied, respectively.

Chapter 5

Conclusion

The contemporary surge in motorization, a byproduct of swift globalization, carries a significant drawback - road safety. Given its profound societal and economic implications, traffic safety has lately risen to prominence as a crucial area of research. While data-driven and statistical models highlight critical parameters and identify high-risk zones, their oversight of the dynamic internal structure of traffic leads to their failure in systematically quantifying network-level collision risk. Additionally, their substantial dependence on historical records of accidents curtails their efficacy as real-time predictive tools, thereby making these models lack comprehensiveness due to their application being significantly influenced by the limitations of their training dataset.

The objective of this thesis was to put forth a physics-informed framework for the safety assessment of roadway networks at various scales, ranging from a road segments to large networks. Building upon the *physics of traffic*, we suggest a metric of accident precursor risk that displays a positive correlation with the actual risk of accidents, offering computational advantages and necessitating only crowdsourced velocity measurements as the singular input. Furthermore, this proposed method not only boasts potential for real-time implementation but also demonstrates the capacity to yield results at both micro and macro scales, bridging the gap between physics and data-driven modelling.

5.1 Summary of Main Findings

In Chapter 2, we focused on the spatial and temporal statistics of traffic, aiming to quantify the inherent patterns that underpin traffic's structural characteristics. Specifically, we explored the system's higher-order statistical properties and identified universal memory effects, existing both in time and space, that poses crucial information pivotal for mapping of traffic properties. Furthermore, by invoking the ergodic theorem from statistical physics, we illustrated that the fluctuations in vehicle speed profiles amid congested flows can capture the statistical properties that govern the collective behavior of traffic as an ensemble. This understanding enabled us to directly access higher-order statistical moments from crowdsourced velocity measurements, shedding light on the internal structure of traffic. The first-, and second-order statistics present a rich repository of information that can be used to decipher the complex interactions within the system, thereby facilitating the quantification of various statistical properties of the system.

In Chapter 3, we introduced a probabilistic accident precursor metric that quantifies the likelihood of observing near-miss events, specifically those featuring decelerations beyond normal operation braking. Utilizing a reduced-unit representation of traffic, we demonstrated a density-dependent peak in near-miss risk within the congested regime, which tapers to zero in both free-flow and jammed regimes. This phase diagram of intrinsic near-miss collision risk applied to both model-generated and crowdsourced vehicle velocity datasets, confirming the framework's robustness independent of any model attributes. We demonstrated that heightened randomness in driver behavior, mirroring phenomena in other systems ranging from fracture risk in random porous materials to the moderation of erratic price shifts in financial markets, results in a safer network at large. Moreover, examining the roadway network of Massachusetts, we juxtaposed the computed near-miss risks with actual collision risks. We found a positive correlation between these two variables, with the peak correlation coinciding with the critical regime, characterized by the highest near-miss and collision risks.

Lastly, in Chapter 4, we highlighted the predictive capabilities and the potential insights that the near-miss framework could offer towards cultivating safer neighborhoods. We demonstrated that the near-miss framework, while instrumental in identifying high-risk road segments or zones, extends its utility beyond such identification by empowering us to assess and comment on broader safety contexts. To this end, we introduced metrics for reliability and robustness to gauge a network's ability to maintain stability under normal conditions, and its capacity to sustain functionality post-disturbance, respectively. Specifically, we applied the near-miss framework across five states - Massachusetts, California, Ohio, Maine, and Oregon - using graph representation of networks (results are summarized in Table 5.1). Our findings revealed an exponential decrease in the reliability of all networks as the destitution gets further away, with California and Massachusetts standing out as the most and least reliable networks respectively. We showed that for relatively short journeys, the expected increase in trip length was linear in relation to the shortest path length of the trip in undisturbed network. Analogous to the reliability observations, Massachusetts and California represented the two extremes in terms of robustness, manifesting the least and most robust networks among the five states studied, respectively.

	α_c	$\beta_c \times 10^4$	L_c [km]	$\mu_p \times 10^4$	Network diameter [km]	Diameter change [km]
California	43.45	0.62	651.13	5.5	525.25	0.022
Maine	54.53	8	302.71	9.4	179.97	0.107
Massachusetts	67.09	20	165.3	44	108.32	0.199
Ohio	56.60	4	340.73	16	207.39	0.080
Oregon	48.02	2	504.78	11	368.97	0.067

Table 5.1: Summary of the network reliability and robustness analyses. α_c serves as an indicator of network reliability, demonstrating the rate at which the transportation network's reliability decreases as a function of trip length. Network robustness is measured by β_c, which represents the rate of increase in trip length as a function of the trip length itself when the network is compromised, and certain edges have failed. μ_p represents the expected near-miss risk for the network. The network diameter, a classic index showcasing network interconnectivity, indicates the network's sensitivity to failures through its variation due to such failures.

5.2 Research Contribution

5.3 Limitations and Future Perspective

The objective of this study was to present a crowdsourced solution designed to enhance roadway safety by understanding the internal structure of traffic. Throughout this investigation, we made several assumptions and explored the situations where these assumptions are valid. However, as part of future work, it would be significant to transcend these assumptions and incorporate greater complexity into the frameworks introduced in this study, thereby fostering a more comprehensive understanding of the statistics of extreme events in traffic across a wider range of scenarios. Specifically,

1. In Chapter 2, we examined the inner workings of traffic dynamics, basing our exploration on the assumption of stationarity and ergodicity. These assumptions, as detailed in the chapter, could hold true given homogeneous traffic on simple roadway networks over properly chosen time/length scales. However, they're not universally applicable. Especially in urban traffic settings, external elements like traffic signals, roundabouts, and school zones with speed bumps can undermine these assumptions. Hence, it becomes crucial to ascertain to what extent these factors are violated and further assess their impact on the results. One possible method is to adopt a multi-scale modeling approach, where these assumptions are true over a certain scale, and then the results are scaled up or down. Additionally, several models can be employed to address the issue of nonstationarity, such as Autoregressive Integrated Moving Average, Autoregressive Conditional Heteroskedasticity, and Generalized Autoregressive Conditional Heteroskedasticity. Furthermore, it's noteworthy that the theory and model calibration presented are restricted to single-lane traffic. Extensions to multiple lane models of the NaSch-type with lane-change probabilities, ramp exits, and corrections for traffic obstacles for inner-city applications, however, are deserving of consideration.

2. In Chapter 3, we introduced a near-miss risk metric based on the reduced-unit representation. While this representation provides a fairly universal approach to traffic dynamics modeling, it does come with certain limitations. It could thus be beneficial to extend these concepts to other traffic models, both continuous and discrete, suggest comparable metrics, and further investigate the correlations with collision risks using historical accident records. Moreover, it's crucial to relax the constraints of stationarity and ergodicity, for example, by contemplating open boundary conditions.

3. In Chapter 4, our reliability and robustness analysis was conducted under the fundamental assumption of independent edge failures. We demonstrated that proper discretization of the roadway network as a graph could reduce the spatial correlation between edge failure probabilities, thus neighboring cells would no longer have a shared cause of failure. However, it could be beneficial to utilize methods like Markov Chains to model failures that are correlated, and further study both the microscopic and macroscopic safety indices while considering the impact of correlation indices.

In summary, the current study has provided valuable insights into traffic dynamics and has proposed a physics-informed framework to obtain an accident precursor metric from the crowdscourced velocity data. Despite the progress made in analyzing traffic dynamics and investigating diverse metrics and models, the potential for future research remains extensive. Continuing to build upon and challenge our assumptions, as well as applying and refining the theoretical frameworks, will enable more comprehensive and nuanced understandings. This will also open up opportunities for tailoring solutions to specific traffic scenarios and thus enhance safety measures more effectively. We eagerly anticipate further progress in this domain and the prospective societal benefits that could emerge from blending physics-informed modeling of traffic safety with the integration of data analytics.

Appendix A

Nagel-Schreckenberg Cellular Automaton Model

Among the many traffic flow models ranging from continuum to agent-driven approaches (for a review, see [3]), the Nagel-Schreckenberg (NaSch) cellular automaton model [35] has emerged as a powerful tool not only to reproduce critical features of traffic flow [68, 69], such as backward-moving shock waves [70] and the fundamental diagram of traffic [72], but foremost to track the internal structure of traffic by means of simulations [64–67, 71, 73, 74]. This includes quantitative insights into the mechanism of jamming from investigations of strength and range of interactions between successive vehicles captured by short-range correlation functions [65, 66], phase transition phenomena from investigation of the correlation length in the velocity-velocity covariance function [67], jamming rate, jam lifetime and size from stability criteria of the NaSch-model close to the critical jamming density [129], and spatial and temporal memory effects [161]. Key in these studies is the apparent simplicity of the

NaSch-model heuristics as a discrete lattice-gas-like model, which we recall:

$$\begin{cases} (i) \quad \text{Acceleration: } v_j^{(1)} = \max(v_k(t), v_{max}) \\ (ii) \quad \text{Deceleration: } v_j^{(2)} = \min(v_j^{(1)}, d_k(t)) \\ (iii) \text{ Randomized braking:} \\ \qquad v_k(t+1) = \begin{cases} v_j^{(2)} \text{ with probability } 1-p \\ \max(v_j^{(2)} - 1, 0) \text{ with probability } p \end{cases} \\ (iv) \quad \text{Update vehicle position:} \\ \qquad x_k(t+1) = x_k(t) + v_k(t+1) \end{cases} \tag{A.1}$$

where $d_k(t)$ represents the headway, and p is the stochasticity parameter which takes into account the drivers' individual freedom to regulate their speed [Fig. A-1]. From a statistical physics point of view, the stochasticity parameter permits random velocity fluctuations and is merely analogous to an external field in ferromagnetism, serving as a conjugated parameter which can destroy the ferromagnetic phase transition [162, 163]. The transition density for the stochastic model is analytically estimated as $\rho_c^p \approx (1-p)/(v_{max}+1-2p)$ from the average free-flow velocity and the analytic upper bound of the jammed dissolution velocity [129].

The classical Monte-Carlo simulations of the NaSch model consist of randomly placing N particles up to a prescribed cell occupation $\rho = N/L$, representative of the traffic density, into the 1-D simulation box of size L with periodic boundary condition. For a given stochasticity parameter p, the NaSch-model heuristics (A.1) is then applied long enough until statistical stationarity is reached [73, 122, 123].

For the special case of deterministic model ($p = 0$), a linear integer programming (IP) formulation can be employed to speed up the simulations. This reformulation is achieved owing to the unique property of velocity in the deterministic model which resembles a backward-moving wave. More specifically, the $(k+1)$th vehicle at time step t imparts its velocity to the kth vehicle at time $t+1$, i.e., $v_k(t+1) = v_{k+1}(t)$. Due to the absence of randomized braking, the particle velocity equals the headway.

Consequently, the ensemble average velocity is equal to the expected value of the headway in the congested flow, i.e., $\bar{v} = 1/\rho - 1$ with $(v_{max}+1)^{-1} \leq \rho \leq 1$. Finally, $v_{k+1}(t) - v_k(t) \leq 1$ always holds true since the maximum acceleration in the NaSch model is bound to one. The heuristics of the deterministic model for the case can be revisited at the system scale via the following linear IP:

$$\max_{v_k(t)} \sum_{k=1}^{N} \theta_k v_k(t), \ \forall k : v_k(t) \in \{0, 1, ...v_{max}\} \tag{A.2}$$

subject to,

$$\begin{cases} v_{k+1}(t) - v_k(t) \leq 1, \quad \forall k \in \{1, ..., N-1\} \\ \\ v_1(t) - v_N(t) \leq 1 \\ \\ \frac{1}{N} \sum_{k=1}^{N} v_k(t) = \bar{v} \end{cases} \tag{A.3}$$

Herein, θ_i is a continuous uniform random variable, $\theta_i \sim U[0,1]$ playing a similar role as the initial configuration of the system required for Monte-Carlo simulations of the NaSch model. Much akin to the negligible impact of the initial configurations, the impact of θ_i on the velocity statistics under stationary conditions is negligible for large-enough system size. The outcome of the linear programming relations are the particle velocities, $v_k(t)$, which are transferred to the succeeding time step through backward-moving wave property.

The outcome of such simulations are velocities of all particles for a sufficiently long time interval $v_k(t)$. These velocity signals serve as the input for calculating the statistical moments of the ensemble, such as the expected velocity, and the autocovariance function from its canonical definition.

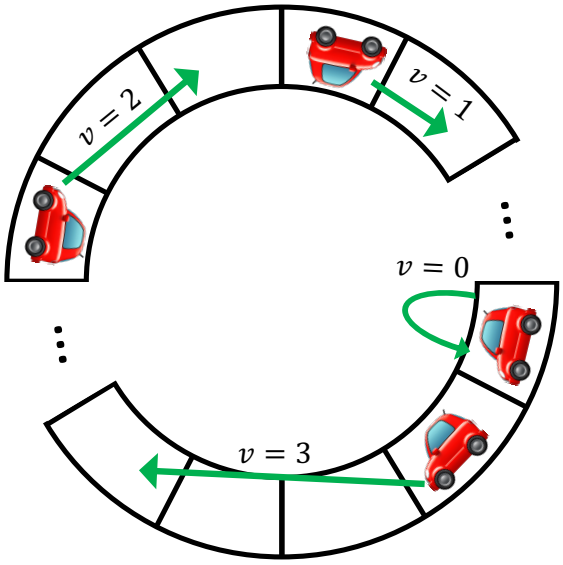

Figure A-1: Schematic of the NaSch model of traffic flow. The number of particles (vehicles) is conserved, with a cyclical/periodic boundary condition. Vehicles either move ahead to unoccupied cells at different speeds ($v > 0$) or stay stationary ($v = 0$), following the prescribed NaSch heuristics.

Appendix B

Statistical Foundations of Stochastic Processes

Consider a random experiment in the probability space (Ω, F, P), with each outcome $\omega \in \Omega$ assigned a real-valued time function $X(t, \omega)$, where $t \in I$, a totally ordered index set. This map of functions indexed by the possible outcomes of Ω forms the foundation of a stochastic process. The index set I could either represent the set of integers $\mathbb{Z}$, denoting a discrete-time process, or the set of real numbers $\mathbb{R}$, characterizing a continuous-time process. It's worth noting that the choice of index set opens up a wide range of possible behaviors for the process, making the study of stochastic processes remarkably versatile and complex. A specific outcome ω results in a deterministic function $X(t, \omega)$, termed a *realization* or *sample path* of the process [Fig. B-1]. The complete statistical properties of the process are deduced from the ensemble of all such realizations, encapsulating a comprehensive overview of the stochastic process's behavior and providing the basis for predictive modeling.

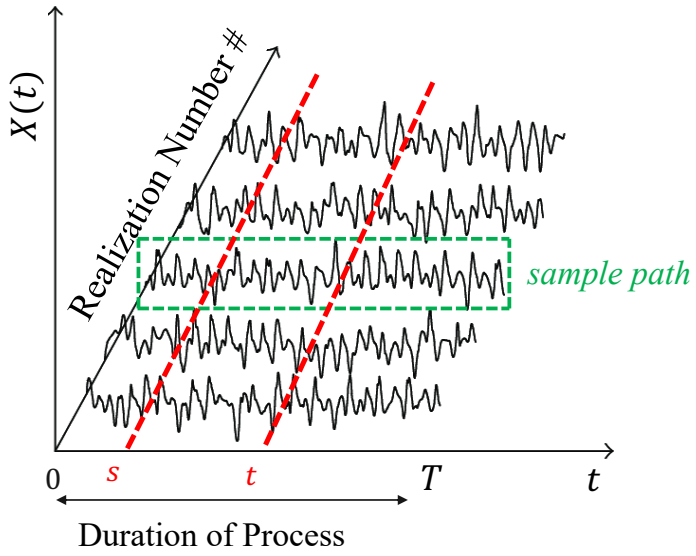

Figure B-1: A simplified illustration of ensemble $X(t, \omega)$. An ensemble represents an array of realizations whose statistics are defined by examining the ensemble at specific time(s) across all the realizations. Assuming stationarity, the statistical moments of the process remain unaffected by time lags [Eq. (B.7)]. Additionally, if ergodicity holds, a single sample path of a stationary process (in the green box) is sufficient to statistically represent the ensemble's statistical moments.

B.1 Characterization

Consider a discrete process with a series of sampling times $t_1, t_2, ..., t_k \in I$. The random variable resulting from establishing the value of the process at each t_i for $i = 1, ..., k$ is denoted by $X_i = X(t_i)$. This forms a vector of random variables $\underline{X} = [X_1, ..., X_k]^T$ for any finite value k. The complete characterization of this vector is achieved through the specification of its joint probability distribution function, offering a robust analytical tool for understanding the process's complex statistical behavior:

$$p_X(x;t) = \mathbb{P}[\omega : X_1(\omega) \leq x_1, ..., X_k(\omega) \leq x_k] = \mathbb{P}[\omega : X(t_1, \omega) \leq x_1, ..., X(t_k, \omega) \leq x_k] \quad \text{(B.1)}$$

Thus, a stochastic process can be visualized as an intricate assemblage of random variables indexed by time, where these variables interact in a complex web of high-order correlations. A simpler situation arises with an independent and identically distributed (i.i.d.) process. In this case, the joint distribution for any sampling times $n_1, ..., n_k$ unfolds as the product of the first order marginal distribution:

$$p_X(x_1, x_2, ...x_n; n_1, ...n_k) = \prod_{i=1}^{k} P_X(x_i) \quad \text{(B.2)}$$

where the first order marginal $p_X(x;t) = p_X(x)$ exists independent of time.

However, the landscape of stochastic processes often reveals interdependencies and correlations. Therefore, the task of specifying the complete set of joint probability distribution functions, as illustrated by Equation (B.1), can be highly complex. It necessitates the elucidation of the properties of all potential subsets of time indices. For practical purposes, analysis is often restricted to first- and second-order moments, offering a simplification that still yields significant insights, given their capacity to capture essential aspects of the process's statistical structure.

B.1.1 First-Order Moment: Mean

A key descriptor of a random process is its first-order moment. This metric illuminates the temporal evolution of the process's expected value. For a discrete-time process, the mean of a random process $X(.)$, denoted by $m_X(t)$, is defined by the time function as follows:

$$\mathbb{E}[X(t)] = m_X(t) = \sum_{x=-\infty}^{\infty} x p_X(x,t) \tag{B.3}$$

This definition can be extended to continuous-time processes:

$$m_X(t) = \int_{-\infty}^{\infty} x p_X(x;t) dx \tag{B.4}$$

The mean operator, which characterizes the first-order statistics of the ensemble, is generally a function of time. The ensemble expected value at a specific time t can be intuitively interpreted as an average of the realization values with similar probabilities of occurrence within the ensemble at time t. This average, $m_X(t)$, represents the mean value of the process at time t, encapsulating a compact description of the process's central tendency.

B.1.2 Second-order Moment: Autocorrelation Function

The second-order moment of the ensemble, otherwise known as the autocorrelation function, elucidates the degree of correlation the process exhibits at two distinct times, s and t. It essentially measures the correlation of the ensemble with itself at these separate points in time. For discrete processes, it is formulated as:

$$R_X(s,t) = \mathbb{E}[X(s)X(t)] = \sum_{x_1=-\infty}^{\infty} \sum_{x_2=-\infty}^{\infty} x_1 x_2 p_X(x_1, x_2; s, t) \tag{B.5}$$

And for continuous-time processes, the autocorrelation function is defined as:

$$R_X(s,t) = \mathbb{E}[X(s)X(t)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 p_X(x_1, x_2; s, t) dx_1 dx_2 \tag{B.6}$$

The second-order moment allows us to explore the temporal dependencies within the process, providing an understanding of how the process evolves and fluctuates over time.

B.2 Stationarity: Strict Sense and Weak Sense

The property of stationarity arises when the statistical properties of a process do not evolve with time. Stationarity can be classified into two types: strict-sense stationarity (SSS) and wide-sense stationarity (WSS), also known as weak-sense stationarity.

A strictly stationary process is one in which the joint distribution of any subset of random variables within the process remains invariant under time shifts. Formally, a random process $X(t)$ is considered strictly stationary if, for any finite set of indices $t_1, t_2, ..., t_k$, the joint distribution function,

$$p_X(x_1, x_2, ...x_k; t_1, t_2, ...t_k) = p_X(x_1, x_2, ...x_k; t_1 + \tau, t_2 + \tau, ...t_k + \tau) \tag{B.7}$$

remains unaltered for all shifts $\tau \in I$. This concept applies to the complete set of time indices, reflecting the invariance of all moments and joint moments of the process. SSS ensures that the process remains consistent throughout, a factor that simplifies its analysis and interpretation.

A weaker form of stationarity is wide-sense stationarity, which concerns the invariance of the first and second moments of the process. A random process $X(t)$ is said to be WSS if the mean $m_X(t)$ is constant, and the autocorrelation function $R_X(s, t)$ depends only on the time difference $\tau = t - s$. Thus, the definitions of the first and second moments for a WSS process are as follows:

$$m_X = \mathbb{E}[X(t)] \tag{B.8}$$

$$R_X(\tau) = \mathbb{E}[X(t + \tau)X(t)] \tag{B.9}$$

The advantage of WSS is that it provides a simplification of the process's characterization while still preserving its essential statistical properties.

B.3 Ergodicity of Stationary Processes

While stationarity implies certain statistical characteristics of a process are time-invariant, it does not ensure those statistics can be computed from a single sample path. This observation brings us to the significant concept of ergodicity. Ergodicity is a vital property in the real-world application of stochastic processes, connecting the theoretical aspects with empirical observations, a bridge between what we predict and what we observe. An ergodic process is one in which the time average equals the ensemble average. In simpler terms, it is a property that ensures a single long-term observation can represent the statistical properties of the entire ensemble. Formally, a process is said to be ergodic if, for any time function $f(.)$,

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T f(x(t))dt = \mathbb{E}[f(x(t))] \tag{B.10}$$

The property of ergodicity is fundamental in the statistical study of stochastic processes. It ensures that the long-term time averages taken over a single realization can be representative of the ensemble averages. Ergodicity thus brings about an interesting connection: It gives us the luxury of studying the ensemble through the lens of a single realization. Consequently, ergodicity bridges the gap between theory and practical measurements, where often only a single sample path can be observed.

Appendix C

Average Headway Estimation from Two-Point Correlation Function

We aim at deriving the average headway from probability considerations. Our starting point is the probability q_{ij} of a cell $r+1$ being in a state i (occupied or empty) conditioned by the state j of cell r:

$$q_{ij} = \mathbb{P}[I_\eta(r+1) = i \mid I_\eta(r) = j] = \frac{\mathbb{P}[I_\eta(r+1) = i \cap I_\eta(r) = j]}{\mathbb{P}[I_\eta(r) = j]} \tag{C.1}$$

where $\mathbb{P}[A \cap B]$ stands for the joint probability. Hence, the conditional probability of a cell $r+1$ to be occupied ($i = 1$) given that cell r is occupied ($j = 1$) is readily obtained when recognizing from Eq. (2.1) that $\mathbb{P}[I_\eta(r+1) = i \cap I_\eta(r) = j] = \mathbb{E}[I_\eta(r+1)I_\eta(r)] = S_2(1)$ and $\mathbb{P}[I_\eta(r) = 1] = S_2(0)$, hence:

$$q_{11} = \mathbb{P}[I_\eta(r+1) = 1 \mid I_\eta(r) = 1] = \frac{S_2(1)}{S_2(0)} \tag{C.2}$$

Given the binary nature of occupation, Eq. (C.2) allows us to determine the probability of cell $r+1$ being empty when cell r is occupied:

$$q_{01} = 1 - q_{11} = \frac{S_2(0) - S_2(1)}{S_2(0)} \tag{C.3}$$

Since $\mathbb{P}[A|B] = \mathbb{P}[A \cap B]/\mathbb{P}[B]$ and $\mathbb{P}[B|A] = \mathbb{P}[B \cap A]/\mathbb{P}[A]$, we readily derive from the expression of q_{10} the probability that cell $r+1$ is occupied when cell r is empty; that is:

$$q_{10} = q_{01}\frac{\mathbb{P}[I_\eta(r) = 1]}{\mathbb{P}[I_\eta(r) = 0]} = q_{01}\frac{S_2(0)}{1 - S_2(0)} \tag{C.4}$$

Finally, the conditional probability of two cells, $r+1$ and r, being empty is obtained from:

$$q_{00} = 1 - q_{10} = \frac{1 - 2S_2(0) + S_2(1)}{1 - S_2(0)} \tag{C.5}$$

With the probabilities in hand, we can determine the average headway, while making use of the found Markov property:

$$\mathbb{P}[I_\eta(r + \delta_r) \mid I_\eta(r)] = \mathbb{P}[I_\eta(r + \delta_r)], \quad \delta_r > 1 \tag{C.6}$$

For illustration, consider a realization I_η of the form:

$$1 - 0 - 0 - 1 \tag{C.7}$$

(i.e., cells 1 and 4 are occupied while cells 2 and 3 are empty). From the Markov property (C.6), we know that occupancy of cell 3 (resp. 4) is independent of cell 1 occupancy (resp. 1 and 2). Otherwise said, headway probability reduces to the pairs $1 - 0$ (cells 1-2), $0 - 0$ (cells 2-3), and $0 - 1$ (cells 3-4) defined by probabilities q_{01}, q_{00} and q_{10}. The probability of observing the realization $1 - 0 - 0 - 1$, which is the headway $d = 2$, is thus:

$$q_{01} \times q_{00} \times q_{10} = \frac{(S_2(0) - S_2(1))^2(1 - 2S_2(0) + S_2(1))}{S_2(0)(1 - S_2(0))^2} \tag{C.8}$$

To generalize, consider that the probability of observing a headway d is $q_{10} \times q_{01} \times q_{00}^{d-1}$; whence the average headway:

$$\overline{d} = q_{10} \times q_{01} \times \sum_{d=1}^{\infty} d \times q_{00}^{d-1} = \frac{1}{S_2(0)} - 1 \tag{C.9}$$

where we used the geometric series development,

$$\sum_{d=1}^{\infty} d \times q_{00}^{d-1} = \left(\sum_{d=1}^{\infty} q_{00}^{d} \right)' \overset{0 \leq q_{00} \leq 1}{=} \frac{1}{(1 - q_{00})^2} \tag{C.10}$$

with $()'$ denoting derivation with respect to q_{00}.

Appendix D

Statics of Velocity: Probability Distribution, Transition Matrix, and Deceleration Probability

In this section, we provide a detailed derivation for the near-miss collision risk from the total law of probability, and show how it can be determined from the state transition matrix, provided statistical stationarity.

Consider the state vector of particles $\vec{s}$ as a vector that includes all the possible integer velocity states from standstill $v = 0$ to maximum velocity $v = v_{max}$:

$$\vec{s} = (0, 1, 2, ..., v_{max})^T \tag{D.1}$$

where the superscript T denotes the transpose operator. Herein, we employ reduced units of the Nagel-Schreckenberg cellular automaton model [35], in which units of cell and velocity coincide. Particle k advances from its current position $x_k(t)$ to the future position as $x_k(t+1) = x_k(t) + v_k(t)$ where $v_k(t)$ is the k^{th} vehicle current velocity state. We coin this reduced system of units as the "NaSch" units in reference to the Nagel-Schreckenberg cellular automaton model [35].

In certain physical systems, such as ideal gases, the velocity of particles follows a discrete distribution, where each velocity in the state vector has a probability of

occurrence. We are interested in the marginal probability of observing a particle at velocity state j, which is given by the law of total probability:

$$\pi_j(t+1) = \sum_{i=0}^{v_{max}} q_{i \to j} \pi_i(t) \tag{D.2}$$

where $\pi_i(t) := \mathbb{P}[v_k(t) = i]$ and $q_{i \to j} := \mathbb{P}[v_k(t+1) = j \mid v_k(t) = i]$. In a more general form:

$$\vec{\pi}(t+\tau) = (\underline{q}^T)^\tau \cdot \vec{\pi}(t) \tag{D.3}$$

where $\vec{\pi}(t)$ and $\vec{\pi}(t+1)$ are, respectively, the discrete velocity probability distribution at time t and $t+\tau$, linked by the transition matrix $\underline{q}$:

$$\vec{\pi}(t) = (\pi_0(t), \pi_1(t), ..., \pi_{v_{max}}(t))^T; \quad \underline{q} \equiv q_{i \to j} \tag{D.4}$$

The generalized law of total probability is readily recognized as the mapping of velocity probability distribution at time t, $\vec{\pi}(t)$, to the velocity probability distribution at time $t+\tau$, $\vec{\pi}(t+\tau)$, by means of the state transition matrix $\underline{q}$, also referred to as the stochastic matrix. Row i and column j of $\underline{q}$ represent the current (t) and future $(t+1)$ states of the particle, respectively. Since every particle must transition into one possible state of $\vec{s}$, each row of the transition matrix adds up to one,

$$\underline{q} \cdot \vec{1} = \vec{1} \tag{D.5}$$

with $\vec{1}$ denoting the unit vector.

D.1 Near-Miss Collision Probability

We define near-miss collision risk from the probability of decelerations that exceed operational decelerations due to factors such as traffic density, headway, and other internal characteristics of the traffic flow. The maximum deceleration occurs when a particle at maximum velocity state $v_k(t) = v_{max}$ transitions into standstill state

$v_k(t+1) = 0$. Evoking the total law of probability, this excessive deceleration risk is part of the marginal probability to find a velocity $v_k(t+1) = 0$ from Eq. (D.2):

$$\pi_0(t+1) = \sum_{i=0}^{v_{max}-1} q_{i\to 0}\pi_i(t) + \mathbb{P}[a_k(t) = -v_{max}] \tag{D.6}$$

where $\mathbb{P}[a_k(t) = -v_{max}]$ is the probability of observing a particle transitioning from $v_k(t) = v_{max}$ to $v_k(t+1) = v_k(t) + a(t) = 0$; that is:

$$\mathbb{P}[a_k(t) = -v_{max}] = \pi_j\, q_{j\to i};\;\; \forall j = v_{max}, i = 0 \tag{D.7}$$

This is the definition of the near-miss collision risk [Eq.(1)] in the main text, which is at the core of our analysis.

D.2 Statistical Stationarity

The homogeneity in the random behavior of particles, along with the closed system behavior of traffic flow (periodic boundary conditions), induces statistical stationarity in the system, thus resulting in a constant velocity probability distribution at all times:

$$\vec{\pi}(t+1) = \vec{\pi}(t) = \vec{\pi} \tag{D.8}$$

In this case, with the help of the Perron–Frobenius theorem, the velocity probability distribution $\vec{\pi}$ is readily obtained as the eigenvector of transposed transition matrix $\underline{q}^T$ associated with the unit eigenvalue:

$$(\underline{q}^T - \underline{I}) \cdot \vec{\pi} = \vec{0} \tag{D.9}$$

with $\underline{I}$ the identity matrix. In other words, the velocity distribution is the eigenstate corresponding to the unit eigenvalue of the transposed state transition matrix, or, equivalently, the stationary distribution of $\underline{q}^T$ [145, 146].

In summary, the two probabilities that define the near-miss collision risk, namely

$\pi_{v_{max}} := \mathbb{P}[v_k(t) = v_{max}]$ and $q_{v_{max} \to 0} := \mathbb{P}[v_k(t+1) = 0 \mid v_k(t) = v_{max}]$ for any t, are entirely defined by the state transition matrix $\underline{q}$ given the stationarity of the system. This is the second pillar of our contribution.

Appendix E

Determination of the State Transition Matrix using the Autocovariance Function

The state transition matrix is at the core of our approach. This section provides a means of determining the state transition matrix from the velocity autocovariance function, which is available from ensemble velocity data.

E.1 Dual Definition of Velocity Autocovariance Function $C_{vv}(\tau)$

The autocovariance function, $C_{vv}(\tau)$, is defined as the ensemble expected value of the product of the deviation of velocity from its mean at two time steps separated by a time lag τ. The canonical definition of C_{vv} reads:

$$C_{vv}(|t_1 - t_2| = \tau) = \mathbb{E}[(v_k(t_1) - \bar{v})(v_k(t_2) - \bar{v})] \tag{E.1}$$

with $\mathbb{E}[.]$ the ensemble expected value operator, and $\bar{v}$ the ensemble velocity average, which can be expressed in function of the velocity probability distribution and the state vector as $\bar{v} = \vec{s}^T \cdot \vec{\pi}$. Conditioning velocity state at time t_2 upon the veloc-

ity state at time t_1, we define the expected velocity vector as $\vec{\nu}(|t_1 - t_2| = \tau) = (\mathbb{E}[v_k(t_2)|v_k(t_1) = 0], \mathbb{E}[v_k(t_2)|v_k(t_1) = 1], ..., \mathbb{E}[v_k(t_2)|v_k(t_1) = v_{max}])^T$ which reads,

$$\vec{\nu}(\tau) = \underline{q}^{\tau} \cdot \vec{s} \tag{E.2}$$

The autocovariance function of this process can therefore be written as $C_{vv}(\tau) = \vec{\nu}(\tau)^T \cdot \text{diag}(\vec{s}) \cdot \vec{\pi} - \bar{v}^2$, which reduces to,

$$C_{vv}(\tau) = \vec{s}^T \cdot (\underline{q}^{\tau})^T \cdot \text{diag}(\vec{s}) \cdot \vec{\pi} - \bar{v}^2 \tag{E.3}$$

That is, expressions (E.1) and (E.3) provide dual definitions of the autocovariance function: the first is the canonical definition of autocovariance function from the velocity of particles, while the second one is expressed in function of the state transition matrix. This dual definition provides a means of determining the state transmission matrix $\underline{q}$ from the autocovariance function or vice versa. For instance, if $\underline{q}$ is known, the application of Eq. (E.3) provides a means of determining autocovariance $C_{vv}(\tau)$ from:

$$C_{vv}(\tau) = \vec{s}^T \cdot (\underline{q}^{\tau})^T \cdot \text{diag}(\vec{s}) \cdot \vec{\pi} - \vec{s}^T \cdot (\vec{\pi}^T \cdot \vec{\pi})^T \cdot \vec{s} \tag{E.4}$$

Conversely, if $C_{vv}(\tau)$ is known from velocity data of particles, minimizing the difference between values provided by the canonical definition and the transition-matrix-based expression of $C_{vv}(\tau)$ allows for obtaining the transition matrix, and hence the near-miss collision risk [Eq. (D.7)]. This is the third contribution of this letter.

E.2 Specification for Ornstein-Uhlenbeck -Type Stochastic Process

One well-known feature of the velocity autocovariance function is the exponential decay of the velocity autocovariance function [161]. This exponential decay is remi-

niscent of an Ornstein-Uhlenbeck (OU) stochastic process [130, 131], for which,

$$\frac{dC_{vv}}{d\tau}|_{\tau=0} = -\frac{\sigma_v^2}{\tau_c} \tag{E.5}$$

where the characteristic decay time τ_c defines the so-called temporal memory, and σ_v^2 is the second-order central moment of the ensemble velocity. Specifically, the autocovariance function of the OU process reads as:

$$C_{vv}^{\exp}(\tau) = \sigma_v^2 \exp(-\tau/\tau_c) \tag{E.6}$$

where $\tau = 0, 1, 2, ..., \infty$ are discrete values of time lag τ in NaSch units. A quadratic minimization can be employed to determine the elements of the state transition matrix from the dual definition of the autocovariance function:

$$\min_{\underline{q}} \sum_{\tau=0}^{\infty} \left[\vec{s}^T \cdot (\underline{q}^\tau)^T \cdot \mathrm{diag}(\vec{s}) \cdot \vec{\pi} - (\bar{v}^2 + \sigma_v^2 \exp(-\tau/\tau_c))\right]^2 \tag{E.7}$$

subject to,

$$\begin{cases} (\underline{q}^T - \underline{I}) \cdot \vec{\pi} = \vec{0} \\ \underline{q} \cdot \vec{1} = \vec{1} \end{cases} \tag{E.8}$$

where, as readily understood, $\bar{v} = \vec{s}^T \cdot \vec{\pi}$ and $\sigma_v^2 = \vec{s}^T \cdot \mathrm{diag}(\vec{s}) \cdot \vec{\pi} - \bar{v}^2$ are ensemble mean and variance of the particles' velocity.

Appendix F

Application to Nagel-Schreckenberg Cellular Automaton Model

We present details on the application of our proposed algorithm to the data set generated via model-based simulations. In the limiting cases of dilute ($\rho << \rho_c^p$) and jammed ($\rho \to 1$) regimes, analytical expression can be derived for the transition matrix, which are shown next. Within these two regimes, the transition matrix allows us to explicitly obtain the near-miss collision risk. We then proceed by presenting the approximate solutions for the statistics of the near-miss events in the congested regime ($\rho_c^p < \rho < 1$) based on the large-scale simulations of the model.

F.1 Dilute Regime: $\rho << \rho_c^p$

Below the transition density $\rho << \rho_c^p$, interactions between particles are negligible allowing for the free flow of vehicles in the system. In this dilute regime, the velocity of every vehicle is an independent random process resembling a Bernoulli process. In other words, velocity of vehicles is independent identically distributed random variables, which permits an analytical determination of the state transition matrix.

Specifically, considering the NaSch heuristics in (A.1), a vehicle at v_{max} either transitions into $v_{max} - 1$ with probability p or remains at its current state with probability $1 - p$, i.e., $\mathbb{P}[v_k(t+1) = v_{max} - 1 \mid v_k(t) = v_{max}] = p$, and $\mathbb{P}[v_k(t+1) =$

$v_{max} \mid v_k(t) = v_{max}] = 1-p$. Similarly, a vehicle at $v_{max}-1$ transitions into state v_{max} with probability $1-p$, or remains at state $v_{max}-1$ with probability p. Moreover, if a particle is at any other state than v_{max} and $v_{max}-1$, it will certainly (with probability one) transition into one higher state, i.e., $\mathbb{P}[v_k(t+1) = i+1 \mid v_k(t) = i < v_{max}-1] = 1$. The state transition matrix is thus of the form:

$$\lim_{\rho/\rho_c^p \to 0} \underline{q} = \begin{bmatrix} 0 & 1 & 0 & ... & 0 & 0 \\ 0 & 0 & 1 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & p & 1-p \\ 0 & 0 & 0 & ... & p & 1-p \end{bmatrix} \tag{F.1}$$

The eigenvector of $\underline{q}^T$, which represents the probability distribution of velocity, is readily obtained:

$$\lim_{\rho/\rho_c^p \to 0} \vec{\pi} = (0, 0, 0, ..., p, 1-p)^T \tag{F.2}$$

Finally, the two probabilities required to determine the near-miss collision risk are $\pi_{v_{max}} = 1 - p$ and $q_{v_{max} \to 0} = 0$ in the dilute regime, whence,

$$\lim_{\rho/\rho_c^p \to 0} \mathbb{P}[a_k(t) = -v_{max}] = 0 \tag{F.3}$$

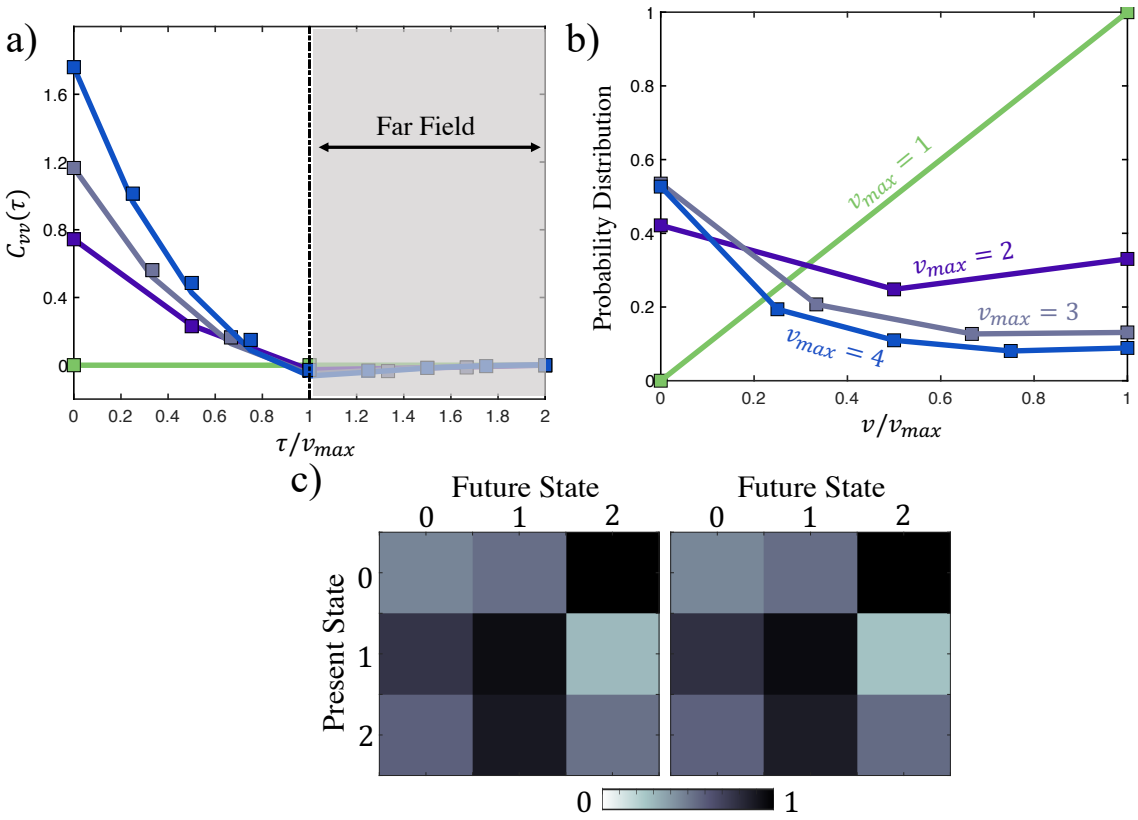

Figure F-1: (a) Autocovariance function for different levels of v_{max} when half of the cells are occupied, $\rho = 1/2$, and (b) their corresponding eigenstates $\vec{\pi}$; the lines are obtained from NaSch simulations and the squares are obtained from the far-field analysis. (c) Transition matrix for $v_{max} = 2$ obtained from the simulations (left) and far-field analysis (right).

F.2 Jammed Regime: $\rho \to 1$

In the jammed regime, the probability of observing a standing car is close to one as most of the cars have zero velocity, i.e., $\mathbb{P}[v_k(t) = 0] = 1 - \epsilon_0$ with $\epsilon_0 << 1$. The probability of observing a particle at other velocity states, $v_k(t) > 0$, is therefore extremely small; that is, $\mathbb{P}[v_k(t) = i] = \epsilon_i$ for $i \geq 1$ with $\epsilon_i << 1$. Since the state probabilities must add up to one, we recognize that $\sum_{i=1}^{v_{max}} \epsilon_i = \epsilon_0$. Invoking Bayes' theorem, the state transition probability $q_{i \to j}$ is related to the inverse transition probability $q_{j \to i}$ via:

$$q_{i \to j} = \frac{\mathbb{P}[v_k(t) = j] q_{j \to i}}{\mathbb{P}[v_k(t) = i]} \tag{F.4}$$

Whereby, considering $i = 0$ and $j \geq 1$:

$$q_{0 \to j>0} = \lim_{\epsilon_0 \to 0} \frac{q_{j \to 0} \epsilon_j}{1 - \epsilon_0} = 0 \tag{F.5}$$

In a similar fashion, it can be shown that $q_{j \geq 0 \to 0} = 1$. Therefore, all entries of the transition matrix in the jammed regime are zero except those in the first column which are $q_{i \to 0} = 1$:

$$\lim_{\rho \to 1} \underline{q} = \begin{bmatrix} 1 & 0 & ... & 0 & 0 \\ 1 & 0 & ... & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & ... & 0 & 0 \\ 1 & 0 & ... & 0 & 0 \end{bmatrix} \tag{F.6}$$

The velocity probability distribution corresponding to $\underline{q}^T$, therefore, reads as

$$\lim_{\rho \to 1} \vec{\pi} = (1, 0, 0, ..., 0)^T. \tag{F.7}$$

The near-miss collision risk in the jammed regime is zero, due to the fact that the probability of finding a vehicle at maximum speed is zero $\pi_{v_{max}} = 0$, whereas the conditional probability that a particle at maximum speed transitions into the standstill

state is one, $q_{v_{max}\to 0} = 1$. Whence, the near-miss collision risk:

$$\lim_{\rho\to 1} \mathbb{P}[a_k(t) = -v_{max}] = 0 \tag{F.8}$$

In summary, the asymptotic behaviors of the NaSch-model, Eqs. (F.3) and (F.8), nicely illustrate the relevance of the proposed definition (D.7) of near-miss collision risk based on the product of two probabilities, and that considering only one of the two (i.e., either maximum velocity probability $\pi_{v_{max}}$, or hard-brake conditional probability $q_{v_{max}\to 0}$ as often done by application of telematic devices) is insufficient to estimate a near-miss collision risk.

F.3 Congested Regime: $\rho_c^p < \rho < 1$

The transition matrix in the congested flow requires consideration of higher-order statistical moments due to the coexistence of free flow and jammed state [129]. Focusing on the far-field behavior and memory effects in, respectively, the deterministic ($p = 0$) and stochastic ($p \neq 0$) models, we proceed by deriving approximate solutions for the transition probability and probability of observing a vehicle at the highest velocity state, as they play a significant role in quantifying the near-miss collision risk.

F.3.1 Deterministic Model

With no randomized braking, velocities at time steps t_1 and t_2 are minimally correlated as long as the time lag $\tau = |t_1 - t_2|$ is greater than v_{max}, i.e., $C_{vv}(\tau \geq v_{max}) \to 0$. Therefore, using the definition of autocovariance function, the far-field behavior of the deterministic model can be enforced via,

$$\vec{s}^T.\left(\left(\underline{q}^\tau\right)^T.\mathrm{diag}(\vec{\pi}) - \vec{\pi}.\vec{\pi}^T\right).\vec{s} \to 0, \quad \forall \tau \geq v_{max} \tag{F.9}$$

Whence,

$$\underline{q}^\tau = \underline{1}.\vec{\pi}^T, \quad \forall \tau \geq v_{max} \tag{F.10}$$

The above equation provides a set of equations which allows us to obtain the transition matrix. Figure F-1 illustrates the autocovariance function and their corresponding eigenstates for different maximum velocities in the congested regime, exhibiting a remarkable agreement between the transition matrices obtained from the canonical definition and the far-field behavior using the transition matrix definition [Eq. (F.9)]. Specifically, we find that the transition probability $q_{v_{max}\to 0}$ is the same as the probability of observing a standing vehicle in the absence of stochastic velocity fluctuations, i.e., $q_{v_{max}\to 0} = \mathbb{P}[v_k(t) = 0]$. The probability of observing a standing vehicle shows a universal pattern with respect to the jamming probability $\bar{\eta}$ [Fig. 3-2.(a)] in the form:

$$q_{v_{max}\to 0} = \mathbb{P}[v_k(t) = 0] = \frac{2\bar{\eta}}{1+\bar{\eta}}, \qquad (p = 0) \tag{F.11}$$

The above equation is a generalization of the deterministic solution for the binary-state system NaSch model [Appendix G]. Similarly, through a power generalization of the deterministic binary-state system, the probability of observing maximum velocity for $v_{max} \geq 1$ can be obtained from:

$$\frac{d\mathbb{P}[v_k(t) = v_{max}|p = 0]}{d\bar{\eta}} = -\frac{-2\gamma}{(1+v_{max}\eta)^{\alpha}} \tag{F.12}$$

where the constants, γ and α, are obtained by simulations satisfying the boundary conditions $\lim_{\bar{\eta}\to 0}\mathbb{P}[v_k(t) = v_{max}] = 1$ and $\lim_{\bar{\eta}\to 1}\mathbb{P}[v_k(t) = v_{max}] = 0$. The solution to the above differential equation is,

$$\mathbb{P}[v_k(t) = v_{max}|p = 0] = 1 - \beta\left((1+v_{max}\bar{\eta})^{1-\alpha} - 1\right) \tag{F.13}$$

with $\beta = (1+v_{max})^{\alpha}/(1+v_{max}-(1+v_{max})^{\alpha})$ and the fitting parameter $\alpha \approx 2 + \tanh\sqrt{(v_{max}-1)/2}$ [Fig. 3-2.(b)].

F.3.2 Stochastic Model

From the binary-state model, we readily realize that the impact of stochasticity on probability of observing maximum velocity depends non-linearly on the density.

Specifically, upon transitioning $\lim_{\bar{\eta}\to 0}\mathbb{P}[v_k(t) = v_{max}] = 1 - \sqrt{p}$, and in the fully-jammed regime stochasticity plays no role since $\lim_{\bar{\eta}\to 1}\mathbb{P}[v_k(t) = v_{max}] = 0$. We decouple the effect of stochasticy and density by approximating the general behavior of $\mathbb{P}[v_k(t) = v_{max}]$ in the congested regime as,

$$\mathbb{P}[v_k(t) = v_{max}|p \neq 0] \approx (1 - \sqrt{p})\left(\mathbb{P}[v_k(t) = v_{max}|p = 0]\right)^{\xi} \tag{F.14}$$

with exponent $\xi = \xi_0 + \xi_1 \tanh(v_{max} - 1)$ where $\xi_1/\xi_0 \approx .85$, and,

$$\begin{aligned}\xi_0 = \underset{\xi}{\operatorname{argmin}} \int_{p=0}^{1}\int_{\rho=1/2}^{1} \Bigg|(1-\sqrt{p})(\frac{1}{\rho}-1)- \\ \frac{1-\sqrt{1-4\rho(1-p)(1-\rho)}}{2\rho}\Bigg|\mathrm{d}\rho\mathrm{d}p\end{aligned} \tag{F.15}$$

resulting in $\xi_0 = 0.77$ [Fig. 3-3.(a)].

We proceed by exploring the intricate interaction between the transition density $q_{v_{max}\to 0}$, and probability of observing a vehicle at velocity states 0 (standing) and v_{max}. In particular, the probabilities of observing a standing vehicle and the transition probability are approximated via power generalization as,

$$\mathbb{P}[v_k(t) = 0] \approx 1 - \mathbb{P}[v_k(t) = v_{max}]^{1-\nu_v} \tag{F.16}$$

$$q_{v_{max}\to 0} \approx 1 - \mathbb{P}[v_k(t) = v_{max}]^{1-\nu_q} \tag{F.17}$$

with $\nu_v = \hat{\nu}_v \tanh(v_{max} - 1)$ and $\nu_q = \hat{\nu}_q \tanh(v_{max} - 1)$ [Fig. 3-3.(b)]. Such generalizations satisfy the boundary conditions and allow us to express the transition density as a function of probability of observing a standing vehicle as,

$$q_{v_{max}\to 0} \approx 1 - (1 - \mathbb{P}[v_k(t) = 0])^{(1-\nu_q)/(1-\nu_v)} \tag{F.18}$$

which degenerates to the exact relation for the binary-state model $q_{v_{max}\to 0} = \mathbb{P}[v_k(t) = 0]$.

Finally, the asymptotic behavior underscores one finding of our paper that the

near-miss collision risk pertains to the congested regime of vehicular traffic flow. This motivates the development of the phase diagram of near-miss collision risk in the NaSch-phase space, $\rho \in]\rho_c^p, 1[$ and $p \in]0, 1[$, shown and discussed in the main text.

Appendix G

Analytical Solutions for the Binary-state NaSch Model

Here, we focus on the special case of binary velocity $v_{max} = 1$ which permits the particles to be either in standing (0) or moving (1) velocity states. The canonical form of the transition matrix for $v_{max} = 1$ reads,

$$\underline{q} = \begin{pmatrix} q_{0\to0} & q_{0\to1} \\ q_{1\to0} & q_{1\to1} \end{pmatrix} \tag{G.1}$$

where $q_{0\to0}+q_{0\to1} = q_{1\to0}+q_{1\to1} = 1$. The eigenstate corresponding to unit eigenvalue of the transposed form of this state transition matrix allows us to obtain the velocity distribution as [Eq. (D.9)],

$$\vec{\pi} = \frac{1}{q_{0\to1} + q_{1\to0}} \begin{pmatrix} q_{1\to0} \\ q_{0\to1} \end{pmatrix} \tag{G.2}$$

We readily realize that the probability of observing standing (zero-state velocity) and moving (unit-state velocity) vehicles in terms of transition probabilities are $\mathbb{P}[v_k(t) = 0] = q_{1\to0}/(q_{0\to1}+q_{1\to0})$ and $\mathbb{P}[v_k(t) = 1] = q_{0\to1}/(q_{0\to1}+q_{1\to0})$, respectively. Furthermore, eigendecompositon allows us to factorize the transition matrix into its

eigenvalues and eigenvectors, as,

$$\underline{q} = \begin{pmatrix} 1 & -q_{01}/q_{10} \\ 1 & 1 \end{pmatrix} . \begin{pmatrix} 1 & 0 \\ 0 & q_{00} - q_{10} \end{pmatrix} . \begin{pmatrix} 1 & -q_{01}/q_{10} \\ 1 & 1 \end{pmatrix}^{-1} \tag{G.3}$$

Two-state system resembles an OU stochastic process as the autocovariance function decays exponentially with $dC_{vv}/d\tau|_{\tau=0} = -\sigma_v^2/\tau_c$ where the variance of the speed $\sigma_v^2 = q_{1\to0}q_{0\to1}/(q_{1\to0}+q_{0\to1})^2$ and the characteristic decay time $\tau_c = -(\ln(q_{0\to0} - q_{1\to0}))^{-1}$.

For $v_{max} = 1$, the probability of observing a particle turn into zero velocity state from current unit velocity state is the same as the probability that a particle remains at zero velocity state, i.e., $q_{0\to0} = q_{1\to0}$. This unique property of binary velocity results in $\ln(q_{0\to0} - q_{1\to0}) \to -\infty$ and, consequently, $\tau_c \to 0$. The transition matrix is therefore fully determined from the velocity distribution as $q_{0\to0} = q_{1\to0} = \mathbb{P}[v_k(t) = 0]$. From exact 2-cluster solution [147],

$$q_{1\to0} = 1 - \frac{1 - \sqrt{1 - 4\rho(1-p)(1-\rho)}}{2\rho} \tag{G.4}$$

and also $\mathbb{P}[v_k(t) = 1] = 1 - q_{1\to0}$. Whence, the probability of observing deceleration in the Binary NaSch model is obtained as,

$$\mathbb{P}[a_k(t) = -1] = q_{1\to0}(1 - q_{1\to0}) \tag{G.5}$$

where the transition probability is given in Eq. (G.4). For a given stochasticity parameter p, the critical density, ρ^p_{crit}, at which the probability of observing maximal deceleration is maximum is obtained through solving the following equation for ρ,

$$\frac{\partial\mathbb{P}[a_k(t) = -1]}{\partial\rho}\Big|_p = \frac{\partial\mathbb{P}[a_k(t) = -1]}{\partial\mathbb{P}[v_k(t) = 0]} . \frac{\partial\mathbb{P}[v_k(t) = 0]}{\partial\rho}\Big|_p = 0 \tag{G.6}$$

thus resulting in,

$$\rho^p_{critic} = \frac{4p - 2}{4p - 3} \geq 0 \tag{G.7}$$

For $p \leq 1/2$, the probability of observing maximal deceleration is therefore $\mathbb{P}[a_k(t) =$

$-1] = 1/4$ regardless of the stochasticity parameter.

Appendix H

From Crowdsourced Velocity Signals to Near-miss Risk

The second data set we use originates from the crowdsourced vehicle velocity data collected by the Carbin Educational Application for smartphones [149]. The focus of this section is to show that the proposed approach is applicable to real-world measurements of traffic independent of specific model attributes. This section shows how velocity data recorded by a single user is employed to determine the near-miss collision risk.

Vehicle velocity data $V(t)$ is collected at 1Hz frequency by sensors (here smartphones) installed in vehicles. The velocity data is transposed into the integer velocity NaSch units, considering

$$v(t) = \lfloor v_{max} \frac{V(t)}{V_{max}} + 0.5 \rfloor \leq v_{max} \tag{H.1}$$

where $\lfloor x \rfloor \leq x < \lfloor x \rfloor + 1$ is the floor function, and V_{max} is the real-life maximum speed limit on the road. Each vehicle velocity signal was split into overlapping trips of 4×10^2 seconds, which is long enough to provide stationarity and, furthermore, ergodicity in the system (for an in-depth discussion of ergodicity for crowdsourced data, see [161]). The concept of ergodicity in this context posits that a vehicle will ultimately explore the entire phase space in a uniform manner over a proper timescale, much akin to the

ergodic theorem in statistical mechanics [135, 164]. As a result, the velocity signal of an individual vehicle is deemed to possess sufficient statistical richness, enabling it to represent the ensemble's moments in their entirety. This allows us to reduce the canonical definition of the autocovariance function to,

$$C_{vv}(\tau) \approx \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} v(t)v(t+\tau) - \bar{v}^2 \tag{H.2}$$

where T is the duration of the signal, and $\bar{v} = (1/T)\sum_{t=1}^{T} v(t)$ is the mean of the velocity signal. With the autocovariance function in hand, we perform the minimization described in Section B to obtain the transition matrix [Fig. H-1], and determine the near-miss collision risk.

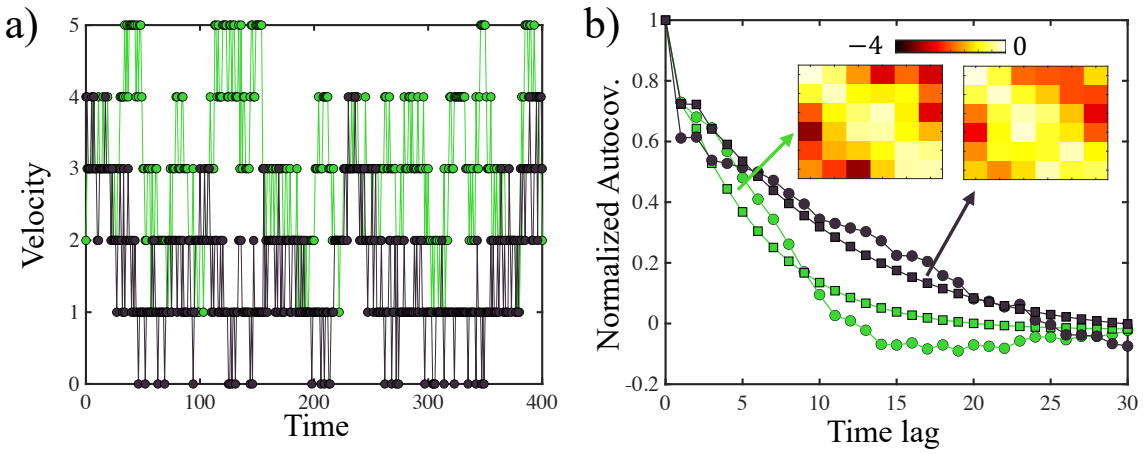

Figure H-1: Sample analysis: (a) the NaSch velocity representation of two trips, and (b) their corresponding normalized autocovariance function, $C_{vv}(\tau)/\sigma_v^2$, obtained from the canonical definition (disks), and transition matrix analysis (squares); the logarithm of transition matrices obtained for each trip are illustrated as the inset. The velocity signal in green and black have an average velocity of $\bar{v} = 3$ and $\bar{v} = 1.5$, respectively, and their corresponding near-miss risks are $10^{-2.6}$ and $10^{-3.3}$, respectively.

Bibliography

[1] Allison Goldberg, Hugh Foley, Brandon Folck, Sue Gallagher, Stephen Hargarten, Dale Herzog, Rubina Ohanian, and Scott Ratzan. *Improving Global Road Safety: Corporate Sector Commitments and Opportunities*. National Academy of Medicine, 2015.

[2] Simiao Chen, Michael Kuhn, Klaus Prettner, and David E Bloom. The global macroeconomic burden of road injuries: estimates and projections for 166 countries. *The Lancet Planetary Health*, 3(9):e390–e398, 2019.

[3] Kai Nagel, Peter Wagner, and Richard Woesler. Still flowing: Approaches to traffic flow and traffic jam modeling. *Operations research*, 51(5):681–710, 2003.

[4] *Carbin Educational Smartphone Application (Android)*, 2021. https://play.google.com/store/apps/details?id=com.carbin.carbin2&hl=en_US&gl=US/ [2021-04-27].

[5] *Carbin Educational Smartphone Application (iOS)*, 2021. https://apps.apple.com/us/app/carbin/id1559648286/ [2021-04-27].

[6] MassDOT. Crash query and visualization, 2022.

[7] Luis Amador-Jimenez and Christopher J Willis. Demonstrating a correlation between infrastructure and national development. *International Journal of Sustainable Development & World Ecology*, 19(3):197–202, 2012.

[8] Marco Percoco. Highways, local economic structure and urban development. *Journal of Economic Geography*, 16(5):1035–1054, 2016.

[9] Peter H Verburg, Erle C Ellis, and Aurelien Letourneau. A global assessment of market accessibility and market influence for global environmental change studies. *Environmental Research Letters*, 6(3):034019, 2011.

[10] World Bank. *World development report 2009: Reshaping economic geography*. The World Bank, 2008.

[11] D.J. Weiss, Andy Nelson, HS Gibson, W Temperley, Stephen Peedell, A Lieber, M Hancher, Eduardo Poyart, Simão Belchior, Nancy Fullman, et al. A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature*, 553(7688):333–336, 2018.

[12] Keywan Riahi, Detlef P Van Vuuren, Elmar Kriegler, Jae Edmonds, Brian C O'neill, Shinichiro Fujimori, Nico Bauer, Katherine Calvin, Rob Dellink, Oliver Fricko, et al. The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global environmental change*, 42:153–168, 2017.

[13] *World Health Organization*, 2022. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries [2022-11-27].

[14] CDC. Road traffic injuries & deaths: A global problem, 11 2019.

[15] Edouard Mathieu, Hannah Ritchie, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Joe Hasell, Bobbie Macdonald, Saloni Dattani, Diana Beltekian, Esteban Ortiz-Ospina, and Max Roser. Coronavirus pandemic (covid-19). *Our World in Data*, 2020. https://ourworldindata.org/coronavirus.

[16] FHWA. Highway safety information system, 2022.

[17] CDC. Deaths: Leading causes for 2019deaths: Leading causes for 2019, 11 2019.

[18] World Health Organization. *Global status report on road safety 2015*. World Health Organization, 2015.

[19] Andres I Vecino-Ortiz, Aisha Jafri, and Adnan A Hyder. Effective interventions for unintentional injuries: a systematic review and mortality impact assessment among the poorest billion. *The Lancet Global Health*, 6(5):e523–e534, 2018.

[20] Dan Chisholm, Huseyin Naci, Adnan Ali Hyder, Nhan T Tran, and Margie Peden. Cost effectiveness of strategies to combat road traffic injuries in sub-saharan africa and south east asia: mathematical modelling study. *Bmj*, 344, 2012.

[21] W Kip Viscusi and Joseph E Aldy. The value of a statistical life: a critical review of market estimates throughout the world. *Journal of risk and uncertainty*, 27:5–76, 2003.

[22] Lawrence Blincoe, Ted R Miller, Eduard Zaloshnja, and Bruce A Lawrence. The economic and societal impact of motor vehicle crashes, 2010 (revised). Technical report, 2015.

[23] Rune Elvik. An analysis of official economic valuations of traffic accident fatalities in 20 motorized countries. *Accident analysis & prevention*, 27(2):237–247, 1995.

[24] Rune Elvik. How much do road accidents cost the national economy? *Accident Analysis & Prevention*, 32(6):849–851, 2000.

[25] Kavi Bhalla, Majid Ezzati, Ajay Mahal, Joshua Salomon, and Michael Reich. A risk-based method for modeling traffic fatalities. *Risk Analysis: An International Journal*, 27(1):125–136, 2007.

[26] Arvind Kumar Gupta and VK Katiyar. Analyses of shock waves and jams in traffic flow. *Journal of Physics A: Mathematical and General*, 38(19):4069, 2005.

[27] Takashi Nagatani. Jamming transition in a two-dimensional traffic flow model. *Physical Review E*, 59(5):4857, 1999.

[28] Boris S Kerner. The physics of traffic. *Physics World*, 12(8):25, 1999.

[29] Takashi Nagatani. The physics of traffic jams. *Reports on progress in physics*, 65(9):1331, 2002.

[30] Gordon Frank Newell. Nonlinear effects in the dynamics of car following. *Operations research*, 9(2):209–229, 1961.

[31] Toshimitsu Musha and Hideyo Higuchi. Traffic current fluctuation and the burgers equation. *Japanese journal of applied physics*, 17(5):811, 1978.

[32] Gerald Beresford Whitham. Exact solutions for a discrete system arising in traffic flow. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 428(1874):49–69, 1990.

[33] Teruhisa S Komatsu and Shin-ichi Sasa. Kink soliton characterizing traffic congestion. *Physical Review E*, 52(5):5574, 1995.

[34] Ofer Biham, A Alan Middleton, and Dov Levine. Self-organization and a dynamical transition in traffic-flow models. *Physical Review A*, 46(10):R6124, 1992.

[35] Kai Nagel and Michael Schreckenberg. A cellular automaton model for freeway traffic. *Journal de physique I*, 2(12):2221–2229, 1992.

[36] Boris S Kerner and Peter Konhäuser. Cluster effect in initially homogeneous traffic flow. *Physical review E*, 48(4):R2335, 1993.

[37] Masako Bando, Katsuya Hasebe, Akihiro Nakayama, Akihiro Shibata, and Yuki Sugiyama. Dynamical model of traffic congestion and numerical simulation. *Physical review E*, 51(2):1035, 1995.

[38] Takashi Nagatani. Density waves in traffic flow. *Physical Review E*, 61(4):3564, 2000.

[39] Takashi Nagatani. Traffic jams induced by fluctuation of a leading car. *Physical Review E*, 61(4):3534, 2000.

[40] Dirk Helbing and Benno Tilch. Generalized force model of traffic dynamics. *Physical review E*, 58(1):133, 1998.

[41] Elad Tomer, Leonid Safonov, and Shlomo Havlin. Presence of many stable nonhomogeneous states in an inertial car-following model. *Physical Review Letters*, 84(2):382, 2000.

[42] Martin Treiber, Ansgar Hennecke, and Dirk Helbing. Congested traffic states in empirical observations and microscopic simulations. *Physical review E*, 62(2):1805, 2000.

[43] Arne Kesting, Martin Treiber, and Dirk Helbing. Enhanced intelligent driver model to access the impact of driving strategies on traffic capacity. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1928):4585–4605, 2010.

[44] Martin Treiber and Arne Kesting. The intelligent driver model with stochasticity-new insights into traffic flow oscillations. *Transportation research procedia*, 23:174–187, 2017.

[45] Anthony D Mason and Andrew W Woods. Car-following model of multispecies systems of road traffic. *Physical Review E*, 55(3):2203, 1997.

[46] Akihiro Nakayama, Yūki Sugiyama, and Katsuya Hasebe. Effect of looking at the car that follows in an optimal velocity model of traffic flow. *Physical Review E*, 65(1):016112, 2001.

[47] Takashi Nagatani. Stabilization and enhancement of traffic flow by the next-nearest-neighbor interaction. *Physical Review E*, 60(6):6395, 1999.

[48] Ihor Lubashevsky, Sergey Kalenkov, and Reinhard Mahnke. Towards a variational principle for motivated vehicle motion. *Physical Review E*, 65(3):036140, 2002.

[49] Luís Vasconcelos, Luís Neto, Sílvia Santos, Ana Bastos Silva, and Álvaro Seco. Calibration of the gipps car-following model using trajectory data. *Transportation research procedia*, 3:952–961, 2014.

[50] Arne Kesting and Martin Treiber. Calibrating car-following models by using trajectory data: Methodological study. *Transportation Research Record*, 2088(1):148–156, 2008.

[51] Carolina Osorio and Vincenzo Punzo. Efficient calibration of microscopic car-following models for large-scale stochastic network simulators. *Transportation Research Part B: Methodological*, 119:156–173, 2019.

[52] Dirk Helbing and Martin Treiber. Gas-kinetic-based traffic model explaining observed hysteretic phase transition. *Physical Review Letters*, 81(14):3042, 1998.

[53] Serge P Hoogendoorn and Piet HL Bovy. Generic gas-kinetic traffic systems modeling with applications to vehicular traffic flow. *Transportation Research Part B: Methodological*, 35(4):317–336, 2001.

[54] Ilya Prigogine and Robert Herman. Kinetic theory of vehicular traffic. Technical report, 1971.

[55] CMJ Tampère, Bart Van Arem, and SP Hoogendoorn. Gas-kinetic traffic flow modeling including continuous driver behavior models. *Transportation research record*, 1852(1):231–238, 2003.

[56] E Ben-Naim and PL Krapivsky. Stationary velocity distributions in traffic flows. *Physical Review E*, 56(6):6680, 1997.

[57] E Ben-Naim and PL Krapivsky. Steady-state properties of traffic flows. *Journal of Physics A: Mathematical and General*, 31(40):8073, 1998.

[58] E Ben-Naim and PL Krapivsky. Maxwell model of traffic flows. *Physical Review E*, 59(1):88, 1999.

[59] I Ispolatov and PL Krapivsky. Phase transition in a traffic model with passing. *Physical Review E*, 62(5):5935, 2000.

[60] R Mahnke and N Pieret. Stochastic master-equation approach to aggregation in freeway traffic. *Physical Review E*, 56(3):2666, 1997.

[61] Serge P Hoogendoorn and Piet HL Bovy. Gas–kinetic model for multilane heterogeneous traffic flow. *Transportation research record*, 1678(1):150–159, 1999.

[62] Dirk Helbing. Modeling multi-lane traffic flow with queuing effects. *Physica A: Statistical Mechanics and its Applications*, 242(1-2):175–194, 1997.

[63] D Ngoduy. Application of gas-kinetic theory to modelling mixed traffic of manual and acc vehicles. *Transportmetrica*, 8(1):43–60, 2012.

[64] HY Lee, H-W Lee, and D Kim. Origin of synchronized traffic flow on highways and its dynamic phase transitions. *Physical Review Letters*, 81(5):1130, 1998.

[65] Hyun Keun Lee and Beom Jun Kim. Dissolution of traffic jam via additional local interactions. *Physica A: Statistical Mechanics and its Applications*, 390(23-24):4555–4561, 2011.

[66] Astrid S de Wijn, DM Miedema, B Nienhuis, and P Schall. Criticality in dynamic arrest: Correspondence between glasses and traffic. *Physical Review Letters*, 109(22):228001, 2012.

[67] Nicolas Bain, Thorsten Emig, Franz-Josef Ulm, and Michael Schreckenberg. Velocity statistics of the nagel-schreckenberg model. *Physical Review E*, 93(2):022305, 2016.

[68] Michael Schreckenberg, Andreas Schadschneider, Kai Nagel, and Nobuyasu Ito. Discrete stochastic models for traffic flow. *Physical Review E*, 51(4):2939, 1995.

[69] Junta Matsukidaira and Katsuhiro Nishinari. Euler-lagrange correspondence of cellular automaton for traffic-flow models. *Physical review letters*, 90(8):088701, 2003.

[70] L Roters, S Lübeck, and KD Usadel. Critical behavior of a traffic flow model. *Physical Review E*, 59(3):2672, 1999.

[71] Dirk Helbing, Ansgar Hennecke, and Martin Treiber. Phase diagram of traffic states in the presence of inhomogeneities. *Physical Review Letters*, 82(21):4360, 1999.

[72] Nikolas Geroliminis and Carlos F Daganzo. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B: Methodological*, 42(9):759–770, 2008.

[73] Andreas Schadschneider. The nagel-schreckenberg model revisited. *The European Physical Journal B-Condensed Matter and Complex Systems*, 10(3):573–582, 1999.

[74] S Lübeck, M Schreckenberg, and KD Usadel. Density fluctuations and phase transition in the nagel-schreckenberg traffic flow model. *Physical Review E*, 57(1):1171, 1998.

[75] Peter Wagner, Kai Nagel, and Dietrich E Wolf. Realistic multi-lane traffic rules for cellular automata. *Physica A: Statistical Mechanics and its Applications*, 234(3-4):687–698, 1997.

[76] Chongyuan Tao and Jian Zhang. A cellular automata simulation on multi-lane traffic flow for designing effective rules. In *2015 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration*, pages 209–212. IEEE, 2015.

[77] Ozan K Tonguz, Wantanee Viriyasitavat, and Fan Bai. Modeling urban traffic: a cellular automata approach. *IEEE Communications Magazine*, 47(5):142–150, 2009.

[78] Ruili Wang and HJ Ruskin. Modelling traffic flow at multi-lane urban roundabouts. *International Journal of Modern Physics C*, 17(05):693–710, 2006.

[79] Simon C Benjamin, Neil F Johnson, and PM Hui. Cellular automata models of traffic flow along a highway containing a junction. *Journal of Physics A: Mathematical and General*, 29(12):3119, 1996.

[80] Chen Rui-Xiong, Bai Ke-Zhao, and Liu Mu-Ren. The ca model for traffic-flow at the grade roundabout crossing. *Chinese Physics*, 15(7):1471, 2006.

[81] Rui Jiang, Qing-Song Wu, and Bing-Hong Wang. Cellular automata model simulating traffic interactions between on-ramp and main road. *Physical Review E*, 66(3):036104, 2002.

[82] BP Hughes, S Newstead, Anna Anund, CC Shu, and Torbjorn Falkmer. A review of models relevant to road safety. *Accident Analysis & Prevention*, 74:250–270, 2015.

[83] Sheila G Klauer, Thomas A Dingus, Vicki L Neale, Jeremy D Sudweeks, David J Ramsey, et al. The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data. Technical report, United States. Department of Transportation. National Highway Traffic Safety . . . , 2006.

[84] Chengcheng Xu, Wei Wang, and Pan Liu. Identifying crash-prone traffic conditions under different weather on freeways. *Journal of safety research*, 46:135–144, 2013.

[85] Reza Tolouei, Mike Maher, and Helena Titheridge. Vehicle mass and injury risk in two-car crashes: a novel methodology. *Accident Analysis & Prevention*, 50:155–166, 2013.

[86] Michael R Crum, Paula C Morrow, Patricia Olsgard, and Philip J Roke. Truck driving environments and their influence on driver fatigue and crash rates. *Transportation research record*, 1779(1):125–133, 2001.

[87] Michael R Crum and Paula C Morrow. The influence of carrier scheduling practices on truck driver fatigue. *Transportation Journal*, pages 20–41, 2002.

[88] Sergio Garbarino, Paolo Durando, Ottavia Guglielmi, Guglielmo Dini, Francesca Bersi, Stefania Fornarino, Alessandra Toletone, Carlo Chiorri, and Nicola Magnavita. Sleep apnea, sleep debt and daytime sleepiness are independently associated with road accidents. a cross-sectional study on truck drivers. *PloS one*, 11(11):e0166262, 2016.

[89] Daniel Mollicone, Kevin Kan, Chris Mott, Rachel Bartels, Steve Bruneau, Matthew van Wollen, Amy R Sparrow, and Hans PA Van Dongen. Predicting performance and safety based on driver fatigue. *Accident Analysis & Prevention*, 126:142–145, 2019.

[90] Zachary E Bowden and Cliff T Ragsdale. The truck driver scheduling problem with fatigue monitoring. *Decision Support Systems*, 110:20–31, 2018.

[91] Kristie Young, Michael Regan, and Mike Hammer. Driver distraction: A review of the literature. *Distracted driving*, 2007:379–405, 2007.

[92] Fernando A Wilson and Jim P Stimpson. Trends in fatalities from distracted driving in the united states, 1999 to 2008. *American journal of public health*, 100(11):2213–2219, 2010.

[93] Rebecca L Olson, Richard J Hanowski, Jeffrey S Hickman, Joseph Bocanegra, et al. Driver distraction in commercial vehicle operations. Technical report, United States. Department of Transportation. Federal Motor Carrier Safety . . . , 2009.

[94] Sheila G Klauer, Feng Guo, Bruce G Simons-Morton, Marie Claude Ouimet, Suzanne E Lee, and Thomas A Dingus. Distracted driving and risk of road crashes among novice and experienced drivers. *New England journal of medicine*, 370(1):54–59, 2014.

[95] Athanasios Theofilatos and George Yannis. A review of the effect of traffic and weather characteristics on road safety. *Accident Analysis & Prevention*, 72:244–256, 2014.

[96] Saman Roshandel, Zuduo Zheng, and Simon Washington. Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis. *Accident Analysis & Prevention*, 79:198–211, 2015.

[97] Chengcheng Xu, Andrew P Tarko, Wei Wang, and Pan Liu. Predicting crash likelihood and severity on freeways with real-time loop detector data. *Accident Analysis & Prevention*, 57:30–39, 2013.

[98] Mohamed M Ahmed, Mohamed Abdel-Aty, and Rongjie Yu. Assessment of interaction of crash occurrence, mountainous freeway geometry, real-time weather, and traffic data. *Transportation research record*, 2280(1):51–59, 2012.

[99] Rongjie Yu, Mohamed Abdel-Aty, and Mohamed Ahmed. Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. *Accident Analysis & Prevention*, 50:371–376, 2013.

[100] Ling Wang, Mohamed Abdel-Aty, Jaeyoung Lee, and Qi Shi. Analysis of real-time crash risk for expressway ramps using traffic, geometric, trip generation, and socio-demographic predictors. *Accident Analysis & Prevention*, 122:378–384, 2019.

[101] Qi Shi and Mohamed Abdel-Aty. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. *Transportation Research Part C: Emerging Technologies*, 58:380–394, 2015.

[102] Anurag Pande and Mohamed Abdel-Aty. Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. *Transportation research record*, 1953(1):31–40, 2006.

[103] Ling Wang, Mohamed Abdel-Aty, Qi Shi, and Juneyoung Park. Real-time crash prediction for expressway weaving segments. *Transportation Research Part C: Emerging Technologies*, 61:1–10, 2015.

[104] Ling Wang, Mohamed Abdel-Aty, and Jaeyoung Lee. Safety analytics for integrating crash frequency and real-time risk modeling for expressways. *Accident Analysis & Prevention*, 104:58–64, 2017.

[105] Dominique Lord and Fred Mannering. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation research part A: policy and practice*, 44(5):291–305, 2010.

[106] Anurag Pande, Abhishek Das, Mohamed Abdel-Aty, and Hany Hassan. Estimation of real-time crash risk: are all freeways created equal? *Transportation research record*, 2237(1):60–66, 2011.

[107] Athanasios Theofilatos, George Yannis, Pantelis Kopelias, and Fanis Papadimitriou. Impact of real-time traffic characteristics on crash occurrence: Preliminary results of the case of rare events. *Accident Analysis & Prevention*, 130:151–159, 2019.

[108] Lei Lin, Qian Wang, and Adel W Sadek. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. *Transportation Research Part C: Emerging Technologies*, 55:444–459, 2015.

[109] Jie Sun and Jian Sun. A dynamic bayesian network model for real-time crash prediction using traffic speed conditions data. *Transportation Research Part C: Emerging Technologies*, 54:176–186, 2015.

[110] Fred L Mannering and Chandra R Bhat. Analytic methods in accident research: Methodological frontier and future directions. *Analytic methods in accident research*, 1:1–22, 2014.

[111] Azad Abdulhafedh et al. Road crash prediction models: different statistical modeling approaches. *Journal of transportation technologies*, 7(02):190, 2017.

[112] Jiří Ambros, Chris Jurewicz, Shane Turner, and Mariusz Kieć. An international review of challenges and opportunities in development and use of crash prediction models. *European transport research review*, 10:1–10, 2018.

[113] George Yannis, Anastasios Dragomanovits, Alexandra Laiou, Thomas Richter, Stephan Ruhl, Francesca La Torre, Lorenzo Domenichini, Daniel Graham, Niovi Karathodorou, and Haojie Li. Use of accident prediction models in road safety management–an international inquiry. *Transportation research procedia*, 14:4257–4266, 2016.

[114] Emily Badger and Alicia Parlapiano. The exceptionally american problem of rising roadway deaths, Nov 2022.

[115] Michael James Lighthill and Gerald Beresford Whitham. On kinematic waves ii. a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 229(1178):317–345, 1955.

[116] H Michael Zhang. A non-equilibrium traffic model devoid of gas-like behavior. *Transportation Research Part B: Methodological*, 36(3):275–290, 2002.

[117] Tobias Jeske. Floating car data from smartphones: What google and waze know about you and how hackers can control traffic. *Proc. of the BlackHat Europe*, pages 1–12, 2013.

[118] Femke van Wageningen-Kessels, Hans van Lint, Kees Vuik, and Serge Hoogendoorn. Genealogy of traffic flow models. *EURO J Transp Logist*, 4:445–473, 12 2015.

[119] Fred L Hall, Brian L Allen, and Margot A Gunter. Empirical analysis of freeway flow-density relationships. *Transportation Research Part A: General*, 20(3):197–210, 1986.

[120] Nino Boccara and Henryk Fuks. Critical behaviour of a cellular automaton highway traffic model. *Journal of Physics A: Mathematical and General*, 33(17):3407, 2000.

[121] Shu-ping Chen and Ding-wei Huang. Noise properties in the nagel-schreckenberg traffic model. *Physical Review E*, 63(3):036110, 2001.

[122] Márton Sasvári and János Kertész. Cellular automata models of single-lane traffic. *Physical Review E*, 56(4):4104, 1997.

[123] B Eisenblätter, L Santen, A Schadschneider, and M Schreckenberg. Jamming transition in a cellular automaton model for traffic flow. *Physical Review E*, 57(2):1309, 1998.

[124] Salvatore Torquato, JD Beasley, and YC Chiew. Two-point cluster function for continuum percolation. *The Journal of Chemical Physics*, 88(10):6540–6547, 1988.

[125] Yang Jiao, FH Stillinger, and SALVATORE Torquato. A superior descriptor of random textures and its predictive capacity. *Proceedings of the National Academy of Sciences*, 106(42):17634–17639, 2009.

[126] P Smith and S Torquato. Computer simulation results for the two-point probability function of composite media. *Journal of Computational Physics*, 76(1):176–191, 1988.

[127] S Torquato. Random heterogeneous media: Microstructure and improved bounds on effective properties. *Applied Mechanics Reviews*, 44(2):37, 1991.

[128] Hadrien Laubie, Siavash Monfared, Farhang Radjai, Roland Pellenq, and Franz-Josef Ulm. Disorder-induced stiffness degradation of highly disordered porous materials. *Journal of the Mechanics and Physics of Solids*, 106:207–228, 2017.

[129] M Gerwinski and J Krug. Analytic approach to the critical density in cellular automata for traffic flow. *Physical Review E*, 60(1):188, 1999.

[130] Daniel T Gillespie. Exact numerical simulation of the ornstein-uhlenbeck process and its integral. *Physical review E*, 54(2):2084, 1996.

[131] F Castro, AD Sanchez, and HS Wio. Reentrance phenomena in noise induced transitions. *Physical review letters*, 75(9):1691, 1995.

[132] Jerzy Łuczka, Peter Hänggi, and Adam Gadomski. Non-markovian process driven by quadratic noise: Kramers-moyal expansion and fokker-planck modeling. *Physical Review E*, 51(4):2933, 1995.

[133] Shri Singh. Phase transitions in liquid crystals. *Physics Reports*, 324(2-4):107–269, 2000.

[134] CF Daganzo. The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transp Res Part B Methodol*, 28:269–287, 4 1994.

[135] Calvin C Moore. Ergodic theorem, ergodic theory, and statistical mechanics. *Proceedings of the National Academy of Sciences*, 112(7):1907–1911, 2015.

[136] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[137] *Traffic Volumes | Boston Region MPO*, 2021. https://www.ctps.org/subjects/traffic-volumes/ [2021-04-27].

[138] Harold Brodsky and A.Shalom Hakkert. Risk of a road accident in rainy weather. *Accident Analysis & Prevention*, 20(3):161–176, 1988.

[139] Yajie Zou, Yue Zhang, and Kai Cheng. Exploring the impact of climate and extreme weather on fatal traffic accidents. *Sustainability*, 13(1), 2021.

[140] Shaw-Pin Miaou and Harry Lum. Modeling vehicle accidents and highway geometric design relationships. *Accident Analysis & Prevention*, 25(6):689–709, 1993.

[141] S. AlKheder, H. Al Gharabally, S. Al Mutairi, and R. Al Mansour. An impact study of highway design on casualty and non-casualty traffic accidents. *Injury*, 53(2):463–474, 2022.

[142] Nelson B. Powell, Kenneth B. Schechtman, Robert W. Riley, Christian Guilleminault, Rayleigh Ping-Ying Chiang, and Edward M. Weaver. Sleepy Driver Near-Misses May Predict Accident Risks. *Sleep*, 30(3):331–342, 03 2007.

[143] Juan Li, Qinglian He, Hang Zhou, Yunlin Guan, and Wei Dai. Modeling driver behavior near intersections in hidden markov model. *International Journal of Environmental Research and Public Health*, 13(12), 2016.

[144] Joseph H. Saleh, Elizabeth A. Saltmarsh, Francesca M. Favarò, and Loïc Brevault. Accident precursors, near misses, and warning signs: Critical review and formal definitions within the framework of discrete event systems. *Reliability Engineering & System Safety*, 114:148–154, 2013.

[145] Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*. John Wiley & Sons, 2009.

[146] Daniel W Stroock. *An introduction to Markov processes*, volume 230. Springer Science & Business Media, 2013.

[147] Andreas Schadschneider and Michael Schreckenberg. Cellular automation models and traffic flow. *Journal of Physics A: Mathematical and General*, 26(15):L679, 1993.

[148] Henrik M Bette, Lars Habel, Thorsten Emig, and Michael Schreckenberg. Mechanisms of jamming in the nagel-schreckenberg model for traffic flow. *Physical Review E*, 95(1):012311, 2017.

[149] Carbin.AI. Carbin educational app, 2021.

[150] MassDOT. Primary and secondary roads state-based shapefile, massachusetts, 2022.

[151] Hadrien Laubie, Farhang Radjai, Roland Pellenq, and Franz-Josef Ulm. Stress transmission and failure in disordered porous media. *Physical review letters*, 119(7):075501, 2017.

[152] Feng Qi, Zhonghuai Hou, and Houwen Xin. Ordering chaos by random shortcuts. *Physical review letters*, 91(6):064102, 2003.

[153] Yu-Zhong Chen, Zi-Gang Huang, and Ying-Cheng Lai. Controlling extreme events on complex networks. *Scientific reports*, 4(1):1–10, 2014.

[154] Vimal Kishore, MS Santhanam, and RE Amritkar. Extreme events on complex networks. *Physical review letters*, 106(18):188701, 2011.

[155] Alessio Emanuele Biondo, Alessandro Pluchino, and Andrea Rapisarda. Modeling financial markets by self-organized criticality. *Physical Review E*, 92(4):042814, 2015.

[156] Michael O Ball, Charles J Colbourn, and J Scott Provan. Network reliability. *Handbooks in operations research and management science*, 7:673–762, 1995.

[157] United States Census Bureau. Tiger/line shapefile, 2019, series information file for the primary and secondary roads state-based shapefile, 2019. Accessed: 2023-06-15.

[158] Federal Highway Administration. Highway performance monitoring system - shapefiles, 2023.

[159] Federico Battiston, Giulia Cencetti, Iacopo Iacopini, Vito Latora, Maxime Lucas, Alice Patania, Jean-Gabriel Young, and Giovanni Petri. Networks beyond pairwise interactions: structure and dynamics. *Physics Reports*, 874:1–92, 2020.

[160] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *nature*, 406(6794):378–382, 2000.

[161] Meshkat Botshekan and Franz-Josef Ulm. Spatial and temporal memory effects in the nagel-schreckenberg model for crowdsourced traffic property determination. *Physical Review E*, 104(4):044102, 2021.

[162] Han Kheng Teoh and Ee Hou Yong. Renormalization-group study of the nagel-schreckenberg model. *Physical Review E*, 97(3):032314, 2018.

[163] André Maurício Conceição de Souza and LCQ Vilar. Traffic-flow cellular automaton: Order parameter and its conjugated field. *Physical Review E*, 80(2):021105, 2009.

[164] Steven Ashley. Core concept: Ergodic theory plays a key role in multiple fields. *Proceedings of the National Academy of Sciences*, 112(7):1914–1914, 2015.