Bounding the Inefficiency of Risk-Averse Selfish Routing with Nonadditive CVaR

Zhaoqi Zang and David Z.W. Wang*
School of Civil and Environmental Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore
zhaoqi.zang@ntu.edu.sg, wangzhiwei@ntu.edu.sg

Abstract

Modern AI-driven systems rely on large populations of autonomous agents that make decentralized routing decisions under uncertainty, where rare but severe tail latency can critically degrade quality of experience, safety, and reliability. To model agents' aversion to such tail latency, we study nonatomic selfish routing games in which agents minimize the Conditional Value-at-Risk (CVaR) of path latency. CVaR explicitly captures both the likelihood and severity of tail latency, but its inherent nonadditivity across network edges poses a fundamental challenge. We address it by identifying a worst-case dependence structure—tail risk concentration—under which tail latency across network edges is synchronized. We show that CVaR penalizes this dependence structure and becomes additive under worst-case tail dependence, which enables tight inefficiency analysis. To quantify the resulting inefficiency induced by risk-aversion, we adopt the price of risk aversion (PRA), defined as the worst-case ratio between the total system cost at a risk-averse equilibrium and that at a risk-neutral equilibrium. We show that, for arbitrary latency functions and general network topologies, the PRA under CVaR admits a tight upper bound that grows linearly with both the network size and the maximum edge-level upper-tail cost. We further prove that this bound is tight by constructing a family of Braess-type networks that achieve a matching lower bound. These results provide the first tight worst-case inefficiency bound for CVaR-based selfish routing and offer insights into how tail-risk-averse decision making by autonomous agents amplifies congestion externalities in large-scale multi-agent systems.

1 Introduction

Modern AI-driven systems increasingly rely on large populations of autonomous agents that make decentralized decisions under uncertainty, including packet routing in communication networks, traffic assignment in intelligent transportation systems, and task allocation in distributed robotic and cloud platforms. In such systems, agents are not only sensitive to average performance but are particularly averse to rare yet severe delays that can critically degrade quality of experience (QoE), safety, or reliability. For example, in communication networks, variability in packet delays—commonly referred to as jitter—disrupts interactive audio, video streaming, and real-time services, leading agents to abandon sessions or reduce quality [Stockhammer, 2011; Jiang *et al.*, 2012; Elbamby *et al.*, 2019]. Similarly, in traffic networks, time variability undermines confidence in arrival time predictions, increasing the risk of arriving late or early, both of which incur substantial costs [Ordónez and Stier-Moses, 2010; Zang *et al.*, 2018a; Zang *et al.*, 2018b; Jian *et al.*, 2024; Zang *et al.*, 2026a].

These rare but severe delays, often referred to as *tail latency*, have been identified as a dominant performance bottleneck in large-scale systems and a key driver of agent behavior. To capture sensitivity to tail latency, a growing body of work in AI, operations research, and economics has adopted risk-averse decision criteria. Among these, Conditional Value-at-Risk (CVaR) has emerged as a principled and widely used measure for modeling tail-risk aversion, with applications in robust optimization, safe reinforcement learning, and risk-aware planning. Unlike mean–variance criteria, CVaR explicitly captures the expected severity of tail events [Rockafellar and Uryasev, 2002], making it well suited for agents seeking to hedge against tail latency rather than symmetric variability.

When risk-averse agents interact in a shared network, their individual attempts to avoid tail latency generate congestion externalities and significantly degrade system efficiency. A central question in multi-agent systems and algorithmic game theory is how to quantify and bound this inefficiency. One such measure is the *price of risk aversion* (PRA) [Lianeas *et al.*, 2019], defined as the worst-case ratio between the total system cost at a risk-averse equilibrium and that at a risk-neutral equilibrium.

Prior work on PRA has largely focused on additive mean–variance objectives, which enable analysis via classical techniques such as alternating paths but fail to capture tail latency and rely on additivity assumptions that do not hold for CVaR, which is inherently nonadditive. In this paper, we study nonatomic selfish routing games in which agents minimize the CVaR of path latency. CVaR admits a clear behavioral

*Corresponding author: David Z.W. Wang.

interpretation for routing under tail risk:

> *How much time should an agent budget for common risks, and how severe should the agent expect delays be under tail risk?*

Because delays in the networks depend endogenously on aggregate traffic through congestion effects, individual routing decisions cannot be analyzed in isolation. Instead, agents' risk-averse shortest path problems are embedded in a noncooperative game in which each agent seeks to minimize its own CVaR-based latency, namely selfish routing behavior. The resulting equilibrium concept extends the classical Wardrop equilibrium [Wardrop, 1952; Beckmann *et al.*, 1956] to stochastic networks with risk-averse agents. Our key insight is that worst-case inefficiency under CVaR is governed by a dependence structure we term *tail risk concentration*, under which extreme delays across network components are synchronized. We show that although CVaR is nonadditive in general, it becomes additive under this worst-case dependence, enabling tight inefficiency analysis.

The main contributions are summarized as follows.

- **Tail-risk-averse routing model.** We study nonatomic selfish routing games in which agents minimize the Conditional Value-at-Risk (CVaR) of path latency, capturing aversion to rare but severe delays that are not adequately modeled by mean–variance or standard deviation rule.
- **Tail risk concentration and CVaR.** We formalize *tail risk concentration* as a worst-case dependence structure under which extreme delays across network components are synchronized. We prove that CVaR penalizes this structure and that tail risk concentration maximizes path-level CVaR aggregation. Leveraging this insight, we show that although CVaR is nonadditive in general, it becomes additive under tail risk concentration. This result enables a surrogate edge-based formulation that coincides with true path-level CVaR, making tight inefficiency analysis feasible.
- **Tight bounds on the price of risk aversion.** We show that for arbitrary latency functions, $\mathrm{PRA} \le 1 + \mathrm{SCVaR}_e^{\max} \cdot \lceil (n-1)/2 \rceil$, implying that the PRA grows linearly with both the maximum edge-level upper-tail cost and the network size. We further establish the tightness of this bound by constructing a family of Braess-type networks with an asymptotically lower bound.
- **Implications for multi-agent systems.** Our results provide worst-case performance guarantees for decentralized, risk-averse routing, clarifying how tail-risk aversion of agents amplifies congestion externalities.

2 Related Work

Increasing efforts have been devoted to selfish routing under risk with risk-averse agents. For example, [Ordóñez and Stier-Moses, 2010] used a constant safety margin to represent agents' risk aversion and analyzed routing games under stochastic delays, establishing connections between risk-averse equilibria and percentile-based equilibria. Focusing on models closer to the present setting, [Nikolova and Stier-Moses, 2014] established the existence of a mean-risk traffic assignment when travel time variability is represented by standard deviations. [Slumbers *et al.*, 2023] developed a game-theoretic framework in which agents trade off expected rewards against variability while accounting for others' actions. Please refer to [Cominetti, 2015; Zang *et al.*, 2022] for comprehensive reviews of risk-averse selfish routing models.

However, the network efficiency of risk-averse selfish routing has received limited attention, despite extensive work on the price of anarchy (PoA), defined as the worst-case ratio between the cost of a Nash equilibrium and that of the socially optimal solution [Roughgarden and Tardos, 2002; Roughgarden, 2003; Koutsoupias and Papadimitriou, 2009]. The PoA has been studied under increasingly general latency functions and network structures [Correa *et al.*, 2004; Roughgarden, 2006; Correa *et al.*, 2008]. [Nikolova and Stier-Moses, 2014; Zhuang *et al.*, 2023] extended this framework to routing games under stochastic delays and risk-aversion of agents. Related but distinct from the PoA, the price of risk captures how perturbations or random player ordering affect equilibrium results [Balcan *et al.*, 2013].

The studies most related to this paper are [Lianeas *et al.*, 2016; Lianeas *et al.*, 2019; Cole *et al.*, 2018], which adopt a mean–variance framework to model risk-averse selfish routing and derive inefficiency bounds using the price of risk aversion (PRA). However, as noted by [Nikolova and Stier-Moses, 2014], combining mean latency with variance or standard deviation is difficult to interpret since they have different units, complicating economic interpretation. Moreover, variance-based measures capture only symmetric fluctuations and therefore fail to consider tail latency, whereas the PRA is inherently concerned with worst-case inefficiency.

3 CVaR-Based Selfish Routing

In this section, we introduce tail risk and its concentration, show that CVaR penalizes tail risk concentration, and formulate a CVaR-based selfish routing model.

We consider a network $G = (V, E)$ with $|V| = n$, total demand d, and W source–sink pairs. For expositional convenience, we set the demand to $d = 1$ in Section 5.2 and set it to be the general case in Section 5.3. Let $\mathcal{R}$ denote the set of all feasible paths. Users' routing decisions are represented by a nonnegative path-flow vector $\mathbf{q} = (q_r)_{r\in\mathcal{R}}$ satisfying $\sum_{r\in\mathcal{R}} q_r = \mathbf{d}$. The induced flow on an edge $e \in E$ is $v_e = \sum_{r\ni e} q_r$. The feasible flow set is $\Omega = \{\mathbf{q} \ge 0 \mid \Lambda\mathbf{q} = \mathbf{d},\ \Delta\mathbf{q} = \mathbf{v}\}$, where $\Delta = [\delta_{er}]$ is the edge–path incidence matrix with $\delta_{er} = 1$ if path r uses edge e and 0 otherwise, and $\Lambda = [\lambda_{rw}]$ is the source–sink–path incidence matrix with $\lambda_{rw} = 1$ if path r connects source–sink pair $w \in W$ and 0 otherwise. Throughout this section, we use the term *latency variability* to encompass travel time, queuing time, and delay variability across various network settings.

3.1 Tail Risk and Tail Risk Concentration

Latency variability in networks does not arise arbitrarily. Rather, it originates from identifiable sources of risks acting on both the supply side and the demand side. This observation applies broadly to road networks [Zang *et al.*, 2022], communication networks [Haque *et al.*, 2015], cloud computing systems [Li *et al.*, 2014], and other congestion-prone service networks [Dean and Barroso, 2013]. Figure 1 conceptually illustrates how these heterogeneous sources of risks jointly induce latency variability. Let Ξ denote the set of all sources of risks contributing to latency variability.

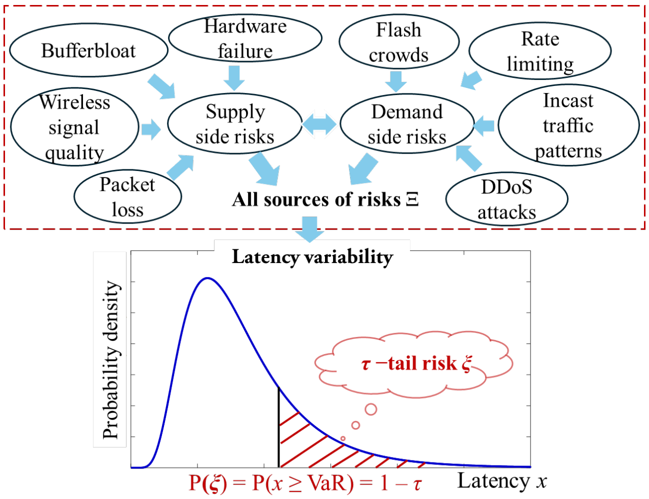

Figure 1: Selected sources of risks inducing latency variability and the τ-tail risk where VaR is value at risk.

Among these risk sources, we formally distinguish τ*-tail risk* in Definition 1, motivated by the notion of tail events developed in the finance and economic literature [Wang and Zitikis, 2021]. While multiple risks may occur with identical probabilities, not all of them are equally consequential. The way we define the τ-tail risk makes it to be the risk leading to the worst upper-tail realizations of latency with probability $1-\tau$, as illustrated in Figure 1. This interpretation is theoretically supported by the following Proposition 1. Potential sources that can generate τ-tail risk include incidents, correlated demand surges, cascading failures, synchronized congestion, and resource contention. We will further elaborate this in Remark 1.

Definition 1 (τ-tail risk)**.** *A* τ*-tail risk of latency* x *is a risk* $\xi \in \Xi$ *with* $\mathbb{P}(\xi) = 1-\tau$ *such that* $x(\omega) \geq x(\omega')$ *holds for all* $\omega \in \xi$ *and* $\omega' \in \xi^c$*, where* ξ^c *denotes the complement of* ξ*.*

A path typically consists of multiple edges, so edge-level tail risk will aggregate to the path level. If a path is properly diversified, agents may benefit from diversification such that path-level tail risk is lower than the sum of edge-level risks. However, consider a path composed of several edges that share a common tail risk ξ. When tail risk ξ occurs—such as congestion, hardware faults, or wireless interference—each edge simultaneously realizes its worst latency. This represents the most problematic form of tail risk concentration in latency variability, as it leads to extreme delays. Such synchronized tail risk can cause severe congestion spillovers and impose substantial losses on both agents and the network as a whole. Definition 2 formalizes this worst-case scenario, in which edge-level tail risks are fully concentrated at the path level, following [Wang and Zitikis, 2021; Zang *et al.*, 2026b].

Definition 2 (Tail risk concentration)**.** *For a path latency* x*, if its all edge latency shares a common* τ*-tail risk, we say it shows tail risk concentration.*

Proposition 1 shows that among all risks with the same probability $1-\tau$, τ-tail risk of Definition 1 leads to the largest possible tail latency, while Lemma 1 characterizes the relationship between path-level tail risk and edge-level tail risk, thereby providing a mechanism through which edge tail risks aggregate to the path level.

Proposition 1 (τ-tail risk maximizes expected tail latency)**.** *Let* x *be a random variable and let* $U_x \sim U[0,1]$ *be such that* $x = F_x^{-1}(U_x)$ *almost surely. For a* τ*-tail risk* ξ_A*, it can be defined as* $\xi_A := \{U_x \geq \tau\}$ *with its indicator* $\mathbf{1}_{\xi_A}$*. Then for every risk* $\xi_B \in \Xi$ *with* $\mathbb{P}(\xi_B) = 1-\tau$*, we have*

$$\mathbb{E}[x\,\mathbf{1}_{\xi_A}] \geq \mathbb{E}[x\,\mathbf{1}_{\xi_B}] \tag{1}$$

Proof. U_x works because any random variable can be generated from a Uniform(0,1) random variable by applying its quantile function. By definition, $\mathbb{E}[\mathbf{1}_{\xi_A} - \mathbf{1}_{\xi_B}] = 0$ and thus we have $\mathbb{E}[x(\mathbf{1}_{\xi_A} - \mathbf{1}_{\xi_B})] = \mathbb{E}[(x-Q)(\mathbf{1}_{\xi_A} - \mathbf{1}_{\xi_B})]$ for all constant $Q \in \mathbb{R}$. Let $Q = F_x^{-1}(\tau)$ which is the τ-quantile of x and recall that we have $x = F_x^{-1}(U_x)$. If $F_x^{-1}(U_x) > Q$, then $U_x > \tau$, $\mathbf{1}_{\xi_A} = 1$ and so $\mathbb{E}[(x-Q)(\mathbf{1}_{\xi_A} - \mathbf{1}_{\xi_B})] \geq 0$; if $F_x^{-1}(U_x) < Q$, then $U_x < \tau$, $\mathbf{1}_{\xi_A} = 0$ and so $\mathbb{E}[(x-Q)(\mathbf{1}_{\xi_A} - \mathbf{1}_{\xi_B})] \geq 0$; if $F_x^{-1}(U_x) = Q$, then $x - Q = 0$ and so $\mathbb{E}[(x-Q)(\mathbf{1}_{\xi_A} - \mathbf{1}_{\xi_B})] = 0$. In summary, we have $\mathbb{E}[(x-Q)(\mathbf{1}_{\xi_A} - \mathbf{1}_{\xi_B})] = 0$. So, $\mathbb{E}[x\,\mathbf{1}_{\xi_A}] \geq \mathbb{E}[x\,\mathbf{1}_{\xi_B}]$. □

Lemma 1 (Common tail risk)**.** *The risk* ξ *is a* τ*-tail risk of path latency if* ξ *is a* τ*-tail risk of its all edge latency.*

Proof. Consider all edges of a path share a common τ-tail risk source ξ_τ. First, assume that path r consists of only two edges, namely edge e_a and edge e_b. We first show that the tail risk ξ_τ is also a tail risk for path r. By definition, we have the following inequality for all scenarios $\omega \in \xi_\tau$ and $\omega' \in \xi_\tau^c$:

$$x_a(\omega) + x_b(\omega) \geq x_a(\omega') + x_b(\omega').$$

The above shows that ξ_τ is an τ-tail risk source of $x_a + x_b$. Next, consider a path r consisting of n edges, denoted by $e_1, \ldots, e_n$, which share a common tail risk. By induction, we conclude that ξ_τ is an τ-tail risk of $x_1 + \cdots + x_n$. Namely, ξ_τ is an τ-tail risk of the path latency. □

3.2 CVaR Minimization Penalizes Tail Risk Concentration

In the context of network latency, the Conditional Value at Risk (CVaR) at the confidence level τ can be defined as the conditional expectation of the latency x exceeding the corresponding Value at Risk (VaR). Namely,

$$\mathit{CVaR}_\tau(x) = \mathbb{E}[x \mid x \geq \mathit{VaR}_\tau(x)] \tag{2}$$

where VaR is defined by a chance constraint at confidence level τ as follows [McNeil *et al.*, 2015]:

$$VaR_\tau(x) = \min\{VaR \mid \Pr(x \le VaR) \ge \tau\}, \tag{3}$$

Then, based on the above Equations (2) and (3) and previous works [Pflug, 2000; Rockafellar *et al.*, 2000; Rockafellar and Uryasev, 2002], the defined CVaR can be equivalently written as the following minimization problem:

$$CVaR_\tau(x) = \min_{c\in\mathbb{R}} \left\{ c + \frac{1}{1-\tau}\mathbb{E}[x-c]_+ \right\}, \tag{4}$$

where $[x]_+ = \max\{0, x\}$ and c is a constant.

Equation (2) shows that CVaR pays particular attention to the tail latency which is caused by τ-tail risk. Specifically, we can rewrite CVaR as follows.

$$CVaR_\tau(x) = VaR_\tau(x) + \frac{1}{1-\tau}\int_\tau^1 \left(F^{-1}(w) - F^{-1}(\tau)\right) dw, \tag{5}$$

where F is the cumulative distribution function (CDF) of x.

The first part, namely $VaR_\tau(x)$, in CVaR of Eq. (5) considers the common latency that travellers could frequently encounter at the confidence level τ, while the second part, $\frac{1}{1-\tau}\int_\tau^1 \left(F^{-1}(w) - F^{-1}(\tau)\right) dw$, is expected value of the tail latency with respect to $VaR_\tau(x)$ where tail latency is actually caused by τ-tail risk. Consequently, for routing agents in networks, CVaR simultaneously addresses both questions: *How much time should an agent budget for common risk, and how severe should the agent expect delays be under tail risk?*

Theorem 2 shows that CVaR minimization penalizes tail risk concentration: for a given path, tail risk concentration across edges maximizes CVaR aggregation, while the subadditivity of CVaR ensures that the path-level CVaR is never greater than the sum of edge-based CVaRs.

Theorem 2. *CVaR minimization penalizes tail risk concentration at path.*

Proof. Consider a path r has n edges with $x_i, i = 1, ..., n$ as their latency and $x_1, ..., x_n$ share a common τ-tail risk. Let ξ_τ be this common τ-tail risk. According to Lemma 1, then ξ_τ is also a τ-tail risk of this path latency x. Based on Proposition 1 and [Embrechts and Wang, 2015], we can derive the following dual representation of CVaR in Equation (4):

$$CVaR_\tau(x) = \max\{\mathbb{E}[x|\xi_B] : \xi_B \in \Xi,\ \mathbb{P}(\xi_B) = 1-\tau\} \tag{6}$$

According to Proposition 1, we have $\mathbb{E}[x\,\mathbf{1}_{\xi_\tau}] \ge \mathbb{E}[x\,\mathbf{1}_{\xi_B}]$ for tail risk ξ_τ with probability $1-\tau$. Consequently, by equation (6), we have $CVaR_\tau(x) = \mathbb{E}[x|\xi_\tau]$. In other words, ξ_τ is the risk that makes Equation (6) hold without the maximization operator. Consequently, we have

$$\begin{aligned} CVaR_\tau(x) &= CVaR_\tau\left(\sum_{i=1}^n x_i\right) \\ &= \mathbb{E}\left[\sum_{i=1}^n x_i \,\middle|\, \xi_\tau\right] = \sum_{i=1}^n \mathbb{E}[x_i \mid \xi_\tau] \\ &= \sum_{i=1}^n CVaR_\tau(x_i) \end{aligned} \tag{7}$$

Recalling the subadditivity of CVaR, we should have the inequality of $CVaR_\tau(x) \le \sum_{i=1}^n CVaR_\tau(x_i)$. However, τ-tail risk concentration of $x_1, ..., x_n$ makes such inequality become equality, thereby maximizing CVaR aggregation. In other words, CVaR minimization penalizes tail risk concentration at path. This complements the proof. □

3.3 Selfish Routing with Nonadditive CVaR

Based on Theorem 2, each agent in a CVaR-based selfish routing model seeks to minimize the CVaR of its experienced path latency. In a network setting, such decentralized CVaR minimization induces a noncooperative routing game. We refer to the resulting equilibrium as a CVaR-based network equilibrium (CNE). Owing to the nonadditivity of CVaR, the CNE generally cannot be characterized through an additive path-cost or convex optimization formulation; instead, it admits a variational inequality (VI) formulation in Proposition 2 based on [Chen and Zhou, 2010].

Proposition 2 (VI formulation of CNE)**.** *A path flow vector* $\mathbf{q}^* \in \Omega$ *is said to be CNE path flow pattern if and only if it solves the following variational inequality problem:*

$$\langle CVaR(\mathbf{q}^*), \mathbf{q} - \mathbf{q}^* \rangle \ge 0, \quad \forall \mathbf{q} \in \Omega. \tag{8}$$

4 Price of Risk Aversion

In this section, we define the price of risk aversion.

4.1 Preliminaries of Routing Equilibria for Agents

We denote a risk-neutral selfish routing instance by (G, d, l), where l specifies the latency functions. This notation represents the class of risk-neutral routing games in which agents minimize expected path latency. Under risk, we denote a corresponding risk-averse selfish routing instance by (G, d, l, Ξ, τ). This notation represents the class of risk-averse routing games in which agents minimize a CVaR-based latency criterion.

For (G, d, l), the path cost is deterministic and can be computed as the expected latency along a path r:

$$l_r(q) = \sum_{e\in r} l_e(v_e). \tag{9}$$

Let $x = \mu + \sigma X$ where X is the standardized latency with CDF $\Phi(X)$; μ and σ represent the mean and standard deviation of x, respectively. We would have $VaR_\tau(x) = \mu + \sigma \cdot \Phi_X^{-1}(\tau)$ according to [Wu and Nie, 2011; Zang *et al.*, 2024]. Consequently, CVaR at Eq. (5) can be rewritten as

$$CVaR_\tau(x) = \mu + \sigma \cdot \frac{1}{1-\tau}\int_\tau^1 \Phi_X^{-1}(w)\, dw, \tag{10}$$

Accordingly, for (G, d, l, Ξ, τ), CVaR of path r is given by

$$CVaR_\tau(q_r) = \sum_{r\in\mathcal{R}} q_r l_r(q_r) + \sum_{r\in\mathcal{R}} \sigma_r \cdot \frac{1}{1-\tau}\int_\tau^1 \Phi_r^{-1}(w)\, dw. \tag{11}$$

Then, based on [Wardrop, 1952] and [Chen and Zhou, 2010], Definition 3 describes risk-neutral network equilibrium (NE) and CVaR-based network equilibrium (CNE).

Definition 3 (Network equilibrium and CVaR-based network equilibrium)**.** *For risk-neutral users, at NE, all used paths have equal and minimal expected latency; for risk-averse users with CVaR-based criterion, at CNE, all used paths have equal and minimal CVaR of path latency.*

4.2 Defining Price of Risk Aversion

To investigate the network inefficiency caused by risk-averse users, we use the total system cost, denoted by C, as the network performance metric to make a fair comparison between NE and CNE. Specifically, total cost is evaluated by

$$C(q) = \sum_{r\in\mathcal{R}} q_r l_r(q) = \sum_{r\in\mathcal{R}} \sum_{e\in r} v_e l_e(v_e). \tag{12}$$

Let q^x and q^z be CNE flow vector and NE flow vector, respectively. Then, we can define the price of risk aversion (PRA) as the worst-case ratio of $C(q^x)$ of CNE to $C(q^z)$ of NE in Definition 4.

Definition 4 (Price of risk aversion in [Lianeas *et al.*, 2019])**.** *For an instance family $\mathcal{F}$ of a routing game under latency variability, the price of risk aversion (PRA) is given by*

$$\mathrm{PRA}(\mathcal{F}, \Xi, \tau) := \sup_{G,d,l,\Xi} \left\{ \frac{C(q^x)}{C(q^z)} =: (G, d, l, \Xi, \tau) \in \mathcal{F} \right\}$$

Such definition of PRA enables us to isolate the inefficiency arising from selfish behavior, as measured by the price of anarchy (PoA) [Roughgarden and Tardos, 2002; Roughgarden, 2003], from those resulted from risk aversion per se. In other words, the ways of defining PRA makes us focus on the network inefficiency due solely to risk aversion.

5 Bounds of PRA

This section bounds the PRA. The main challenge is due to the nonadditivity of CVaR and we tackle it by proving that CVaR is τ-additive for tail risk concentration in Theorem 4.

5.1 Alternating Path

The derivation of the upper bound of PRA depends on the *alternating path* constructed by [Lianeas *et al.*, 2019].

Consider CNE flows q^x and NE flows q^z for instances (G, d, l, Ξ, τ) and (G, d, l), respectively. Then, for an edge e of G, we have the following terminologies:

- (q^x, q^z)-light if $q^x_e \le q^z_e$;
- $q^z_e > 0$; (q^x, q^z)-heavy if $q^x > q^z_e$;
- (q^x, q^z)-null if $q^x_e = q^z_e = 0$.

Note that we do not need to consider the scenario of (q^x, q^z)-null for an edge as it can be removed from the network G due to no flow on both CNE and NE. Then, define two set of edges to partition of the edge set E as follows:

$$A := \{ e \in E \mid q^x_e \le q^z_e \},\ B := \{ e \in E \mid q^z_e < q^x_e \}. \tag{13}$$

Under the assumption that every edge of the network at least carries positive flow in one of the equilibria q^x and q^z, we can have $q^z_e > 0$ for each $e \in A$ and $q^x_e > 0$ for each $e \in B$. With these terminologies, we can have the definition of alternating path in Definition 5. Alternating paths allow us to exploit network structure to bound the PRA. Besides, there exists a alternating path for any instance (G, d, l, Ξ, τ) due to flow conservation and definitions of light and heavy edges.

Definition 5 ([Lianeas *et al.*, 2019])**.** *An undirected path is called an (q^x, q^z)-alternating path if it consists only of forward light edges and backward heavy edges. For any instance (G, d, l, Ξ, τ), there is a (q^x, q^z)-alternating path.*

5.2 Upper Bound

The following Lemma 3 is a rough bound to say that $C(q^x)$ of CNE is bounded by any path CVaR cost of this network.

Lemma 3 (Rough bound of $C(q^x)$)**.** *Let $r \in \mathcal{R}$ be any path, regardless of whether it carries positive flow at equilibrium. The total system cost at CNE, denoted by $C(q^x)$, admits an upper bound determined by the CVaR cost of path r. Namely,*

$$C(q^x) \ \le\ \left(1 + \mathit{SCVaR}_r\right) l_r(q^x), \tag{14}$$

where $\mathit{SCVaR}_r = \mathrm{CV}_r \cdot \frac{1}{1-\tau} \int_\tau^1 \Phi_r^{-1}(p)\, dp$ and CV denotes the coefficient of variation of the latency of path r.

Proof. At equilibrium, we have that $\mathrm{CVaR}_{r'} \le \mathrm{CVaR}_r$ for all $r' \in \mathcal{R}$ that carry positive flow according to Definition 3. This means that $l_{r'}(q^x) + \frac{1}{1-\tau} \int_\tau^1 \Phi_{r'}^{-1}(w)\, dw \cdot \sigma_{r'}(q^x) \le l_r(q^x) + \frac{1}{1-\tau} \int_\tau^1 \Phi_r^{-1}(w)\, dw \cdot \sigma_r(q^x)$ based on Eq.(10). Thus,

$$\begin{aligned} C\left(q^x\right) &= \sum_{r'\in\mathcal{R}} q^x_{r'} l_{r'}\left(q^x\right) \\ &\le \sum_{r'\in\mathcal{R}} q^x_{r'} \left(l_r\left(q^x\right) + \frac{\sigma_r\left(q^x\right)}{1-\tau} \int_\tau^1 \Phi_r^{-1}\left(w\right) dw \right) \\ &= d\mathrm{CVaR}_r\left(q^x\right) \end{aligned}$$

According to the inequality for l_1-norm and l_2-norm, we have

$$\begin{aligned} \mathrm{CVaR}_r(q^x) &= \sum_{e\in r} l_e(q^x) + \frac{\sqrt{\sum_{e\in r} \sigma_e^2(q^x)}}{1-\tau} \int_\tau^1 \Phi_r^{-1}(w)\, dw \\ &\le \sum_{e\in r} l_e(q^x) + \frac{\sum_{e\in r} \sigma_e(q^x)}{1-\tau} \int_\tau^1 \Phi_r^{-1}(w)\, dw \\ &= \left(1 + \mathrm{SCVaR}_r\right) l_r(q^x). \end{aligned}$$

where the first inequality is by l_2-norm $\le l_1$-norm. ☐

To make further bounds, we have to use the additivity of risk measure, while we know the CVaR is not additive. To this end, Theorem 4 finds that edge-based CVaR naturally serves as a surrogate CVaR as this surrogate does coincide with the true path-level CVaR when tail risk of edges are concentrated at the path. Its proof is based on equation (7) in the proof of Theorem 2. Now, we can formulate CNE as a edge-based mathematical programming based on [Xu *et al.*, 2017; Zang *et al.*, 2026b] in Corollary 4.1.

Theorem 4 (τ-additive CVaR for common tail risk)**.** *The path CVaR is additive if all its edges share a common tail risk.*

Corollary 4.1 (MP of CNE)**.** *The edge-based CNE can be formulated as a mathematical programming (MP):*

$$\min_{\mathbf{v}}\ Z(\mathbf{v}) = \sum_{e\in E} \int_0^{v_e} \mathit{CVaR}_e(w)\, dw,\ s.t.\ \mathbf{v} \in \Omega \tag{15}$$

Although the condition in Theorem 4 may initially appear strong, it is appropriate in the context of the PRA, as by definition (i.e., Definition 4) the PRA's focus is on worst-case ratio rather than typical or average ratio. As discussed and justified around Definition 2, tail risk concentration is the most problematic form of edge-level tail risk aggregated to path level. Therefore, bound the inefficiency using the PRA should consider such problematic form.

Remark 1 (Common tail risks in networks)**.** *In transportation networks, tail risks arise from adverse weather, synchronized signal or control failures, and major incidents in which crashes or lane blockages propagate extreme delays upstream through spillback and shockwaves [Horn and Wang, 2017]. A recent example is the malfunction of multiple Waymo autonomous vehicles that simultaneously blocked streets and disrupted traffic at several intersections in San Francisco [Abhirup and David, 2025]. In communication networks, analogous tail risks include correlated traffic surges or flash crowds that overload multiple links, synchronized packet losses or retransmissions due to buffer overflows, control-plane events such as routing updates or protocol failures, and hardware or software faults at shared infrastructure that generate network-wide delay spikes [Zhao* et al.*, 2023]. In both domains, these risks concentrate extreme delays across the network rather than affecting edges only.*

The additive form of the CVaR is quite important to use alternating path to bound the PRA. That is, with Lemma 3 and Theorem 4, we can have Lemma 5 to bound $C(q^x)$ of CNE and $C(q^z)$ of NE and further derive the final Theorem 6 based on the results in [Lianeas *et al.*, 2016].

Lemma 5 (Tighter bound based on alternating path)**.** *Let π denote an alternating path, then total cost of a CNE satisfies*

$$C(q^x) \leq (1 + \textit{SCVaR}_e^{\max}) \sum_{e \in A \cap \pi} l_e(q_e^x) - \sum_{e \in B \cap \pi} l_e(q_e^x).$$

As for NE, total cost of a NE q^z satisfies

$$C(q^z) \geq \sum_{e \in A \cap \pi} l_e(q_e^z) - \sum_{e \in B \cap \pi} l_e(q_e^z).$$

Proof. The main idea of proving upper bound of $C(q^x)$ is to leverage τ-additivity of CVaR in Theorem 4 and alternating-path decomposition to compare equilibrium paths with carefully chosen alternatives, allowing equilibrium conditions to telescope into a tight upper bound on $C(q^x)$ that depends only on edge costs along the alternating path and $\mathrm{SCVaR}_e^{\max}$. The detailed proof can refer to [Lianeas *et al.*, 2016] where the main difference is that we use CVaR as path cost while they use mean-variance cost. For the lower bound of $C(q^x)$, please refer to Lemma 4.5 of [Nikolova and Stier-Moses, 2015]. □

Remark 2 (Interpretation of SCVaR and $\mathrm{SCVaR}_e^{\max}$)**.** *In Lemma 3, we introduce the scaled standardized CVaR, namely SCVaR, based on Eq.*(10) *as*

$$\textit{SCVaR}_\tau = \mathrm{CV} \cdot \frac{1}{1-\tau} \int_\tau^1 \Phi_X^{-1}(p)\, dp. \tag{16}$$

Recalling that $x = \mu + \sigma X$, we should have $\textit{CVaR}_\tau(x) = \mu + \sigma \textit{SCVaR}_\tau(X)$. Then, under CVaR-based selfish routing, $\textit{SCVaR}_\tau(X)$ represents the expected standardized latency in the worst $(1-\tau)$ fraction of latency and depends only on the shape of the upper tail of the latency distribution. Larger values of $\textit{SCVaR}_\tau(X)$ indicate more severe tail latency and therefore it actually represents the the upper-tail cost added to the mean latency for agents adopting CVaR choice rule.

Consequently, $\textit{SCVaR}_e^{\max} := \max_{e \in E} \textit{SCVaR}_\tau(X_e)$ represents the maximum edge-level upper-tail cost in the network.

Theorem 6. *Consider an arbitrary instance (G, d, l, Ξ, τ), the upper bound of the PRA is $1 + \textit{SCVaR}_e^{\max} \cdot \lceil (n-1)/2 \rceil$.*

Proof. The claim follows from the following chain of inequalities:

$$\begin{aligned} C(q^x) &\leq (1 + \mathrm{SCVaR}_e^{\max}) \sum_{e \in A \cap \pi} l_e(q_e^x) - \sum_{e \in B \cap \pi} l_e(q_e^x) \\ &\leq (1 + \mathrm{SCVaR}_e^{\max}) \sum_{e \in A \cap \pi} l_e(q_e^z) - \sum_{e \in B \cap \pi} l_e(q_e^z) \\ &\leq C(q^z) + \mathrm{SCVaR}_e^{\max} \sum_{e \in A \cap \pi} l_e(q_e^z) \\ &\leq C(q^z) + \mathrm{SCVaR}_e^{\max} \cdot m C(q^z) \\ &= (1 + \mathrm{SCVaR}_e^{\max} \cdot m)\, C(q^z). \end{aligned}$$

where m denote the number of disjoint forward subpaths. The first and third inequalities are because of Lemma 5, while second inequality is based on how we define A and B in Eq.(13). The fourth inequality is derived by the fact that $\sum_{e \in A \cap \pi} l_e(q_e^z) \leq m\, C(q^z)$. It holds because, for every forward subpath $A_k \in \pi$, all its edges have $q_e^z > 0$. Consequently, for an appropriate flow decomposition, there exists for each k a path r_k with $q_{r_k}^z > 0$ that contains the subpath A_k, implying $c_{A_k}(q^z) \leq l_{r_k}(q^z) = C(q^z)$. The equality follows from $d = 1$. In the worst case the path alternates edge by edge, implying that the maximum number of disjoint forward subpaths is $\lceil (n-1)/2 \rceil$. Namely $m \leq \lceil (n-1)/2 \rceil$. □

One may expect that the resulting bound should be governed by two intuitive elements: CVaR-related cost $\mathrm{SCVaR}_e^{\max}$ and network size. Somewhat unexpectedly, this dependence remains linear even under arbitrary latency functions. With Remark 2, $\mathrm{SCVaR}_e^{\max}$ helps us identify the most critical edge of a network affecting the PRA. Section 5.3 further shows that the upper bound in Theorem 6 is tight.

5.3 Tight Lower Bound

To further derive the lower bound, we have the following Lemma 7 based on [Lianeas *et al.*, 2016]. The construction of graphs in the proof of Lemma 7 ensures each path contains at most a single edge with positive variability. Therefore, it works for both additive and nonadditive risk models including CVaR, as it eliminates the need to aggregate risks across multiple edges in which nonadditive measures fail.

The main idea of the proof is that by constructing instances based on the Braess-like network as shown in Figure 2 (modified based on [Lianeas *et al.*, 2016; Lianeas *et al.*, 2019]) where the base case is left network and induction case is right network. Them, we can apply mathematical induction to show that $C(q^x) = (1 + 2^i \mathrm{SCVaR}_e^{\max})\, d_{\mathrm{CNE}}^i$ and

$C(q^z) = dr^i_{\mathrm{NE}}$ hold for each positive integer i and each demand pair $d^i_{\mathrm{CNE}}, d^i_{\mathrm{NE}} \in \mathbb{R}_{\geq 0}$ such that $2^i d^i_{\mathrm{CNE}} > (2^i - 1)d^i_{\mathrm{NE}}$. d_{CNE} and d_{NE} are demands of CNE (i.e., with risk-averse users) and NE (i.e., with risk-neutral users), respectively. The instance for index i is obtained recursively by constructing a Braess-type network using the network constructed for the $(i-1)$th case. For each step of the construction, the key is to select an appropriate latency function such that the properties stated in Lemma 7 are satisfied. Interested readers can refer to Theorem 2 in [Lianeas *et al.*, 2016] for more details.

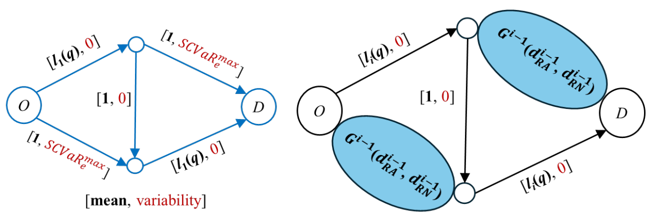

Figure 2: The base case $G^1(d^1_{\mathrm{CNE}}, d^1_{AN})$ and recursive construction.

Lemma 7. *For each integer $i \geq 1$ and for every demand pair, $d^i_{CNE}, d^i_{NE} \in \mathbb{R}_{\geq 0}$ such that $2^i d^i_{CNE} > (2^i - 1)d^i_{NE}$, there exists a graph instance $G^i(d^i_{CNE}, d^i_{NE})$ with 2^{i+1} nodes that satisfies*

$$\frac{C(q^x)}{C(q^z)} = \frac{\left(1 + 2^i SCVaR^{\max}_e\right) d^i_{CNE}}{d^i_{NE}} \tag{17}$$

With Lemma 7, we can show that our established upper bound in Theorem 6 is tight. This observation is refined in Theorem 8, which indicates that the supremum in the definition of PRA is achieved by explicitly constructed Braess-type instances as visually shown by Figures 2 and 3.

Theorem 8. *One can construct Braess-type graphs of arbitrary size such that*

$$\mathrm{PRA} = \frac{C(q^x)}{C(q^z)} \geq 1 + SCVaR^{\max}_e \left\lceil \frac{n-1}{2} \right\rceil. \tag{18}$$

Thus, the upper bound of the PRA in Theorem 6 is tight.

Proof. Fix an arbitrary demand d and invoke Lemma 7 by setting $d^i_{\mathrm{CNE}} = d^i_{\mathrm{NE}} = d$ and choosing $i = \min\left\{j \in \mathbb{N} : n_0 \leq 2^j\right\}$. This yields a constructed instance $G^i(d^i_{\mathrm{CNE}}, d^i_{\mathrm{NE}})$. For this network, the corresponding CNE flow q^x and NE flow q^z satisfy

$$\frac{C(q^x)}{C(q^z)} = \frac{\left(1 + 2^i \mathrm{SCVaR}^{\max}_e\right) d}{d} = 1 + \mathrm{SCVaR}^{\max}_e \frac{n}{2}, \tag{19}$$

since the graph $G^i(d^i_{\mathrm{CNE}}, d^i_{\mathrm{NE}})$ contains 2^{i+1} vertices by construction. As n is therefore a power of two, Eq.(19) is true.

The inductive construction in the proof of Lemma 7 builds the graph $G^i(d^i_{\mathrm{CNE}}, d^i_{\mathrm{NE}})$ by composing two copies of $G^{i-1}(d^{i-1}_{\mathrm{CNE}}, d^{i-1}_{\mathrm{NE}})$ in a Braess-type structure, resulting in a graph with 2^{i+1} vertices. However, the validity of the induction does not depend on exact graph symmetry or on the number of vertices, but rather on an equilibrium-cost invariant: each inductively constructed subgraph must induce a CNE path cost of $1 + 2^{i-1}\mathrm{SCVaR}^{\max}_e$ and a NE path cost of 1. Consequently, for any graph size n satisfying $2^i < n < 2^{i+1}$, one may modify the Braess-based construction by replacing the upper-right edge with a copy of $G^{i-1}(d^{i-1}_{\mathrm{CNE}}, d^{i-1}_{\mathrm{NE}})$ and the lower-left edge with a graph constructed recursively with $n - 2^i$ vertices that satisfies the same equilibrium-cost properties as shown in Figure 3. Since NE depends solely on perceived path costs and not on internal network structure, such asymmetric substitutions preserve the equilibrium behavior of the construction. As a result, the recursive argument and the linear growth of the PRA extend to networks of arbitrary size, including those with an odd number of vertices. □

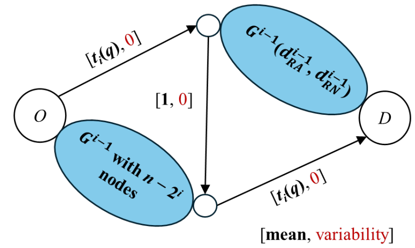

Figure 3: The recursive construction of $G^i(d^i_{\mathrm{CNE}}, d^i_{\mathrm{NE}})$ based on $G^{i-1}(d^{i-1}_{\mathrm{CNE}}, d^{i-1}_{\mathrm{NE}})$ where the bottom left edge and top right edge are replaced by the graph G^{i-1} with $n - 2^i$ nodes and $G^{i-1}(d^{i-1}_{\mathrm{CNE}}, d^{i-1}_{\mathrm{NE}})$.

6 Conclusion

This paper establishes the first tight worst-case bounds on the inefficiency of CVaR-based selfish routing in stochastic networks. By analyzing the price of risk aversion (PRA), we show that even under arbitrary latency functions, the inefficiency induced by tail-risk-averse behavior is linearly bounded by network size and the maximum edge-level tail cost, and this bound can be tight. A key conceptual insight underlying our results is that worst-case inefficiency under CVaR is governed by *tail risk concentration*, a dependence structure in which extreme delays across network components are synchronized. Although CVaR is nonadditive in general, we show that it becomes effectively additive under this worst-case dependence, enabling tight inefficiency analysis despite the failure of classical additive techniques. This clarifies the role of tail dependence in amplifying congestion externalities under risk-averse decision making.

From a multi-agent systems perspective, our results provide worst-case performance for decentralized autonomous agents that optimize tail-risk-averse objectives and thus offer a theoretical bridge between individual risk-averse objectives and collective system performance in large-scale systems.

Ethical Statement

There are no ethical issues.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2022-031).

References

[Abhirup and David, 2025] Roy Abhirup and Shepardson David. California examines incidents of stalled waymo robotaxis after san francisco power outage. *Reuters*, 2025. Accessed January 2026.

[Balcan *et al.*, 2013] Maria-Florina Balcan, Avrim Blum, and Yishay Mansour. The price of uncertainty. *ACM Transactions on Economics and Computation (TEAC)*, 1(3):1–29, 2013.

[Beckmann *et al.*, 1956] Martin J. Beckmann, C. B. McGuire, and C. B. Winsten. *Studies in the Economics of Transportation*. Yale University Press, New Haven, CT, 1956.

[Chen and Zhou, 2010] Anthony Chen and Zhong Zhou. The α-reliable mean-excess traffic equilibrium model with stochastic travel times. *Transportation Research Part B: Methodological*, 44(4):493–513, 2010.

[Cole *et al.*, 2018] Richard Cole, Thanasis Lianeas, and Evdokia Nikolova. When does diversity of agent preferences improve outcomes in selfish routing? In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 173–179. International Joint Conferences on Artificial Intelligence Organization, 2018.

[Cominetti, 2015] Roberto Cominetti. Equilibrium routing under uncertainty. *Mathematical Programming*, 151(1):117–151, 2015.

[Correa *et al.*, 2004] José R Correa, Andreas S Schulz, and Nicolás E Stier-Moses. Selfish routing in capacitated networks. *Mathematics of Operations Research*, 29(4):961–976, 2004.

[Correa *et al.*, 2008] José R Correa, Andreas S Schulz, and Nicolás E Stier-Moses. A geometric approach to the price of anarchy in nonatomic congestion games. *Games and Economic Behavior*, 64(2):457–469, 2008.

[Dean and Barroso, 2013] Jeffrey Dean and Luiz André Barroso. The tail at scale. *Communications of the ACM*, 56(2):74–80, 2013.

[Elbamby *et al.*, 2019] Mohammed S Elbamby, Cristina Perfecto, Chen-Feng Liu, Jihong Park, Sumudu Samarakoon, Xianfu Chen, and Mehdi Bennis. Wireless edge computing with latency and reliability guarantees. *Proceedings of the IEEE*, 107(8):1717–1737, 2019.

[Embrechts and Wang, 2015] Paul Embrechts and Ruodu Wang. Seven proofs for the subadditivity of expected shortfall. *Dependence Modeling*, 3(1):126–140, 2015.

[Haque *et al.*, 2015] Md E Haque, Yong Hun Eom, Yuxiong He, Sameh Elnikety, Ricardo Bianchini, and Kathryn S McKinley. Few-to-many: Incremental parallelism for reducing tail latency in interactive services. *ACM Sigplan Notices*, 50(4):161–175, 2015.

[Horn and Wang, 2017] Berthold KP Horn and Liang Wang. Wave equation of suppressed traffic flow instabilities. *IEEE Transactions on Intelligent Transportation Systems*, 19(9):2955–2964, 2017.

[Jian *et al.*, 2024] Zihan Jian, Xiangdong Xu, and Zhaoqi Zang. Estimating a route travel time distribution function with segment correlations using segment-level travel time data only: A moment-based method. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):18815–18830, 2024.

[Jiang *et al.*, 2012] Junchen Jiang, Vyas Sekar, and Hui Zhang. Improving fairness, efficiency, and stability in http-based adaptive video streaming with festive. In *Proceedings of the 8th international conference on Emerging networking experiments and technologies*, pages 97–108, 2012.

[Koutsoupias and Papadimitriou, 2009] Elias Koutsoupias and Christos Papadimitriou. Worst-case equilibria. *Computer Science Review*, 3(2):65–69, 2009.

[Li *et al.*, 2014] Jialin Li, Naveen Kr Sharma, Dan RK Ports, and Steven D Gribble. Tales of the tail: Hardware, os, and application-level sources of tail latency. In *Proceedings of the ACM Symposium on Cloud Computing*, pages 1–14, 2014.

[Lianeas *et al.*, 2016] Thanasis Lianeas, Evdokia Nikolova, and Nicolas E Stier-Moses. Asymptotically tight bounds for inefficiency in risk-averse selfish routing. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 338–344, 2016.

[Lianeas *et al.*, 2019] Thanasis Lianeas, Evdokia Nikolova, and Nicolas E Stier-Moses. Risk-averse selfish routing. *Mathematics of Operations Research*, 44(1):38–57, 2019.

[McNeil *et al.*, 2015] Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton university press, 2015.

[Nikolova and Stier-Moses, 2014] Evdokia Nikolova and Nicolás E Stier-Moses. A mean-risk model for the traffic assignment problem with stochastic travel times. *Operations Research*, 62(2):366–382, 2014.

[Nikolova and Stier-Moses, 2015] Evdokia Nikolova and Nicolas E Stier-Moses. The burden of risk aversion in mean-risk selfish routing. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 489–506, 2015.

[Ordónez and Stier-Moses, 2010] Fernando Ordónez and Nicolás E Stier-Moses. Wardrop equilibria with risk-averse users. *Transportation science*, 44(1):63–86, 2010.

[Pflug, 2000] Georg Ch. Pflug. Some remarks on the value-at-risk and the conditional value-at-risk. In *Probabilistic Constrained Optimization: Methodology and Applications*, pages 272–281. Springer, Boston, MA, 2000.

[Rockafellar and Uryasev, 2002] R. Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

[Rockafellar *et al.*, 2000] R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

[Roughgarden and Tardos, 2002] Tim Roughgarden and Éva Tardos. How bad is selfish routing? *Journal of the ACM (JACM)*, 49(2):236–259, 2002.

[Roughgarden, 2003] Tim Roughgarden. The price of anarchy is independent of the network topology. *Journal of Computer and System Sciences*, 67(2):341–364, 2003.

[Roughgarden, 2006] Tim Roughgarden. On the severity of braess's paradox: Designing networks for selfish users is hard. *Journal of Computer and System Sciences*, 72(5):922–953, 2006.

[Slumbers *et al.*, 2023] Oliver Slumbers, David Henry Mguni, Stefano B Blumberg, Stephen Marcus Mcaleer, Yaodong Yang, and Jun Wang. A game-theoretic framework for managing risk in multi-agent systems. In *International Conference on Machine Learning*, pages 32059–32087. PMLR, 2023.

[Stockhammer, 2011] Thomas Stockhammer. Dynamic adaptive streaming over http– standards and design principles. In *Proceedings of the second annual ACM conference on Multimedia systems*, pages 133–144, 2011.

[Wang and Zitikis, 2021] Ruodu Wang and Ričardas Zitikis. An axiomatic foundation for the expected shortfall. *Management Science*, 67(3):1413–1429, 2021.

[Wardrop, 1952] John Glen Wardrop. Road paper. some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers*, 1(3):325–362, 1952.

[Wu and Nie, 2011] Xing Wu and Yu Marco Nie. Modeling heterogeneous risk-taking behavior in route choice. *Transportation Research Part A: Policy and Practice*, 45(9):896–915, 2011.

[Xu *et al.*, 2017] Xiangdong Xu, Anthony Chen, Lin Cheng, and Chao Yang. A link-based mean-excess traffic equilibrium model under uncertainty. *Transportation Research Part B: Methodological*, 95:53–75, 2017.

[Zang *et al.*, 2018a] Zhaoqi Zang, Xiangdong Xu, Chao Yang, and Anthony Chen. A closed-form estimation of the travel time percentile function for characterizing travel time reliability. *Transportation Research Part B: Methodological*, 118:228–247, 2018.

[Zang *et al.*, 2018b] Zhaoqi Zang, Xiangdong Xu, Chao Yang, and Anthony Chen. A distribution-fitting-free approach to calculating travel time reliability ratio. *Transportation Research Part C: Emerging Technologies*, 89:83–95, 2018.

[Zang *et al.*, 2022] Zhaoqi Zang, Xiangdong Xu, Kai Qu, Ruiya Chen, and Anthony Chen. Travel time reliability in transportation networks: A review of methodological developments. *Transportation Research Part C: Emerging Technologies*, 143:103866, 2022.

[Zang *et al.*, 2024] Zhaoqi Zang, Richard Batley, Xiangdong Xu, and David ZW Wang. On the value of distribution tail in the valuation of travel time variability. *Transportation Research Part E: Logistics and Transportation Review*, 190:103695, 2024.

[Zang *et al.*, 2026a] Zhaoqi Zang, Richard Batley, David Z.W. Wang, and Hong K Lo. Reliability premium: A generic conceptual framework for evaluating the cost of travel time variability. *Transportation Research Part B: Methodological*, 205:103408, 2026.

[Zang *et al.*, 2026b] Zhaoqi Zang, Xiangdong Xu, Anthony Chen, David Z.W. Wang, and Shaojun Liu. Tail uncertainty concentration from link to route: Unveiling the preference-based foundation of mean-excess travel time in stochastic traffic networks. *Working paper*, 2026.

[Zhao *et al.*, 2023] Kevin Zhao, Prateesh Goyal, Mohammad Alizadeh, and Thomas E Anderson. Scalable tail latency estimation for data center networks. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 685–702, 2023.

[Zhuang *et al.*, 2023] Ziyu Zhuang, Kai Wang, and Zihe Wang. Price of anarchy for first price auction with risk-averse bidders. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1363–1369. International Foundation for Autonomous Agents and Multiagent Systems, 2023.