Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Risk-Averse Selfish Routing

Thanasis Lianeas, Evdokia Nikolova, Nicolas E. Stier-Moses

informs
http://pubsonline.informs.org/journal/moor/

MATHEMATICS OF OPERATIONS RESEARCH
Articles in Advance, pp. 1–20
ISSN 0364-765X (print), ISSN 1526-5471 (online)

Risk-Averse Selfish Routing

Thanasis Lianeas,[a] Evdokia Nikolova,[a] Nicolas E. Stier-Moses[b]

[a] University of Texas at Austin, Austin, Texas 78705; [b] Facebook Core Data Science, Menlo Park, California 94303

Contact: lianeas@corelab.ntua.gr (TL); nikolova@austin.utexas.edu (EN); nstier@utdt.edu, http://orcid.org/0000-0001-6380-1516 (NES-M)

Received: August 27, 2016
Revised: April 22, 2017; September 16, 2017
Accepted: September 29, 2017
Published Online in Articles in Advance: September 6, 2018

MSC2010 Subject Classification: Primary: 91A15; secondary: 91A43, 90B20, 90B15, 05C21
OR/MS Subject Classification: Primary: stochastic: games/group decisions; secondary: flow algorithms: networks/graphs

https://doi.org/10.1287/moor.2017.0913

Abstract. We consider a nonatomic selfish routing model with independent stochastic travel times for each edge, represented by mean and variance latency functions that depend on edge flows. This model can apply to traffic in the Internet or in a road network. Variability negatively impacts packets or drivers by introducing jitter in transmission delays, which lowers quality of streaming audio or video, or by making it more difficult to predict the arrival time at destination. At equilibrium, agents may select paths that do not minimize the expected latency so as to obtain lower variability. A social planner, who is likely to be more risk neutral than agents because it operates at a longer time scale, quantifies social cost with the total expected delay along routes. From that perspective, agents may make suboptimal decisions that degrade long-term quality. We define the *price of risk aversion* (PRA) as the worst-case ratio of the social cost at a risk-averse Wardrop equilibrium to that where agents are risk neutral. This inefficiency metric captures the degradation of system performance caused by variability and risk aversion.

For networks with general delay functions and a single source–sink pair, we first show upper bounds for the PRA that depend linearly on the agents' risk tolerance and on the degree of variability present in the network. We call these bounds *structural*, as they depend on the structure of the network. To get this result, we rely on a combinatorial proof that employs alternating paths that are reminiscent of those used in max-flow algorithms. For *series-parallel* graphs, the PRA becomes independent of the network topology and its size. Next, we provide tight and asymptotically tight lower bounds on the PRA by showing a family of *structural* lower bounds, which grow linearly with the number of nodes in the graph and players' risk aversion. These are tight for graph sizes that are powers of 2. After that, by focusing on restricting the set of allowable mean latency and variance functions, we derive *functional* bounds on the PRA that are asymptotically tight and depend on the allowed latency functions but not on the topology. The functional bounds match the price-of-anarchy bounds for congestion games multiplied by an extra factor that accounts for risk aversion. Finally, we turn to the mean-standard deviation user objective—a much more complex model of risk aversion because the cost of a path is nonadditive over edge costs—and provide tight bounds for instances that admit alternating paths with one or two forward subpaths.

Funding: Part of this work was done while the first two authors were visiting the Simons Institute for the Theory of Computing at Berkeley and the third author had a faculty position at Universidad Torcuato Di Tella in Buenos Aires, Argentina. This research was supported by National Science Foundation [Grants CCF-1216103, CCF-1350823, and CCF-1331863], National Scientific and Technical Research Council (CONICET), Argentina [Grant PIP 112-201201-00450CO], and Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT), Argentina [Grant PICT-2012-1324].

Keywords: stochastic Wardrop game • risk aversion • price of anarchy

1. Introduction

A central question in decision making is how to make good decisions under uncertainty, particularly when decision makers are risk averse. Applications of crucial national importance, including alleviating congestion in transportation networks, as well as improving telecommunications, robotics, security, and others, all face pervasive uncertainty and often require finding *reliable* or *risk-minimizing solutions*. Those applications have motivated the development of algorithms that incorporate risk primitives and the inclusion of risk aversion in questions related to algorithmic game theory. While risk has been extensively studied in the fields of finance and operations, among others, in comparison, there is relatively little prior research in the network optimization community devoted to this issue. One of the goals of this paper is to inspire more work devoted to understanding and mitigating risk in networked systems.

Capturing uncertainty and risk aversion in traditional combinatorial problems often reduces to nonlinear or nonconvex optimization over combinatorial feasible sets for which no efficient algorithms are known. Possibly as a result of the difficulty in writing the ensuing problems in simple terms, at present, we lack a systematic understanding of how risk considerations can be successfully incorporated into classic combinatorial problems. Doing so would necessitate new techniques for analyzing risk-minimizing combinatorial structures rigorously.

Within the fields of algorithms and algorithmic game theory, routing has proved to be a pervasive source of important questions. Indeed, many fundamental questions on risk-averse routing are still open, including several intriguing cases where the complexity is unknown. For example, in a network with uncertain edge delays, what is the complexity of finding the path with highest probability of reaching the destination within a given deadline? What is the path that minimizes the mean delay plus the standard deviation along that path? The best-known algorithms for both of these questions have a polynomial average and a polynomial smoothed complexity but a subexponential worst-case running time (Nikolova [40], Nikolova et al. [45]). Hence, these problems are unlikely to be NP-hard, yet polynomial algorithms have so far been elusive. There are also a myriad of other possible risk objectives, yielding nonlinear and nonconvex optimization problems over the path polytope, that are open from complexity, algorithmic, and approximation points of view.

Consequently, we are at the very beginning of understanding risk and designing appropriate models in routing games, which have been instrumental in the development of algorithmic game theory in the past two decades. Routing games capture decision making by multiple agents in a networking context, internalizing the congestion externalities generated by self-minded agents. Externalities are traditionally captured by considering that edge delays are functions of the edge flow. In addition to the technical challenges involving the characterization and computation of equilibria, a key insight brought by the study of these games was that equilibria are not extremely inefficient. The *price of anarchy*—by now a widely studied concept used in a variety of problems and applications—represents the worst-case ratio between the cost of a Nash equilibrium and the cost of the socially optimal solution. This ratio, which quantifies the degradation of system performance due to selfish behavior, was first defined in the context of routing and applied to a network with parallel links (Koutsoupias and Papadimitriou [29]). It was subsequently analyzed for general networks and different types of players (Correa et al. [14, 15], Roughgarden [53], Roughgarden and Tardos [56]). With few exceptions, this stream of work has assumed that delays are deterministic, while almost every practical situation in which such games could be useful presents uncertainty. For instance, there are uncertain delays in a transportation network as a result of weather, accidents, traffic lights, etc., and in telecommunications networks as a result of changing demand, hardware failures, interference, packet retransmissions, etc.

1.1. Risk Model

A generalization of the classic selfish routing model (Beckmann et al. [4]) to the case of uncertain delays is to associate every edge to a random variable whose distribution depends on the edge flow. The problem faced by a risk-averse agent becomes more involved since it is not merely finding the shortest path with respect to delays. To find the best route, agents must consider both the expected delay and the variability along all possible choices, leading them to solve stochastic shortest path problems. For instance, it is common that commuters add a buffer to the expected travel time for their trip to maximize the chance of arriving on time to an important meeting or to a flight at the airport. A classic model in finance that captures the trade-off between mean and variability is Markowitz's mean-risk framework (Markowitz [33]). It considers an agent that optimizes a linear combination of mean and risk, weighted by a *risk-aversion coefficient* γ that quantifies the degree of risk aversion of that agent (i.e., the agent utility is mean $+ \gamma \cdot$ risk). In the context of routing, Nikolova and Stier-Moses [42] adapted that framework to Wardrop equilibria, showed the existence of equilibria, commented on their uniqueness, and computed their worst-case inefficiency as captured by the price of anarchy. Henceforth, we now focus on studying the impact of risk aversion.

Of course, there are multiple ways to capture risk. The expected utility theory (e.g., Neumann and Morgenstern [38]), which is prevalent in economics, captures risk-averse preferences using concave utility functions. This theory has been criticized because of unrealistic assumptions such as independence of irrelevant alternatives so other theories have been proposed (e.g., Tversky and Kahneman [59]). The theory of coherent risk measures, proposed in the late 1990s, takes an axiomatic approach to risk (Krokhmal et al. [30], Rockafellar [52]). Cominetti and Torrico [12] adapted these ideas to the context of network routing and concluded that the mean-variance objective has benefits over other risk measures in being additively consistent, i.e., any subpath of an optimal path remains optimal. Furthermore, the mean-variance objective also arises from constant absolute risk aversion (CARA) expected utilities for exponentials with normally distributed uncertainty. For instance, working out the expectation of the CARA utility if the distribution X is normal, one gets $\mathbb{E}(\exp(\gamma X)) = \exp(\gamma \cdot \text{mean} + \frac{1}{2}\gamma^2 \cdot \text{variance})$, which has the same structure as the mean-variance objective after a monotonic transformation.

In finance, the mean risk and other traditional risk measures have been criticized for leading to paradoxes such as preferring stochastically dominated solutions. Nevertheless, different risk postulates may be relevant in the context of transportation and telecommunications. For instance, stochastically dominated routes may be admissible if certainty is more valued than a stochastically dominant solution with large variance. Indeed, one may choose a larger latency path rather than routing along variable paths that introduce jitter in real-time communications. Following on previous work on risk-averse congestion games (Nikolova and Stier-Moses [42]), in this paper, we consider the mean-variance and mean-standard deviation objectives for risk-averse routing.

It is important to note that risk aversion may induce agents to choose longer routes to reduce risk, effectively trading off mean with variability. Hence, a natural question to ask for a network game with uncertain delays and risk-averse agents is *how much of the degradation in system performance can be attributed to the agents' risk aversion*. We refer to this degradation by the *price of risk aversion* (PRA), which we formally define as the worst-case ratio of the social cost of the equilibrium (with risk-averse agents) to that of an equilibrium if agents were risk neutral. The reason for choosing this particular ratio is that we want to disentangle the effects caused by selfish behavior, captured by the price of anarchy, from those caused by risk aversion per se. The social cost is considered with respect to average delays because a central planner would typically care about a long-term perspective and minimize average agent delays and average pollutant emissions.

The mean-variance objective for a path is expressed as the mean travel time of the path plus the risk-aversion coefficient γ times the variance of travel time of the path (i.e., mean $+\gamma\cdot$ variance). Using the variance of delay along a route as a risk indicator leads to models that satisfy natural and intuitive optimality conditions for routes; namely, a subpath of an optimal path remains optimal (called the *additive consistency* property). Indeed, the mean-variance objective is additive along paths (the cost of a path is the sum of the cost of its edges). It thus lends itself to tractable algorithms in terms of computing equilibria, at least as long as delays are pairwise independent across edges. On the other hand, comparing a Wardrop equilibrium with risk-averse players to a standard Wardrop equilibrium is far from straightforward and requires new techniques for understanding how the two differ.

The measurement units of the terms that appear in the mean-variance objective function deserve further discussion. The mean is expressed in time units, the variance in time units squared, the risk-aversion coefficient γ in time units inverted. Hence, the objective ends up being measured in time units. Let us see through an example that the objective is invariant to change of units: consider a path with mean 1 hour, variance 0.25 hours squared, and $\gamma = 2\ \text{hr}^{-1}$. The mean-variance objective is $1\ \text{hr} + 2\ \text{hr}^{-1}\cdot 0.25\ \text{hr}^2 = 1.5\ \text{hr}$. Changing the units to minutes does not change the end result: $60\,\text{min} + 2(1/60)\,\text{min}^{-1}\cdot 0.25\cdot 3{,}600\,\text{min}^2 = 90\,\text{min}$.

The mean-variance objective can also be expressed as mean$(1+\gamma\text{VMR})$, where VMR is the variance-to-mean ratio of the path.[1] Here, the VMR is also expressed in time units. We compute the price of risk aversion with respect to an upper bound to the VMR, which captures the maximum allowed variation in travel time. We refer to this bound as κ and note that it is also measured in time units. Our results end up depending on the product $\kappa\gamma\eta$, where η is a unitless topological metric of the network. Putting the factors together, this expression is also unitless and, therefore, also robust to unit changes.

Alternatively, one could set the risk indicator to be the standard deviation of delays. A big advantage is that the mean-standard deviation objective can be thought of as a quantile of delay, easily justifying the buffer time that commuters consider when selecting the departure time of the trip. The disadvantages are that additive consistency is lost and, technically, that to compute the standard deviation one must take a square root, which makes the objective nonseparable and nonconvex. For more details, we refer the reader to Nikolova and Stier-Moses [42], where these pros and cons are discussed in further detail.

1.2. Our Results

We define a new concept, the price of risk aversion (PRA), as the worst-case ratio of the social cost (total expected delay) of a risk-averse Wardrop equilibrium (RAWE) to that of a risk-neutral Wardrop equilibrium (RNWE). Our first result, presented in Section 3, is a bound on the price of risk aversion for arbitrary graphs with a single origin–destination (OD) pair and symmetric players who minimize their mean-variance objective. We provide a bound of $1+\gamma\kappa\eta$, where γ is the risk-aversion coefficient, κ is the maximum possible variability (variance-to-mean ratio) of all edges when the prevailing traffic conditions are those under the equilibrium, and η is a topological parameter that captures how many flow-bearing paths are needed to cover a special structure called an alternating path. The resulting bound is appealing in that it depends on the three factors that one would have expected (risk aversion, variability, and network size) but perhaps unexpectedly does so in a linear way and for arbitrary delay functions. The parameter η strongly depends on the topology and is at most half the number of nodes in the network, $\lceil (n-1)/2\rceil$. From a graph topology perspective, η captures the number of forward (maximal) subpaths of some specific undirected path that initiates at the source; terminates at the sink; and traverses edges in their actual or their reversed direction in between, using forward or backward edges, respectively. This specific path additionally has the property that in its forward edges, the flow under an RAWE is less than or equal to the flow under an RNWE, and in the backward edges, the opposite inequality holds.

The proof of our first result is based on the construction of an *alternating path*, which, as referred to in the previous paragraph, allows us to compare the risk-averse and risk-neutral equilibrium flows. Technically, we prove three key lemmas that show that (a) an alternating path always exists, (b) the cost of an RAWE is upper bounded by an inflated total mean delay along forward edges minus the total mean delay along backward edges (Lemma 3), and (c) the cost of an RNWE is lower bounded by the total mean delay along forward edges minus the total mean delay along backward edges (Lemma 4). Steps (a) and (c) are proved independently of the choice of risk model. Step (b) is more subtle: it constructs a series of subpaths that connect different parts of the alternating path to the source and the sink, and it uses the equilibrium conditions to provide partial bounds for subpaths of the alternating path. The lemma then exploits the linearity of the mean-variance objective to get a telescopic sum that simplifies precisely to the total delay along the alternating path.

Theorem 1 puts the lemmas together and upper bounds the total mean delay of the forward subpaths in the alternating path by the cost of the RNWE times the number of such forward paths, obtaining the factor $\eta \leq \lceil (n-1)/2 \rceil$ in the worst case, as mentioned above. We prove that this bound is tight for series-parallel (SP) graphs (alternatively, graphs not containing the Braess graph as a minor), as it turns out that there must exist an alternating path that consists of only forward edges (i.e., $\eta = 1$), which implies that the price of risk aversion for those topologies is exactly $1 + \gamma\kappa$ (Corollary 2 and Example 1).

In Section 4, we provide lower bounds to the price of risk aversion, which are tight or asymptotically tight—that is, a lower bound to the price of risk aversion of $1 + \kappa\gamma n/2$, with number of vertices n that are powers of 2 (Theorem 3). This bound essentially closes the gap to the upper bounds of Section 3—that is, Corollary 1 and Theorem 1 (since for the constructed instance, $\eta = n/2$)—and characterizes the exact price of risk aversion for an infinite number of graph sizes.

Our construction of the worst-case graph family involves finding an instance in which an alternating path goes through every vertex in the graph and alternates between forward and backward edges at every internal vertex of the path. We achieve this by inductively defining a graph family with appropriate mean and variance functions for each edge (Theorem 2).

We call the above bounds *structural*, since they depend only on the network structure and not on the mean latency and variance functions used. Meir and Parkes [35] define biased smoothness and provide a technique that can be used to compare an equilibrium under modified cost functions to the social optimum of the original game. As an example of this technique, they derive an upper bound on the price of risk aversion of $(1 + \kappa\gamma)(1 - \mu)^{-1}$ when cost functions are $(1, \mu)$-smooth. As we discuss in this paper, this upper bound and the ones in Sections 3 and 4 are of a different type; that is, they are *functional* (based on the latency function classes) versus *structural* (based on the network structure), which is why they cannot be compared directly. One of our conceptual contributions is to put forward both perspectives in relation to the price of risk aversion, demonstrating that bounds can be given either in terms of the network structure or in terms of the class of edge latency functions that are allowed.

In Section 5, we present a new lower bound of the above functional form, as well as a simpler proof of the functional upper bound of Meir and Parkes that relies on a generalization of the earlier price of anarchy proof based on variational inequalities (Correa et al. [15]). Consequently, in addition to being direct and simpler, this method makes possible a more direct comparison with the existing literature including the traditional price of anarchy proofs. Our asymptotically tight functional lower bound (Theorem 5) follows from the same inductively defined family of graphs defined for the structural lower bound but with different appropriately chosen mean and variance functions. In Theorem 4, we give the new proof of the *functional* upper bound $(1 + \kappa\gamma)(1 - \mu)^{-1}$ for $(1, \mu)$-smooth latency functions via a variational inequality characterization of equilibria. We remark that for unrestricted functions, the functional upper bounds become vacuous since $\mu = 1$, which provides further support for the structural analysis of Sections 3 and 4.

As mentioned above, many of the results for the mean-variance risk model extend to the mean-standard deviation objective. In particular, the only piece missing to prove a general theorem is an equivalent of Lemma 3, which bounds the cost of a RAWE by an expression of the edge delays along the alternating path. The difficulty in extending our current proof to general graphs is the nonlinearity of the mean-standard deviation (mean-stdev) cost function, which, in turn, puts a restriction on the equilibrium flow in that its edge-flow representation cannot be decomposed arbitrarily to a path-flow representation.[2]

Circumventing the nonlinearity challenge, in Section 5, we provide upper bounds for instances that admit alternating paths with one or two forward subpaths and at most one backward edge. Both our structural and functional lower bounds for the mean-variance (mean-var) model readily provide corresponding lower bounds in the mean-stdev model. Showing a structural and functional upper bound for general graphs in that model remains an open problem.

1.3. Related Work

In this work, we consider how having stochastic delays and risk-averse users influence the traditional competitive network game introduced by Wardrop [62]. He postulated that the prevailing traffic conditions can be determined from the assumption that users jointly select shortest routes, and the mathematics that go with this idea were formalized in Beckmann et al. [4]. These models find applications in various application domains such as in transportation (Sheffi [57]) and telecommunications (Altman et al. [1]). In the last decade, these types of models have received renewed attention with many studies aimed at understanding existence, uniqueness, computation, and efficiency of equilibria (Correa and Stier-Moses [13], Nisan et al. [46]). The route choice model in this paper consists of users that select the path that minimizes the mean plus a multiple of the variability of travel time (captured by either variance or standard deviation). Exact algorithms and fully polynomial approximation schemes have been proposed for this problem and for a more general risk-averse combinatorial framework (Nikolova [40], Nikolova et al. [45]). See also Nikolova et al. [44], Swamy [58], and Li and Deshpande [31] for approximation algorithms for related frameworks.

The variance terms on the edges in our mean-variance model can be interpreted as tolls and thus our work can be seen as related to the vast literature on tolls, of which the most closely related work is that by Cole et al. [10], Bonifaci et al. [5], and Karakostas and Kolliopoulos [26]. A related stream of literature investigates how the topology of the graph affects total latency, specifically under biases or tolls (Chen and Kempe [9], Epstein et al. [18], Fotakis [21], Fotakis and Spirakis [22], Meir and Parkes [36], Milchtaich [37]).

Our alternating path construction is similar to an alternating cycle concept used in the context of tolls by Bonifaci et al. [5] and an alternating path concept used in the proof of Braess phenomena by Roughgarden [54]. Despite the notion of alternations, the proofs for these three different contexts seem unrelated. It is an interesting open question to discover deeper connections between the three models and results.

There is a growing literature on stochastic congestion games with risk-averse players. Ordóñez and Stier-Moses [47] introduce a game with uncertain delays and risk-averse users and study the relations between its solutions and *percentile equilibria*, which are flows under which percentiles of delays along flow-bearing paths are equal. Similar to the present work, they compare risk-averse equilibria to those with risk-neutral players. Nie [39] presents additional results on percentile equilibria. More closely related to the model considered here, Nikolova and Stier-Moses [42] prove existence and price-of-anarchy results (see the next paragraph), when the variability is captured by the standard deviations of delays. Piliouras et al. [51] consider the sensitivity of the price of anarchy to several risk-averse user objectives, in a different routing game model with atomic players and affine delay functions. Angelidakis et al. [2] also focus on atomic congestion games with uncertainty induced by stochastic players or stochastic delays and characterize when equilibria can be computed efficiently. Meir and Parkes [34] study a congestion game where agents have uncertainty over the routes used by other agents, which leads to the consideration of a range of users choosing each edge. Fotakis et al. [23] consider games with heterogeneous risk-averse players and show how uncertainty may and can be used to improve a network's performance.

For general congestion games, a series of papers in the last decade have studied the inefficiency introduced by self-minded behavior. To quantify that inefficiency, Koutsoupias and Papadimitriou [29] computed the supremum over all problem instances of the ratio of the equilibrium cost to the social optimum cost, which has been called the *price of anarchy* (POA) (Papadimitriou [48]). The POA has been characterized for increasingly more general assumptions (Chau and Sim [8], Correa et al. [14, 15], Perakis [49], Roughgarden [53], Roughgarden and Tardos [56]). Nikolova and Stier-Moses [42] extended that notion to the case of stochastic delays with risk-averse players. A different concept, the price of uncertainty, was considered in congestion games in reference to how best-response dynamics change under randomness introduced by an adversary and random ordering of players (Balcan et al. [3]). Risk aversion in the algorithmic game theory literature has been considered recently in the context of general games (e.g., Fiat and Papadimitriou [20]) and mechanism design (e.g., Dughmi [16], Dughmi and Peres [17], Fu et al. [24]). In transportation, the mean-variance model has been considered in the context of congestion pricing (Boyles et al. [7]) and network flow (Boyles and Waller [6]), and the mean-standard deviation model has been considered in the context of shortest paths (Khani and Boyles [27]).

The asymptotically tight functional bounds we present here were inspired by the recent work of Meir and Parkes [35]. In their paper, they prove a result that compares an equilibrium when players consider a modified cost function to the social optimum of the original game. As a corollary, they indirectly derive an upper bound on the price of risk aversion of $(1+\kappa\gamma)(1-\mu)^{-1}$ when cost functions are $(1,\mu)$-smooth. As we establish in this paper, this upper bound and that of Nikolova and Stier-Moses [43] are of a different type (i.e., functional versus topological), which is why they cannot be compared directly. Our proof of the upper bound relies on a simpler approach that is a straightforward generalization of the earlier price of anarchy proof based on variational inequalities put forward by Correa et al. [15]. Consequently, the method allows for an easier comparison and consistency with the traditional price of anarchy proofs. We also provide an asymptotically matching functional lower bound, which follows from the same graph construction as our topological lower bound.

Building on results that appeared in the conference versions of this work (Lianeas et al. [32], Nikolova and Stier-Moses [43]), Kleer and Schäfer [28] proved matching upper and lower bounds for the case of single-source (or single-sink) multicommodity networks. For that, they first defined a general model where the latencies perceived by the players are changed by adding an extra factor (positive or negative) to the original latencies. Note that this model can capture the one where players consider the mean-variance objective. Then, they compared the social cost at equilibrium under these perturbed latency functions to the social cost of the equilibrium under the original functions. To derive structural upper bounds for single-source multicommodity networks, they generalize the approach presented here by obtaining alternating paths for all commodities (an alternating path tree) and prove a similar bound. To derive structural lower bounds, they provide a family of single-commodity networks (different from our recursively constructed family) for which the PRA matches that upper bound. Note that this family of single-commodity networks (as well as ours) can be used for obtaining tight bounds for single-source multicommodity networks as one can pick any number of networks from that family and identify all the sources as one common source. By contrast, for general multicommodity networks, they proved that an exponential dependency of the PRA to the number of nodes is inevitable. They also improved recent smoothness results to bound the price of risk aversion. For that, they generalized the variational inequality technique of Correa et al. [15] but in a slightly different way from what we do here.

Finally, we mention again that this paper is part of a relatively new and growing literature exploring the effect of risk aversion on network equilibria in routing games (Angelidakis et al. [2], Cominetti and Torrico [12], Nie [39], Nikolova and Stier-Moses [41, 42, 43], Ordóñez and Stier-Moses [47], Piliouras et al. [51]). We refer the reader to the recent paper by Nikolova and Stier-Moses [43] for a more comprehensive review of additional related work, as well as a detailed discussion on the pros and cons of the risk-averse models considered here. We also refer the reader to the recent survey by Cominetti [11] for a more extensive review of equilibrium routing under uncertainty.

2. Model and Preliminaries

We consider a directed graph $G=(V,E)$ with a single source–sink pair (s,t) and an aggregate demand of d units of flow that need to be routed from s to t. For simplicity, and without loss of generality, we assume that $d=1$ in Section 3. Afterward, we switch back and use a general d, as it helps to better present the lower bound construction. We let $\mathcal{P}$ be the set of all feasible paths between s and t. We encode the players' decisions as a flow vector $f=(f_p)_{p\in\mathcal{P}}\in\mathbb{R}_+^{|\mathcal{P}|}$ over all paths. Such a flow is feasible when demand is satisfied, as given by the constraint $\sum_{p\in\mathcal{P}} f_p = d$. For notational simplicity, we denote the flow on an edge e by $f_e=\sum_{p\ni e} f_p$. When we need multiple flow variables, we use x, x_p, x_e, and z, z_p, z_e.

The network is subject to congestion, modeled with stochastic delay functions $l_e(f_e)+\xi_e(f_e)$ for each edge $e\in E$. Here, the deterministic function $l_e(f_e)$ measures the expected delay when the edge has flow f_e, and $\xi_e(f_e)$ is a random variable that represents a noise term on the delay. Functions $l_e(\cdot)$, generally referred to as *latency functions*, are assumed continuous and nondecreasing. The expected latency along a path p is given by $l_p(f):=\sum_{e\in p} l_e(f_e)$. Random variables $\xi_e(f_e)$ are pairwise independent and have expectation zero and standard deviation $\sigma_e(f_e)$ for arbitrary continuous functions $\sigma_e(\cdot)$. For the variational inequality characterization used in Section 5, we further assume that the mean-variance objective of users, defined below, is nondecreasing. The variance along a path equals $v_p(f)=\sum_{e\in p}\sigma_e^2(f_e)$, and the standard deviation is $\sigma_p(f)=(v_p(f))^{1/2}$.

We consider the *nonatomic* version of the routing game where infinitely many players control an infinitesimal amount of flow each so that the path choice of a single player does not unilaterally affect the costs experienced by other players.

Players are risk averse and strategically choose paths taking into account the variability of delays by considering a *mean-var* objective $Q_p^{\gamma}(f)=l_p(f)+\gamma v_p(f)$. (See the introduction for a discussion about units of measurement.) We refer to this objective simply as the *path cost* (as opposed to latency). Here, $\gamma\geq 0$ is a constant that quantifies the risk aversion of the players, which we assume to be homogeneous. The special case of $\gamma=0$

corresponds to risk neutrality. In the last section, we consider the *mean-stdev* objective where the variance in the objective is replaced with the standard deviation of the path.

In summary, an instance of the problem is given by the tuple (G, d, l, v, γ), which represents the topology, demand, latency functions, variability functions, and degree of player risk aversion. The following definition captures that at equilibrium, players route flow along paths with minimum cost.

Definition 1 (Equilibrium)**.** A *γ-equilibrium* of a stochastic nonatomic routing game is a flow f such that for every path $p \in \mathcal{P}$ with positive flow, the path cost $Q_p^\gamma(f) \le Q_q^\gamma(f)$ for any other path $q \in \mathcal{P}$. For a fixed risk-aversion parameter γ, we refer to a γ-equilibrium as an RAWE, denoted by x. For $\gamma = 0$, we call the equilibrium an RNWE and usually denote it by z.

Notice that since the variance decomposes as a sum over all the edges on the path, the previous definition represents a standard Wardrop equilibrium with respect to modified costs $l_e(f_e) + \gamma v_e(f_e)$. For the existence of the equilibrium, it is sufficient that the modified cost functions are increasing.

Our goal is to investigate the effect that risk-averse players have on the quality of equilibria. The quality of a solution that represents collective decisions can be quantified by the cost of equilibria with respect to expected delays since, over time, different realizations of delays average out to the mean by the law of large numbers. For this reason, a social planner, who is concerned about the long term, is typically risk neutral. Furthermore, the social planner may aim to reduce long-term emissions, which would be better captured by the total expected delay of all users.

Definition 2. The *social cost* of a flow f is defined as the sum of the expected latencies of all players: $C(f) := \sum_{p \in \mathcal{P}} f_p l_p(f) = \sum_{e \in E} f_e l_e(f_e)$.

Although one could have measured total cost as the weighted sum of the costs $Q_p^\gamma(f)$ of all users, this captures users' utilities but not the system's benefit. Such a cost function was previously considered to compute the price of anarchy (Nikolova and Stier-Moses [42]). By contrast, our goal is to compare across different values of risk aversion so we want the various flow costs to be compared apples to apples, which requires using the same cost function.

The variability of delays is usually not too large with respect to the expected latency. By the reasoning that follows Example 1 at the end of this section, we assume that $v_e(x_e)/l_e(x_e)$ is bounded from above by a fixed constant κ for all $e \in E$ at the equilibrium flow of interest $x_e \in \mathbb{R}_+$. This means that the variance cannot be larger than κ times the expected latency in any edge at equilibrium. The next definition captures the increase in social cost at equilibrium introduced by user risk aversion, compared with the social cost if users were risk neutral.

Definition 3 (Nikolova and Stier-Moses [43])**.** Considering an instance family $\mathcal{F}$ of a routing game with uncertain delays, the *price of risk aversion* (PRA) associated with γ and κ (the risk-aversion coefficient and the variance-to-mean ratio) is defined by

$$\mathrm{PRA}(\mathcal{F}, \gamma, \kappa) := \sup_{G,d,l,v} \left\{ \frac{C(x)}{C(z)} : (G, d, l, v, \gamma) \in \mathcal{F}, \text{ and } v(x) \le \kappa l(x) \right\},$$

where x and z are the risk-averse and the risk-neutral Wardrop equilibria of the corresponding instance.

This supremum depends on $\mathcal{F}$, which may be defined in terms of the network topology (as, e.g., general, series-parallel, or Braess networks), the number of vertices, or the set of allowed latency functions (as, e.g., affine or quadratic polynomials). Different results will be with respect to different families $\mathcal{F}$, with Sections 3, 4, and 6 focusing on structural definitions and Section 5 focusing on sets of allowed mean latency and variance functions. For the sake of brevity, we will typically write just PRA, and the parameters $\mathcal{F}$, γ, and κ will be clear from the context. Although we do not specify it explicitly in each result for brevity, all our results work for arbitrary values of $\gamma \ge 0$ and $\kappa \ge 0$.

We present the following example to motivate the form of the bound to the PRA, which is linear in $\gamma\kappa$. The example is based on a simple network with two edges, usually referred to as the *Pigou network* (Pigou [50], Roughgarden and Tardos [56]).

Example 1. Consider an instance with two nodes connected by two parallel edges with latencies equal to $(1 + \gamma\kappa)x$ and 1, respectively; variances equal to $v_1(\cdot) = 0$ and $v_2(\cdot) = \kappa$; and $d = 1$. Computing equilibria, the RNWE flow routes $1/(1 + \gamma\kappa)$ units along the first edge and $\gamma\kappa/(1 + \gamma\kappa)$ along the second. This gives a total cost of 1. Instead, the RAWE flow routes all the flow along the first edge, which gives a total cost of $1 + \gamma\kappa$. Dividing, we get that $\mathrm{PRA} \ge 1 + \gamma\kappa$.

The previous example motivates the need of imposing an upper bound on the variability of delays. Taking $\kappa \to \infty$, it follows that if one does not constrain variability of delays, the price of risk aversion is unbounded. Having bounded variability is a reasonable assumption in real-life networks since the variability is never too many times larger than the expected latency of an edge (see, e.g., Federal Highway Administration [19], Vicroads [61, p. 14]). In the following section, we shall prove that $1+\gamma\kappa$ is a matching upper bound for instances on series-parallel networks. Indeed, we will see that this will be a special case of a result for general topologies.

3. Structural Upper Bounds

We start by introducing bounds on the latency of the RAWE, which we will use to find the PRA. Throughout this section, we assume that $d = 1$ without loss of generality. We let z denote the RNWE and let x denote the RAWE. It is well known that, by definition, the social cost $C(z)$ of an RNWE can be upper bounded by the latency $l_p(z)$ of an arbitrary path $p \in \mathcal{P}$, and the bound is tight if the path carries flow. We now extend that argument to an RAWE. We prove that its social cost is bounded by the cost $Q_p^\gamma(x)$ of an arbitrary path $p \in \mathcal{P}$. As a corollary, $C(x)$ is also bounded by the expected latency of an arbitrary path, blown up by a constant that depends on the risk-aversion coefficient γ and the maximum coefficient of variation κ.

Lemma 1. *Letting $p \in \mathcal{P}$ denote an arbitrary path (potentially not carrying flow at equilibrium), the social cost of an RAWE $C(x)$ is upper bounded by the path cost $Q_p^\gamma(x)$. In addition, if variance functions satisfy that the variance-to-mean ratio at equilibrium is bounded by κ, then $C(x) \le (1+\gamma\kappa) l_p(x)$.*

Proof. From the equilibrium conditions, we have that $l_q(x) + \gamma v_q(x) \le l_p(x) + \gamma v_p(x)$ for all paths $q \in \mathcal{P}$ that carry positive flow. Therefore,

$$\begin{aligned} C(x) = \sum_{q\in\mathcal{P}} x_q l_q(x) &\le \sum_{q\in\mathcal{P}} x_q [l_p(x) + \gamma v_p(x) - \gamma v_q(x)] \\ &= l_p(x) + \gamma v_p(x) - \gamma \sum_{q\in\mathcal{P}} x_q v_q(x) \\ &\le l_p(x) + \gamma v_p(x) = Q_p^\gamma(x). \end{aligned}$$

Here, we have used the equilibrium condition and removed a negative term.

If variance functions satisfy that the variance-to-mean ratio at equilibrium is bounded by κ, then the mean-var cost of a path is bounded as follows:

$$Q_p^\gamma(x) = l_p(x) + \gamma \sum_{e\in p} v_e(x_e) \le l_p(x) + \gamma \sum_{e\in p} \kappa l_e(x_e) \le l_p(x)(1+\gamma\kappa),$$

by the assumption that $v_e(x_e) \le \kappa l_e(x_e)$ for all edges $e \in E$ at the equilibrium x, and the lemma follows. □

We will assume from here on that all the edges of the network carry flow under at least one of x or z since edges that carry no flow in both equilibria (i.e., $z_e = x_e = 0$) can be removed from the graph without loss of generality. We proceed to bound the price of risk aversion on a general graph by an appropriate construction of an alternating path that contains edges from the following two sets, which form a partition of the edges in E:

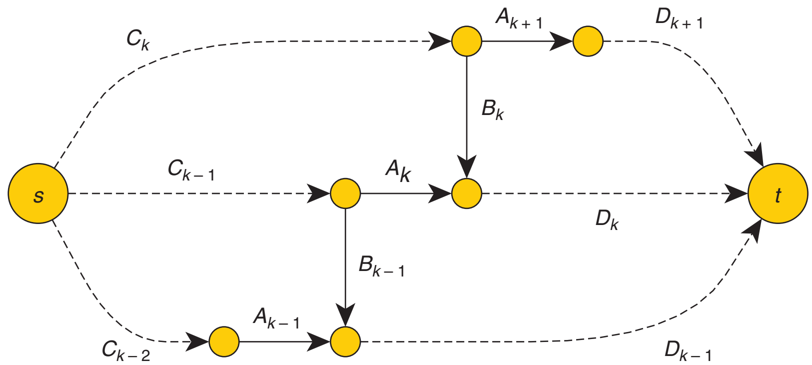
$$A = \{e \in E \mid x_e \le z_e\} \quad \text{and} \quad B = \{e \in E \mid z_e < x_e\}.$$

From the assumption that all network edges carry flow in at least one of x or z, we have that $z_e > 0$ for all $e \in A$ and $x_e > 0$ for all $e \in B$. If there is a full s–t path π contained in the set A, then it is not too hard to prove that $C(x) \le (1+\gamma\kappa)C(z)$. In other words, this would give the lowest possible PRA bound of $1+\gamma\kappa$ (recall Example 1). We now prove that this bound can be extended to *alternating paths* in G, which are s–t paths consisting of edges in A plus reversed edges in B. We shall refer to edges on the alternating path that belong to A as forward edges and those in B as backward edges.

Definition 4. A generalized s–t path $\pi = A_1$–B_1–A_2–B_2–$\cdots$–A_t–B_t–A_{t+1}, composed of a sequence of subpaths, is an *alternating path* when, for every i, every edge in $A_i \subseteq A$ is directed in the direction of the path and every edge in $B_i \subseteq B$ is directed in the opposite direction from the path. In other words, the subpaths A_i's and the subpaths B_i's with the directions reversed are actual paths in the underlying graph. We say that π has $t+1$ disjoint forward subpaths and t *alternations*.

Figure 1 provides an illustration of the alternating path definition where reversing edges in B creates a feasible path. The existence of an alternating path follows from flow conservation and the definitions of sets A and B.

Figure 1. (Color online) Part of an alternating path. Labels denote the names of subpaths used in this section. The C_i's are paths that route flow from s to A_{i+1} or B_i paths, and the D_i's are paths that route flow from A_i or B_i to t.



Lemma 2. *For any instance* $\mathcal{G} = (G, d, l, v, \gamma)$*, an alternating path exists.*

Proof. We give a constructive proof that such a path exists by first showing that any s–t cut in G must have a forward edge in A or a backward edge in B and then, starting from a cut that only contains s, repeatedly exploiting this property to expand the cut until t enters inside the cut. This construction will stop when an alternating path is found, as needed.

To start, consider any s–t cut defined by $S \subset V$ with $s \in S$. We first prove that we can cross the cut with an edge in A or a reverse edge in B. To get a contradiction, suppose that all edges incoming to S are in A and all edges outgoing from S are in B. Denote by x_A and z_A the total incoming flow into S corresponding to flow vectors x and z, respectively, and by x_B and z_B, the total outgoing flows from S, respectively. The definition of set A implies that $x_A \leq z_A$. Since conservation of flow imposes that $x_B - x_A = z_B - z_A$, we have $x_B \leq z_B$. Furthermore, from the definition of B, $x_B > z_B$ (note that we removed edges with $x_e = z_e = 0$), which is a contradiction. Now, starting with the cut $(S, G\backslash S)$, where $S = \{s\}$, we find an appropriate edge crossing the cut (i.e., an outgoing or incoming edge that belongs to A or B, respectively) and move both of its endpoints to S. Thus, we add nodes to S one by one until $t \in S$. At this point, we will have a tree of forward and backward edges containing s and t. Consequently, this tree yields an alternating path from the source to the destination. □

We use the alternating path to provide an upper bound on the PRA that depends on the number of times the alternating path switches from A to B. To get there, we need two lemmas. The first lemma extends Lemma 1, which applies to (standard) paths, to the case of alternating paths. Note that it allows us to tighten the previous bound by subtracting the latencies of the backward edges in the alternating path. The lemma provides an upper bound on the social cost of the RAWE x by exploiting the equilibrium conditions on the subpaths B_i on the alternating path with respect to the risk-averse objective.

Lemma 3. *Consider a graph with variance functions that satisfy that the variance-to-mean ratio at equilibrium is bounded by* κ*. Letting* π *be an alternating path, the social cost of a risk-averse Wardrop equilibrium* x *satisfies*

$$C(x) \leq (1 + \gamma\kappa) \sum_{e \in A \cap \pi} l_e(x_e) - \sum_{e \in B \cap \pi} l_e(x_e).$$

Proof. Let us assume that the alternating path consists of subpaths A_1–B_1–A_2–$\cdots$–$A_{\eta-1}$–$B_{\eta-1}$–A_η, where each subpath is in the corresponding set A or B. Since by definition each edge e in B_k carries flow ($x_e > 0$) for any k, e must belong to an s–t path that carries flow under RAWE x. Selecting a decomposition where the whole subpath B_k is on the same path (we have the freedom to do that since this is a standard Wardrop model with respect to the mean-variance objective), there must be a flow-carrying path that consists of subpaths C_k–B_k–D_k, where C_k originates at the source node and D_k terminates at the destination node (see Figure 1 for an illustration). We define $C_0 = D_\eta = \emptyset$. To simplify notation, only for the proof of this lemma will we refer to the mean-variance cost of subpath P also by $P = \sum_{e \in P}(l_e(x_e) + \gamma v_e(x_e))$.

We next use the equilibrium conditions to derive bounds on C_k and D_k. Since the subpath C_k–B_k carries flow, and the subpath C_{k-1}–A_k is an alternative route between the endpoints of C_k–B_k, we have that $C_k + B_k \leq C_{k-1} + A_k$ for all k.[3] Note that here and in what follows, we critically use the additivity of the mean-variance cost. Therefore,

$$C_k \leq C_{k-1} + A_k - B_k \leq \cdots \leq (A_1 + A_2 + \cdots + A_k) - (B_1 + B_2 + \cdots + B_k). \tag{1}$$

Similarly, since B_k–D_k carries flow, and A_{k+1}–D_{k+1} is an alternative route between the same endpoints, we have that

$$D_k \leq (A_{k+1} + A_{k+2} + \cdots + A_\eta) - (B_k + B_{k+1} + \cdots + B_{\eta-1}). \tag{2}$$

Then, for path $q = C_k$–B_k–D_k for any k, we have that

$$\begin{aligned}
C(x) &= \sum_p x_p l_p(x) \\
&\leq \sum_p x_p(l_q(x) + \gamma v_q(x) - \gamma v_p(x)) && \text{since either } x_p = 0 \text{ or } Q_p^\gamma(x) \leq Q_q^\gamma(x) \\
&\leq C_k B_k D_k && \text{after neglecting the negative term} \\
&\leq (A_1 + \cdots + A_\eta) - (B_1 + \cdots + B_{\eta-1}) && \text{using inequalities (1) and (2)} \\
&\leq \sum_{i=1}^{\eta} \sum_{e \in A_i} (l_e(x_e) + \gamma v_e(x_e)) - \sum_{i=1}^{\eta-1} \sum_{e \in B_i} l_e(x_e) && \text{neglecting variances in negative term} \\
&\leq (1 + \gamma\kappa) \sum_{e \in A \cap \pi} l_e(x_e) - \sum_{e \in B \cap \pi} l_e(x_e).
\end{aligned}$$

The last inequality follows by applying the variability bound on the variances. □

The previous result provided an upper bound for the RAWE x. Now, we complement it with a lower bound for the RNWE z. Again, to get the result, we exploit the equilibrium conditions, now with respect to $l(\cdot)$.

Lemma 4. *Letting π be an alternating path, the social cost of an RNWE z satisfies $C(z) \geq \sum_{e \in A \cap \pi} l_e(z_e) - \sum_{e \in B \cap \pi} l_e(z_e)$.*

Proof. Since $z_e > 0$ for any $e \in A_k$, there must be a subpath C_{k-1} that brings flow to A_k (this C_{k-1} need not be the same as that used in the proof of Lemma 3). Then, there is a flow decomposition in which the subpath C_{k-1}–A_k is used by z. Because subpath C_k–B_k is an alternative route from s to the node at the end of A_k, we must have that $l_{C_{k-1}}(z) + l_{A_k}(z) \leq l_{C_k}(z) + l_{B_k}(z)$. Summing the previous inequalities for all k (where C_0 is defined as an empty path), we get $l_{C_{\eta-1}}(z) \geq \sum_{k=1}^{\eta-1}(l_{A_k}(z) - l_{B_k}(z))$. This proves the lemma because $C(z) = l_{C_{\eta-1}}(z) + l_{A_\eta}(z)$, since $C_{\eta-1}$–A_η is a flow-carrying s–t path for z, all flow-carrying s–t paths have the same cost, and the total demand is $d = 1$. □

With the previous two lemmas that provided bounds for x and z, and the sets A and B that allow us to compare both flows, the proof of the main result consists of just chaining the inequalities.

Theorem 1. *Consider a general instance that has variance functions with variance-to-mean ratio at equilibrium bounded by κ. Letting π be an alternating path, the price of risk aversion is upper bounded by $1 + \gamma\kappa\eta$, where η is the number of disjoint forward subpaths in the alternating path π.*

Proof. The result follows from

$$\begin{aligned}
C(x) &\leq (1 + \gamma\kappa) \sum_{e \in A \cap \pi} l_e(x_e) - \sum_{e \in B \cap \pi} l_e(x_e) && \text{by Lemma 3} \\
&\leq (1 + \gamma\kappa) \sum_{e \in A \cap \pi} l_e(z_e) - \sum_{e \in B \cap \pi} l_e(z_e) && \text{by definitions of } A \text{ and } B \\
&\leq C(z) + \gamma\kappa \sum_{e \in A \cap \pi} l_e(z_e) && \text{by Lemma 4} \\
&\leq C(z) + \gamma\kappa\eta C(z) = (1 + \gamma\kappa\eta) C(z).
\end{aligned}$$

In the last inequality, we have used that $\sum_{e \in A \cap \pi} l_e(z) \leq \eta C(z)$. This holds because for all forward subpaths $A_k \in \pi$, their edges satisfy $z_e > 0$. Hence, under some decomposition for every k, there is some path q_k with $z_{q_k} > 0$ that includes the subpath A_k, implying $l_{A_k}(z) \leq l_{q_k}(z) = C(z)$. The equality holds because $d = 1$. □

The parameter η, referred to in the introduction, is the maximum possible number of disjoint forward subpaths. By way of construction, an alternating path goes through every node at most once and the number of forward subpaths is maximized when the path consists of alternating forward and backward edges, for a total of at most $n - 1$ edges. Therefore $\eta \leq \lceil (n-1)/2 \rceil$.

Corollary 1. *The price of risk aversion in a general graph is upper bounded by $1 + \gamma\kappa\lceil (n-1)/2 \rceil$.*

The bound depends on the three factors that one would expect (risk aversion, variability, and network size) but perhaps unexpectedly does so in a linear way and for arbitrary delay and variance functions. The tightness of the bounds of Theorem 1 and Corollary 1 is proved in the next section—althhough, for graphs with alternating paths that consist of only forward edges, or graphs with two vertices, this tightness is already proved via Example 1.

Next, as another corollary of Theorem 1, we upper bound the price of risk aversion in series-parallel graphs to be at most $1+\gamma\kappa$, independently of the size of the network. Given the lower bound provided by Example 1 (a Pigou graph is series-parallel), this bound must be tight. Series-parallel graphs are those formed recursively by subdividing an edge in two subedges or replacing an edge by two parallel edges. A noteworthy alternative characterization is that a graph is series-parallel if and only if it does not contain a Braess subgraph as an induced minor (Valdes et al. [60]).

Corollary 2. *The price of risk aversion among all series-parallel instances is exactly* $1+\gamma\kappa$.

Proof. We are going to prove that there exists an alternating path π consisting only of forward edges, so $\pi \subseteq A$. Let us consider a minimal (cardinality-wise) alternating path with a backward edge. The key property of series-parallel graphs is that after taking a reverse edge e^-, where $e=(i,j)\in E$, π has to either come back to node j or close a loop with itself. If that did not happen, it would imply that a Braess graph is embedded in the instance, which is not possible. Hence, there is an alternating path π' without the reverse edge e^-, which is a contradiction to the minimality of π. □

4. Structural Lower Bounds

In this section, we present two lower bounds on the price of risk aversion that match the upper bounds of Theorem 1 and Corollary 1, respectively. The first bound for PRA is with respect to the minimum number of alternations among all alternating paths, while the second bound is with respect to the number of vertices in the graph. In fact, the same bounds hold in the mean-standard deviation model, but we defer that discussion to Section 6.

To prove the lower bounds, we first prove a more general result that shows how instances with a high price of risk aversion can be constructed. Specifically, we will show that the price of risk aversion is $1+2\gamma\kappa$ in the Braess graph G^1 with 2^2 nodes in Figure 2. Then, we will inductively prove that the price of risk aversion is $1+2^i\gamma\kappa$ in a bigger graph G^i with 2^{i+1} nodes, which is constructed from two copies of graph G^{i-1}, connected in a Braess-like fashion, as shown in Figure 3. In the inductive step, we will argue that the RAWE and RNWE flows are of a certain form—the RAWE flow will take "zigzag" paths (paths using vertical edges), and the RNWE flow will take paths that do not use vertical edges. Since the flow going into subgraphs will be broken down into smaller and smaller fractions, in Theorem 2, we state and analyze the equilibria for such smaller amounts of flow entering the graph G^i, denoted by r_A^i and r_N^i for the RAWE flow and RNWE flows, respectively.

Theorem 2. *For every positive integer i and every demand pair $r_A^i, r_N^i \in \mathbb{R}_{>0}$ such that $2^i r_A^i > (2^i-1)r_N^i$, there exists a graph instance $G^i(r_A^i, r_N^i)$ with 2^{i+1} nodes that satisfies the following two properties:*

- *If r_A^i risk-averse players are routed through $G^i(r_A^i, r_N^i)$, then the mean-var cost along used paths at the RAWE flow x, as well as the expected latency along used paths, is $1+2^i\gamma\kappa$. The social cost is $C(x)=(1+2^i\gamma\kappa)r_A^i$.*
- *If r_N^i risk-neutral players are routed through $G^i(r_A^i, r_N^i)$, then the expected latency along used paths at the RNWE flow z is 1. The social cost is $C(z)=r_N^i$.*

The proof is by induction on i. We will recursively construct the instance for i by forming a Braess instance with the graph resulting for the $i-1$ case. At each step, we will need to find a mean latency function that makes the properties in the statement work.

Figure 2. (Color online) The base case $G^1(r_A^1, r_N^1)$ is a Braess graph.

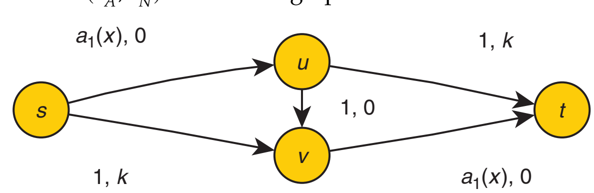
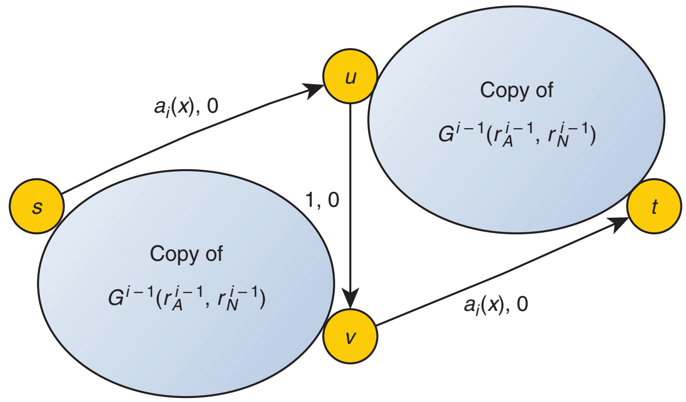


Figure 3. (Color online) The recursive construction of $G^i(r_A^i, r_N^i)$ forms a Braess graph topology using components of the earlier step. For the bottom left, the source vertex is identified with s and the sink vertex is identified with v, and for the top right, the source vertex is identified with u and the sink vertex is identified with t.



Proof. For the base case $i = 1$, we let $G^1(r_A^1, r_N^1)$ be the Braess graph, shown in Figure 2. Indeed, consider any r_A^1, r_N^1 such that $r_A^1 > r_N^1/2$, as indicated in the statement of the result. We define the mean latency function $a_1(x)$ to be any function that is strictly increasing for $x \geq r_N^1/2$ and such that $a_1(r_N^1/2) = 0$ and $a_1(r_A^1) = \gamma\kappa$. Note that for $a_1(x)$ to be strictly increasing, it is necessary that $r_A^1 > r_N^1/2$, which holds by hypothesis.

The flow x that routes the r_A^1 risk-averse players along the zigzag path is a risk-averse equilibrium. Observe that under this flow, the upper left and lower right edges have mean-var cost, as well as expected latency, equal to $\gamma\kappa$ each. Consequently, the mean-var cost along each of the three possible paths is $1 + 2\gamma\kappa$, confirming that flow x is an RAWE and $C(x) = (1 + 2\gamma\kappa)r_A^1$. Instead, the RNWE flow z routes the r_N^1 risk-neutral players along the top and bottom paths, half and half. Hence, the cost for each player is 1 and $C(z) = r_N^1$, proving the base case.

Let us consider the inductive step where we assume that we have an instance satisfying the properties for $i-1$ and construct the instance for step i. Starting from r_A^i and r_N^i satisfying the condition in the statement for case i, we set $r_A^{i-1} = (2^i r_A^i - r_N^i)/2^{i+1}$ and $r_N^{i-1} = r_N^i/2$. We first verify that these values satisfy the hypothesis for the case $i-1$. Indeed, $r_A^{i-1} > ((2^{i-1}-1)/2^{i-1})r_N^{i-1}$ because, by hypothesis, $r_A^i > ((2^i - 1)/2^i)r_N^i \Leftrightarrow r_A^i/2 > ((2^i-1)/2^{i+1})r_N^i$, which implies that $r_A^i/2 - r_N^i/2^{i+1} > ((2^i-1)/2^{i+1})r_N^i - (1/2^{i+1})r_N^i = ((2^{i-1}-1)/2^{i-1})(r_N^i/2)$.

Using the graph G^{i-1} from step $i-1$ and the values of r_A^{i-1} and r_N^{i-1} specified above, we construct graph $G^i(r_A^i, r_N^i)$ from two copies of graph G^{i-1} connected in a Braess-like fashion, as shown in Figure 3. We define the mean latency function $a_i(x)$ to be any function that is strictly increasing for $x \geq r_N^i/2$ and such that $a_1(r_N^i/2) = 0$ and $a_i(r_A^i/2 + r_N^i/2^{i+1}) = 2^{i-1}\gamma\kappa$. Note that for $a_i(x)$ to be strictly increasing, it is necessary that $r_A^i/2 + r_N^i/2^{i+1} > r_N^i/2$, which actually holds because, by hypothesis, $r_A^i > ((2^i-1)/2^i)r_N^i \Leftrightarrow r_A^i/2 > ((2^i-1)/2^{i+1})r_N^i$.

The RAWE flow x routes the r_A^i risk-averse players as follows: $r_A^i/2 - r_N^i/2^{i+1}$ units along the upper path, $r_N^i/2^i$ units along the zigzag path, and $r_A^i/2 - r_N^i/2^{i+1}$ units along the lower path. The mean-var objective of the upper left and the lower right edges, as well as the mean latency, will each be $2^{i-1}\gamma\kappa$ since the flow through them is equal to $r_A^i/2 + r_N^i/2^{i+1}$. The flow inside each of the copies of $G^{i-1}(r_A^{i-1}, r_N^{i-1})$ is an RAWE for which we know, by induction, that all players perceive a path cost of $1 + 2^{i-1}\gamma\kappa$, which additionally, by induction, is the mean latency of all used paths. Thus, the path cost that players perceive in $G^i(r_A^i, r_N^i)$ under the RAWE flow x is $1 + 2^i\gamma\kappa$, which additionally is the mean latency of all used paths, and the social cost is $C(x) = (1 + 2^i\gamma\kappa)r_A^i$.

The RNWE flow z routes the r_N^i risk-neutral players along the top and bottom paths, half and half. Hence, the path cost perceived by each player is 1, as the mean-var objective in the upper left and lower right edges is equal to 0, and, by induction, passing through either of both copies of $G^{i-1}(r_A^{i-1}, r_N^{i-1})$ has a mean-var objective of 1. This implies that $C(z) = r_N^i$, which completes the proof. □

The previous result provides a constructive way to generate instances with high price of risk aversion. Notice that the paths of the instances resulting from these constructions have at most one edge with nonzero variance. This fact is useful to extend our lower bounds to the mean-stdev model, since in that case, summing and taking square roots is not needed.

Another useful observation is that the prevailing value for mean latency functions a_j under the RNWE flow z is 0, and under the RAWE flow x is 2^{j-1}. This can be easily proved by induction and will be used when establishing functional lower bounds on the PRA in the next section.

We now use the previous result to get lower bounds for PRA matching the upper bound specified earlier.

Corollary 3. *For any $n_0 \in \mathbb{N}$, there is an instance on a graph G with $n \geq n_0$ vertices such that its equilibria satisfy $C(x) \geq (1+\gamma\kappa\lceil (n-1)/2\rceil)C(z)$.*

Proof. Consider an arbitrary demand d, and apply Theorem 2 with $r_A^i = r_N^i = d$ and $i = \min\{j \in \mathbb{N}: n_0 \leq 2^j\}$ to get instance $G^i(r_A^i, r_N^i)$. Consequently, the RAWE flow x and the RNWE flow z satisfy that

$$\frac{C(x)}{C(z)} = \frac{(1+2^i\gamma\kappa)d}{d} = 1 + \gamma\kappa\frac{n}{2},$$

because $G^i(r_A^i, r_N^i)$ has 2^{i+1} vertices by construction. Finally, the result holds because n is a power of 2. □

The previous lower bound together with the upper bound given by Corollary 1 implies that the PRA with respect to the set of instances on graphs with up to n vertices is *exactly* equal to $1+\gamma\kappa\lceil (n-1)/2\rceil$ when n is a power of 2. From there, the bound is tight infinitely often. Although for other values of n the bounds are not tight, they are close together, so these results provide an understanding of the asymptotic growth of the PRA. We now refine this observation to the bound in Corollary 1.

Theorem 3. *The upper bound for the price of risk aversion shown in Corollary 1 and the lower bound shown in Corollary 3 coincide for graphs of size that is a power of 2. Otherwise, the gap between them is less than 2.*

Proof. For an arbitrary $i > 1$, we consider the instance with $2^{i+1} = 2\eta$ vertices constructed in Corollary 3. In that instance, the only alternating path has exactly $2^i = \eta$ disjoint forward subpaths. Indeed, using Figure 4 as an example of the representation of the graph, we define an alternating path by (1) recursively choosing the lower component, (2) the reverse vertical edge, and (3) recursively choosing the upper component. By expanding both recursions, it is not hard to see that the alternating path covers all 2η vertices, and its η nonvertical edges are disjoint forward subpaths, as required. According to the equilibrium flows computed in Corollary 3, the nonvertical edges in the alternating paths belong to A, while the rest of the edges belong to B. Hence, the alternating path is compatible with the definitions of A and B, as required.

For graph sizes n that are not a power of 2, there is a rounding error. For the lower bound, we need to consider the maximum power of 2 smaller than n. The relative gap satisfies

$$\frac{UB}{LB} \leq \frac{1+\gamma\kappa\lceil (n-1)/2\rceil}{1+\gamma\kappa 2^{\lfloor \log_2(n)\rfloor}} < 2. \quad \square$$

Remark 1. In fact, a stronger version of Theorem 2 could be proved, which would then make the upper bound for the price of risk aversion shown in Corollary 1 and the lower bound shown in Corollary 3 coincide for all graph sizes (not just those of size that is a power of 2). In the proof of Theorem 2, we inductively used two copies of $G^{i-1}(r_A^{i-1}, r_N^{i-1})$ to create $G^i(r_A^i, r_N^i)$, a graph with 2^{i+1} vertices. If one wanted to create a graph of size $2^i < n < 2^{i+1}$ with properties for the RAWE and RNWE similar to those of the graphs of Theorem 2, then by an inductive construction that uses the Braess graph, the upper right edge could be replaced by graph $G^{i-1}(r_A^i, r_N^i)$ and the lower left edge could be replaced by an inductively constructed graph of size $n-2^i$ with similar properties to those of the graphs of Theorem 2.

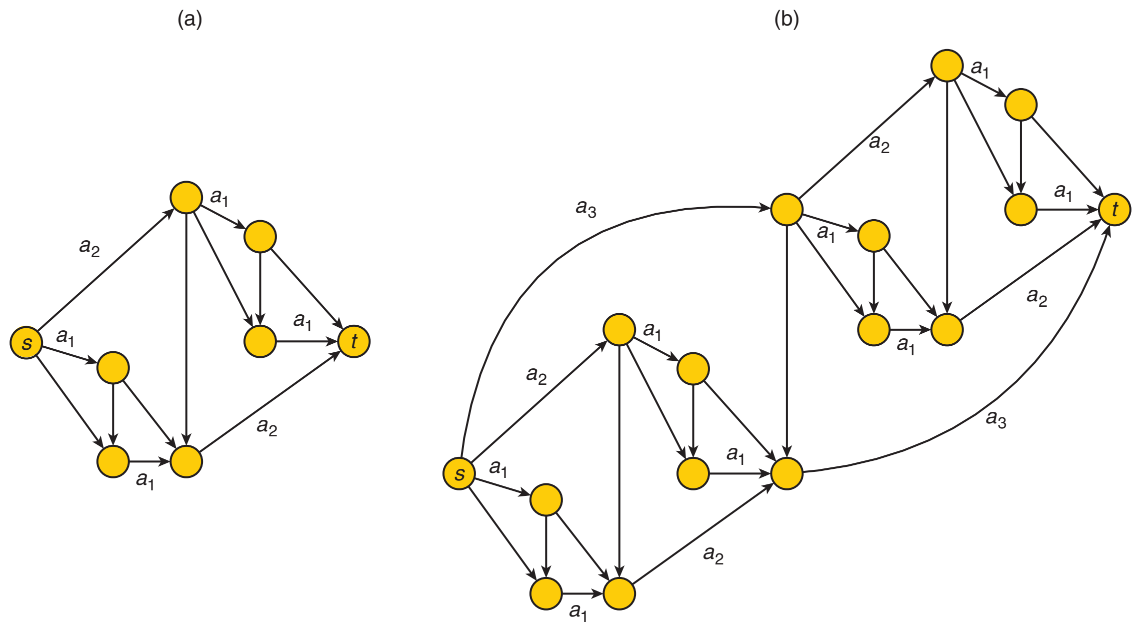
In conclusion, to match the bound of Theorem 1, $\text{PRA} = 1 + \eta\gamma\kappa$ when the family of instances is defined as graphs with arbitrary mean and variance functions that admit alternating paths with up to η disjoint forward subpaths, for η equal to a power of 2. We have equality because the supremum in the definition of PRA is attained by the instance constructed previously.

5. Functional Bounds

In this section, we turn our attention to instances with mean latency functions restricted to be in a certain family (as, e.g., affine functions). We prove upper and lower bounds for the PRA that are asymptotically tight as $\gamma\kappa$ increases. The results rely on the variational inequality approach that was first used by Correa et al. [15] to prove price of anarchy (POA) bounds for fixed families of functions. This approach was based on the properties of the allowed functions. Since then, these properties have been successively refined (Harks [25], Roughgarden and Schoppmann [55]), and they are now usually referred to as the *local smoothness* property. Although not really needed for the results here, we use the latter terminology since it has become standard by now. To characterize a family of mean latency functions, we rely on the smoothness property, defined below.

Definition 5 (Roughgarden and Schoppmann [55]). A function $l: \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is said to be *(λ, μ)-smooth around $x \in \mathbb{R}_{\geq 0}$* if $yl(x) \leq \lambda yl(y) + \mu xl(x)$ for all $y \in \mathbb{R}_{\geq 0}$.

Figure 4. (Color online) The resulting graphs G^i for $i=1$ on the left and $i=2$ on the right. Edges labeled with a_j have the mean latency function with equal name and 0 variance. Vertical edges have mean latency functions equal to 1 and variance equal to 0. Finally, the rest of the edges have mean latency functions equal to 1 and variance equal to κ.



Using the previous definition, we construct an upper bound for the PRA when mean latency functions $\{l_e\}_{e\in E}$ are $(1,\mu)$-smooth around the RAWE flow x_e for all edges $e\in E$. Meir and Parkes [35] prove a similar bound using a related approach in which they generalize the smoothness definition to *biased smoothness*, which holds with respect to a modified latency function. In our case, the modified latency function would be $l_e+\gamma v_e$. One advantage of our approach is its simplicity; it is a straightforward generalization of the POA proof given in Correa et al. [15]. We provide a proof corresponding to our assumptions, matching what is needed to get our asymptotically tight lower bounds.

Theorem 4. *Consider the set of general instances with mean latency functions $\{l_e\}_{e\in E}$ that are $(1,\mu)$-smooth around any RAWE flow x_e for all $e\in E$. Then, with respect to that set of instances,* $\mathrm{PRA}\le(1+\gamma\kappa)(1/(1-\mu))$.

Proof. We consider an instance within the family, a corresponding RAWE flow x, and an RNWE flow z. Furthermore, we let $A=\{e\in E\mid x_e\le z_e\}$ and $B=\{e\in E\mid z_e<x_e\}$. Using a variational inequality formulation for the RAWE (Nikolova and Stier-Moses [42]), we have that

$$\sum_{e\in E}x_e(l_e(x_e)+\gamma v_e(x_e))\le\sum_{e\in E}z_e(l_e(x_e)+\gamma v_e(x_e)).$$

In the next three steps, we partition the sum over E at both sides of the previous inequality into sums over A and B, we subtract the inequality

$$\sum_{e\in A}x_e\gamma v_e(x_e)+\sum_{e\in B}x_e\gamma v_e(x_e)\ge\sum_{e\in B}z_e\gamma v_e(x_e) \tag{3}$$

from it, and we further bound $v_e(x_e)$ by $\kappa l_e(x_e)$:

$$\sum_{e\in A}x_e(l_e(x_e)+\gamma v_e(x_e))+\sum_{e\in B}x_e(l_e(x_e)+\gamma v_e(x_e))\le\sum_{e\in A}z_e(l_e(x_e)+\gamma v_e(x_e))+\sum_{e\in B}z_e(l_e(x_e)+\gamma v_e(x_e))$$
$$\Downarrow$$
$$\sum_{e\in A}x_el_e(x_e)+\sum_{e\in B}x_el_e(x_e)\le\sum_{e\in A}z_e(l_e(x_e)+\gamma v_e(x_e))+\sum_{e\in B}z_el_e(x_e)$$
$$\Downarrow$$
$$C(x)\le\sum_{e\in A}(1+\gamma\kappa)z_el_e(x_e)+\sum_{e\in B}z_el_e(x_e).$$

Inequality (3) follows from the nonnegativeness of the flow and the variance, and from the definition of B. Applying the definition of A to the first term in the right-hand side of the last inequality and the $(1,\mu)$-smoothness condition to the second term, we upper bound the cost, and the result follows:

$$C(x) \le \sum_{e\in A}(1+\gamma\kappa)z_e l_e(z_e) + \sum_{e\in B}(z_e l_e(z_e) + \mu x_e l_e(x_e)) \le (1+\gamma\kappa)C(z) + \mu C(x). \quad \square$$

The bound in the previous result is similar to that for the POA for nonatomic games with no uncertainty. Indeed, the result there is that $\text{POA} \le (1-\mu)^{-1}$, the same without the $1+\gamma\kappa$ factor. The values of $(1-\mu)^{-1}$ have been computed for different families of functions in previous work. For example, it is equal to $4/3$ for affine latency functions and approximately equal to 1.626, 1.896, and 2.151 for quadratic, cubic, and quartic polynomial latency functions, respectively. For unrestricted functions this value is infinite so the bound becomes vacuous in that case, which provides support for the structural analysis of Section 4.

To evaluate the tightness of our upper bounds, we now propose lower bounds for the PRA. More specifically, we provide a family of instances indexed by i whose latency functions are $(1,\mu_i)$-smooth, for $\mu_i = 1-2^{-i}$, for which the bound is approximately tight. These instances imply lower bounds equal to $1+\gamma\kappa(1-\mu_i)^{-1} = 1+\gamma\kappa 2^i$. First, notice that although the lower and upper bounds do not match, they are similar. The difference is whether the 1 is or is not multiplied by the μ factor. When the $\gamma\kappa$ term is large, both bounds are essentially equal. Second, notice that for large values of i, by necessity, the number of alternations of the longest alternating path must grow exponentially large to simultaneously match the structural upper bound presented in Theorem 1.

Theorem 5. *For any* $i > 0$, *letting* $\mu_i = 1-2^{-i}$, $\text{PRA} \ge 1+\gamma\kappa(1-\mu_i)^{-1}$ *for the family of instances satisfying the* $(1,\mu_i)$-*smoothness property.*

To get the result, we use the recursive construction of Theorem 2 but with cost functions satisfying the $(1,\mu_i)$-smoothness condition around the RAWE. For the given i, we construct a graph G^i that implies that $\text{PRA} \ge 1+\gamma\kappa 2^i = 1+\gamma\kappa(1-\mu_i)^{-1}$. A brief road map of the proof is as follows. We specify the instance and determine the RNWE flow z and the RAWE flow x with their costs. From there, we conclude that $\text{PRA} \ge 1+\gamma\kappa(1-\mu_i)^{-1}$. Finally, we prove that l_e is $(1,\mu_i)$-smooth around x_e for all e.

Proof. We consider the graph G^i constructed in Theorem 2 with $r_A^i = r_N^i = 1$ (see Figures 2 and 3) but with alternative functions $a_j(\cdot)$. For the functions defined at level $1 \le j \le i$, we set $a_j(x) = 0$ for $x \le 2^{j-1}/2^i$, whereas for larger x, $a_j(x)$ increases linearly to attain $2^{j-1}\gamma\kappa$ at $x = 2^{j-1}/2^i - 1$. Mathematically,

$$a_j(x) = \max\left\{0, \frac{2^{j-1}\gamma\kappa}{2^{j-1}/2^i - 1 - 2^{j-1}/2^i}\left(x - \frac{2^{j-1}}{2^i}\right)\right\}.$$

To simplify notation, we refer to edges that have cost function $a_j(\cdot)$ as a_j. Figure 4 illustrates the construction for $i = 2$ and $i = 3$. As an example, we specify the resulting mean latency functions for $i = 2$: a_1 is such that $a_1(\frac{1}{4}) = 0$ and $a_1(\frac{1}{3}) = \gamma\kappa$, and a_2 satisfies $a_2(\frac{1}{2}) = 0$ and $a_2(\frac{2}{3}) = 2\gamma\kappa$. The RNWE flow splits the unit flow equally along the four paths not containing any vertical edge. Instead, the RAWE flow splits the unit flow equally along the three paths that contain a vertical edge. Evaluating those functions, $a_j(z) = 0$ and $a_j(x) = 2^{j-1}\gamma\kappa$, from where $\text{PRA} \ge C(x)/C(z) = (1+4\gamma\kappa)/1 = 1+4\gamma\kappa$ for the family of functions that are $(1, 3/4)$-smooth.

We refer to paths in G_i not containing any vertical edge in representations such as that of Figure 4 as *parallel paths*. The rest of the paths, containing a single vertical edge, are referred to as *zigzag paths*. It is not hard to see using an inductive proof on the construction of G^i that there are 2^i parallel paths and there are $2^i - 1$ zigzag paths.

Generalizing what we saw in the example for $i = 2$, the RNWE flow z splits the unit flow equally along the 2^i parallel paths. To verify that z is at equilibrium, observe that for each a_j edge, for $1 \le j \le i$, there are 2^{j-1} parallel paths passing through it. Consequently, each a_j will get $2^{j-1}/2^i$ units of flow, implying that their costs are 0. Path costs under z are thus the same as those in Theorem 2, which implies that z is indeed an RNWE and that $C(z) = 1$.

Furthermore, the RAWE flow x splits the unit flow equally along the $2^i - 1$ zigzag paths. To verify that x is at equilibrium, observe that for each a_j edge, for $1 \le j \le i$, there are 2^{j-1} zigzag paths passing through it. Consequently, each a_j will get $2^{j-1}/(2^i - 1)$ units of flow, implying that their costs are $2^{j-1}\gamma\kappa$. Path costs under x are thus the same as those in Theorem 2, which implies that x is indeed the RAWE and that $C(x) = 1 + 2^i\gamma\kappa$. From there, the bound for PRA in the statement of the theorem follows.

What remains to be shown is that for any chosen G^i, functions a_j are $(1,\mu_i)$-smooth around x_j, where $x_j = 2^{j-1}/(2^i - 1)$ is the RAWE flow at edge a_j. The other cost functions are constant so they trivially satisfy the smoothness properties. To prove this, let us consider $1 \le j \le i$. From the definition of smoothness, we need to show that $ya_j(x_j) \le ya_j(y) + \mu_i x_j a_j(x_j)$ for all $y \in \mathbb{R}_{\ge 0}$. Equivalently, we can show that

$$1 - 2^{-i} = \mu_i \ge \frac{\max_{y \in \mathbb{R}_{\ge 0}} y(a_j(x_j) - a_j(y))}{x_j a_j(x_j)}.$$

First, note that the maximum is attained in the interval $[2^{j-1}/2^i, x_j]$ since $a_j(y) = 0$ to the left of the interval, and the argument of the maximum becomes negative to the right. Since a_j is linear in that interval, we solve the maximum problem by extending it linearly to the whole domain. Since additive constants are irrelevant because we maximize the difference of a_j evaluated in two points, we modify the linearized function and add a constant so it evaluates to 0 at 0. For linear functions that cross the origin, the maximizer of the problem is $x_j/2$ (see, e.g., Correa et al. [15]). Because $x_j/2$ is to the left of the interval where the maximizer must be, the maximizer with respect to a_j is $y^* = 2^{j-1}/2^i$, and $a_j(y^*) = 0$. From there,

$$\frac{\max_{y \in \mathbb{R}_{\ge 0}} y(a_j(x_j) - a_j(y))}{x_j a_j(x_j)} = \frac{y^*}{x_j} = \frac{2^{j-1}/2^i}{2^{j-1}/(2^i - 1)} = \frac{2^i - 1}{2^i} = \mu_i. \quad \square$$

6. The Mean-Stdev Model

In this section, we turn to the mean-standard deviation model and prove upper and lower structural bounds on PRA, now assuming that κ is the maximum coefficient of variation $CV_e(f_e)$ among edges, where CV is defined as the ratio between the standard deviation and the mean. The lower bounds follow from the instances used to prove the lower bounds in the mean-variance case in Section 4. Considering families of graphs with up to τ forward disjoint subpaths and general mean latency and standard deviation functions, we prove upper bounds for the cases $\tau = \{1, 2\}$ and lower bounds for arbitrary τ. For $\tau \le 2$, both bounds coincide, so the PRA for the standard deviation case gets characterized exactly.

Since we are now dealing with standard deviations, we redefine $Q_p^\gamma(f) = l_p(f) + \gamma\sigma_p(f)$, where $\sigma_p(f) = (\sum_{e \in p} \sigma_e^2(f_e))^{1/2}$. Given an instance based on a graph G, we refer to an RNWE flow by z and to an RAWE flow by y. Furthermore, we denote the path cost perceived by players at the RAWE y by $Q^\gamma(y)$ and the expected latencies perceived by players at the RNWE z by $Q^0(z)$. The definition for the price of risk aversion is analogous to that given in Section 2.

Our first result provides inequalities that relate the social cost at equilibrium with the perceived utilities. The first part is known and the second is a generalization.

Proposition 1. *For an arbitrary instance, $C(z) = d \cdot Q^0(z)$ and $C(y) \le d \cdot Q^\gamma(y)$, where d is the traffic demand.*

Proof Using that at equilibria all used paths have equal path costs, we get

1. $C(z) = \sum_{p \in \mathcal{P}} z_p l_p(z) = d \cdot Q^0(z)$,
2. $C(y) = \sum_{p \in \mathcal{P}} y_p l_p(y) \le \sum_{p \in \mathcal{P}} y_p(l_p(y) + \gamma\sigma_p(y)) = d \cdot Q^\gamma(y)$. $\square$

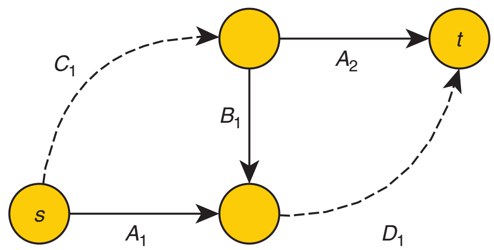
As in previous sections, we partition the edge set E into two: $A = \{e \in E\colon y_e \le z_e\}$ and $B = \{e \in E\colon z_e < y_e\}$. We assume that all edges in E are used by either flow y or z, which is without loss of generality because unused edges can be deleted without any consequence. The definition of B implies that $y_e > 0$ for all $e \in B$, while the assumption implies that $z_e > 0$ for all $e \in A$. To prove an upper bound on PRA, we rely again on alternating paths, the existence of which is guaranteed by Lemma 2. The next result bounds the PRA for graphs that admit simple alternating paths (i.e., actual s–t paths) or alternating paths with a single alternation.

Lemma 5. *For $\tau \in \{1, 2\}$, considering the set of instances on general topologies with arbitrary mean latency and standard deviation functions that admit alternating paths with τ forward subpaths and at most one backward edge,* PRA $\le 1 + \tau\gamma\kappa$.

Proof. For $\tau = 1$, we let π be an alternating path consisting of just edges in A (i.e., it is an actual s–t path). Let p be any used path under the RAWE y. Using the equilibrium conditions, the relationship between 1-norms and 2-norms, and the coefficient of variation bound κ,

$$\begin{aligned} Q^\gamma(y) &= \sum_{e \in p} l_e(y) + \gamma\sqrt{\sum_{e \in p} \sigma_e^2(y)} \le \sum_{e \in \pi} l_e(y) + \gamma\sqrt{\sum_{e \in \pi} \sigma_e^2(y)} \\ &\le \sum_{e \in \pi} l_e(y) + \gamma\sum_{e \in \pi} \sigma_e(y) \le (1 + \gamma\kappa)\sum_{e \in \pi} l_e(y). \end{aligned} \tag{4}$$

Figure 5. (Color online) A subgraph of a graph admitting an alternating path with a single alternation. Since the edges of the alternating path receive flow, there must exist paths C_1 and D_1 that bring flow from the source to the edges in A_2 or B_1 and take flow from A_1 or B_1 and deliver it to the sink, respectively.



Using the monotonicity of the mean latency functions, $l_e(y) \le l_e(z)$ for $e \in \pi$, from where $Q^\gamma(y) \le (1+\gamma\kappa) \cdot \sum_{e\in\pi} l_e(z)$. Since $z_e > 0$ for all $e \in \pi$ because of the property stated earlier, a flow decomposition can be found where π is used under the RNWE z, implying that $C(z) = d \cdot \sum_{e\in\pi} l_e(z)$. Finally, using Proposition 1, $C(y) \le d \cdot Q^\gamma(y) \le d \cdot (1+\gamma\kappa)\sum_{e\in\pi} l_e(z) = (1+\gamma\kappa)C(z)$, from where we get the bound on PRA.

For $\tau = 2$, we let $\pi = A_1$–B_1–A_2 be the alternating path with two disjoint forward subpaths, A_1 and A_2, and reverse subpath B_1, where these subpaths belong to A or B correspondingly. Figure 5 illustrates the topology of these subpaths. Consider an RAWE flow y and a flow-carrying path C_1–B_1–D_1 under y. Such a path must exist because there is only one edge in B_1 and it carries flow under y, as it was mentioned earlier. Using the equilibrium conditions for y,

$$l_{C_1}(y) + l_{B_1}(y) + l_{D_1}(y) + \gamma\sqrt{\sigma^2_{C_1}(y) + \sigma^2_{B_1}(y) + \sigma^2_{D_1}(y)} \le l_{A_1}(y) + l_{D_1}(y) + \gamma\sqrt{\sigma^2_{A_1}(y) + \sigma^2_{D_1}(y)}. \tag{5}$$

Let us first assume that $\sigma^2_{C_1}(y) + \sigma^2_{B_1}(y) + \sigma^2_{D_1}(y) \le \sigma^2_{A_1}(y) + \sigma^2_{D_1}(y)$. For $\alpha \le \beta$ and $\delta \ge 0$, it can be proved that $\sqrt{\beta+\delta} - \sqrt{\alpha+\delta} \le \sqrt{\beta} - \sqrt{\alpha}$. Letting $\alpha = \sigma^2_{C_1}(y) + \sigma^2_{B_1}(y)$, $\beta = \sigma^2_{A_1}(y)$, and $\delta = \sigma^2_{D_1}(y)$, (5) implies that

$$l_{C_1}(y) + l_{B_1}(y) + \gamma\sqrt{\sigma^2_{C_1}(y) + \sigma^2_{B_1}(y)} \le l_{A_1}(y) + \gamma\sqrt{\sigma^2_{A_1}(y)} \le (1+\gamma\kappa)l_{A_1}(y). \tag{6}$$

If $\sigma^2_{C_1}(y) + \sigma^2_{B_1}(y) + \sigma^2_{D_1}(y) > \sigma^2_{A_1}(y) + \sigma^2_{D_1}(y)$, (5) implies that $l_{C_1}(y) + l_{B_1}(y) \le l_{A_1}(y)$, from where

$$l_{C_1}(y) + l_{B_1}(y) + \gamma\sqrt{\sigma^2_{C_1}(y) + \sigma^2_{B_1}(y)} \le (1+\gamma\kappa)(l_{C_1}(y) + l_{B_1}(y)) \le (1+\gamma\kappa)l_{A_1}(y). \tag{7}$$

Hence, in both cases, the same inequality holds.

Using a similar argument for the path C_1–A_2 instead of A_1–D_1, we get $l_{D_1}(y) + l_{B_1}(y) + \gamma(\sigma^2_{D_1}(y) + \sigma^2_{B_1}(y))^{1/2} \le (1+\gamma\kappa)l_{A_2}(y)$. From the monotonicity of the square root, we further derive that

$$l_{D_1}(y) + \gamma\sigma_{D_1}(y) \le l_{D_1}(y) + \gamma\sqrt{\sigma^2_{D_1}(y) + \sigma^2_{B_1}(y)} \le (1+\gamma\kappa)l_{A_2}(y) - l_{B_1}(y). \tag{8}$$

Since the path is used under y, $Q^\gamma(y) = l_{C_1}(y) + l_{B_1}(y) + l_{D_1}(y) + \gamma(\sigma^2_{C_1}(y) + \sigma^2_{B_1}(y) + \sigma^2_{D_1}(y))^{1/2}$. Using again the norm-1, norm-2 inequality, together with (6) (or (7)) and (8), and the definitions of A and B, we can upper bound the previous expression with

$$\begin{aligned} Q^\gamma(y) &= l_{C_1}(y) + l_{B_1}(y) + l_{D_1}(y) + \gamma\sqrt{\sigma^2_{C_1}(y) + \sigma^2_{B_1}(y) + \sigma^2_{D_1}(y)} \\ &\le l_{C_1}(y) + l_{B_1}(y) + \gamma\sqrt{\sigma^2_{C_1}(y) + \sigma^2_{B_1}(y)} + l_{D_1}(y) + \gamma\sigma_{D_1}(y) \\ &\le (1+\gamma\kappa)l_{A_1}(y) + (1+\gamma\kappa)l_{A_2}(y) - l_{B_1}(y) \\ &\le (1+\gamma\kappa)(l_{A_1}(z) + l_{A_2}(z)) - l_{B_1}(z). \end{aligned} \tag{9}$$

Now, we derive a related bound for the RNWE flow z. For a potentially different D_1, the path A_1–D_1 must carry flow under z. Such a path must exist because $z_e > 0$ for $e \in A_1$. Furthermore, by the equilibrium conditions for z applied to path A_2, which also carries flow under z, $l_{A_2}(z) \le l_{B_1}(z) + l_{D_1}(z)$. Putting both remarks together, $Q^0(z) = l_{A_1}(z) + l_{D_1}(z) \ge l_{A_1}(z) + l_{A_2}(z) - l_{B_1}(z)$. Combining the last inequality with (9), and the fact that $Q^0(z)$ is an upper bound for both $l_{A_1}(z)$ and $l_{A_2}(z)$, we get

$$\begin{aligned} Q^\gamma(y) &\le (1+\gamma\kappa)(l_{A_1}(z) + l_{A_2}(z)) - l_{B_1}(z) \\ &\le l_{A_1}(z) + l_{A_2}(z) - l_{B_1}(z) + \gamma\kappa(l_{A_1}(z) + l_{A_2}(z)) \\ &\le Q^0(z) + 2\gamma\kappa Q^0(z) \le (1+2\gamma\kappa)Q^0(z). \end{aligned}$$

That implies the result since Proposition 1 yields $C(y) \le dQ^\gamma(y) \le d(1+2\gamma\kappa)Q^0(z) = (1+2\gamma\kappa)C(z)$. □

Lemma 5 implies that $\mathrm{PRA} \le 1 + \tau\gamma\kappa$ is valid for the set of instances admitting alternating paths in which the number of disjoint forward subpaths is not more than $\tau = 1, 2$, and the alternating path contains at most one backward edge. Although this does not fully generalize Theorem 1 to the standard deviations case, it is a first step in that direction.

To provide matching lower bounds for the results of this section, we note that all paths in the proof of Theorem 2 have at most one edge with nonzero variance. Thus, all the lower bounds in Section 4 work by reinterpreting the variances as standard deviations and the variance-to-mean ratios as coefficients of variation. This is summarized below.

Corollary 4. *The upper bounds* $\mathrm{PRA} \le 1 + \gamma\kappa$ *and* $\mathrm{PRA} \le 1 + 2\gamma\kappa$ *corresponding to graphs that admit alternating paths with one forward path, or two disjoint forward subpaths and at most one backward edge, respectively, are tight.*

7. Conclusion

We have considered the effect of risk-averse players on selfish routing with stochastic travel times, captured by mean and variance functions of flow, following the mean-var and mean-stdev risk models in Nikolova and Stier-Moses [43].

Our main conceptual contribution is a new perspective and understanding of efficiency loss as a result of risk-averse behavior in terms of *structural* versus *functional* measures, the first one depending on the topology of the network and independent of the expected latency functions and the second depending on the class of allowed latency functions and independent of the network topology, similar to previous price of anarchy analysis.

Our main technical contributions are (i) establishing an upper bound on the ratio of the cost of the risk-averse equilibrium to that of the risk-neutral one, for users who aim to minimize the mean variance of their route in a general network and (ii) the inductive construction of a family of graphs that can be adapted with appropriate mean and variance functions to yield both structural and functional lower bounds. We also show how to generalize the previous price of anarchy analysis for deterministic congestion games based on variational inequalities (Correa et al. [15]) to provide a functional upper bound here. Our results may, in turn, inspire a reinvestigation of the classic price of anarchy results in deterministic settings through the lens of the structural analysis.

We leave open whether there is a deeper connection between the alternating paths in our upper bound on the price of risk aversion and those in the proof on the severity of the Braess paradox (Roughgarden [54]), as well as the alternating (negative) cycles in the study of tolls of Bonifaci et al. [5].

In applications, it is reasonable to expect correlated costs, although this leads to models with nonseparable path costs, which are difficult to handle (similar to the mean-stdev model). Studying correlations is an interesting open direction. We remark that finding stochastic shortest paths can extend to locally correlated edge delays via a polynomial graph transformation, as mentioned in Nikolova and Stier-Moses [42].

Some other immediate open questions include whether the general upper bound can be extended to mean-stdev and other risk objectives and whether the bounds or analysis can be extended to heterogeneous risk profiles and multiple origin–destination pairs. Kleer and Schäfer [28] have recently shed light on extending the results for games with homogeneous users and multiple origin–destination pairs, although the case of heterogeneous risk profiles is left unhandled even in the single-commodity case. The approach for the upper bounds fails in that case because we do not know the risk-aversion factors of the players using each of the C_i–B_i–D_i paths in the alternating path construction, and consequently, we cannot take advantage of the equilibrium conditions to express the RAWE and RNWE costs through the costs of the edges of the alternating path.

A key challenge is to develop a better understanding and technical approaches to nonadditive risk models, such as the mean-stdev model, which have so far resisted fully general upper-bound analysis for arbitrary graphs—a step in that direction is our fourth technical contribution on the mean-stdev model for a family of graphs that contain and generalize series-parallel graphs and the Braess graphs. Moreover, although the construction used for the various lower bounds can also provide functional lower bounds for the standard deviation case, functional upper bounds for the standard deviation case so far remain elusive and will be the subject of future research.

Another interesting direction is to characterize how much risk can help rather than hurt the quality of equilibrium and social welfare. For a different model of uncertainty and user objectives, Meir and Parkes [34] show that moderate uncertainty can improve the social welfare. More closely related to our model, Fotakis et al. [23] characterize when a socially optimal flow can be enforced as an equilibrium of risk-averse users. This is also related to the question of what flow can be enforced as an equilibrium under restricted tolls (Bonifaci et al. [5]). In light of these recent results and our current work, it would be especially interesting to understand when risk-averse attitudes can be leveraged in mechanism design instead of tolls to reduce congestion.

Acknowledgments

The authors thank Roberto Cominetti and José Correa for excellent discussions and comments about the material in this paper. The authors are also indebted to an anonymous referee who pointed out a problem in one of the proofs for the standard deviation objective. Finally, the authors also thank the editorial team for insightful remarks and questions that allowed them to improve their results and the presentation of the paper.

Endnotes

[1] The risk factor VMR, which is the variance version of the coefficient of variation, is prevalent in science and goes under several different names, such as index of dispersion, dispersion index, coefficient of dispersion, relative variance, and Fano factor. This factor is often used in a number of fields, including finance, statistics, particle physics, and neuroscience.

[2] The latter leads to an interesting open problem, posed by Nikolova and Stier-Moses [42]: Is there an efficient algorithm that converts a given equilibrium edge-flow vector into an equilibrium path-flow decomposition? That reference shows that a succinct path-flow decomposition that uses polynomially many paths exists.

[3] The actual inequality, without using the shorthand expression, is $\sum_{e\in C_k}(l_e(x_e)+\gamma v_e(x_e))+\sum_{e\in B_k}(l_e(x_e)+\gamma v_e(x_e))\le \sum_{e\in C_{k-1}}(l_e(x_e)+\gamma v_e(x_e))+\sum_{e\in A_k}(l_e(x_e)+\gamma v_e(x_e))$.

References

[1] Altman E, Boulogne T, El-Azouzi R, Jiménez T, Wynter L (2006) A survey on networking games in telecommunications. *Comput. Oper. Res.* 33(2):286–311.
[2] Angelidakis H, Fotakis D, Lianeas T (2013) Stochastic congestion games with risk-averse players. Vöcking B, ed. *Algorithmic Game Theory—Proc. 6th Internat. Sympos.*, Lecture Notes in Computer Science, Vol. 8146 (Springer, Berlin), 86–97.
[3] Balcan MF, Blum A, Mansour Y (2013) The price of uncertainty. *ACM Trans. Econ. Comput.* 1(3):article 15; http://dx.doi.org/10.1145/2509413.2509415.
[4] Beckmann MJ, McGuire CB, Winsten CB (1956) *Studies in the Economics of Transportation* (Yale University Press, New Haven, CT).
[5] Bonifaci V, Salek M, Schafer G (2011) Efficiency of restricted tolls in non-atomic network routing games. Persiano G, ed. *Algorithmic Game Theory—Proc. 4th Internat. Sympos.*, Lecture Notes in Computer Science, Vol. 6982 (Springer, Berlin), 302–313.
[6] Boyles SD, Waller ST (2010) A mean-variance model for the minimum cost flow problem with stochastic arc costs. *Networks* 56(3): 215–227.
[7] Boyles SD, Kockelman KM, Waller ST (2010) Congestion pricing under operational, supply-side uncertainty. *Transportation Res. Part C: Emerging Tech.* 18(4):519–535.
[8] Chau CK, Sim KM (2003) The price of anarchy for non-atomic congestion games with symmetric cost maps and elastic demands. *Oper. Res. Lett.* 31(5):327–334.
[9] Chen PA, Kempe D (2008) Altruism, selfishness, and spite in traffic routing. *Proc. 9th ACM Conf. Electronic Commerce* (ACM Press, New York), 140–149, http://dx.doi.org/10.1145/1386790.1386816.
[10] Cole R, Dodis Y, Roughgarden T (2006) How much can taxes help selfish routing? *J. Comput. System Sci.* 72(3):444–467.
[11] Cominetti R (2015) Equilibrium routing under uncertainty. *Math. Programming* 151(1):117–151.
[12] Cominetti R, Torrico A (2016) Additive consistency of risk measures and its application to risk-averse routing in networks. *Math. Oper. Res.* 41(4):1510–1521.
[13] Correa JR, Stier-Moses NE (2011) Wardrop equilibria. Cochran JJ, ed. *Encyclopedia of Operations Research and Management Science* (John Wiley & Sons, Hoboken, NJ).
[14] Correa JR, Schulz AS, Stier-Moses NE (2004) Selfish routing in capacitated networks. *Math. Oper. Res.* 29(4):961–976.
[15] Correa JR, Schulz AS, Stier-Moses NE (2008) A geometric approach to the price of anarchy in nonatomic congestion games. *Games Econom. Behav.* 64(2):457–469.
[16] Dughmi S (2014) On the hardness of signaling. *Proc. IEEE 55nd Annual Sympos. Foundations Comput. Sci.* (IEEE Computer Society, Washington, DC), 354–363. http://dx.doi.org/10.1109/FOCS.2014.45.
[17] Dughmi S, Peres Y (2012) Mechanisms for risk averse agents, without loss. *Workshop Risk Aversion Algorithmic Game Theory Mechanism Design, Valencia, Spain*.
[18] Epstein A, Feldman M, Mansour Y (2009) Efficient graph topologies in network routing games. *Games Econom. Behav.* 66(1):115–125.
[19] Federal Highway Administration (2017) Traffic congestion and reliability: Trends and advanced strategies for congestion mitigation. Final report, U.S. Department of Transportation, Washington, DC. https://ops.fhwa.dot.gov/congestion_report/executive_summary.htm.
[20] Fiat A, Papadimitriou C (2010) When the players are not expectation maximizers. Kontogiannis S, Koutsoupias E, Spirakis PG, eds. *Algorithmic Game Theory—Proc. 3rd Internat. Sympos.*, Lecture Notes in Computer Science, Vol. 6386 (Springer, Berlin), 1–14.
[21] Fotakis D (2010) Congestion games with linearly independent paths: Convergence time and price of anarchy. *Theory Comput. Systems* 47(1):113–136.
[22] Fotakis D, Spirakis PG (2008) Cost-balancing tolls for atomic network congestion games. *Internet Math.* 5(4):343–363.
[23] Fotakis D, Kalimeris D, Lianeas T (2015) Improving selfish routing for risk-averse players. Markakis E, Schäfer G, eds. *Web Internet Econom.—Proc. 11th Internat. Conf.*, Lecture Notes in Computer Science, Vol. 9470 (Springer, Berlin), 328–342.
[24] Fu H, Hartline J, Hoy D (2013) Prior-independent auctions for risk-averse agents. *Proc. 14th ACM Conf. Electronic Commerce* (ACM Press, New York), 471–488.
[25] Harks T (2011) Stackelberg strategies and collusion in network games with splittable flow. *Theory Comput. Systems* 48(4):781–802.
[26] Karakostas G, Kolliopoulos SG (2005) The efficiency of optimal taxes. López-Ortiz A, Hamel A, eds. *Combinatorial Algorithmic Aspects Networking* (*CAAN'04*), Lecture Notes in Computer Science, Vol. 3405 (Springer, Berlin), 3–12.
[27] Khani A, Boyles SD (2015) An exact algorithm for the mean-standard deviation shortest path problem. *Transportation Res. Part B* 81(1):252–266.

[28] Kleer P, Schäfer G (2016) The impact of worst-case deviations in non-atomic network routing games. Gairing M, Savani R, eds. *Algorithmic Game Theory—Proc. 9th Internat. Sympos.*, Lecture Notes in Computer Science, Vol. 9928 (Springer, Berlin), 129–140.
[29] Koutsoupias E, Papadimitriou CH (2009) Worst-case equilibria. *Comput. Sci. Rev.* 3(2):65–69.
[30] Krokhmal P, Zabarankin M, Uryasev S (2011) Modeling and optimization of risk. *Surveys Oper. Res. Management Sci.* 16(2):49–66.
[31] Li J, Deshpande A (2011) Maximizing expected utility for stochastic combinatorial optimization problems. *Proc. IEEE 52nd Annual Sympos. Foundations Comput. Sci.* (IEEE Computer Society, Washington, DC), 797–806.
[32] Lianeas T, Nikolova E, Stier-Moses NE (2016) Asymptotically tight bounds for inefficiency in risk-averse selfish routing. Kambhampati S, ed. *Proc. 25th Internat. Joint Conf. Artificial Intelligence* (*IJCAI 2016*) (IJCAI/AAAI Press, Menlo Park, CA), 338–344.
[33] Markowitz H (1952) Portfolio selection. *J. Finance* 7(1):77–98.
[34] Meir R, Parkes D (2015) Congestion games with distance-based strict uncertainty. *Proc. 29th AAAI Conf. Artificial Intelligence* (*AAAI'*15) (AAAI Press, Menlo Park, CA), 986–992.
[35] Meir R, Parkes D (2015) Playing the wrong game: Smoothness bounds for congestion games with behavioral biases. *ACM SIGMETRICS Perform. Eval. Rev.* 43(3):67–70.
[36] Meir R, Parkes DC (2016) When are marginal congestion tolls optimal? Presentation, 9th International Workshop on Agents in Traffic and Transportation (ATT 2016), July 10, CEUR Workshop, New York.
[37] Milchtaich I (2016) Internalization of social cost in congestion games. Working paper, Bar-Ilan University, Ramat Gan, Israel. https://faculty.biu.ac.il/~milchti/papers/continuum.pdf.
[38] Neumann JV, Morgenstern O (1944) *Theory of Games and Economic Behavior* (Princeton University Press, Princeton, NJ).
[39] Nie Y (2011) Multi-class percentile user equilibrium with flow-dependent stochasticity. *Transportation Res. Part B* 45(10):1641–1659.
[40] Nikolova E (2010) Approximation algorithms for reliable stochastic combinatorial optimization. Serna M, Shaltiel R, Jansen K, Rolim J, eds. *Approximation, Randomization, Combinatorial Optimization: Algorithms Techniques—Proc. 13th Internat. Conf. Approximation* (*APPROX/RANDOM'10*), Lecutre Notes in Computer Science, Vol. 6302 (Springer, Berlin), 338–351.
[41] Nikolova E, Stier-Moses NE (2011) Stochastic selfish routing. Persiano G, ed. *Algorithmic Game Theory—Proc. 4th Internat. Sympos.*, Lecture Notes in Computer Science, Vol. 6982 (Springer, Berlin), 314–325.
[42] Nikolova E, Stier-Moses NE (2014) A mean-risk model for the traffic assignment problem with stochastic travel times. *Oper. Res.* 62(2):366–382.
[43] Nikolova E, Stier-Moses NE (2015) The burden of risk aversion in mean-risk selfish routing. Roughgarden T, Feldman M, Schwarz M, eds. *Proc. 16th ACM Conf. Econom. Comput.* (ACM Press, New York), 489–506.
[44] Nikolova E, Brand M, Karger DR (2006) Optimal route planning under uncertainty. Long D, Smith SF, Borrajo D, McCluskey L, eds. *Proc. 16th Internat. Conf. Automated Planning Scheduling* (*ICAPS*) (AAAI Press, Menlo Park, CA), 131–141.
[45] Nikolova E, Kelner JA, Brand M, Mitzenmacher M (2006) Stochastic shortest paths via quasi-convex maximization. *Algorithms—ESA 2006*, Lecture Notes in Computer Science, Vol. 4168 (Springer-Verlag, Berlin), 552–563.
[46] Nisan N, Roughgarden T, Tardos É, Vazirani VV (2007) *Algorithmic Game Theory* (Cambridge University Press, Cambridge, UK).
[47] Ordóñez F, Stier-Moses NE (2010) Wardrop equilibria with risk-averse users. *Transportation Sci.* 44(1):63–86.
[48] Papadimitriou CH (2001) Algorithms, games, and the Internet. *Proc. 33rd Annual ACM Sympos. Theory Comput.* (*STOC*) (ACM Press, New York), 749–753.
[49] Perakis G (2007) The "price of anarchy" under nonlinear and asymmetric costs. *Math. Oper. Res.* 32(3):614–628.
[50] Pigou AC (1920) *The Economics of Welfare* (Macmillan, London).
[51] Piliouras G, Nikolova E, Shamma JS (2013) Risk sensitivity of price of anarchy under uncertainty. *Proc. 14th ACM Conf. Electronic Commerce* (ACM Press, New York), 715–732.
[52] Rockafellar RT (2007) Coherent approaches to risk in optimization under uncertainty. Klastorin T, ed. *OR Tools and Applications: Glimpses of Future Technologies*, Tutorials in Operations Research (INFORMS, Hanover, MD), 38–61.
[53] Roughgarden T (2003) The price of anarchy is independent of the network topology. *J. Comput. System Sci.* 67(2):341–364.
[54] Roughgarden T (2006) On the severity of Braess's paradox: Designing networks for selfish users is hard. *J. Comput. System Sci.* 72(5): 922–953.
[55] Roughgarden T, Schoppmann F (2015) Local smoothness and the price of anarchy in splittable congestion games. *J. Econom. Theory* 156(March):317–342.
[56] Roughgarden T, Tardos É (2002) How bad is selfish routing? *J. ACM* 49(2):236–259.
[57] Sheffi Y (1985) *Urban Transportation Networks* (Prentice-Hall, Englewood, NJ).
[58] Swamy C (2011) Risk-averse stochastic optimization: Probabilistically-constrained models and algorithms for black-box distributions. *Proc. 22nd Annual ACM-SIAM Sympos. Discrete Algorithms* (SIAM, Philadelphia), 1627–1646.
[59] Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. *Science* 211(4481):453–458.
[60] Valdes J, Tarjan RE, Lawler EL (1979) The recognition of series parallel digraphs. *Proc. 11th Annual ACM Sympos. Theory Comput.* (*STOC*) (ACM Press, New York), 1–12.
[61] Vicroads (2014) Traffic monitor 2012–13. Last accessed October 2017, https://www.vicroads.vic.gov.au/~/media/files/documents/traffic%20and%20road%20use/trafficmonitorreport20122013.ashx.
[62] Wardrop JG (1952) Some theoretical aspects of road traffic research. *Proc. Institute of Civil Engrg.*, Part II, Vol. 1, 325–378.