Mathematical Programming
https://doi.org/10.1007/s10107-022-01853-0

FULL LENGTH PAPER

Series A

A convergence analysis of the price of anarchy in atomic congestion games

Zijun Wu[1] · Rolf H. Möhring[1,2] · Chunying Ren[3] · Dachuan Xu[3]

Received: 10 November 2020 / Accepted: 15 June 2022

Abstract

We analyze the convergence of the price of anarchy (PoA) of Nash equilibria in atomic congestion games with growing total demand T. When the cost functions are polynomials of the same degree, we obtain explicit rates for a rapid convergence of the PoAs of pure and mixed Nash equilibria to 1 in terms of $1/T$ and d_{max}/T, where d_{max} is the *maximum* demand controlled by an individual. Similar convergence results carry over to the random inefficiency of the random flow induced by an arbitrary mixed Nash equilibrium. For arbitrary polynomial cost functions, we derive a related convergence rate for the PoA of pure Nash equilibria (if they exist) when the demands fulfill certain regularity conditions and d_{max} is bounded as $T \to \infty$. In this general case, also the PoA of mixed Nash equilibria converges to 1 as $T \to \infty$ when d_{max} is bounded. Our results constitute the first convergence analysis for the PoA in atomic congestion games and show that selfish behavior is well justified when the total demand is large.

Keywords Atomic congestion games · Pure and mixed Nash equilibria · Price of anarchy · Inefficiency of equilibria

✉ Rolf H. Möhring
rolf.moehring@tu-berlin.de

Zijun Wu
wuzj@hfuu.edu.cn

Chunying Ren
renchunying@emails.bjut.edu.cn

Dachuan Xu
xudc@bjut.edu.cn

1 Present Address: Department of Artificial Intelligence and Big Data, Hefei University, Jinxiu 99, Hefei 230601, Anhui, China

2 Kombinatorische Optimierung und Graphenalgorithmen (COGA), Fakultät II–Mathematik und Naturwissenschaften, Institut für Mathematik, Sekr. MA 5–1, Technische Universität Berlin, Strasse des 17. Juni 136, Berlin 10623, Germany

3 Department of Operations Research and Information Engineering, Beijing University of Technology, Pingleyuan 100, Beijing 100124, China

Mathematics Subject Classification 91A10 · 91A14 · 91A68 · 05C57 · 90B06 · 90B20

1 Introduction

The price of anarchy (PoA, [35]) is an important notion in *algorithmic game theory* ([32]) and has been investigated intensively during the last two decades in *congestion games* ([15, 37]), starting with the pioneering paper of Roughgarden and Tardos [42] on the PoA of pure Nash equilibria in *non-atomic congestion games* ([15]) with affine linear cost functions. Much of this work has then been devoted to worst-case upper bounds of the PoA for different types of cost functions $\tau_a(\cdot)$, and the influence of the network topology on these upper bounds, see, e.g., [32] for an overview.

Much less attention has been paid to the evolution of the PoA as a function of the growing total demand, although this is quite important for traffic and transportation networks in which the demands tend to be high. Only recently, it has been shown empirically ([29, 34, 48]) and analytically ([8–10, 47]) for *non-atomic* congestion games that the PoA of pure Nash equilibria actually converges to 1 with growing total demand for a large class of cost functions that includes all polynomials.

Non-atomic congestion games have the special feature that every *individual* user (player) is *infinitesimal* and controls a *negligible* amount of demand, and so has a *negligible influence* on the performance of the whole game. This can be stated alternatively as that the demands are *arbitrarily splittable*. Prototypical non-atomic congestion games are traffic networks in which each (travel) origin-destination pair has an arbitrarily splittable traffic demand that need to be distributed on paths connecting the origin and the destination. A direct consequence is the *essential uniqueness* ([42]) of pure Nash equilibria in non-atomic congestion games, which plays a pivotal role in the convergence analysis of the PoA of pure Nash equilibria by Colini-Baldeschi et al. [8–10] and Wu et al. [47].

In general, demands may not be arbitrarily splittable or even may not be split at all. This is captured by *atomic congestion games* ([37]). A prototypical such game is a transportation network in which each user wants to transport a certain unsplittable demand of a good along a single path of that network. In this case, the congestion game is *finite* ([30]), and each individual user is no longer infinitesimal and has a *non-negligible influence* on the whole game, and thus in particular on the existence and other properties of Nash equilibria. When the game is *unweighted,* i.e., users have equal demands, then pure Nash equilibria exist, but may have *different* cost and so are *not* essentially unique, see, e.g., [37, 44]. When the game is *weighted*, i.e., users have unequal demands, then pure Nash equilibria need *not exist* and one has to resort to *mixed* Nash equilibria except for *particular* cases, see, e.g., [16–18, 30].

This raises an important question if and how much the *non-negligible role* of individuals in atomic congestion games may influence the total (transportation) inefficiency for growing total (transportation) demand compared to their *negligible role* in non-atomic congestion games. This asks for a convergence analysis of the PoA of both, *pure* and *mixed*, Nash equilibria for growing total demands in atomic congestion games.

1.1 Our contribution

To address this question, we study the evolution of the PoA for growing total demand T in atomic congestion games with *unsplittable* demands and *polynomial* cost functions. While our results hold for arbitrary atomic congestion games, we will mostly use the notation of transportation networks, since they are more intuitive.

Our analysis covers the PoAs for both, pure and mixed, Nash equilibria. When pure Nash equilibria exist, then we call the ratio of their worst-case cost over the social optimum cost the *atomic PoA*, see (2.8). This distinguishes it from the PoA of pure Nash equilibria in non-atomic congestion games, which is called the *non-atomic PoA* in this paper, see (2.9). Since mixed Nash equilibria are probability distributions, they induce random flows on the transportation network. We then call the ratio of the worst-case *expected* cost of these random flows induced by mixed Nash equilibria over the social optimum cost the *mixed PoA*, see (2.10), and call the ratio of the *random* cost of the random flow induced by a specific mixed Nash equilibrium over the social optimum cost the *random PoA of that mixed Nash equilibrium*, see (2.11).

The atomic PoA measures the inefficiency of selfish *deterministic* choices, while the mixed and random PoAs quantify the inefficiency of selfish *random* choices in *expectation* and as a *stochastic variable*, respectively. They are thus different. In particular, the random PoA is a random variable, and the atomic PoA is bounded by the mixed PoA, since pure Nash equilibria in atomic congestion games can be considered as particular mixed Nash equilibria that result in deterministic choices of users.

We first derive upper bounds on the atomic, mixed and random PoAs for polynomial cost functions of the *same degree*, which cover *BPR cost functions* ([5]) that are of the form $\xi_a \cdot x^{\beta} + \gamma_a$. In this analysis, we apply the technique of *scaling* that was used implicitly in Colini-Baldeschi et al. [10] and formalized and extended in Wu et al. [47] and Wu and Möhring [46].

Using this technique, we show that the atomic PoA is $1 + O(\frac{1}{T}) + O(\sqrt{\frac{d_{max}}{T}})$ when pure Nash equilibria exist, see Theorem 1. Here, T is the *total demand* and d_{max} is the *maximum demand over all individuals* (simply, *maximum individual demand*), which reflects to a certain extent the possible influence of an individual. Moreover, we show that the mixed PoA is $1 + O(\frac{1}{T}) + O(\frac{d_{max}^{1/6}}{T^{1/6}})$, see Theorem 2b. These upper bounds converge quickly to 1 as $T \to \infty$ and $\frac{d_{max}}{T} \to 0$. We also explore the probability distribution of the random PoA of an arbitrary mixed Nash equilibrium and obtain with Chebyshev's inequalities in Theorem 2a that the random PoA is bounded from above by $1 + O(\frac{1}{T}) + O(\frac{d_{max}^{1/6}}{T^{1/6}})$ with an overwhelming probability of $1 - O(\frac{d_{max}^{1/3}}{T^{1/3}})$. This shows that an arbitrary mixed Nash equilibrium is also efficient as a random variable. We further illustrate that both conditions $T \to \infty$ and $\frac{d_{max}}{T} \to 0$ are necessary for these convergence results, see Examples 2 and 3.

We then investigate conditions for the convergence of the atomic PoA and the mixed PoA for arbitrary polynomial cost functions. We demonstrate first that the conditions $T \to \infty$ and $\frac{d_{max}}{T} \to 0$ are *no longer* sufficient for the convergence of the atomic PoA to 1, since the cost functions may have different degrees and the (transportation) origin-destination pairs may have asynchronous demand growth rates. This may result in significantly discrepant influences of different origin-destination pairs on the limits of the PoAs, see Example 4 or Wu et al. [47].

To capture these discrepant influences, we employ the *asymptotic decomposition* technique introduced by Wu et al. [47]. We show for arbitrary polynomial cost functions that both, the mixed PoA and the worst-case ratio of the total cost of the *expected* flow of a mixed Nash equilibrium over the social optimum cost, converge to 1 as $T \to \infty$, when the maximum individual demand d_{max} is *bounded* from above by a constant independent of the growth of T, see Theorem 3a–b. Note that the total cost of the expected flow of a mixed Nash equilibrium need not coincide with the *expected cost* of the random flow of that mixed Nash equilibrium, which is used in the definition of the mixed PoA, and that the condition "d_{max} is bounded from above" is necessary for these convergence results, see Example 4. To obtain these results, we have coupled the asymptotic decomposition technique with Chernorff-Hoeffding inequalities, see Appendices A.6–A.7.

Hence, the atomic PoA converges also to 1 in this general case when pure Nash equilibria exist. To analyze its convergence speed, we show, with a result by Colini-Baldeschi et al. [10] for the convergence rate of the non-atomic PoA and with a result by Wu and Möhring [46] for the sensitivity of the non-atomic PoA, that the atomic PoA (if pure Nash equilibria exist) converges to 1 at a rate of $O(T^{-\frac{1}{2\cdot\beta_{max}}})$, when $\beta_{max} = \max_{a\in A} \beta_a > 0$ is the maximum of the degrees β_a of the polynomial cost functions τ_a, the maximum individual demand d_{max} is bounded from above, and the ratio $\frac{d_k}{T}$ of the total demand d_k of each origin-destination pair k over T is bounded away from 0, see Theorem 3c.

In summary, this paper presents for atomic congestion games with growing total demands the *first* convergence analysis of the atomic and mixed PoAs, and the *first* probabilistic analysis of the random PoA. While individual users have a non-negligible role in atomic congestion games, our convergence results show that this does not significantly increase the total transportation inefficiency for a large total demand T when the maximum individual demand d_{max} is very small compared to T. Our convergence results imply, in addition to Colini-Baldeschi et al. [8–10] and Wu et al. [47], that pure Nash equilibria, mixed Nash equilibria and social optima of an atomic congestion game with a large total demand are almost equally efficient, and even as efficient as the social optima of the corresponding non-atomic congestion games, see (A.28)–(A.31) in Appendix A.6.

Thus, both pure Nash equilibria and mixed Nash equilibria in congestion games with a large total demand need not be bad. The selfish choice of strategies leads then to an almost optimal behavior, regardless whether users employ mixed or pure strategies, and whether their transportation demands are splittable or not. Users may then restrict to pure strategies and need not consider mixed strategies.

Although that need not lead to an equilibrium, it simplifies their decisions, and benefits both their own cost and the total cost of the whole transportation network.

1.2 Related work

1.2.1 Existence of equilibria

The existence of equilibria in atomic congestion games was obtained in, e.g., [16–18, 37] and others. Rosenthal [37] showed that an arbitrary *unweighted* atomic congestion game has a pure Nash equilibrium. Fotakis et al. [16] showed that an arbitrary *weighted* atomic congestion game Γ with affine linear cost functions is a *potential game* ([28]) and thus has a pure Nash equilibrium. Moreover, Harks et al. [18] proved that if $\mathcal{C}$ is a class of cost functions such that every weighted atomic congestion game Γ with cost functions in $\mathcal{C}$ is a potential game, then $\mathcal{C}$ contains only affine linear functions. The existence of pure Nash equilibria in weighted atomic congestion games was further studied by Harks and Klimm [17]. Beyond these cases, we have to consider mixed Nash equilibria in atomic congestion games, as Nash [30] has shown that every finite game has a mixed Nash equilibrium.

1.2.2 Worst-case upper bounds on the price of anarchy

Koutsoupias and Papadimitriou [24] proposed to quantify the inefficiency of equilibria in arbitrary congestion games from a worst-case perspective. This resulted in the concept of the *price of anarchy* (PoA) that is usually defined as the ratio of the worst-case cost of (pure or mixed) Nash equilibria over the social optimum cost, see [35].

A wave of research has been started with the pioneering paper of Roughgarden and Tardos [42] on the PoA of pure Nash equilibria in non-atomic congestion games with affine linear cost functions. Examples are Roughgarden [38–41], Roughgarden and Tardos [42, 43], Christodoulou and Koutsoupias [7], Correa et al. [12, 13], Perakis [36] and others. They investigated the worst-case upper bounds of the PoA of pure Nash equilibria in both atomic and non-atomic congestion games for different types of cost functions $\tau_a(\cdot)$, and analyzed the influence of the network topology on these bounds. For non-atomic congestion games, this upper bound is $\frac{4}{3}$ for affine linear cost functions ([42]), and $\Theta(\frac{\beta}{\ln \beta})$ for polynomial cost functions of degree at most β ([43]). For unweighted atomic congestion games, Christodoulou and Koutsoupias [7] showed that this upper bound is $\frac{5}{2}$ for affine linear cost functions, and $\beta^{\Theta(\beta)}$ for polynomial cost functions of degree at most β. Hence, the non-atomic PoA is not larger than the atomic PoA in general. Moreover, these upper bounds are independent of the network topology, see, e.g., [39]. Roughgarden [39, 41] also developed a (λ, μ)-smooth method by which one can obtain a *tight* and *robust* worst-case upper bound. This method was then reproved by Correa et al. [13] from a geometric perspective. Besides, Perakis [36] generalized the analysis to non-atomic congestion games with non-separable and asymmetric cost maps.

1.2.3 Convergence of the price of anarchy

Recent papers have empirically studied the PoA of pure Nash equilibria in non-atomic congestion games with *BPR cost functions* ([5]) of the same degree $\beta > 0$ and real traffic demands. Youn et al. [48] observed that the empirical PoA of pure Nash equilibria depends crucially on the total demand. Starting from 1, it grows with some oscillations, and ultimately becomes 1 again as the total demand increases. A similar observation was made by O'Hare et al. [34]. They even conjectured that the PoA of pure Nash equilibria in non-atomic congestion games with BPR cost functions of the same degree $\beta > 0$ converges to 1 at a rate of $O\big(T^{-2\cdot\beta}\big)$ when the total demand T becomes large. Monnot et al. [29] showed that traffic choices of commuting students in Singapore are near-optimal and that the empirical PoA of pure Nash equilibria is much smaller than known worst-case upper bounds. Similar observations have been reported by Jahn et al. [23].

These observations have been recently confirmed by Colini-Baldeschi et al. [8–10] and Wu et al. [47]. Colini-Baldeschi et al. [8–10] were the first to theoretically analyze the convergence of the PoA of pure Nash equilibria in non-atomic congestion games with growing total demand.

Colini-Baldeschi et al. [8] showed that the PoA of pure Nash equilibria converges to 1 as the total demand $T \to \infty$ when the non-atomic congestion game has a *single* origin-destination pair and *regularly varying* ([2]) cost functions. This convergence result was then substantially extended by Colini-Baldeschi et al. [9] to *multiple* origin-destination pairs for both the case $T \to 0$ and the case $T \to \infty$, when the ratio of the demand of each origin-destination pair over the total demand T remains a positive constant as $T \to 0$ or ∞. Colini-Baldeschi et al. [10] further extended these results to the cases where the demands and the cost functions together fulfill certain *tightness and salience conditions* that allow the ratios of demands to vary in a certain pattern as $T \to 0$ or ∞. Moreover, Colini-Baldeschi et al. [10] illustrated by an example that the PoA of pure Nash equilibria in non-atomic congestion games need not converge to 1 as $T \to \infty$ when the cost functions are not regularly varying. In addition, they showed that the PoA of pure Nash equilibria in non-atomic congestion games with polynomial cost functions converges to 1 at a rate of $O(\frac{1}{T})$ when the ratio of the demand of each origin-destination pair over the total demand T remains a positive constant as $T \to 0$ or ∞.

Wu et al. [47] generalized the work of Colini-Baldeschi et al. [8–10] for growing total demand. They formalized the *scaling technique* used implicitly in Colini-Baldeschi et al. [8–10], proposed a *limit notion* for a sequence of games with growing total demand, and developed a general technical framework, called *asymptotic decomposition*, for the convergence analysis of the PoA. With this framework, they showed for non-atomic congestion games with *arbitrary* regularly varying cost functions that the PoA of pure Nash equilibria converges to 1 as the total demand tends to ∞ *regardless of the growth pattern of the demands.* In particular, they proved a convergence rate of $o(T^{-\beta})$ for BPR cost functions of degree β and illustrated by examples that the conjecture proposed by O'Hare et al. [34] need not hold.

Wu and Möhring [46] extended the techniques of Wu et al. [47] to a sensitivity analysis of the PoA. For an arbitrary non-atomic congestion game Γ with Lipschitz continuous cost functions on $[0, T]$, they proved that the cost of an ϵ-approximate equilibrium of Γ deviates at most by $O(\sqrt{\epsilon})$ from that of a pure Nash equilibrium of Γ, and that $O(\sqrt{\epsilon})$ is a tight upper bound of this deviation. Moreover, they defined a metric $||\Gamma_1, \Gamma_2||$ for two arbitrary games in a set of non-atomic congestion games with the same *combinatorial structure*. That metric induces a topological space of such games and permits to consider *continuous real-valued maps* and the *limit of a sequence of non-atomic congestion games*. Wu and Möhring [46] used these notions for a comprehensive analysis of the *Hölder continuity* of the PoA map of pure Nash equilibria in that topological space. They showed that the PoA map is *point-wise* continuous, but neither Lipschitz continuous, nor *uniformly* Hölder continuous. However, it is *point-wise* Hölder continuous with Hölder exponent $\frac{1}{2}$ on a *dense* subspace, i.e., $|\rho_{nat}(\Gamma_1) - \rho_{nat}(\Gamma_2)| \in O(\sqrt{||\Gamma_1, \Gamma_2||})$ for any two non-atomic congestion games Γ_1 and Γ_2 of that subspace, where $\rho_{nat}(\Gamma_i)$ denotes the PoA value of pure Nash equilibria of the game Γ_i, $i = 1, 2$. This results in an approximate computation of the PoA $\rho_{nat}(\cdot)$, meaning that one can approximate $\rho_{nat}(\Gamma)$ for irregular cost functions with $\rho_{nat}(\Gamma')$ for relatively simpler polynomial cost functions when the polynomial cost functions of Γ' are sufficiently close to the irregular cost functions of Γ.

As a byproduct of the above Hölder continuity analysis, Wu and Möhring [46] showed that the total cost difference between Nash equilibria of two non-atomic congestion games Γ_1 and Γ_2 is in $O(\sqrt{\|\Gamma_1 - \Gamma_2\|})$ when Γ_1 and Γ_2 have the same Lipschitz continuous cost functions. Moreover, when the two non-atomic congestion games Γ_1 and Γ_2 have the same demands but different Lipschitz continuous cost functions, they proved a similar upper bound on the total cost difference between their Nash equilibria. These results together with the convergence rate of Colini-Baldeschi et al. [10] will help us to obtain an explicit convergence rate of the atomic PoA for polynomial cost functions of different degres, see Theorem 3c and its proof in Appendix A.6.

Conditions implying the convergence of mixed Nash equilibria in atomic congestion games to pure Nash equilibria in non-atomic congestion games have also been studied in, e.g., [11, 19, 21, 22, 26], and others.

Among these papers, Cominetti et al. [11] is the closest to our work. They showed that mixed Nash equilibria of an atomic congestion game with *strictly increasing* cost functions converge *in distribution* to pure Nash equilibria of a limit non-atomic congestion game, when the total demand T converges to a constant $T_0 \in (0, \infty)$, the maximum individual demand d_{max} converges to 0, and the number of users converges to ∞. Moreover, they showed that this convergence happens at a rate of $O(\sqrt{d_{max}})$ when the cost functions have *strictly positive first-order derivatives.* Consequently, the PoA of mixed Nash equilibria (i.e., the mixed PoA) in such an atomic congestion game converges also to that of pure Nash equilibria in a "limit non-atomic congestion game" under these conditions.

The results of Cominetti et al. [11] are inspiring and seminal. They confirm the intuition that atomic congestion games can be thought of as non-atomic congestion games when d_{max} is *tiny*, the number of users is *huge*, and T is *moderate*, i.e., neither too small nor too large. Our convergence result for the mixed PoA actually generalizes those of Cominetti et al. [11] to the case that $T \to \infty$. This is a non-trivial generalization, since it does not require the existence of the limit non-atomic congestion game, which is a premise in the analysis of Cominetti et al. [11].

Our work also extends the convergence results for the PoA of pure Nash equilibria in non-atomic congestion games that were obtained recently by Colini-Baldeschi et al. [8–10] and Wu et al. [47] to convergence results for pure and mixed Nash equilibria in atomic congestion games. This implies that selfishness is also good in "atomic congestion". In particular, our results show for arbitrary congestion games with a large total demand that selfish choice of users is almost as efficient as social optima, regardless whether demands are splittable or not, and whether users use pure strategies or mixed strategies.

1.3 Outline of the paper

The paper is organized as follows. We develop our results for arbitrary atomic congestion games. These and their relevant concepts are introduced in Sect. 2. We analyze the convergence of the PoAs for atomic congestion games in Sect. 3. Sect. 3.1 then presents our convergence results for polynomial cost functions with the same degree. Subsequently, Sect. 3.2 presents our convergence results for arbitrary polynomial cost functions. We conclude with a short summary and discussion in Sect. 4. To improve readability, all proofs have been moved to an Appendix.

2 Model and preliminaries

Our study involves both atomic and non-atomic congestion games. To facilitate the discussion, we introduce a unified notation in Sect. 2.1, and distinguish games implicitly by properties of their strategy profiles, see Sect. 2.2.

2.1 Atomic and non-atomic congestion games

We define an arbitrary atomic congestion game with the notation of transportation games (see, e.g., [32, 37]), since this is more intuitive and closer to practice. An *atomic congestion game* Γ is thus associated with a transportation network $G = (V, A)$, and represented symbolically by a tuple $(\mathcal{K}, \mathcal{P}, \tau, \mathcal{U}, d)$ with components defined in (G1)–(G5).

- **(G1)** $\mathcal{K}$ is a finite *non-empty* set of (transportation) *origin-destination* (O/D) pairs $(o_k, t_k) \in V \times V$ with $o_k \neq t_k$. We will denote an O/D pair (o_k, t_k) simply by its index k when this is not ambiguous.
- **(G2)** $\mathcal{P} = \cup_{k\in\mathcal{K}}\mathcal{P}_k$ with each $\mathcal{P}_k \subseteq 2^A \setminus \{\emptyset\}$ denotes the non-empty set of all *paths* from the origin o_k to the destination t_k. Here, a path is a non-empty subset of the arc set A. Then $\mathcal{P}_k \cap \mathcal{P}_{k'} = \emptyset$ for $k, k' \in \mathcal{K}$ with $k \neq k'$.
- **(G3)** $\tau = (\tau_a)_{a\in A}$ is a cost function vector, s.t. $\tau_a : [0, \infty) \to [0, \infty)$ is *non-negative*, *continuous* and *non-decreasing* and denotes the flow-dependent *latency* or *cost* of arc $a \in A$. We assume that no arc can be used for free, i.e., $\tau_a(x) > 0$ for all pairs $(a, x) \in A \times (0, \infty)$.
- **(G4)** Associated with each O/D pair $k \in \mathcal{K}$ is a finite non-empty set $\mathcal{U}_k$ of *agents* that are individual users or players. Then $\mathcal{U} = \cup_{k\in\mathcal{K}}\mathcal{U}_k$ is the agent set of Γ. We assume that $\mathcal{U}_k \cap \mathcal{U}_{k'} = \emptyset$ for all $k, k' \in \mathcal{K}$ with $k \neq k'$.
- **(G5)** $d = (d_{k,i})_{k\in\mathcal{K}, i\in\mathcal{U}_k}$ is a *demand vector,* where $d_{k,i} > 0$ denotes an *unsplittable* demand to be transported by agent $i \in \mathcal{U}_k$. So Γ has the *total (transportation) demand* $T = T(\mathcal{U}, d) := \sum_{k\in\mathcal{K}} d_k$, where $d_k := \sum_{i\in\mathcal{U}_k} d_{k,i}$ is the *demand* of

O/D pair $k \in \mathcal{K}$. We call $d_{max} := \max_{i\in\mathcal{U}_k,k\in\mathcal{K}} d_{k,i}$ the *maximum individual demand* of Γ. Note that Γ is *unweighted* if $d_{k,i} \equiv \upsilon$ for all $k \in \mathcal{K}$ and all $i \in \mathcal{U}_k$, for a constant $\upsilon > 0$. Otherwise, Γ is *weighted*.

To unify notation, we view a non-atomic congestion game as a *variant* of an atomic congestion game, in which each agent $i \in \mathcal{U}_k$ is no longer an individual user, but a *population of infinitesimal users,* who together have the demand $d_{k,i}$. Hence, the demands $d_{k,i}$ can be split arbitrarily over paths in $\mathcal{P}_k$ when Γ is non-atomic. This differs from an atomic congestion game, in which the demands $d_{k,i}$ cannot be split. With a little abuse of notation, we denote a non-atomic congestion game again by the same tuple $\Gamma = (\mathcal{K}, \mathcal{P}, \tau, \mathcal{U}, d)$. We will simply call a tuple Γ a congestion game, and distinguish atomic and non-atomic congestion games by their *atomic* and *non-atomic profiles* in Sect. 2.2.

The tuple $(\mathcal{K}, \mathcal{P})$ together with the transportation network G constitutes the *combinatorial structure* of Γ. For ease of notation, we may fix an arbitrary network G and an arbitrary tuple $(\mathcal{K}, \mathcal{P})$, and denote Γ simply by $(\tau, \mathcal{U}, d)$. Viewed as a general congestion game, the arcs $a \in A$ and the paths $p \in \mathcal{P}$ correspond to resources and (pure) strategies, see, e.g., [15] and Rosenthal [37]. Although we use the nomenclature of transportation networks, the analysis and results below are independent of this view and carry over to arbitrary congestion games.

2.2 Atomic, non-atomic and mixed profiles

Users distribute their demands *simultaneously* and *independently* on paths in $\mathcal{P}$. This results in a *strategy profile* or simply *profile* $\Pi = (\Pi_i)_{i\in\mathcal{U}} = (\Pi_i)_{i\in\mathcal{U}_k,k\in\mathcal{K}} = (\Pi_{i,p})_{i\in\mathcal{U}_k,p\in\mathcal{P}_k,k\in\mathcal{K}}$ satisfying the condition (2.1),

$$\sum_{p'\in\mathcal{P}_k} \Pi_{i,p'} = 1 \text{ and } \Pi_{i,p} \geq 0 \quad \forall i \in \mathcal{U}_k \ \forall p \in \mathcal{P}_k \ \forall k \in \mathcal{K}. \tag{2.1}$$

We put $\Pi_{i,p} = 0$ when $i \in \mathcal{U}_k$ and $p \in \mathcal{P}_{k'}$ for some $k, k' \in \mathcal{K}$ with $k \neq k'$. This extends a profile Π naturally to a vector $(\Pi_{i,p})_{i\in\mathcal{U},p\in\mathcal{P}}$ with components $\Pi_{i,p}$ satisfying condition (2.1).

A profile Π is called *atomic* if Π is *binary*. In this case, $\Pi_{i,p} \in \{0, 1\}$, $i \in \mathcal{U}_k$, $p \in \mathcal{P}_k$, $k \in \mathcal{K}$, indicates whether path p is used by i, i.e., $\Pi_{i,p} = 1$, or not, i.e., $\Pi_{i,p} = 0$. Condition (2.1) then means that each $i \in \mathcal{U}_k$ satisfies his demand $d_{k,i}$ by a *single* path $p \in \mathcal{P}_k$ in an atomic profile Π. So a congestion game Γ with only atomic profiles is indeed an atomic congestion game whose demands $d_{k,i}$ cannot be split.

In a non-atomic congestion game, each agent $i \in \mathcal{U}_k$ is a population of infinitesimal users and can split the demand $d_{k,i}$ arbitrarily, i.e., agents $i \in \mathcal{U}_k$ can send their demands $d_{k,i}$ along several paths $p \in \mathcal{P}_k$. This is captured by *non-atomic* profiles. The components $\Pi_{i,p}$ are then *fractions* of the demands $d_{k,i}$ deposited by agents $i \in \mathcal{U}_k$ on paths $p \in \mathcal{P}_k$, i.e., agents i totally allocate $d_{k,i} \cdot \Pi_{i,p}$ units of demands to paths p. Hence, these $\Pi_{i,p}$ can take arbitrary values in $[0, 1]$ when Π is non-atomic. Condition (2.1) is then a *feasibility* constraint for non-atomic profiles that ensures that *all* demands are satisfied. Clearly, a congestion game is non-atomic when it has only non-atomic profiles.

In a *mixed* profile Π, each $\Pi_i = (\Pi_{i,p})_{p\in\mathcal{P}_k}$ is a *probability distribution* over the set $\mathcal{P}_k$ for all $i \in \mathcal{U}_k$ and all $k \in \mathcal{K}$. Then the decisions are random, and every agent $i \in \mathcal{U}_k$ delivers his demand $d_{k,i}$ on a single *random path* $p_{k,i}(\Pi_i)$ drawn *independently* from $\Pi_i = (\Pi_{i,p})_{p\in\mathcal{P}_k}$, where $\Pi_{i,p} \in [0,1]$ is the probability of the random event "$p_{k,i}(\Pi_i) = p$". Note that we consider mixed profiles only for atomic congestion games, although we use a unified notation for both atomic and non-atomic congestion games. Note also that an atomic profile is a particular mixed profile with $\{0,1\}$-probabilities.

2.3 Multi-commodity flows and their cost

Each profile Π induces a *multi-commodity flow* $f = (f_p)_{p\in\mathcal{P}} = (f_p)_{p\in\mathcal{P}_k, k\in\mathcal{K}}$. When Π is atomic or non-atomic, then f is *deterministic* with *flow value* $f_p := \sum_{i\in\mathcal{U}_k} d_{k,i}\cdot\Pi_{i,p}$ for all $p\in\mathcal{P}_k$ and all $k\in\mathcal{K}$. We then call f *atomic* and *non-atomic*, respectively. There are only finitely many atomic flows, as the number $|\mathcal{U}| = \sum_{k\in\mathcal{K}} |\mathcal{U}_k|$ of agents is finite and the demands $d_{k,i}$ cannot be split in an atomic flow.

When Π is mixed, then the flow $f = (f_p)_{p\in\mathcal{P}}$ is a *random* vector in which each component f_p is a weighted sum $\sum_{i\in\mathcal{U}_k} d_{k,i}\cdot\mathbb{1}_{\{p\}}\big(p_{k,i}(\Pi_i)\big)$ of mutually independent *Bernoulli* random variables $\mathbb{1}_{\{p\}}\big(p_{k,i}(\Pi_i)\big)$, where $p_{k,i}(\Pi_i)$ is the *random path* draw from the distribution Π_i by agent i of O/D pair k, and $\mathbb{1}_{\{p\}}(\cdot)$ is the indicator function of the membership of the singleton $\{p\}$. Then

$$\begin{aligned}\mathbb{E}_\Pi(f_p) &= \sum_{i\in\mathcal{U}_k} d_{k,i}\cdot\Pi_{i,p},\\ \mathbb{VAR}_\Pi(f_p) &= \sum_{i\in\mathcal{U}_k} d_{k,i}^2\cdot\Pi_{i,p}\cdot(1-\Pi_{i,p})\end{aligned} \tag{2.2}$$

for all $p\in\mathcal{P}_k$ and all $k\in\mathcal{K}$. Here, we used that agents choose their paths *mutually independently*, that $\mathbb{E}_\Pi[\mathbb{1}_{\{p\}}(p_{k,i}(\Pi_i))] = \Pi_{i,p}$ and $\mathbb{VAR}_\Pi[\mathbb{1}_{\{p\}}(p_{k,i}(\Pi_i))] = \Pi_{i,p}\cdot(1-\Pi_{i,p})$, and that every agent $i\in\mathcal{U}_k$ transports his demand $d_{k,i}$ entirely on the single random path $p_{k,i}(\Pi_i)\in\mathcal{P}_k$. We will write $\mathbb{E}_\Pi(f) := \big(\mathbb{E}_\Pi(f_p)\big)_{p\in\mathcal{P}}$, and call $\mathbb{E}_\Pi(f)$ and $f = (f_p)_{p\in\mathcal{P}}$ the *expected flow* and the *random flow* of the mixed profile Π, respectively.

The expected flow $\mathbb{E}_\Pi(f)$ is a non-atomic flow, and an arbitrary non-atomic flow is the expected flow of a mixed profile. Moreover, an atomic flow f is a particular random flow, in which the random flow values f_p have a variance of zero. Note that each state of a random flow is an atomic flow, and the finite set of all atomic flows is the *state space* of random flows, i.e., $\sum_{f' \text{ is an atomic flow}} \mathbb{P}_\Pi[f = f'] = 1$ for a mixed profile Π with random flow f.

An arbitrary flow f induces an *arc flow* $(f_a)_{a\in A}$ in which component $f_a := \sum_{p\in\mathcal{P}: a\in p} f_p$ is the flow value on arc $a\in A$. When Π is atomic or non-atomic, then f_a is again deterministic for all $a\in A$. When Π is mixed, then each f_a is *random*, and has the *expectation* and *variance* in (2.3),

$$\begin{aligned}\mathbb{E}_{\Pi}(f_a) = \sum_{k\in\mathcal{K}} \sum_{p\in\mathcal{P}_k:a\in p} \mathbb{E}_{\Pi}(f_p) = \sum_{k\in\mathcal{K}} \sum_{i\in\mathcal{U}_k} d_{k,i}\cdot \\ \sum_{p\in\mathcal{P}_k:a\in p} \Pi_{i,p} := \sum_{k\in\mathcal{K}} \sum_{i\in\mathcal{U}_k} d_{k,i}\cdot \Pi_{i,a}, \\ \mathbb{VAR}_{\Pi}(f_a) = \sum_{k\in\mathcal{K}} \sum_{i\in\mathcal{U}_k} d_{k,i}^2 \cdot \Pi_{i,a}\cdot (1-\Pi_{i,a}).\end{aligned} \tag{2.3}$$

Here, $\Pi_{i,a} := \sum_{p\in\mathcal{P}_k:a\in p} \Pi_{i,p} \in [0,1]$ is the probability that agent i of O/D pair $k \in \mathcal{K}$ uses arc $a \in A$. Then (2.3) follows since agents use an arc $a \in A$ *mutually independently*, and only if the arc a belongs to one of their random paths $p_{k,i}(\Pi_i)$.

For a non-atomic flow, we need only to specify the *O/D pair demand vector* $(d_k)_{k\in\mathcal{K}}$ with $d_k = \sum_{i\in\mathcal{U}_k} d_{k,i}$, since the demands $d_{k,i}$ are arbitrarily splittable, and two congestion games have the same set of non-atomic flows if and only if they have the same $(d_k)_{k\in\mathcal{K}}$. Nonetheless, the demand vector $d = (d_{k,i})_{k\in\mathcal{K},i\in\mathcal{U}_k}$ need to be specified for atomic and random flows, as the demands $d_{k,i}$ can then be not split.

Given a flow f, an arc $a \in A$ has the cost $\tau_a(f_a)$, and a path $p \in \mathcal{P}$ has the cost $\tau_p(f) := \sum_{a\in p} \tau_a(f_a)$. When f is atomic or non-atomic, then these cost values are deterministic. Every $i \in \mathcal{U}_k$ then has the deterministic cost

$$C_{k,i}(f,\Gamma) = C_{k,i}(f,\tau,\mathcal{U},d) := \sum_{p\in\mathcal{P}_k} d_{k,i}\cdot \Pi_{i,p}\cdot \tau_p(f),$$

and all agents together have the (deterministic) *total cost*

$$C(f,\Gamma) := \sum_{k\in\mathcal{K}} \sum_{i\in\mathcal{U}_k} C_{k,i}(f,\tau,\mathcal{U},d) = \sum_{p\in\mathcal{P}} f_p\cdot \tau_p(f) = \sum_{a\in A} f_a\cdot \tau_a(f_a).$$

Note that the cost $C_{k,i}(f,\Gamma)$ can be expressed equivalently as $C_{k,i}(f,\Gamma) = d_{k,i}\cdot \tau_{p_{k,i}(f)}(f)$ when f is *atomic* and $p_{k,i}(f) \in \mathcal{P}_k$ is the single path used by agent i in f.

The cost values $\tau_a(f_a)$ and $\tau_p(f)$ are *random* when f is the random flow of a mixed profile Π. Then each $i \in \mathcal{U}_k$ has the *random* cost

$$C_{k,i}(f,\Gamma) := d_{k,i}\cdot \tau_{p_{k,i}(\Pi_i)}(f) = \sum_{p\in\mathcal{P}_k} d_{k,i}\cdot \mathbb{1}_{\{p\}}(p_{k,i}(\Pi_i))\cdot \tau_p(f),$$

where $p_{k,i}(\Pi_i)$ is again the random path of agent $i \in \mathcal{U}_k$. The *random total cost* is then $C(f,\Gamma) := \sum_{k\in\mathcal{K}} \sum_{i\in\mathcal{U}_k} C_{k,i}(f,\Gamma)$. Consequently, all agents together have the *expected total cost*

$$\begin{aligned}\mathbb{E}_{\Pi}[C(f,\Gamma)] &= \sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{U}_k}\mathbb{E}_{\Pi}[C_{k,i}(f,\Gamma)]\\ &= \sum_{a\in A}\mathbb{E}_{\Pi}\big[f_a\cdot\tau_a(f_a)\big]\\ &= \sum_{p\in\mathcal{P}}\mathbb{E}_{\Pi}\big[f_p\cdot\tau_p(f)\big].\end{aligned}$$

The expected total cost $\mathbb{E}_{\Pi}[C(f,\Gamma)]$ of a random flow f need not equal the total cost $C(\mathbb{E}_{\Pi}(f),\Gamma)$ of its expected flow $\mathbb{E}_{\Pi}(f)$. But they coincide when Π is atomic.

We denote *atomic*, *non-atomic* and *random* flows by $f_{at} = (f_{at,p})_{p\in\mathcal{P}}$, $f_{nat} = (f_{nat,p})_{p\in\mathcal{P}}$ and $f_{ran} = (f_{ran,p})_{p\in\mathcal{P}}$, respectively, and will not refer explicitly to the corresponding profiles since they are clear from the context.

2.4 Social optima and equilibria

Consider an arbitrary congestion game Γ. An atomic flow f_{at}^* is an *atomic system optimum* (atomic SO), if $C(f_{at}^*,\Gamma)\leq C(f_{at},\Gamma)$ for every atomic flow f_{at}. Similarly, a non-atomic flow f_{nat}^* is a *non-atomic SO* if $C(f_{nat}^*,\Gamma)\leq C(f_{nat},\Gamma)$ for every non-atomic flow f_{nat}, and a random flow f_{ran}^* is a *mixed SO* if $\mathbb{E}_{\Pi^*}[C(f_{ran}^*,\Gamma)]\leq\mathbb{E}_{\Pi}[C(f_{ran},\Gamma)]$ for each random flow f_{ran}, where Π^* and Π are the mixed profiles of f_{ran}^* and f_{ran}, respectively.

The expected total cost of an arbitrary mixed SO flow coincides with that of an arbitrary atomic SO flow, since the set of atomic flows is the state space of random flows and every atomic flow is a random flow with zero variance. Moreover, the total cost of an atomic SO flow is not smaller than that of a non-atomic SO flow, since every atomic SO flow is also a non-atomic flow. We summarize this in Lemma 1.

Lemma 1 *Consider an arbitrary congestion game Γ with a mixed SO flow f_{ran}^* of a mixed profile Π^*, an atomic SO flow f_{at}^*, and a non-atomic SO flow f_{nat}^*. Then $\mathbb{E}_{\Pi^*}[C(f_{ran}^*,\Gamma)] = C(f_{at}^*,\Gamma)\geq C(f_{nat}^*,\Gamma)$.*

Similar to the different types of SO flows in Lemma 1, congestion games admit also *Nash equilibrium flows* of different types. In each of them an individual does *not* benefit from unilaterally changing his strategy. Hence, a Nash equilibrium flow is essentially a steady-state of the network that is stable under unilateral selfish behavior. Since we consider three types of flows, i.e., atomic, non-atomic and random flows, we define their Nash equilibria separately.

An atomic flow $\tilde{f}_{at} = (\tilde{f}_{at,p})_{p\in\mathcal{P}}$ is an atomic *(pure) Nash equilibrium* (NE), if $C_{k,i}(\tilde{f}_{at},\Gamma) = d_{k,i}\cdot\tau_{p_{k,i}(\tilde{f}_{at})}(\tilde{f}_{at})\leq C_{k,i}(f'_{at},\Gamma) = d_{k,i}\cdot\tau_{p'}(f'_{at})$ for all $k\in\mathcal{K}$, all $i\in\mathcal{U}_k$ and all $p'\in\mathcal{P}_k$, where $p_{k,i}(\tilde{f}_{at})\in\mathcal{P}_k$ is the path used by agent $i\in\mathcal{U}_k$ in atomic flow $\tilde{f}_{at}$, and $f'_{at} = (f'_{at,p})_{p\in\mathcal{P}}$ is an atomic flow with components $f'_{at,p}$

defined in (2.4).

$$f'_{at,p} = \begin{cases} \tilde{f}_{at,p} & \text{if } p \notin \{p_{k,i}(\tilde{f}_{at}), p'\}, \\ \tilde{f}_{at,p} - d_{k,i} & \text{if } p = p_{k,i}(\tilde{f}_{at}), \\ \tilde{f}_{at,p} + d_{k,i} & \text{if } p = p', \end{cases} \qquad \forall p \in \mathcal{P}. \tag{2.4}$$

Clearly, f'_{at} is the atomic flow obtained by only moving i from $p_{k,i}(\tilde{f}_{at})$ to p', and so differs slightly from $\tilde{f}_{at}$ when d_{max} is tiny. Rosenthal [37] has shown the existence of atomic NE flows for unweighted atomic congestion games. Weighted atomic congestion games usually do not have atomic NE flows, except for particular cases, e.g., affine linear cost functions, see Harks et al. [18] and Harks and Klimm [17].

Since the cost functions $\tau_a(\cdot)$ are non-decreasing, non-negative and continuous, and since each agent in a non-atomic flow is a population of infinitesimal users, non-atomic (pure) NE are identical to *Wardrop equilibria* (WE, [45]), see, e.g., [1, 42]. Thus a non-atomic flow $\tilde{f}_{nat} = (\tilde{f}_{nat,p})_{p\in\mathcal{P}}$ is a *non-atomic NE* if and only if it fulfills *Wardrop's first principle*, i.e., $\tau_p\big(\tilde{f}_{nat}\big) \le \tau_{p'}\big(\tilde{f}_{nat}\big)$ for any two paths $p, p' \in \mathcal{P}_k$ with $\tilde{f}_{nat,p} > 0$ for each $k \in \mathcal{K}$. Here, we note that the cost of each path does not change when an infinitesimal user unilaterally changes his path. Hence a path $p \in \mathcal{P}_k$ is used, i.e., $\tilde{f}_{nat,p} > 0$, in a non-atomic NE flow $\tilde{f}_{nat}$ only if $\tau_p\big(\tilde{f}_{nat}\big) = \min_{p'\in\mathcal{P}_k} \tau_{p'}\big(\tilde{f}_{nat}\big)$. Dafermos [14] has shown that non-atomic NE flows always exist, and can be characterized equivalently by the *variational inequality* (2.5),

$$\sum_{a\in A} \tau_a\big(\tilde{f}_{nat,a}\big) \cdot \big(f_{nat,a} - \tilde{f}_{nat,a}\big) \ge 0, \tag{2.5}$$

for all non-atomic flows f_{nat}. Moreover, Roughgarden and Tardos [42] have shown that non-atomic NE flows are *essentially unique*, i.e., $\tau_a(\tilde{f}_{nat,a}) = \tau_a(\tilde{f}'_{nat,a})$ for each $a \in A$ for two arbitrary non-atomic NE flows $\tilde{f}_{nat}$ and $\tilde{f}'_{nat}$. Clearly, atomic and non-atomic NE flows differ. Nonetheless, both of them are pure Nash equilibria.

Mixed NE flows directly generalize atomic NE flows by considering random flows of mixed profiles. Formally, a random flow $\tilde{f}_{ran}$ is a *mixed NE flow* if, for each $i \in \mathcal{U}_k$ and each $k \in \mathcal{K}$,

$$\mathbb{E}_{\tilde{\Pi}}[C_{k,i}(\tilde{f}_{ran}, \Gamma)] = \mathbb{E}_{\tilde{\Pi}_{-i}}\big[d_{k,i} \cdot \tau_p(\tilde{f}_{ran|i,p})\big] \le \mathbb{E}_{\tilde{\Pi}_{-i}}\big[d_{k,i} \cdot \tau_{p'}(\tilde{f}_{ran|i,p'})\big] \tag{2.6}$$

when $p, p' \in \mathcal{P}_k$ are two arbitrary paths with $\tilde{\Pi}_{i,p} > 0$, $\tilde{\Pi} = (\tilde{\Pi}_j)_{j\in\mathcal{U}}$ is the mixed profile of $\tilde{f}_{ran}$, and $\tilde{\Pi}_{-i} = (\tilde{\Pi}_j)_{j\in\mathcal{U}\setminus\{i\}}$ is the mixed profile of all agents other than i in $\tilde{\Pi}$, see also Cominetti et al. [11]. Herein, $\tilde{f}_{ran|i,p} = (\tilde{f}_{ran,p''|i,p})_{p''\in\mathcal{P}}$ is the random flow in which agent i uses the fixed path p and the others still follow the mixed profile $\tilde{\Pi}_{-i}$, i.e.,

$$\tilde{f}_{ran,p''|i,p} = \begin{cases} \tilde{f}_{ran,p''} & \text{if } p'' \in \cup_{k''\in\mathcal{K}\setminus\{k\}}\mathcal{P}_{k''}, \\ d_{k,i} + \sum_{j\in\mathcal{U}_k\setminus\{i\}} d_{k,j} \cdot \mathbb{1}_{\{p''\}}(p_{k,j}(\tilde{\Pi}_j)) & \text{if } p'' = p, \\ \sum_{j\in\mathcal{U}_k\setminus\{i\}} d_{k,j} \cdot \mathbb{1}_{\{p''\}}(p_{k,j}(\tilde{\Pi}_j)) & \text{if } p'' \in \mathcal{P}_k \setminus \{p\}, \end{cases}$$

for all paths $p'' \in \cup_{k''\in\mathcal{K}}\mathcal{P}_{k''}$. Inequality (2.6) then means that each support of the mixed strategy $\tilde{\Pi}_i$ of an agent $i \in \mathcal{U}$ is the best response to the mixed profile $\tilde{\Pi}_{-i}$ of his opponents when $\tilde{f}$ is a mixed NE flow with mixed profile $\tilde{\Pi}$. Hence, no agent can reduce his (expected) cost by unilaterally changing his mixed strategy when the random flow is a mixed NE. Since atomic congestion games equipped with only atomic profiles are finite games, mixed NE flows always exist, see [30]. Note that atomic NE flows are mixed NE flows with zero variance, but mixed NE flows need not be atomic NE flows, see, e.g., [32].

Remark 1 (The mixed Wardrop equilibria) Note that one may consider also random flows f_{ran} in which all paths with *positive* expected flow values have minimum expected cost, i.e.,

$$\mathbb{E}_{\Pi}[\tau_p(f_{ran})] \leq \mathbb{E}_{\Pi}[\tau_{p'}(f_{ran})] \tag{2.7}$$

for two arbitrary paths $p, p' \in \mathcal{P}_k$ with $\mathbb{E}_{\Pi}(f_{ran,p}) > 0$ for each $k \in \mathcal{K}$, where Π is the mixed profile of f_{ran}. Such random flows then generalize WE flows of non-atomic congestion games in atomic congestion games. We thus call them *mixed WE flows*. Using *Brouwer's fixed point theorem* ([3]) and an argument similar to that in Dafermos [14] for the existence of WE flows in non-atomic congestion games, we can show easily that mixed WE flows always exist in atomic congestion games, see Lemma 7 in Appendix A.1. The convergence results presented in this paper carry also over to the inefficiency of mixed WE flows. In fact, we can even view mixed NE flows as mixed WE flows in the convergence analysis of the PoA of mixed NE, since mixed NE flows approximate mixed WE flows when $\frac{d_{max}}{T}$ is tiny, see, e.g., (2.6)–(2.7), (A.11) in Appendix A.5, and Appendix A.6. Nonetheless, we will not go deeper into the discussion of mixed WE flows, so as to save space.

Example 1 Consider the congestion game Γ with one O/D pair (o, t) (i.e., $\mathcal{K} = \{1\}$) and two parallel paths (arcs) shown in Fig. 1. We label the upper and lower arcs as u and ℓ, respectively. Γ has cost functions $\tau_u(x) = x^2$ and $\tau_\ell(x) \equiv 2$, and two agents with O/D pair (o, t) and demand 2 each. Then Γ has a unique atomic NE flow $\tilde{f}_{at} = (\tilde{f}_{at,u}, \tilde{f}_{at,\ell}) = (0, 4)$, since an agent using the upper arc u has a cost of at least $4 > \tau_\ell(x) \equiv 2$ and can always benefit by moving to the lower arc ℓ. Moreover, Γ has the unique non-atomic NE flow $\tilde{f}_{nat} = (\sqrt{2}, 4 - \sqrt{2})$, since demands can be arbitrarily split in a non-atomic flow, and a non-atomic NE flow fulfills Wardrop's first principle. So the sets of atomic and non-atomic NE flows of Γ do not overlap.

Clearly, $\tilde{f}_{at}$ is also the unique mixed NE flow, since the expected cost of the upper arc u is always larger than the constant cost of the lower arc ℓ when either of the two agents uses the upper arc. Hence, neither the set of mixed NE flows nor the set of their expectations need to intersect the set of non-atomic NE flows. Moreover, by a little calculation, one can also see that neither the set of mixed WE flows (Remark 1) nor the set of their expectations intersects the sets of atomic and non-atomic NE flows in this

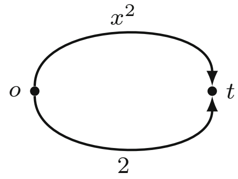

Fig. 1 An example of atomic, non-atomic, and mixed NE flows

Example. This means that these equilibrium notions are mutually different, although atomic and mixed NE flows coincide in this Example.

2.5 The price of anarchy

Since we consider non-atomic, atomic and mixed NE flows, we define four PoAs in (2.8)–(2.11), in which $\tilde{f}_{nat}$, f^*_{nat} and f^*_{at} are an arbitrary non-atomic NE flow, an arbitrary non-atomic SO flow and an arbitrary atomic SO flow, respectively. We call $\rho_{at}(\Gamma)$ the *atomic* PoA, $\rho_{nat}(\Gamma)$ the *non-atomic* PoA, $\rho_{mix}(\Gamma)$ the *mixed* PoA, and $\rho(\tilde{f}_{ran}, \Gamma)$ the *random* PoA of the mixed NE flow $\tilde{f}_{ran}$. Here, we recall that non-atomic NE flows are essentially unique.

$$\rho_{at}(\Gamma) := \max \left\{ \frac{C(\tilde{f}_{at}, \Gamma)}{C(f^*_{at}, \Gamma)} : \tilde{f}_{at} \text{ is an atomic NE flow of } \Gamma \right\} \tag{2.8}$$

$$\rho_{nat}(\Gamma) := \frac{C(\tilde{f}_{nat}, \Gamma)}{C(f^*_{nat}, \Gamma)} \tag{2.9}$$

$$\begin{aligned} \rho_{mix}(\Gamma) &:= \max \left\{ \frac{\mathbb{E}_{\tilde{\Pi}}[C(\tilde{f}_{ran}, \Gamma)]}{\mathbb{E}_{\Pi^*}[C(f^*_{ran}, \Gamma)]} : \tilde{f}_{ran}, f^*_{ran} \text{ are mixed NE and SO flows} \right\} \\ &= \max \left\{ \frac{\mathbb{E}_{\tilde{\Pi}}[C(\tilde{f}_{ran}, \Gamma)]}{C(f^*_{at}, \Gamma)} : \tilde{f}_{ran} \text{ is a mixed NE flow of } \Gamma \right\} \end{aligned} \tag{2.10}$$

$$\rho(\tilde{f}_{ran}, \Gamma) := \frac{C(\tilde{f}_{ran}, \Gamma)}{C(f^*_{at}, \Gamma)} \tag{2.11}$$

Note that $\rho(\tilde{f}_{ran}, \Gamma)$ is a random variable and thus differs from the deterministic values $\rho_{at}(\Gamma)$, $\rho_{nat}(\Gamma)$ and $\rho_{mix}(\Gamma)$. Moreover, $\rho_{nat}(\Gamma)$ differs from $\rho_{at}(\Gamma)$ and $\rho_{mix}(\Gamma)$, see Example 1, in which $\rho_{nat}(\Gamma) = \frac{18}{18-\sqrt{6}} > \rho_{at}(\Gamma) = \rho_{mix}(\Gamma) = 1$. Although $\rho_{at}(\Gamma)$ and $\rho_{mix}(\Gamma)$ coincide in Example 1, they differ in general, and $\rho_{mix}(\Gamma) \geq \rho_{at}(\Gamma)$. In particular, neither $\rho_{nat}(\Gamma) \geq \rho_{at}(\Gamma)$ nor $\rho_{nat}(\Gamma) \geq \rho_{mix}(\Gamma)$ holds in general, see, e.g., Christodoulou and Koutsoupias [7]. Thus the known convergence results of the non-atomic PoA in Colini-Baldeschi et al. [8–10] and Wu et al. [47] do not naturally carry over to random, atomic and mixed PoAs.

Due to the "no free arc" assumption in (G3), all PoAs are different from $\frac{0}{0}$, and take values in $[1, \infty)$. This follows from Lemma 1, and the fact that the non-atomic SO cost is strictly positive, see [46].

3 Convergence results of the PoAs in atomic congestion games

We now analyze the convergence of the PoAs for atomic congestion games with *polynomial* cost functions, i.e., all $\tau_a(\cdot)$ have the form

$$\tau_a(x) = \sum_{l=0}^{\beta_a} \eta_{a,l} \cdot x^{\beta_a - l}, \quad \forall x \in [0, \infty), \tag{3.1}$$

where $\beta_a \geq 0$ is an *integer* degree, and $\eta_{a,l}$, $l = 0, \ldots, \beta_a$, $a \in A$, are the coefficients. Since all $\tau_a(\cdot)$ are nondecreasing and no arc can be used for free, see (G3), all leading coefficients $\eta_{a,0}$, $a \in A$, are strictly *positive*. We assume, w.l.o.g., that all other coefficients $\eta_{a,l}$ are also *non-negative*. This will simplify our analysis. Note that this is not restrictive, and our results carry over to arbitrary polynomial cost functions. We will come back to this later in Sects. 3.1.1, 3.1.2 and 3.2, respectively.

3.1 Convergence results for polynomial cost functions of the same degree

We consider first polynomial cost functions $\tau_a(\cdot)$ of the same degree $\beta_a \equiv \beta \geq 0$, i.e., they have the form (3.2)

$$\tau_a(x) = \sum_{l=0}^{\beta} \eta_{a,l} \cdot x^{\beta - l} \quad \forall x \geq 0 \; \forall a \in A. \tag{3.2}$$

This covers BPR cost functions, which are of the simpler form $\eta_{a,0} \cdot x^\beta + \eta_{a,\beta}$ and frequently used in urban traffic to model travel latency, see [5].

With these cost functions, the total cost of a non-atomic SO flow is at least $\frac{T^{\beta+1} \cdot \eta_{0,\min}}{|\mathcal{P}|^{\beta+1}} > 0$ when $T > 0$, where $\eta_{0,\min} := \min_{a \in A} \eta_{a,0} > 0$, see [46]. Note that there is at least one path with a flow value of at least $\frac{T}{|\mathcal{P}|}$ in an arbitrary non-atomic SO flow. Note also, that $x \cdot \tau_a(x) \geq \eta_{0,\min} \cdot x^{\beta+1}$ for all $a \in A$ and all $x \geq 0$.

3.1.1 An upper bound for the atomic PoA

Theorem 1 presents an upper bound for the atomic PoA in congestion games with polynomial cost functions of the same degree, see (3.2). Here, $\eta_{\max} := \max\{\eta_{a,l} : a \in A, l = 0, \ldots, \beta\} \geq \eta_{0,\min} > 0$, and $\kappa := \beta \cdot \eta_{\max} \cdot \left(1 + \sum_{l=1}^{\beta} \frac{1}{T^l}\right) > 0$, which is a *Lipschitz bound* for the Lipschitz continuous functions $\frac{\tau_a(T \cdot x)}{T^\beta}$ on the compact interval $[0, 1]$, i.e., κ satisfies the condition that $|\frac{\tau_a(T \cdot x)}{T^\beta} - \frac{\tau_a(T \cdot y)}{T^\beta}| \leq \kappa \cdot |x - y|$ for all $x, y \in [0, 1]$ and all $a \in A$.

Theorem 1 *Consider an arbitrary congestion game* $\Gamma = (\tau, \mathcal{U}, d)$ *with cost functions* $\tau_a(\cdot)$ *of the form* (3.2). *If* Γ *has atomic NE flows, then*

$$\rho_{at}(\Gamma) \le 1 + \frac{\beta \cdot \eta_{\max} \cdot |\mathcal{P}|^{\beta+1}}{\eta_{0,\min}} \cdot \sum_{l=1}^{\beta} \frac{1}{T^l} + \frac{|A| \cdot \kappa \cdot |\mathcal{P}|^{\beta+1}}{\eta_{0,\min}} \cdot \sqrt{|\mathcal{P}| \cdot |A| \cdot \frac{d_{max}}{T}} + \frac{|A| \cdot \kappa \cdot |\mathcal{P}|^{\beta+2}}{\eta_{0,\min}} \cdot \frac{d_{max}}{T}.$$

Here, we use the convention that $\sum_{l=1}^{\beta} \frac{1}{T^l} = 0$ *when* $\beta = 0$.

The upper bound holds for all T and d_{max}, and converges to 1 at a rate of $O(\frac{1}{T}) + O(\sqrt{\frac{d_{max}}{T}})$ as $T \to \infty$ and $\frac{d_{max}}{T} \to 0$. So the atomic PoA decays to 1 quickly when Γ has atomic NE flows. Examples 2–3 show that the conditions "$T \to \infty$" and "$\frac{d_{max}}{T} \to 0$" are necessary for this convergence.

Example 2 Consider an unweighted congestion game Γ with the network of Fig. 1, but cost functions x and $x + 1$ for the upper and lower arc, respectively.

Assume that Γ has $|\mathcal{U}| = 4 \cdot n$ agents with $\frac{1}{4 \cdot n}$ demand each. Then $T \equiv 1$ and $d_{max} = \frac{1}{4 \cdot n}$. Clearly, Γ has only one atomic NE flow $\tilde{f}_{at}$, in which all agents use the upper arc. So $C(\tilde{f}_{at}, \Gamma) = 1$. Γ has also a unique atomic SO flow f^*_{at}, in which $3 \cdot n$ agents use the upper arc and the remaining n agents use the lower arc. Then $C(f^*_{at}, \Gamma) = \frac{7}{8}$, and $\rho_{at}(\Gamma) = \frac{8}{7}$ for all n, which does not converge to 1 when only $\frac{d_{max}}{T} = d_{max} \to 0$.

Example 3 Consider an unweighted congestion game Γ again with the network of Fig. 1, but now with cost functions x and $2 \cdot x$ for the upper and lower arc, respectively. Assume that there are two agents with demand n each. Then $T = 2 \cdot n$, which tends to ∞ as $n \to \infty$. However, $\frac{d_{max}}{T} \to \frac{1}{2} > 0$ as $n \to \infty$. Obviously, Γ has only one atomic SO flow f^*_{at}, in which one agent uses the upper and the other the lower arc. So $C(f^*_{at}, \Gamma) = 3 \cdot n^2$. However, Γ has two atomic NE flows. One atomic NE flow is just the unique SO flow. In the other atomic NE flow, both agents use the upper arc, and its total cost is $4 \cdot n^2$. Consequently, $\rho_{at}(\Gamma) = \frac{4}{3} \not\to 1$ as $T = 2 \cdot n \to \infty$.

We now prove Theorem 1 with the technique of scaling from Colini-Baldeschi et al. [10] and Wu et al. [47].

Definition 1 (Scaled games, Wu et al. [47]) Consider an arbitrary congestion game $\Gamma = (\tau, \mathcal{U}, d)$ with arbitrary cost functions, and an arbitrary constant $g > 0$. The *scaled game* of Γ w.r.t. *scaling factor* g is the congestion game $\Gamma^{[g]} = \big(\tau^{[g]}, \mathcal{U}, \bar{d}\big)$ whose cost function vector $\tau^{[g]} := (\tau_a^{[g]})_{a \in A}$ has a component $\tau_a^{[g]}(x) := \frac{\tau_a(x \cdot T)}{g}$ for each pair $(a, x) \in A \times [0, 1]$, and whose demand vector $\bar{d} = (\bar{d}_{k,i})_{i \in \mathcal{U}_k, k \in \mathcal{K}}$ has a component $\bar{d}_{k,i} := \frac{d_{k,i}}{T}$ for each $i \in \mathcal{U}_k$ and each $k \in \mathcal{K}$.

Lemma 2 shows that scaling does not change the four PoAs. We omit the straightforward proof. Note that a flow f of Γ corresponds to a flow $f^{[g]} := \frac{f}{T}$ of $\Gamma^{[g]}$, and $C(f, \Gamma) = C(f^{[g]}, \Gamma^{[g]}) \cdot g \cdot T$.

Lemma 2 *Consider an arbitrary congestion game Γ, an arbitrary mixed NE flow $\tilde{f}_{ran}$ of Γ, and an arbitrary scaling factor $g > 0$. Let $\Gamma^{[g]}$ be the scaled game with factor g. Then $\rho_{at}(\Gamma^{[g]}) = \rho_{at}(\Gamma)$, $\rho_{nat}(\Gamma^{[g]}) = \rho_{nat}(\Gamma)$, and $\rho_{mix}(\Gamma^{[g]}) = \rho_{mix}(\Gamma)$. Moreover, $\tilde{f}_{ran}^{[g]} := \frac{\tilde{f}_{ran}}{T}$ is a mixed NE flow of the scaled game $\Gamma^{[g]}$, and $\rho(\tilde{f}_{ran}^{[g]}, \Gamma^{[g]}) = \rho(\tilde{f}_{ran}, \Gamma)$.*

Lemma 2 enables us to prove Theorem 1 by bounding $\rho_{at}(\Gamma^{[g]})$ instead of $\rho_{at}(\Gamma)$. We can thus purely concentrate on the influence of $\frac{d_{max}}{T}$ on the convergence, as the total demand of $\Gamma^{[g]}$ is $\bar{T} = T(\mathcal{U}, \bar{d}) := \sum_{k\in\mathcal{K}, i\in\mathcal{U}_k} \bar{d}_{k,i} = 1$. However, the scaling factor g must be chosen carefully, so as to ensure that the total cost in $\Gamma^{[g]}$ is moderate, i.e., neither too large nor too small. Following [47], we use $g := T^\beta$ for polynomial cost functions of the same degree β. Then $\Gamma^{[g]}$ has the *scaled* cost function

$$\tau_a^{[g]}(x) = \frac{\sum_{l=0}^{\beta} \eta_{a,l} \cdot (T \cdot x)^{\beta-l}}{g} = \eta_{a,0} \cdot x^\beta + \sum_{l=1}^{\beta} \frac{\eta_{a,l}}{T^l} \cdot x^{\beta-l} \tag{3.3}$$

for arc $a \in A$, the *bounded* demand $\bar{d}_{k,i} = \frac{d_{k,i}}{T} \in [0, 1]$ for $i \in \mathcal{U}_k$, and the *bounded* demand $\bar{d}_k := \frac{d_k}{T} \in [0, 1]$ for $k \in \mathcal{K}$. Consequently, each flow $f^{[g]}$ of $\Gamma^{[g]}$ has *bounded* arc flow values $f_a^{[g]} \in [0, 1]$, and $C(f^{[g]}, \Gamma^{[g]}) \geq \frac{\eta_{0,\min}}{|\mathcal{P}|^{\beta+1}}$.

Definition (2.8) of the atomic PoA and Lemma 1 together imply that

$$\rho_{at}(\Gamma) = \rho_{at}(\Gamma^{[g]}) \leq \rho_{nat}(\Gamma^{[g]}) + \frac{|\max_{\tilde{f}_{at}^{[g]}} C(\tilde{f}_{at}^{[g]}, \Gamma^{[g]}) - C(\tilde{f}_{nat}^{[g]}, \Gamma^{[g]})|}{C(f_{nat}^{*[g]}, \Gamma^{[g]})}, \tag{3.4}$$

where $\tilde{f}_{nat}^{[g]}$ and $f_{nat}^{*[g]}$ are arbitrary non-atomic NE and SO flows of $\Gamma^{[g]}$, respectively, and the maximization is taken over all atomic NE flows $\tilde{f}_{at}^{[g]}$ of $\Gamma^{[g]}$. With (3.4), we can then prove Theorem 1 by upper bounding

$$|\max_{\tilde{f}_{at}^{[g]}} C(\tilde{f}_{at}^{[g]}, \Gamma^{[g]}) - C(\tilde{f}_{nat}^{[g]}, \Gamma^{[g]})| \tag{3.5}$$

and $\rho_{nat}(\Gamma^{[g]})$, respectively. Here, we observe that $C(f_{nat}^{*[g]}, \Gamma^{[g]}) \geq \frac{\eta_{0,\min}}{|\mathcal{P}|^{\beta+1}} > 0$. To that end, we need the notion of ϵ*-approximate non-atomic NE flow* and a result from Wu and Möhring [46].

Definition 2 We call an arbitrary non-atomic flow f_{nat} of Γ an ϵ*-approximate non-atomic NE flow* for a constant $\epsilon > 0$ if $\sum_{a\in A} \tau_a(f_{nat,a}) \cdot (f_{nat,a} - f'_{nat,a}) \leq \epsilon$ for an arbitrary non-atomic flow f'_{nat} of Γ.

Wu and Möhring [46] have shown that the total cost difference between ϵ-approximate and accurate non-atomic NE flows is in $O(\sqrt{\epsilon})$, see Lemma 3.

Lemma 3 (Wu and Möhring [46]) *Consider an arbitrary congestion game $\Gamma = (\tau, \mathcal{U}, d)$ with a total demand of* 1 *and an arbitrary ϵ-approximate non-atomic NE* *flow* $\tilde{f}^{\epsilon}_{nat}$. *If all cost functions are Lipschitz continuous (or Lipschitz bounded) on* $[0,1]$ *with a Lipschitz constant* $\kappa > 0$, *i.e.,* $|\tau_a(x) - \tau_a(y)| \le \kappa \cdot |x - y|$ *for all* $(a, x, y) \in A \times [0,1]^2$, *then* $|C(\tilde{f}_{nat}, \Gamma) - C(\tilde{f}^{\epsilon}_{nat}, \Gamma)| \le |A| \cdot \sqrt{\kappa \cdot \epsilon} + \epsilon$, *and* $|\tau_a(\tilde{f}_{nat,a}) - \tau_a(\tilde{f}^{\epsilon}_{nat,a})| \le \sqrt{\kappa \cdot \epsilon}$ *for all* $a \in A$ *and all non-atomic NE flows* $\tilde{f}_{nat}$.

Lemma 4 below shows that $\tilde{f}^{[g]}_{at}$ is an $O(\frac{d_{max}}{T})$-approximate non-atomic NE flow of $\Gamma^{[g]}$. Then Lemma 3 yields a desired upper bound for (3.5), see Lemma 4c. We move the proof of Lemma 4 to Appendix A.3.

Lemma 4 *Consider an arbitrary congestion game* Γ *as in Theorem 1. Let* $\Gamma^{[g]}$ *be its scaled game with factor* $g = T^{\beta}$, *and let* $\tilde{f}^{[g]}_{at}$ *and* $\tilde{f}^{[g]}_{nat}$ *be arbitrary atomic and non-atomic NE flows, respectively. Then:*

(a) $\tau^{[g]}_{p}\left(\tilde{f}^{[g]}_{at}\right) \le \tau^{[g]}_{p'}\left(\tilde{f}^{[g]}_{at}\right) + \frac{|A| \cdot \kappa \cdot d_{max}}{T}$ *for all* $k \in \mathcal{K}$ *and all* $p, p' \in \mathcal{P}_k$ *with* $f^{[g]}_{at,p} > 0$.

(b) $\tilde{f}^{[g]}_{at}$ *is a* $\frac{|\mathcal{P}| \cdot |A| \cdot \kappa \cdot d_{max}}{T}$*-approximate non-atomic NE flow of* $\Gamma^{[g]}$.

(c) $|C(\tilde{f}^{[g]}_{at}, \Gamma^{[g]}) - C(\tilde{f}^{[g]}_{nat}, \Gamma^{[g]})| \le |A| \cdot \kappa \cdot \sqrt{|\mathcal{P}| \cdot |A| \cdot \frac{d_{max}}{T}} + |\mathcal{P}| \cdot |A| \cdot \kappa \cdot \frac{d_{max}}{T}$.

Lemma 5 yields an upper bound for $\rho_{nat}(\Gamma^{[g]})$, which results in a convergence rate of $O(\frac{1}{T})$. Note that Wu et al. [47] have shown a stronger convergence rate of $o(\frac{1}{T^{\beta}})$ for BPR cost functions, and that Colini-Baldeschi et al. [10] have shown a similar rate as in Lemma 5 for arbitrary polynomial cost functions under the condition that $\frac{d_k}{T} \ge \xi_k > 0$ for some constant ξ_k independent of T for each $k \in \mathcal{K}$. We move the proof of Lemma 5 to Appendix A.4.

Lemma 5 *Consider an arbitrary congestion game* Γ *as in Theorem 1. Let* $\Gamma^{[g]}$ *be the scaled game with scaling factor* $g = T^{\beta}$. *Then* $\rho_{nat}(\Gamma) = \rho_{nat}(\Gamma^{[g]}) \le 1 + \frac{\beta \cdot \eta_{\max} \cdot |\mathcal{P}|^{\beta+1}}{\eta_{0,\min}} \cdot \sum_{l=1}^{\beta} \frac{1}{T^{l}}$.

Theorem 1 then follows from Lemma 2, (3.4), Lemmas 4c and 5.

The above proofs build essentially on inequality (3.4), Lemma 3 and the Lipschitz continuity of the scaled cost functions $\tau^{[g]}_a(\cdot)$ on $[0,1]$, but not on the sign of the coefficients $\eta_{a,l}$, $l = 1, \ldots, \beta$, $a \in A$. Thus Theorem 1 indeed carries over to arbitrary polynomial cost functions of the same degree $\beta \ge 0$.

When $\eta_{a,l} < 0$ for some terms $l = 1, \ldots, \beta$ and some arcs $a \in A$, then $\frac{\eta_{0,\min}}{|\mathcal{P}|^{\beta+1}}$ may be larger than $C(f^{*[g]}_{nat}, \Gamma^{[g]})$. Instead, $C(f^{*[g]}_{nat}, \Gamma^{[g]})$ can be bounded from below by $\min_{a \in A} \frac{1}{|\mathcal{P}|} \cdot \tau^{[g]}_a(\frac{1}{|\mathcal{P}|}) \in \Theta(1)$. The Lipschitz bound for the scaled cost functions is still $\kappa = \beta \cdot \eta_{\max} \cdot \left(1 + \sum_{l=1}^{\beta} \frac{1}{T^{l}}\right) > 0$, but with $\eta_{\max} := \{|\eta_{a,l}| : a \in A, l = 0, 1, \ldots, \beta\}$. Lemma 4 then still holds, since its proof in Appendix A.3 does not involve the sign of coefficients $\eta_{a,l}$, but only the Lipschitz continuity of the scaled cost functions on $[0,1]$. Although the proof of Lemma 5 in Appendix A.4 does involve the sign of coefficients $\eta_{a,l}$, it can be adapted accordingly.

3.1.2 Upper bounds for the mixed PoA and the random PoA

Theorem 2 below proves similar upper bounds for $\rho(\tilde{f}_{ran}, \Gamma)$ and $\rho_{mix}(\Gamma)$, respectively, in terms of T, $\frac{d_{max}}{T}$ and constants M_i, $i = 1, \ldots, 5$. We hide the detailed values of these constants M_i in Theorem 2, since they are complicated expressions. Interested readers may find their values in the proof. When $T \to \infty$ and $\frac{d_{max}}{T} \to 0$, these upper bounds converge (with an overwhelming probability for $\rho(\tilde{f}_{ran}, \Gamma)$) to 1 at a rate of $O(\frac{1}{T}) + O(\frac{d_{max}^{1/6}}{T^{1/6}})$. Note that $\rho_{mix}(\Gamma)$ converges more slowly than $\rho_{at}(\Gamma)$ since $\rho_{at}(\Gamma) \le \rho_{mix}(\Gamma)$.

Theorem 2 *Consider the same congestion game Γ as in Theorem 1. Let $\tilde{f}_{ran}$ be an arbitrary mixed NE flow of Γ. Then the following statements hold.*

(a) The random event "$\rho(\tilde{f}_{ran}, \Gamma) \le 1 + M_1 \cdot \frac{1}{T} + M_2 \cdot \frac{d_{max}^{1/6}}{T^{1/6}}$" occurs with a probability of at least $1 - M_3 \cdot \frac{d_{max}^{1/3}}{T^{1/3}}$.

(b) $\rho_{mix}(\Gamma) \le 1 + M_4 \cdot \frac{1}{T} + M_5 \cdot \frac{d_{max}^{1/6}}{T^{1/6}}$

Herein, $M_i > 0$, $i = 1, \ldots, 5$, are constants independent of d_{max} and T.

We also prove Theorem 2 with the scaled game $\Gamma^{[g]}$ and Lemma 2. Let $\tilde{f}_{nat}^{[g]}$ and $f_{nat}^{*[g]}$ be an arbitrary non-atomic NE flow and an arbitrary non-atomic SO flow of $\Gamma^{[g]}$, respectively. We obtain by Lemma 1, (2.10) and (2.11) that

$$\rho_{mix}(\Gamma^{[g]}) \le \rho_{nat}(\Gamma^{[g]}) + \frac{|\max_{\tilde{f}_{ran}^{[g]}} \mathbb{E}_{\tilde{\Pi}}[C(\tilde{f}_{ran}^{[g]}, \Gamma^{[g]})] - C(\tilde{f}_{nat}^{[g]}, \Gamma^{[g]})|}{C(f_{nat}^{*[g]}, \Gamma^{[g]})}, \quad (3.6)$$

and that

$$\rho(\tilde{f}_{ran}^{[g]}, \Gamma^{[g]}) \le \rho_{nat}(\Gamma^{[g]}) + \frac{|C(\tilde{f}_{ran}^{[g]}, \Gamma^{[g]}) - C(\tilde{f}_{nat}^{[g]}, \Gamma^{[g]})|}{C(f_{nat}^{*[g]}, \Gamma^{[g]})}, \quad (3.7)$$

where $\tilde{f}_{ran}^{[g]}$ is an arbitrary mixed NE flow of $\Gamma^{[g]}$. Using Lemma 5, we now only need to derive upper bounds for the numerators of the two fractions in (3.6) and (3.7), respectively.

Lemma 6a below shows that the expected flow $\mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran}^{[g]})$ of a mixed NE $\tilde{f}_{ran}^{[g]}$ is an ϵ-approximate non-atomic NE flow with $\epsilon \in O(\frac{d_{max}^{1/3}}{T^{1/3}})$. Lemma 3 then yields $|C(\mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran}^{[g]}), \Gamma^{[g]}) - C(\tilde{f}_{nat}^{[g]}, \Gamma^{[g]})| \in O(\frac{d_{max}^{1/6}}{T^{1/6}})$. Then Lemma 6b–c upper bound the total cost difference between a mixed NE flow $\tilde{f}_{ran}^{[g]}$ and its expected flow $\mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran}^{[g]})$ both in expectation and as a random variable. Moreover, Lemma 6 together with Lemma 5 and (3.6)–(3.7) prove Theorem 2.

We move the detailed proof of Lemma 6 to Appendix A.5.

Lemma 6 *Consider the congestion game Γ in Theorem 2, and the scaling factor $g = T^{\beta}$. Let $\Gamma^{[g]}$ be the scaled game with factor g, and let $\tilde{f}_{ran}^{[g]}$ be an arbitrary mixed NE flow of $\Gamma^{[g]}$ with mixed profile $\tilde{\Pi}$.*

(a) *When $\beta > 0$, then the expected flow $\mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran}^{[g]})$ is an ϵ-approximate non-atomic NE flow with $\epsilon = 3 \cdot |\mathcal{P}| \cdot \kappa \cdot |A| \cdot \big(1 + \frac{|A|}{4 \cdot \beta}\big) \cdot \big(\frac{d_{max}}{T}\big)^{1/3}$, and $|C(\tilde{f}_{nat}^{[g]}, \Gamma^{[g]}) - C\big(\mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran}^{[g]}), \Gamma^{[g]}\big)| \leq |A| \cdot \sqrt{\kappa \cdot \epsilon} + \epsilon \in O(\frac{d_{max}^{1/6}}{T^{1/6}})$ for an arbitrary non-atomic NE flow $\tilde{f}_{nat}^{[g]}$ of $\Gamma^{[g]}$. When $\beta = 0$, then $\mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran}^{[g]})$ is a non-atomic NE flow of $\Gamma^{[g]}$.*

(b) *Consider an arbitrary constant $\delta \in (0, 1/2)$. The event "$|C(\tilde{f}_{ran}^{[g]}, \Gamma^{[g]}) - C\big(\mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran}^{[g]}), \Gamma^{[g]}\big)| \leq |A| \cdot \big(\kappa + \eta_{\max} \cdot \sum_{l=0}^{\beta} \frac{1}{T^l}\big) \cdot \big(\frac{d_{max}}{T}\big)^{\delta}$" occurs with a probability of at least $1 - \frac{|A|}{4} \cdot \big(\frac{d_{max}}{T}\big)^{1-2\cdot\delta}$.*

(c) $\Big|\mathbb{E}_{\tilde{\Pi}}\big[C(\tilde{f}_{ran}^{[g]}, \Gamma^{[g]})\big] - C\big(\mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran}^{[g]}), \Gamma^{[g]}\big)\Big| \leq |A| \cdot \big(\kappa + \big(1 + \frac{|A|}{4}\big) \cdot \eta_{\max} \cdot \sum_{l=0}^{\beta} \frac{1}{T^l}\big) \cdot \big(\frac{d_{max}}{T}\big)^{1/3}$.

Similar to the proof for Lemma 4 in Appendix A.3, the proof of Lemma 6 in Appendix A.5 does neither involve the sign of the coefficients $\eta_{a,l}$, but only the Lipschitz continuity of the scaled cost functions on [0, 1] and the finite upper bound $\max_{a \in A} \tau_a^{[g]}(1) \in \Theta(1)$. Hence, Lemma 6 carries also over to arbitrary polynomial cost functions of the same degree, and so does Theorem 2.

Note that Cominetti et al. [11] have shown that the mixed NE flow $\tilde{f}_{ran}$ of an atomic congestion game Γ converges in distribution to a non-atomic NE flow $\tilde{f}_{nat}$ of a limit non-atomic congestion game $\Gamma^{(\infty)}$ when the cost functions τ_a are strictly increasing, $T \to T_0$ for a constant $T_0 > 0$, $d_{max} \to 0$, and the number $|\mathcal{U}|$ of agents tends to ∞. Combined with the scaling technique, this may imply also that the mixed PoA in the scaled game $\Gamma^{[g]}$ converges to 1 for polynomial cost functions of the same degree when $\frac{d_{max}}{T} \to 0$ as $T \to \infty$, although the cost functions of the atomic congestion games in the analysis of Cominetti et al. [11] are fixed and equal those of the limit non-atomic congestion game, and although the scaled cost functions $\tau_a^{[g]}$ here depend on T and vary with the growth of T. While implying a similar convergence, we aim at upper bounding the mixed and random PoAs, and so have results for arbitrary demand vectors d, i.e., neither need $T \to \infty$ nor need $\frac{d_{max}}{T} \to 0$ in the proofs. Moreover, the results of Cominetti et al. [11] do not imply the convergence of the mixed PoA in atomic congestion games with arbitrary polynomial cost functions for growing total demand, since then the atomic congestion games cannot be scaled to have a unified limit non-atomic congestion game for all O/D pairs, see [47].

3.2 Concergence results for polynomial cost functions with arbitrary degrees

We consider now polynomial cost functions with arbitrary degrees, i.e., $\beta_a \neq \beta_{a'}$ may hold for some arcs $a \neq a'$. Example 4 below shows that the conditions "$\frac{d_{max}}{T} \to 0$" and "$T \to \infty$" are no longer sufficient for the convergence of $\rho_{mix}(\Gamma)$ and $\rho_{at}(\Gamma)$ in this case.

Example 4 Consider a congestion game Γ with the network of Fig. 2. Γ has two non-overlapping O/D pairs (o_1, t_1) and (o_2, t_2), and both of them have two parallel arcs. Assume that (o_1, t_1) has $2 \cdot \sqrt{n}$ agents with each a demand of $\sqrt{n}$, and that (o_2, t_2)

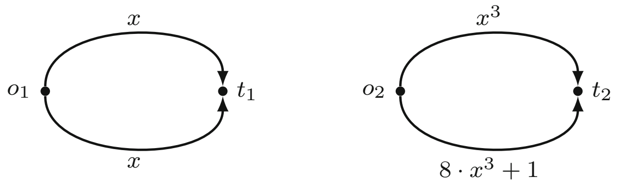

Fig. 2 The PoA need not converge to 1

has 2 agents with the same demand of $\sqrt{n}$ each. So $d_{max} = \sqrt{n}$. Then, as $n \to \infty$, $T = 2 \cdot n + 2 \cdot \sqrt{n} \to \infty$ and $\frac{d_{max}}{T} = \frac{\sqrt{n}}{2 \cdot n + 2 \cdot \sqrt{n}} \to 0$.

However, $\rho_{at}(\Gamma) \to \frac{16}{9} > 1$ as $n \to \infty$. This follows since Γ has the worst-case total cost of $2 \cdot n + 16 \cdot n^2$ for atomic NE flows, and the total cost of $2 \cdot n + 9 \cdot n^2 + \sqrt{n}$ for atomic SO flows when n is large.

While the game Γ in Example 4 is artificial, it shows that the convergence of the PoAs can be ruined by O/D pairs with small demands but polynomial cost functions of higher degrees, since they may dominate the PoAs completely when $T \to \infty$ and d_{max} is unbounded. To ensure the convergence of the PoAs for polynomial cost functions of arbitrary degrees, we may thus need to impose a stronger condition that d_{max} is bounded when $T \to \infty$. Theorem 3 below confirms this.

Theorem 3 *Consider an arbitrary congestion game Γ with cost functions $\tau_a(\cdot)$ defined in (3.1). Assume that d_{max} is bounded from above by a constant $\upsilon > 0$ independent of T. Then the following statements hold.*

(a) $\frac{\max_{\tilde{f}_{ran}} C\left(\mathbb{E}_{\tilde{\Pi}}\left(\tilde{f}_{ran}\right), \Gamma\right)}{C(f^*_{nat}, \Gamma)} \to 1$ *as* $T \to \infty$, *where the maximization in the numerator is taken over all possible mixed NE flows* $\tilde{f}_{ran}$ *of* Γ.

(b) $\rho_{mix}(\Gamma) \to 1$ *as* $T \to \infty$.

(c) If Γ *has atomic NE flows for all demand vectors* d, *if* $\beta_{max} = \max_{a \in A} \beta_a > 0$, *and if* $\frac{d_k}{T} = \frac{\sum_{i \in \mathcal{U}_k} d_{k,i}}{T} \geq \xi_k > 0$ *for all* $k \in \mathcal{K}$ *and some constants* $\xi_k > 0$ *independent of* T, *then* $\rho_{at}(\Gamma) = 1 + O(T^{-\frac{1}{2 \cdot \beta_{max}}})$.

Theorem 3a states that the expected flow $\mathbb{E}_{\tilde{\Pi}}[\tilde{f}_{ran}]$ of a mixed NE flow $\tilde{f}_{ran}$ is as efficient as a non-atomic SO flow for large T when the polynomial cost functions have arbitrary degrees and d_{max} is bounded. Theorem 3b then shows that $\rho_{mix}(\Gamma)$ converges to 1 for growing T in this more general case. Hence, if the atomic NE flows exist, then $\rho_{at}(\Gamma) \to 1$ as $T \to \infty$, since $\rho_{at}(\Gamma) \leq \rho_{mix}(\Gamma)$. In addition to the pure convergence in Theorems 3a–b, 3c shows that $\rho_{at}(\Gamma)$ converges at a rate of $O(T^{-\frac{1}{2 \cdot \beta_{max}}})$. This demonstrates how fast the convergence of the PoAs can be in this more general case, when each O/D pair demand d_k has a positive ratio $\frac{d_k}{T}$ as $T \to \infty$. So far, we are unable to remove this restrictive condition, as we do not see a way to compute a concrete upper bound in terms of $\frac{1}{T}$ for $\rho_{at}(\Gamma)$ when the cost functions have *different* degrees and the O/D pairs have significantly *asynchronous* demand growth rates.

Theorem 3c can be proved by a scaling technique similar to the proofs of Theorem 1 and Theorem 2. However, due to the absence of a unified scaling factor, similar arguments will not be applicable in the proofs of Theorem 3a–b, for which we need a more sophisticated technique called *asymptotic decomposition* developed by Wu et al. [47]. In fact, Example 4 has shown that different O/D pairs $k \in \mathcal{K}$ may have significantly *discrepant* influences on the limits of the PoAs for polynomial cost functions with arbitrary degrees. These discrepant influences are caused by the different degrees of polynomial cost functions and the asynchronous growth rates of the demands of the O/D pairs. The asymptotic decomposition technique enables us to capture these discrepant influences from different O/D pairs $k \in \mathcal{K}$. It puts O/D pairs $k \in \mathcal{K}$ with a similar influence on the limits of the PoAs together to form a "subgame", then analyzes the resulting subgames independently and combines the convergence results for these subgames to a convergence result for the whole game Γ. Interested readers may refer to Wu et al. [47] for a detailed introduction of this general technique. We move a description of the asymptotic decomposition and the very long proof of Theorem 3 to Appendix A.6 in order to save space and improve readability.

Although we have assumed at the beginning of Sect. 3 that the polynomial cost functions have only non-negative coefficients, the proof of Theorem 3a–b in Appendix A.6 is essentially independent of this condition. The proof of Theorem 3c uses the non-negativity of the coefficients to obtain explicit lower and upper bounds of the scaled cost function values on the domain [0, 1], which carries also over to polynomials of arbitrary degrees when we slightly adapt the constants in those bounds. Hence, the convergence results in Theorem 3 hold for arbitrary polynomial cost functions, even with non-negative real-valued exponents.

With the asymptotic decomposition, the convergence results for the non-atomic PoA in Wu et al. [47], and Lemma 1, we can actually show in the proof that all the flows, $\tilde{f}_{ran}$, $\mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran})$, $\tilde{f}_{nat}$, f^*_{at}, f^*_{nat}, are equally efficient when $T \to \infty$ and d_{max} is bounded, see (A.29) in Appendix A.6. In particular, to obtain the convergence results in Theorem 3a–b, we have considered a mixed NE flow as an approximate mixed WE flow (see Remark 1) in the proof, and so these convergence results carry also over to the "PoA" of mixed WE flows. Hence, we need not distinguish between atomic and non-atomic congestion games for quantifying the inefficiency of selfish choices of users, when the cost functions are polynomials, the total demand T is large, and the individual maximum demand d_{max} is bounded.

4 Summary

We have studied the inefficiency of both pure and mixed Nash equilibria in atomic congestion games with unsplittable demands.

When the cost functions are polynomials of the same degree, we derive upper bounds for the atomic, mixed and random PoAs, respectively. These upper bounds tend to 1 quickly as $T \to \infty$ and $\frac{d_{max}}{T} \to 0$.

When the cost functions are polynomials of arbitrary degrees and d_{max} is bounded, we show that the mixed PoA converges again to 1 as $T \to \infty$. Moreover, we illustrate that this need not hold when d_{max} is unbounded. To demonstrate the convergence rates in this more general case, we show in addition that the atomic PoA converges to 1 at a rate of $O(T^{-\frac{1}{2\cdot\beta_{max}}})$ under the relatively restrictive condition that all O/D pairs have demand proportions $\frac{d_k}{T}$ that do not vanish when $T \to \infty$. However, it is still open and challenging to obtain concrete convergence rates without this condition.

Nevertheless, our results already imply, under rather mild conditions, that pure and mixed Nash equilibria in atomic congestion games with large unsplittable demands need not be bad. This, together with studies of Colini-Baldeschi et al. [8–10] and Wu et al. [47], indicates that the selfish choice of strategies leads to a near-optimal behavior in arbitrary congestion games with large total demands, regardless whether users choose mixed or pure strategies, and whether the demands are splittable or not.

The convergence rate of the PoAs for arbitrary polynomial cost functions under arbitrary demand growth pattern remains an important future research topic. It is a crucial step for further bounding the PoAs in a congestion game with a high demand and arbitrary analytic cost functions. Note that analytic cost functions can be approximated with polynomials, and that the Hölder continuity results in Wu and Möhring [46] seem to indicate that this approximation of analytic cost functions may also be used for the PoAs.

While pure Nash equilibria need not exist in arbitrary finite games, Nash [30] has shown that every finite game has a mixed Nash equilibrium. Since the user choices in a mixed Nash equilibrium are random, the probability distribution of the random PoA might be a more suitable measure for the inefficiency of mixed Nash equilibria. Our analysis of the random PoA for atomic congestion games with polynomial cost functions of the same degree has already provided the first positive evidence in that direction, which may apply also to finite games of other types. Thus another important future research topic is to generalize the probabilistic analysis of the random PoA to finite games of other types.

In our study, we have assumed that the cost functions are *separable*, i.e., each arc $a \in A$ has a cost function depending only on its own flow value f_a. However, it may happen also that the cost of some arc $a \in A$ depends not only on f_a, but also on flow values f_b of other arcs $b \in A$. Then the cost functions are called *non-separable*, see, e.g., [36]. A convergence analysis of atomic, mixed and non-atomic PoAs for congestion games with *non-separable* cost functions would also be an interesting future research topic, as worst-case upper bounds of the non-atomic PoA in such games have already been obtained by Chau and Sim [6] and Perakis [36]. In fact, the expected flow $\mathbb{E}_{\Pi}[f_{ran}]$ of a mixed WE flow f_{ran} introduced in Remark 1 is essentially a non-atomic NE flow of a congestion game with the expected cost $\mathbb{E}_{\Pi}[\tau_a(f_{ran,a})]$ as non-separable cost when viewed as a non-atomic flow of that congestion game. Hence, the proof of Theorem 3 has already provided a first positive example for a convergence analysis of the PoAs for non-separable cost functions, although the expected cost is still rather simple compared with general non-separable cost functions.

Acknowledgements The first author acknowledges support from the National Natural Science Foundation of China with grant No. 61906062, support from the Natural Science Foundation of Anhui Province of China with grant No. 1908085QF262, and support from the Talent Foundation of Hefei University with grant No. 1819RC29. The second author acknowledges support from the National Natural Science Foundation of China with grant No. 12131003. The fourth author acknowledges support from the National Natural Science Foundation of China with grants No. 12131003 and No. 11871081.

Funding Open Access funding enabled and organized by Projekt DEAL.

A Detailed proofs

A.1 The existence of mixed WE flows

Lemma 7 *Every congestion game $\Gamma = (\tau, \mathcal{U}, d)$ has a mixed WE flow.*

Proof of Lemma 7 We use Brouwer's fixed point theorem and an argument similar to that in Dafermos [14]. Inequality (2.7) is equivalent to the variational inequality that

$$\sum_{k\in\mathcal{K}}\sum_{p\in\mathcal{P}_k}\mathbb{E}_{\Pi}[\tau_p(f_{ran})]\cdot(\mathbb{E}_{\Pi}[f_{ran,p}]-\mathbb{E}_{\Pi'}[f'_{ran,p}])\le 0$$

for an arbitrary mixed profile Π' with random flow f'_{ran}. Brouwer's fixed theorem implies that there is a fixed point Π^α of the continuous map

$$D_\alpha(\Pi') := \arg\min_{\Pi''}\sum_{k\in\mathcal{K}}\sum_{p\in\mathcal{P}_k}\left|\mathbb{E}_{\Pi''}\left[f''_{ran,p}\right]-\mathbb{E}_{\Pi'}\left[f'_{ran,p}\right]+\alpha\cdot\mathbb{E}_{\Pi'}\left[\tau_p(f'_{ran})\right]\right|^2$$

for an arbitrary $\alpha > 0$. This follows since $D_\alpha(\cdot)$ maps the space of all mixed profiles continuously into a subspace, and since the space of all mixed profiles is convex and compact. This fixed point Π^α fulfills the condition that

$$\begin{aligned}&\sum_{k\in\mathcal{K}}\sum_{p\in\mathcal{P}_k}\mathbb{E}_{\Pi^\alpha}[\tau_p(f^\alpha_{ran})]\cdot\Big(\mathbb{E}_{\Pi^\alpha}[f^\alpha_{ran,p}]-\mathbb{E}_{\Pi''}[f''_{ran,p}]\Big)\\&\quad\le\frac{1}{2\cdot\alpha}\cdot\sum_{k\in\mathcal{K}}\sum_{p\in\mathcal{P}_k}\Big(\mathbb{E}_{\Pi^\alpha}[f^\alpha_{ran,p}]-\mathbb{E}_{\Pi''}[f''_{ran,p}]\Big)^2\end{aligned}\tag{A.1}$$

for an arbitrary mixed profile Π'' with random flow f''_{ran}, where f^α_{ran} is the random flow of Π^α. Since the mixed profile sequence $(\Pi^\alpha)_{\alpha\in(0,\infty)}$ is bounded, there is an infinite subsequence $(\alpha_n)_{n\in\mathbb{N}}$ such that $\alpha_n\to\infty$ and that $(\Pi^{\alpha_n})_{n\in\mathbb{N}}$ converges to a limit mixed profile Π, as $n\to\infty$. This limit mixed profile Π has a mixed WE flow f_{ran}, since inequality (A.1) holds for an arbitrary $\alpha > 0$ and an arbitrary mixed profile Π''. Here, we used that $\lim_{n\to\infty} \mathbb{E}_{\Pi^{\alpha_n}}[f_{ran,p}^{\alpha_n}] = \mathbb{E}_{\Pi}[f_{ran,p}]$ and $\lim_{n\to\infty} \mathbb{E}_{\Pi^{\alpha_n}}[\tau_p(f_{ran}^{\alpha_n})] = \mathbb{E}_{\Pi}[\tau_p(f_{ran})]$ as $n \to \infty$, when $\Pi^{\alpha_n} \to \Pi$ as $n \to \infty$. This proves the existence of mixed WE flows. □

A.2 Stochastic inequalities

Our proofs will use Markov's inequality, Chebyshev's inequality and Jensen's inequality. We summarize them in Lemma 8 below.

Lemma 8 *Let X be a non-negative random variable whose expectation $\mathbb{E}(X)$ exists, and let $\Delta > 0$ be an arbitrary constant. Then*

a) (Markov's inequality, see, e.g., [31]) $\mathbb{P}(X \geq \Delta) \leq \frac{\mathbb{E}(X)}{\Delta}$.
b) (Chebyshev's inequality, see, e.g., [4]) $\mathbb{P}(|X - \mathbb{E}(X)| \geq \Delta) \leq \frac{\mathbb{VAR}(X)}{\Delta^2}$.
c) (Jensen's inequality, see, e.g., [25]) $\mathbb{E}(h(X)|\mathcal{E}) \geq h(\mathbb{E}(X|\mathcal{E}))$ for every convex function $h : \mathbb{R} \to \mathbb{R}$ and an arbitrary random event $\mathcal{E}$.

A.3 Proof of lemma 4

Note that Lemma 4 holds trivially if the integer degree $\beta = 0$, since all cost functions $\tau_a^{[g]}(\cdot)$ are then positive constants, and so the total cost of atomic and non-atomic NE flows coincide. We thus assume that $\beta \geq 1$.

Proof of Lemma 4a Consider now an arbitrary $k \in \mathcal{K}$ and an arbitrary $i \in \mathcal{U}_k$. Lemma 4a follows if

$$\tau^{[g]}_{p_{k,i}(\tilde{f}_{at}^{[g]})}\big(\tilde{f}_{at}^{[g]}\big) \leq \tau_{p'}^{[g]}\big(\tilde{f}_{at}^{[g]}\big) + |A| \cdot \kappa \cdot \frac{d_{max}}{T} \tag{A.2}$$

for all paths $p' \in \mathcal{P}_k$, where we recall that $p_{k,i}(\tilde{f}_{at}^{[g]})$ is the path of agent i and that $\kappa = \beta \cdot \eta_{\max} \cdot \big(1 + \sum_{l=1}^{\beta} \frac{1}{T^l}\big) > 0$ is the Lipschitz constant of scaled cost functions $\tau_a^{[g]}$ on $[0, 1]$.

To prove (A.2), we consider an arbitrary path $p' \in \mathcal{P}_k$. Since $\tilde{f}_{at}^{[g]}$ is an atomic NE flow, we obtain

$$C_{k,i}(\tilde{f}_{at}^{[g]}, \Gamma^{[g]}) = \frac{d_{k,i}}{T} \cdot \tau^{[g]}_{p_{k,i}(\tilde{f}_{at}^{[g]})}(\tilde{f}_{at}^{[g]}) \leq C_{k,i}(f_{at}^{[g]'}, \Gamma^{[g]}) = \frac{d_{k,i}}{T} \cdot \tau_{p'}^{[g]}(f_{at}^{[g]'}), \tag{A.3}$$

where $f_{at}^{[g]'}$ is an atomic flow of $\Gamma^{[g]}$ as defined in (2.4), i.e., $f_{at}^{[g]'}$ is the resulting flow obtained by moving i from $p_{k,i}(\tilde{f}_{at}^{[g]})$ to p' in the atomic NE flow $\tilde{f}_{at}^{[g]}$. (A.3) implies further that

$$\begin{aligned} \tau^{[g]}_{p_{k,i}(\tilde{f}_{at}^{[g]})}\big(\tilde{f}_{at}^{[g]}\big) &= \sum_{a \in p_{k,i}(\tilde{f}_{at}^{[g]})} \tau_a^{[g]}\big(\tilde{f}_{at,a}^{[g]}\big) \leq \tau_{p'}^{[g]}\big(f_{at}^{[g]'}\big) \\ &= \sum_{a \in p'} \tau_a^{[g]}\big(f_{at,a}^{[g]'}\big). \end{aligned} \tag{A.4}$$

Note that the atomic flows $\tilde{f}_{at}^{[g]}$ and $\tilde{f}_{at}^{[g]'}$ differ only in the choice of i. Note also that i controls an amount $\bar{d}_{k,i} \leq \frac{d_{max}}{T}$ of demand in $\Gamma^{[g]}$. So we obtain for all $a \in A$ that $|\tilde{f}_{at,a}^{[g]} - f_{at,a}^{[g]'}| \leq \bar{d}_{k,i} \leq \frac{d_{max}}{T}$, where we recall that i uses only a single path in any atomic flow. This and (3.3) imply that

$$\left|\tau_a^{[g]}\left(\tilde{f}_{at,a}^{[g]}\right) - \tau_a^{[g]}\left(f_{at,a}^{[g]'}\right)\right| \leq \kappa \cdot |\tilde{f}_{at,a}^{[g]} - f_{at,a}^{[g]'}| = \kappa \cdot \bar{d}_{k,i} \leq \kappa \cdot \frac{d_{max}}{T} \quad \forall a \in A. \tag{A.5}$$

Here, we used that $\tau_a^{[g]}(x)$ is Lipschitz bounded on $[0, 1]$ with the constant κ, and that all arc flow values of $\Gamma^{[g]}$ are in $[0, 1]$. Then (A.5) and (A.4) imply that $\tau^{[g]}_{p_{k,i}(\tilde{f}_{at}^{[g]})}\left(\tilde{f}_{at}^{[g]}\right) \leq \tau_{p'}^{[g]}\left(\tilde{f}_{at}^{[g]}\right) + |A| \cdot \kappa \cdot \frac{d_{max}}{T}$, which proves (A.2) due to arbitrary choice of $p' \in \mathcal{P}_k$. This completes the proof of Lemma 4a. □

Proof of Lemma 4b Lemma 4a yields that $\max_{p\in\mathcal{P}_k:\, \tilde{f}_{at,p}^{[g]}>0} \tau_p^{[g]}(\tilde{f}_{at}^{[g]}) \leq \min_{p\in\mathcal{P}_k} \tau_p^{[g]}(\tilde{f}_{at}^{[g]}) + \frac{|A|\cdot\kappa\cdot d_{max}}{T}$ for each $k \in \mathcal{K}$. This in turn implies for an arbitrary non-atomic flow $f_{nat}^{[g]}$ that

$$\begin{aligned}
\sum_{a\in A} \tau_a^{[g]}\left(\tilde{f}_{at,a}^{[g]}\right) \cdot \left(f_{nat,a}^{[g]} - \tilde{f}_{at,a}^{[g]}\right) &= \sum_{p\in\mathcal{P}} \tau_p^{[g]}\left(\tilde{f}_{at}^{[g]}\right) \cdot \left(f_{nat,p}^{[g]} - \tilde{f}_{at,p}^{[g]}\right) \\
&= \sum_{k\in\mathcal{K}} \sum_{p\in\mathcal{P}_k} \left(\tau_p^{[g]}(\tilde{f}_{at}^{[g]}) - \tau_{p_k^*}^{[g]}(\tilde{f}_{at}^{[g]})\right) \cdot \left(f_{nat,p}^{[g]} - \tilde{f}_{at,p}^{[g]}\right) \\
&\geq \sum_{k\in\mathcal{K}} \sum_{p\in\mathcal{P}_k:\tilde{f}_{at,p}^{[g]}>0} \left(\tau_p^{[g]}(\tilde{f}_{at}^{[g]}) - \tau_{p_k^*}^{[g]}(\tilde{f}_{at}^{[g]})\right) \cdot \left(f_{nat,p}^{[g]} - \tilde{f}_{at,p}^{[g]}\right) \\
&\geq -|\mathcal{P}| \cdot |A| \cdot \kappa \cdot \frac{d_{max}}{T}.
\end{aligned} \tag{A.6}$$

Here, we used that the total demand of $\Gamma^{[g]}$ is $\bar{T} = 1$, and that

$$\begin{aligned}
&\sum_{k\in\mathcal{K}} \sum_{p\in\mathcal{P}_k} \tau_{p_k^*}^{[g]}\left(\tilde{f}_{at}^{[g]}\right) \cdot \left(f_{nat,p}^{[g]} - \tilde{f}_{at,p}^{[g]}\right) \\
&\quad = \sum_{k\in\mathcal{K}} \tau_{p_k^*}^{[g]}\left(\tilde{f}_{at}^{[g]}\right) \cdot \sum_{p\in\mathcal{P}_k} \left(f_{nat,p}^{[g]} - \tilde{f}_{at,p}^{[g]}\right) = 0,
\end{aligned}$$

where p_k^* is the least costly path in $\mathcal{P}_k$ w.r.t. the atomic NE flow $\tilde{f}_{at}^{[g]}$. By Definition 2, $\tilde{f}_{at}^{[g]}$ is an ϵ-approximate non-atomic NE flow of $\Gamma^{[g]}$ with $\epsilon := \frac{|\mathcal{P}|\cdot|A|\cdot\kappa\cdot d_{max}}{T}$.

In the sequel, we will use without further proof that a flow f is a $|\mathcal{P}|\cdot\epsilon$-approximate non-atomic NE flow when it satisfies the condition that

$$\max_{p\in\mathcal{P}_k:f_p>0} \tau_p^{[g]}(f) \leq \min_{p\in\mathcal{P}_k} \tau_p^{[g]}(f) + \epsilon \quad \forall k \in \mathcal{K}.$$

This can be justified by an argument similar to that in (A.6). □

Proof of Lemma 4c: It follows immediately from Lemma 4b and Lemma 3. □

A.4 Proof of Lemma 5

Let $\tilde{f}^{[g]}$ and $f^{*[g]}$ be non-atomic NE and SO flows of the scaled game $\Gamma^{[g]}$, respectively. Then $\rho_{nat}(\Gamma^{[g]}) = \frac{\sum_{a\in A} \tau_a^{[g]}(\tilde{f}_a^{[g]})\cdot \tilde{f}_a^{[g]}}{\sum_{a\in A} \tau_a^{[g]}(f_a^{*[g]})\cdot f_a^{*[g]}}$. Note that $\tilde{f}^{[g]}$ is an optimal solution of the *non-linear program* (NLP) (A.7),

$$\begin{aligned} \min \quad & \Phi(y) := \sum_{a\in A} \int_0^{y_a} \tau_a^{[g]}(x)dx \\ \text{s.t.} \quad & \sum_{p\in\mathcal{P}_k} y_p = \bar{d}_k = \sum_{i\in\mathcal{U}_k} \frac{d_{k,i}}{T} = \frac{d_k}{T} \quad \forall k \in \mathcal{K}, \\ & y_p \geq 0 \quad \forall p \in \mathcal{P}, \end{aligned} \tag{A.7}$$

see, e.g., [42, 44]. So $\Phi(\tilde{f}^{[g]}) \leq \Phi(f^{*[g]})$.

As the scaled cost functions $\tau_a^{[g]}(\cdot)$ have the form (3.3), we obtain that

$$\begin{aligned} \int_0^{y_a} \tau_a^{[g]}(x)dx &= \frac{1}{\beta+1} \cdot \eta_{a,0} \cdot y_a^{\beta+1} + \sum_{l=1}^{\beta} \frac{\eta_{a,l}}{(\beta-l+1)\cdot T^l} \cdot y_a^{\beta-l+1} \\ &= \frac{1}{\beta+1} \cdot \tau_a^{[g]}(y_a) \cdot y_a + \sum_{l=1}^{\beta} \frac{l\cdot\eta_{a,l}}{(\beta-l+1)\cdot(\beta+1)\cdot T^l} \cdot y_a^{\beta-l+1} \end{aligned} \tag{A.8}$$

for all $a \in A$ and all $y_a \in [0, 1]$. So

$$0 \leq \int_0^{y_a} \tau_a^{[g]}(x)dx - \frac{1}{\beta+1} \cdot \tau_a^{[g]}(y_a) \cdot y_a \leq \frac{\beta\cdot\eta_{\max}}{\beta+1} \cdot \sum_{l=1}^{\beta} \frac{1}{T^l}$$

for all $a \in A$ and all $y_a \in [0, 1]$. Here, we employ the convention that $\sum_{l=1}^{\beta} \frac{1}{T^l} = 0$ when $\beta = 0$. We thus obtain that

$$\begin{aligned} \sum_{a\in A} \tau_a^{[g]}(\tilde{f}_a^{[g]}) \cdot \tilde{f}_a^{[g]} &\leq (\beta+1)\cdot\Phi(\tilde{f}^{[g]}) \leq (\beta+1)\cdot\Phi(f^{*[g]}) \\ &\leq \sum_{a\in A} \tau_a^{[g]}(f_a^{*[g]}) \cdot f_a^{*[g]} + \beta\cdot\eta_{\max}\cdot\sum_{l=1}^{\beta} \frac{1}{T^l}, \end{aligned}$$

which in turn implies that $\rho_{nat}(\Gamma^{[g]}) \leq 1 + \frac{\beta\cdot\eta_{\max}\cdot|\mathcal{P}|^{\beta+1}}{\eta_{0,\min}} \cdot \sum_{l=1}^{\beta} \frac{1}{T^l}$. Here, we recall that $\eta_{0,\min} = \min_{a\in A} \eta_{a,0} > 0$, and that the total cost $\sum_{a\in A} \tau_a^{[g]}(f_a^{*[g]}) \cdot f_a^{*[g]}$ is bounded from below by $\frac{\eta_{0,\min}}{|\mathcal{P}|^{\beta+1}}$. This completes the proof of Lemma 5. □

A.5 Proof of Lemma 6

Recall that β is the common degree of the polynomial cost functions, and is thus a non-negative integer. When $\beta = 0$, then the scaled cost functions $\tau_a^{[g]}(\cdot)$ are positive constants, and Lemma 6 holds trivially. We thus assume $\beta \geq 1$.

Consider now an arbitrary mixed NE flow $f_{ran}^{[g]}$ of $\Gamma^{[g]}$. Chebyshev's inequality, see Lemma 8b, implies that

$$\begin{aligned}
\mathbb{P}_{\tilde{\Pi}}\left[\left|\tilde{f}_{ran,a}^{[g]} - \mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran,a}^{[g]})\right| > \left(\frac{d_{max}}{T}\right)^{\delta}\right] &\leq \left(\frac{T}{d_{max}}\right)^{2\cdot\delta} \cdot \mathbb{VAR}_{\tilde{\Pi}}(\tilde{f}_{ran,a}^{[g]}) \\
&\leq \left(\frac{T}{d_{max}}\right)^{2\cdot\delta} \cdot \sum_{i\in\mathcal{U}_k, k\in\mathcal{K}} \frac{d_{k,i}^2}{4\cdot T^2} \\
&\leq \frac{1}{4}\cdot\left(\frac{d_{max}}{T}\right)^{1-2\cdot\delta} \qquad \forall a \in A\ \forall \delta \in (0, \frac{1}{2}).
\end{aligned} \tag{A.9}$$

Here, we used $\mathbb{VAR}_{\tilde{\Pi}}(\tilde{f}_{ran,a}^{[g]}) = \sum_{i\in\mathcal{U}_k,k\in\mathcal{K}} \frac{d_{k,i}^2}{T^2}\cdot\tilde{\Pi}_{i,a}\cdot(1-\tilde{\Pi}_{i,a})$. This follows since $\tilde{\Pi}_{i,a} = \sum_{p\in\mathcal{P}_k:a\in p}\tilde{\Pi}_{i,p}$ is the probability that agent $i \in \mathcal{U}_k$ uses arc a, since the demand of agent $i \in \mathcal{U}_k$ is $\frac{d_{k,i}}{T}$ in the scaled game $\Gamma^{[g]}$ and since $\Gamma^{[g]}$ has total demand $\bar{T} = \sum_{k\in\mathcal{K},i\in\mathcal{U}_k}\frac{d_{k,i}}{T} = 1$.

We now show that the mixed NE flow $\tilde{f}_{ran}^{[g]}$ is an *approximate* mixed WE flow (see Remark 1). Consider an arbitrary $k \in \mathcal{K}$ and an arbitrary $p \in \mathcal{P}_k$ with $\tilde{\Pi}_{i,p} > 0$ for some $i \in \mathcal{U}_k$.

Note that $|\tilde{f}_{ran,a|i,p''}^{[g]} - \tilde{f}_{ran,a|i,p'}^{[g]}| \leq \frac{d_{max}}{T}$. for all $a \in A$ and all $p', p'' \in \mathcal{P}_k$. Here, we recall that $\tilde{f}_{ran,a|i,p'}^{[g]}$ is the random flow of arc a when agent i uses the fixed path p' and the other agents $j \in \mathcal{U}\setminus\{i\}$ still follow their random paths drawn from $\tilde{\Pi}_j$. Then

$$\begin{aligned}
&\left|\mathbb{E}_{\tilde{\Pi}}\left[\tau_{p'}^{[g]}\left(\tilde{f}_{ran}^{[g]}\right)\right] - \mathbb{E}_{\tilde{\Pi}_{-i}}\left[\tau_{p'}^{[g]}\left(\tilde{f}_{ran|i,p''}^{[g]}\right)\right]\right| \\
&\quad= \left|\sum_{p'''\in\mathcal{P}_k}\left[\mathbb{E}_{\tilde{\Pi}_{-i}}\left[\tau_{p'}^{[g]}\left(\tilde{f}_{ran|i,p'''}^{[g]}\right)\right] - \mathbb{E}_{\tilde{\Pi}_{-i}}\left[\tau_{p'}^{[g]}\left(\tilde{f}_{ran|i,p''}^{[g]}\right)\right]\right]\cdot\tilde{\Pi}_{i,p'''}\right| \\
&\quad\leq \sum_{p'''\in\mathcal{P}_k}\tilde{\Pi}_{i,p'''}\cdot\sum_{a\in p'}\mathbb{E}_{\tilde{\Pi}_{-i}}\left|\tau_a^{[g]}(\tilde{f}_{ran,a|i,p'''}^{[g]}) - \tau_a^{[g]}(\tilde{f}_{ran,a|i,p''}^{[g]})\right| \\
&\quad\leq \kappa\cdot\sum_{p'''\in\mathcal{P}_k}\tilde{\Pi}_{i,p'''}\cdot\sum_{a\in p'}\mathbb{E}_{\tilde{\Pi}_{-i}}\left|\tilde{f}_{ran,a|i,p'''}^{[g]} - \tilde{f}_{ran,a|i,p''}^{[g]}\right| \\
&\quad\leq |A|\cdot\kappa\cdot\frac{d_{max}}{T}, \qquad \forall p', p'' \in \mathcal{P}_k,
\end{aligned} \tag{A.10}$$

since each $\tau_a^{[g]}$ is Lipschitz bounded on [0, 1] with Lipschitz constant κ. (A.10) implies that $\tilde{f}_{ran}^{[g]}$ is an approximate mixed WE flow, i.e.,

$$\begin{aligned}\mathbb{E}_{\tilde{\Pi}}\left[\tau_p^{[g]}\left(\tilde{f}_{ran}^{[g]}\right)\right] = \mathbb{E}_{\tilde{\Pi}_{-i}}\left[\tau_p^{[g]}\left(\tilde{f}_{ran|i,p}^{[g]}\right)\right] \le \mathbb{E}_{\tilde{\Pi}_{-i}}\left[\tau_{p'}^{[g]}\left(\tilde{f}_{ran|i,p'}^{[g]}\right)\right] \\ \le \mathbb{E}_{\tilde{\Pi}}\left[\tau_{p'}^{[g]}\left(\tilde{f}_{ran}^{[g]}\right)\right] + |A| \cdot \kappa \cdot \frac{d_{max}}{T}\end{aligned} \tag{A.11}$$

for an arbitrary $p' \in \mathcal{P}_k$. This follows since $\tilde{f}_{ran}^{[g]}$ is a mixed NE flow and $\tilde{\Pi}_{i,p} > 0$.

We now show with (A.9) and (A.11) that the expected flow $\mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran}^{[g]})$ is an ϵ-approximate non-atomic NE flow with ϵ tending to 0 as $\frac{d_{max}}{T} \to 0$.

(A.9) implies

$$\begin{aligned}\mathbb{P}_{\tilde{\Pi}}\left[\forall a \in A : \left|\tilde{f}_{ran,a}^{[g]} - \mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran,a}^{[g]})\right| \le \left(\frac{d_{max}}{T}\right)^{\delta}\right] \\ \ge 1 - \frac{|A|}{4} \cdot \left(\frac{d_{max}}{T}\right)^{1-2\cdot\delta} = 1 - \mathbf{P}_{\delta},\end{aligned} \tag{A.12}$$

where $\mathbf{P}_{\delta} := \frac{|A|}{4} \cdot \left(\frac{d_{max}}{T}\right)^{1-2\cdot\delta}$. Consequently,

$$\mathbb{P}_{\tilde{\Pi}}\left[\forall a \in A :\ \left|\tau_a^{[g]}(\tilde{f}_{ran,a}^{[g]}) - \tau_a^{[g]}\left(\mathbb{E}_{\tilde{\Pi}}\left(\tilde{f}_{ran,a}^{[g]}\right)\right)\right| \le \kappa \cdot \left(\frac{d_{max}}{T}\right)^{\delta}\right] \ge 1 - \mathbf{P}_{\delta}, \tag{A.13}$$

again since the scaled cost functions are Lipschitz continuous on [0, 1] with the Lipschitz constant κ.

Note that $|\tau_a^{[g]}(\tilde{f}_{ran,a}^{[g]}) - \tau_a^{[g]}(\mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran,a}^{[g]}))| \le \tau_a^{[g]}(1) \le \frac{\kappa}{\beta} = \sum_{l=0}^{\beta} \frac{\eta_{\max}}{T^l}$ with probability 1 for all $a \in A$. This, together with (A.13), implies that

$$\begin{aligned}&|\mathbb{E}_{\tilde{\Pi}}\left(\tau_a^{[g]}(\tilde{f}_{ran,a}^{[g]})\right) - \tau_a^{[g]}\left(\mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran,a}^{[g]})\right)| \le \mathbb{E}_{\tilde{\Pi}}\left(|\tau_a^{[g]}(\tilde{f}_{ran,a}^{[g]}) - \tau_a^{[g]}(\mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran,a}^{[g]}))|\right) \\ &\quad \le (1 - \mathbf{P}_{\delta}) \cdot \kappa \cdot \left(\frac{d_{max}}{T}\right)^{\delta} + \mathbf{P}_{\delta} \cdot \frac{\kappa}{\beta} \\ &\quad = \kappa \cdot \left(\frac{d_{max}}{T}\right)^{\delta} + \frac{|A|}{4} \cdot \left(\frac{d_{max}}{T}\right)^{1-2\cdot\delta} \cdot \frac{\kappa}{\beta} \quad \forall a \in A\ \forall \delta \in (0, \frac{1}{2}).\end{aligned} \tag{A.14}$$

(A.14) uses that the random event "$\kappa \cdot \left(\frac{d_{max}}{T}\right)^{\delta} < |\tau_a^{[g]}(\tilde{f}_{ran,a}^{[g]}) - \tau_a^{[g]}(\mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran,a}^{[g]}))| \le \frac{\kappa}{\beta}$" occurs with a probability of *at most* $\mathbf{P}_{\delta}$, since the random event of (A.13) occurs with a probability of at least $1 - \mathbf{P}_{\delta}$.

Putting $\delta = \frac{1}{3}$ in (A.14), we obtain that

$$\left|\mathbb{E}_{\tilde{\Pi}}\left(\tau_a^{[g]}(\tilde{f}_{ran,a}^{[g]})\right) - \tau_a^{[g]}\left(\mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran,a}^{[g]})\right)\right| \le \kappa \cdot \left(\frac{d_{max}}{T}\right)^{1/3} + \frac{|A|}{4} \cdot \left(\frac{d_{max}}{T}\right)^{1/3} \cdot \frac{\kappa}{\beta}$$
$$= \kappa \cdot \left(1 + \frac{|A|}{4 \cdot \beta}\right) \cdot \left(\frac{d_{max}}{T}\right)^{1/3} \quad \forall a \in A. \tag{A.15}$$

(A.15) in turn implies that

$$\left|\mathbb{E}_{\tilde{\Pi}}\left(\tau_p^{[g]}\left(\tilde{f}_{ran}^{[g]}\right)\right) - \tau_p^{[g]}\left(\mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran}^{[g]})\right)\right|$$
$$\le \kappa \cdot |A| \cdot \left(1 + \frac{|A|}{4 \cdot \beta}\right) \cdot \left(\frac{d_{max}}{T}\right)^{1/3} \quad \forall p \in \mathcal{P}. \tag{A.16}$$

(A.16) and (A.11) together then yield

$$\tau_p^{[g]}\left(\mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran}^{[g]})\right) \le \mathbb{E}_{\tilde{\Pi}}\left(\tau_p^{[g]}(\tilde{f}_{ran}^{[g]})\right) + \kappa \cdot |A| \cdot \left(1 + \frac{|A|}{4 \cdot \beta}\right) \cdot \left(\frac{d_{max}}{T}\right)^{1/3}$$
$$\le \mathbb{E}_{\tilde{\Pi}}\left(\tau_{p'}^{[g]}\left(\tilde{f}_{ran}^{[g]}\right)\right) + \kappa \cdot |A| \cdot \left(1 + \frac{|A|}{4 \cdot \beta}\right) \cdot \left(\frac{d_{max}}{T}\right)^{1/3} + |A| \cdot \kappa \cdot \frac{d_{max}}{T}$$
$$\le \tau_{p'}^{[g]}\left(\mathbb{E}_{\tilde{\Pi}}\left(\tilde{f}_{ran}^{[g]}\right)\right) + 2 \cdot \kappa \cdot |A| \cdot \left(1 + \frac{|A|}{4 \cdot \beta}\right) \cdot \left(\frac{d_{max}}{T}\right)^{1/3} + |A| \cdot \kappa \cdot \frac{d_{max}}{T}$$
$$\le \tau_{p'}^{[g]}\left(\mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran}^{[g]})\right) + 3 \cdot \kappa \cdot |A| \cdot \left(1 + \frac{|A|}{4 \cdot \beta}\right) \cdot \left(\frac{d_{max}}{T}\right)^{1/3} \tag{A.17}$$

for all $k \in \mathcal{K}$, and any two paths $p, p' \in \mathcal{K}$ with expected flow value $\mathbb{E}_{\tilde{\Pi}}(\tilde{f}_{ran,p}^{[g]}) > 0$.

(A.17) yields with a similar argument as in the proof of Lemma 4b that the expected non-atomic flow $\mathbb{E}_{\tilde{\Pi}}[\tilde{f}_{ran}^{[g]}]$ is an ϵ-approximate non-atomic NE flow with $\epsilon := 3 \cdot |\mathcal{P}| \cdot \kappa \cdot |A| \cdot \left(1 + \frac{|A|}{4 \cdot \beta}\right) \cdot \left(\frac{d_{max}}{T}\right)^{1/3}$. Lemma 6a then follows immediately from Lemma 3.

Lemma 6b then follows from (A.12) and (A.13), since they together imply that the random event

$$\forall a \in A : \left|\tilde{f}_{ran,a}^{[g]} \cdot \tau_a^{[g]}\left(\tilde{f}_{ran,a}^{[g]}\right) - \mathbb{E}_{\tilde{\Pi}}\left(\tilde{f}_{ran,a}^{[g]}\right) \cdot \tau_a^{[g]}\left(\mathbb{E}_{\tilde{\Pi}}\left(\tilde{f}_{ran,a}^{[g]}\right)\right)\right|$$
$$\le \left(\kappa + \eta_{\max} \cdot \sum_{l=0}^{\beta} \frac{1}{T^l}\right) \cdot \left(\frac{d_{max}}{T}\right)^{\delta} \tag{A.18}$$

occurs with a probability of *at least* $1 - \mathbf{P}_\delta$. Here, we use that $\tau_a^{[g]}(\cdot)$ is Lipschitz bounded in $[0, 1]$ with Lipschitz constant κ, that $|x \cdot \tau_a^{[g]}(x) - y \cdot \tau_a^{[g]}(y)| \le x \cdot |\tau_a^{[g]}(x) - \tau_a^{[g]}(y)| + \tau_a^{[g]}(y) \cdot |x - y| \le (\kappa + \tau_a^{[g]}(y)) \cdot |x - y| \le (\kappa + \tau_a^{[g]}(1)) \cdot |x - y|$ for all $x, y \in [0, 1]$, that $\Gamma^{[g]}$ has arc flow values in $[0, 1]$, and that $\max_{a \in A} \max_{[0,1]} \tau_a^{[g]}(x) \le \sum_{l=0}^{\beta} \frac{\eta_{\max}}{T^l}$.

(A.12) and (A.18) yield

$$\begin{aligned}
&\left|\mathbb{E}_{\tilde{\Pi}}\left(\tilde{f}^{[g]}_{ran,a}\cdot\tau^{[g]}_a(\tilde{f}^{[g]}_{ran,a})\right)-\mathbb{E}_{\tilde{\Pi}}\left(\tilde{f}^{[g]}_{ran,a}\right)\cdot\tau^{[g]}_a\left(\mathbb{E}_{\tilde{\Pi}}(\tilde{f}^{[g]}_{ran,a})\right)\right|\\
&\quad\le\mathbb{E}_{\tilde{\Pi}}\left(\left|\tilde{f}^{[g]}_{ran,a}\cdot\tau^{[g]}_a(\tilde{f}^{[g]}_{ran,a})-\mathbb{E}_{\tilde{\Pi}}\left(\tilde{f}^{[g]}_{ran,a}\right)\cdot\tau^{[g]}_a\left(\mathbb{E}_{\tilde{\Pi}}\left(\tilde{f}^{[g]}_{ran,a}\right)\right)\right|\right)\\
&\quad\le\left(\kappa+\eta_{\max}\cdot\sum_{l=0}^{\beta}\frac{1}{T^l}\right)\cdot\left(\frac{d_{max}}{T}\right)^{\delta}+\mathbf{P}_{\delta}\cdot\eta_{\max}\cdot\sum_{l=0}^{\beta}\frac{1}{T^l}\\
&\quad=\left(\kappa+\eta_{\max}\cdot\sum_{l=0}^{\beta}\frac{1}{T^l}\right)\cdot\left(\frac{d_{max}}{T}\right)^{\delta}+\frac{|A|}{4}\cdot\left(\frac{d_{max}}{T}\right)^{1-2\cdot\delta}\cdot\eta_{\max}\cdot\sum_{l=0}^{\beta}\frac{1}{T^l}
\end{aligned}\tag{A.19}$$

for all $a\in A$ and all $\delta\in(0,\frac{1}{2})$. Here, we use that $\max_{a\in A}\max_{x\in[0,1]}x\cdot\tau^{[g]}_a(x)\le\sum_{l=0}^{\beta}\frac{\eta_{\max}}{T^l}$, and that the random event (A.18) occurs with a probability of at least $1-\mathbf{P}_{\delta}$, and so the complement event of (A.18) occurs with a probability of at most $\mathbf{P}_{\delta}$. Lemma 6c then follows immediately from (A.19) when we put $\delta=\frac{1}{3}$. □

A.6 Proof of Theorem 3

We first show Theorem 3c, and then prove Theorem 3a–b with the technique of asymptotic decomposition proposed by Wu et al. [47].

Proof of Theorem 3c We define $\beta=\max_{k\in\mathcal{K}}\min_{p\in\mathcal{P}_k}\max_{a\in p}\beta_a$, and put the scaling factor $g:=T^{\beta}$. Here, we recall that the degree $\beta_a\ge 0$ of arc $a\in A$ is an integer. We call a path $p\in\mathcal{P}=\cup_{k\in\mathcal{K}}\mathcal{P}_k$ with $\max_{a\in p}\beta_a\le\beta$ a *tight* path, and an arc $a\in A$ with $\beta_a\le\beta$ a *tight* arc. Clearly, each O/D pair $k\in\mathcal{K}$ has at least one tight path $p\in\mathcal{P}_k$. We denote by $\Gamma^{[g]}$ the resulting scaled game with scaling factor g. This has a total demand of 1.

Let $\tilde{f}^{[g]}_{nat}$ be an abitrary non-atomic NE flow of $\Gamma^{[g]}$, and let $\tilde{f}^{[g]}_{at}$ be an arbitrary atomic NE flow of $\Gamma^{[g]}$.

Colini-Baldeschi et al. [10] have shown that $\rho_{nat}(\Gamma)=\rho_{nat}(\Gamma^{[g]})=1+O(\frac{1}{T})$ under the condition of Theorem 3c, i.e., $\frac{d_k}{T}\ge\xi_k$ for each $k\in\mathcal{K}$ for constants $\xi_k>0$ independent of T. To obtain the convergence rate of the atomic PoA $\rho_{at}(\Gamma)=\rho_{at}(\Gamma^{[g]})$, we again need to upper bound only the cost difference $|C(\tilde{f}^{[g]}_{at},\Gamma^{[g]})-C(\tilde{f}^{[g]}_{nat},\Gamma^{[g]})|$ because of inequality (3.4). Here, we observe that non-atomic SO flows of $\Gamma^{[g]}$ have a cost of $\Omega(1)$, since every O/D pair $k\in\mathcal{K}$ has a total demand of $\frac{d_k}{T}\in\Theta(1)$ in $\Gamma^{[g]}$, and since there is at least one O/D pair $k\in\mathcal{K}$ with $\min_{p\in\mathcal{P}_k}\max_{a\in p}\beta_a=\beta$.

When all arcs are tight, i.e., $\beta_a\le\beta$ for all $a\in A$, then all the scaled polynomial cost functions $\tau^{[g]}_a(x)$ of $\Gamma^{[g]}$ have bounded coefficients and degrees smaller than β, and are thus Lipschitz continuous on [0, 1] with a Lipschitz constant independent of T. Moreover, with arguments similar to those for Theorem 1, we obtain immediately that $|C(\tilde{f}_{at}^{[g]}, \Gamma^{[g]}) - C(\tilde{f}_{nat}^{[g]}, \Gamma^{[g]})| \in O(\sqrt{\frac{1}{T}})$ and so $\rho_{at}(\Gamma^{[g]}) = 1 + O(\sqrt{\frac{1}{T}})$ by inequality (3.4). Here, note that the maximum individual demand in Γ is $d_{\max} \leq \upsilon$ for a constant $\upsilon > 0$ independent of T.

Now assume that there are non-tight arcs $a \in A$, i.e., arcs $a \in A$ with $\beta_a > \beta$. Then the scaled cost functions $\tau_a^{[g]}(\cdot)$ of these non-tight arcs $a \in A$ need not be Lipschitz continuous on [0, 1], since their coefficients may tend to ∞ with growing T. A natural idea here is to remove the influence of these non-tight arcs in the analysis.

Since each O/D pair $k \in \mathcal{K}$ has at least one tight path $p \in \mathcal{P}_k$, we obtain that

$$\eta_{0,\min} \cdot T^{\beta_a - \beta} \cdot (\tilde{f}_{at,a}^{[g]})^{\beta_a} \leq \tau_a^{[g]}(\tilde{f}_{at,a}^{[g]}) \leq \eta_{\max} \cdot (\beta + 1) \cdot |A|, \quad \forall a \in A. \tag{A.20}$$

Here, we used that a tight path p has a scaled cost of at most $\eta_{\max} \cdot (\beta + 1) \cdot |A|$ in an arbitrary flow, as it contains at most $|A|$ many arcs, and has a flow value of at most 1 in an arbitrary flow of the scaled game $\Gamma^{[g]}$. Moreover, by the definition of atomic NE flows, the scaled cost $\frac{d_{k,i}}{T} \cdot \tau_{p'}^{[g]}(\tilde{f}_{at}^{[g]})$ of an arbitrary individual $i \in \mathcal{U}_k$ with an arbitrary "pure strategy" $p' \in \mathcal{P}_k$ will not decrease, even that individual unilaterally moves from path p' to a tight path $p \in \mathcal{P}_k$.

Hence, we obtain for each non-tight arc $a \in A$ that

$$\tilde{f}_{at,a}^{[g]} \leq \theta_a(T) := \frac{\eta_{\max} \cdot |A| \cdot (\beta + 1)}{\eta_{0,\min}} \cdot T^{-\frac{\beta_a - \beta}{\beta_a}} \in o(1). \tag{A.21}$$

Similarly, $\tilde{f}_{nat,a}^{[g]} \leq \theta_a(T)$ for each non-tight arc $a \in A$. Moreover, inequality (A.21) implies for each $k \in \mathcal{K}$ and each non-tight path $p \in \mathcal{P}_k$, i.e., $\max_{a \in p} \beta_a > \beta$, that

$$\begin{aligned} &\tilde{f}_{at,p}^{[g]} \leq \theta_p(T) \\ &:= \min_{a \in p: \beta_a > \beta} \theta_a(T) \in \Theta(T^{-\max_{a \in p: \beta_a > \beta} \frac{\beta_a - \beta}{\beta_a}}) \quad \text{and} \quad \tilde{f}_{nat,p}^{[g]} \leq \theta_p(T), \end{aligned} \tag{A.22}$$

since the flow value of a path is not larger than the minimum flow value of arcs contained in that path.

Inequalities (A.20)–(A.22) actually indicate that we can ignore all non-tight arcs $a \in A$ and all non-tight paths $p \in \mathcal{P}$ in the analysis. In particular, we have

$$\begin{aligned} |C(\tilde{f}_{at}^{[g]}, \Gamma^{[g]}) - C(\tilde{f}_{nat}^{[g]}, \Gamma^{[g]})| \leq |&\sum_{a \in A: \beta_a \leq \beta} \tilde{f}_{at,a}^{[g]} \cdot \tau_a^{[g]}(\tilde{f}_{at,a}^{[g]}) \\ &- \sum_{b \in A: \beta_b \leq \beta} \tilde{f}_{nat,b}^{[g]} \cdot \tau_b^{[g]}(\tilde{f}_{nat,b}^{[g]})| \\ &+ 2 \cdot \eta_{\max} \cdot (\beta + 1) \cdot |A| \cdot \sum_{a \in A: \beta_a > \beta} \theta_a(T). \end{aligned} \tag{A.23}$$

This provides a very good basis for further upper bounding the cost difference in this general case.

For each O/D pair $k \in \mathcal{P}_k$, we denote by $\mathcal{P}_k' = \{p \in \mathcal{P}_k :\ \max_{a\in p} \beta_a \le \beta\}$ the subset of all tight paths $p \in \mathcal{P}_k$, and put $\mathcal{P}' := \cup_{k\in\mathcal{K}} \mathcal{P}_k'$. Moreover, we denote by $A' = \{a \in A :\ \beta_a \le \beta\}$ the subset of all tight arcs $a \in A$.

For each tight arc $a \in A'$, we define an auxiliary cost function

$$\sigma_{1,a}(x) = \tau_a^{[g]} \left(x + \sum_{p\in\mathcal{P}\setminus\mathcal{P}':a\in p} \tilde{f}_{at,p}^{[g]} \right), \quad \forall x \ge 0,$$

where $\sum_{p\in\mathcal{P}\setminus\mathcal{P}':a\in p} \tilde{f}_{at,p}^{[g]} = 0$ when a is not included in any non-tight path, i.e., $a \notin p$ for each $p \in \mathcal{P} \setminus \mathcal{P}'$. Then the restricted flow $\tilde{f}_{at|\mathcal{P}'}^{[g]} = (\tilde{f}_{at,p}^{[g]})_{p\in\mathcal{P}_k',k\in\mathcal{K}}$ is an atomic NE flow w.r.t. these tight paths $p \in \mathcal{P}'$ and w.r.t. the arc cost functions $\sigma_{1,a}(\cdot)$ of these tight arcs $a \in A'$. This follows since $\tau_p^{[g]}(\tilde{f}_{at}^{[g]}) = \sigma_{1,p}(\tilde{f}_{at|\mathcal{P}'}^{[g]})$ for all $p \in \mathcal{P}'$. Here, we note that $\sigma_{1,p}(\tilde{f}_{at|\mathcal{P}'}^{[g]}) = \sum_{a\in p} \sigma_{1,a}(\tilde{f}_{at,a|\mathcal{P}'}^{[g]}) = \sum_{a\in p} \tau_a^{[g]}(\tilde{f}_{at,a}^{[g]})$ for each $p \in \mathcal{P}'$, and $\tilde{f}_{at,a|\mathcal{P}'}^{[g]} = \sum_{k\in\mathcal{K}} \sum_{p\in\mathcal{P}_k':a\in p} \tilde{f}_{at,p}^{[g]}$ for each $a \in A'$, and that the flow values $\tilde{f}_{at,p}^{[g]}$ on non-tight paths $p \in \mathcal{P} \setminus \mathcal{P}'$ are constant parameters of the auxiliary arc cost functions $\sigma_{1,a}(\cdot)$.

We denote by $\Gamma_1^{[g]}$ the resulting "reduced" scaled game that ignores all non-tight paths $p \in \mathcal{P}\setminus\mathcal{P}'$ together with their "demands" $\tilde{f}_{at,p}^{[g]}$, and, moreover, has the auxiliary functions $\sigma_{1,a}(\cdot)$ as the cost functions of the tight arcs $a \in A'$. Then the total cost $C(\tilde{f}_{at|\mathcal{P}'}^{[g]}, \Gamma_1^{[g]})$ of $\tilde{f}_{at|\mathcal{P}'}^{[g]}$ satisfies the condition that

$$\begin{aligned}\sum_{a\in A'} \tilde{f}_{at,a}^{[g]} \cdot \tau_a^{[g]}(\tilde{f}_{at,a}^{[g]}) \ge C(\tilde{f}_{at|\mathcal{P}'}^{[g]}, \Gamma_1^{[g]}) = \sum_{a\in A'} \tilde{f}_{at,a|\mathcal{P}'}^{[g]} \cdot \sigma_{1,a}(\tilde{f}_{at,a|\mathcal{P}'}^{[g]}) \\ \ge \sum_{a\in A'} \tilde{f}_{at,a}^{[g]} \cdot \tau_a^{[g]}(\tilde{f}_{at,a}^{[g]}) - \eta_{\max} \cdot |A|^2 \cdot (\beta+1) \cdot \sum_{p\in\mathcal{P}\setminus\mathcal{P}'} \theta_p(T), \qquad \text{(A.24)}\end{aligned}$$

where the quantity $\theta_p(T)$ defined in inequality (A.22) is an upper bound of the atomic flow value $\tilde{f}_{at,p}^{[g]}$ on a non-tight path $p \in \mathcal{P} \setminus \mathcal{P}'$. Here, we used inequalities (A.20), (A.22), $|A'| \le |A|$, and the fact that

$$0 \le \tilde{f}_{at,a}^{[g]} - \tilde{f}_{at,a|\mathcal{P}'}^{[g]} \le \sum_{p\in\mathcal{P}\setminus\mathcal{P}'} \tilde{f}_{at,p}^{[g]}$$

for each $a \in A'$.

Let $\tilde{f}_{1,nat}^{[g]}$ be a non-atomic NE flow of $\Gamma_1^{[g]}$. Since $\Gamma_1^{[g]}$ ignores all non-tight arcs $a \in A \setminus A'$, all its cost functions $\sigma_{1,a}(\cdot)$ have coefficients bounded from above by a constant independent of T, and are thus Lipschitz continuous on [0, 1]. While $\Gamma_1^{[g]}$ ignores all demands $\tilde{f}_{at,p}^{[g]}$ of non-tight paths $p \in \mathcal{P} \setminus \mathcal{P}'$, inequality (A.22) implies that $\Gamma_1^{[g]}$ has a total demand tending to 1 as $T \to \infty$. Hence, we obtain again by arguments similar to those for Theorem 1 that

$$|C(\tilde{f}^{[g]}_{1,nat},\Gamma^{[g]}_1) - C(\tilde{f}^{[g]}_{at|\mathcal{P}'},\Gamma^{[g]}_1)| \in O(\sqrt{\frac{1}{T}}). \tag{A.25}$$

Here, we note that $\tilde{f}^{[g]}_{at|\mathcal{P}'}$ is an atomic NE flow of $\Gamma^{[g]}_1$.

We proceed similarly with the non-atomic NE flow $\tilde{f}^{[g]}_{nat}$, and consider its restriction $\tilde{f}^{[g]}_{nat|\mathcal{P}'} = (\tilde{f}^{[g]}_{nat,p})_{p\in\mathcal{P}'_k,k\in\mathcal{K}}$ to tight paths $p \in \mathcal{P}'$. We define the auxiliary cost functions $\sigma_{2,a}(\cdot)$ for each tight arc $a \in A'$ and the resulting reduced scaled game $\Gamma^{[g]}_2$ by using non-atomic flow values $\tilde{f}^{[g]}_{nat,p}$ instead of atomic flow values $\tilde{f}^{[g]}_{at,p}$ in the above definitions. Then we obtain also that $\tilde{f}^{[g]}_{nat|\mathcal{P}'}$ is a non-atomic NE flow of $\Gamma^{[g]}_2$, and, moreover,

$$\begin{aligned}
&\sum_{a\in A'} \tilde{f}^{[g]}_{nat,a}\cdot\tau^{[g]}_a(\tilde{f}^{[g]}_{nat,a}) \geq C(\tilde{f}^{[g]}_{nat|\mathcal{P}'},\Gamma^{[g]}_2)\\
&\quad= \sum_{a\in A'} \tilde{f}^{[g]}_{nat,a|\mathcal{P}'}\cdot\sigma_{2,a}(\tilde{f}^{[g]}_{nat,a|\mathcal{P}'})\\
&\quad\geq \sum_{a\in A'} \tilde{f}^{[g]}_{nat,a}\cdot\tau^{[g]}_a(\tilde{f}^{[g]}_{nat,a}) - \eta_{\max}\cdot|A|^2\cdot(\beta+1)\cdot\sum_{p\in\mathcal{P}\setminus\mathcal{P}'}\theta_p(T).
\end{aligned} \tag{A.26}$$

Inequalities (A.23)–(A.26) yield that

$$\begin{aligned}
&|C\left(\tilde{f}^{[g]}_{at},\Gamma^{[g]}\right) - C\left(\tilde{f}^{[g]}_{nat},\Gamma^{[g]}\right)|\\
&\quad\leq |C\left(\tilde{f}^{[g]}_{1,nat},\Gamma^{[g]}_1\right) - C\left(\tilde{f}^{[g]}_{nat|\mathcal{P}'},\Gamma^{[g]}_2\right)|\\
&\qquad+ O\left(\sqrt{\frac{1}{T}}\right) + O\left(\sum_{a\in A\setminus A'}\theta_a(T)\right) + O\left(\sum_{p\in\mathcal{P}\setminus\mathcal{P}'}\theta_p(T)\right).
\end{aligned} \tag{A.27}$$

Note that $\Gamma^{[g]}_1$ and $\Gamma^{[g]}_2$ share the same path set $\mathcal{P}'$ and the same arc set A'. In particular, inequality (A.22) yields that the respective total demands of an arbitrary O/D pair $k \in \mathcal{K}$ in $\Gamma^{[g]}_1$ and $\Gamma^{[g]}_2$ deviate from each other by at most $O(\sum_{p\in\mathcal{P}\setminus\mathcal{P}'}\theta_p(T))$, and that $|\sigma_{1,a}(x) - \sigma_{2,a}(x)| \in O(\sum_{p\in\mathcal{P}\setminus\mathcal{P}'}\theta_p(T))$ for all $x \in [0,1]$ and all $a \in A'$. Hence, viewed as non-atomic congestion games, the distance $\|\Gamma^{[g]}_1 - \Gamma^{[g]}_2\|$ between $\Gamma^{[g]}_1$ and $\Gamma^{[g]}_2$ w.r.t. the metric defined in Wu and Möhring [46] is $O(\sum_{p\in\mathcal{P}\setminus\mathcal{P}'}\theta_p(T))$. Here, to save space, we recommend readers to **(author?)** [46] for a detailed definition of that metric.

Let $\Gamma^{[g]'}_1$ be the non-atomic congestion game that has the same components as $\Gamma^{[g]}_1$, but with the original scaled cost functions $\tau^{[g]}_a$ for each arc $a \in A'$. Similarly, let $\Gamma^{[g]'}_2$ be the non-atomic congestion game with all components of $\Gamma^{[g]}_2$, but again with the original scaled cost functions $\tau^{[g]}_a$ for each arc $a \in A'$. Then we obtain also that $\|\Gamma_1^{[g]} - \Gamma_1^{[g]'}\| \in O(\sum_{p\in\mathcal{P}\setminus\mathcal{P}'} \theta_p(T))$, $\|\Gamma_2^{[g]} - \Gamma_2^{[g]'}\| \in O(\sum_{p\in\mathcal{P}\setminus\mathcal{P}'} \theta_p(T))$, and $\|\Gamma_1^{[g]'} - \Gamma_2^{[g]'}\| \in O(\sum_{p\in\mathcal{P}\setminus\mathcal{P}'} \theta_p(T))$.

Since $\Gamma_1^{[g]}$ and $\Gamma_1^{[g]'}$ differ only in their cost functions, Lemma 10d of [46] then yields that the total cost difference between the respective non-atomic NE flows of $\Gamma_1^{[g]}$ and $\Gamma_1^{[g]'}$ is in $O(\sqrt{\sum_{p\in\mathcal{P}\setminus\mathcal{P}'} \theta_p(T)})$. Here, we observe that the cost functions of both $\Gamma_1^{[g]}$ and $\Gamma_1^{[g]'}$ are Lipschitz bounded by a constant independent of T on [0, 1]. Similarly, the cost difference between the respective non-atomic NE flows of $\Gamma_2^{[g]}$ and $\Gamma_2^{[g]'}$ is also in $O(\sqrt{\sum_{p\in\mathcal{P}\setminus\mathcal{P}'} \theta_p(T)})$. Moreover, as $\Gamma_1^{[g]'}$ and $\Gamma_2^{[g]'}$ differ only at demands, Lemma 11a of [46] implies that the cost difference between their non-atomic NE flows is again in $O(\sqrt{\sum_{p\in\mathcal{P}\setminus\mathcal{P}'} \theta_p(T)})$. In summary, we have that

$$|C\left(\tilde{f}_{1,nat}^{[g]}, \Gamma_1^{[g]}\right) - C\left(\tilde{f}_{nat|\mathcal{P}'}^{[g]}, \Gamma_2^{[g]}\right)| \in O\left(\sqrt{\sum_{p\in\mathcal{P}\setminus\mathcal{P}'} \theta_p\left(T\right)}\right),$$

which, combined with inequality (A.27), yields that

$$|C\left(\tilde{f}_{at}^{[g]}, \Gamma^{[g]}\right) - C\left(\tilde{f}_{nat}^{[g]}, \Gamma\right)| \in O\left(T^{-\frac{1}{2\cdot\max_{a\in A}\beta_a}}\right).$$

Here, we note that both $\theta_a(T)$ and $\theta_p(T)$ are $O(-\frac{1}{\max_{b\in A}\beta_b})$ for all non-tight arcs $a \in A$ and all non-tight paths $p \in \mathcal{P}$, that $\tilde{f}_{1,nat}^{[g]}$ is a non-atomic NE flow of $\Gamma_1^{[g]}$, and that $\tilde{f}_{nat|\mathcal{P}'}^{[g]}$ is a non-atomic NE flow of $\Gamma_2^{[g]}$. Then Lemma 10 and Lemma 11 of Wu and Möhring [46] apply here, since they bound the non-atomic NE cost difference from above by the square root of the metric with constant multipliers in terms of the total demands, of the arc cost function values at the maximum feasible arc flows w.r.t. the total demands, and of the Lipschitz constants of the cost functions, each of which is bounded from above by a constant independent of T in the four games $\Gamma_1^{[g]}$, $\Gamma_1^{[g]'}$, $\Gamma_2^{[g]}$ and $\Gamma_2^{[g]'}$. Again, to save space, we recommend the readers to Wu and Möhring [46] for details.

This completes the proof of Theorem 3c. □

Proof of Theorem 3a–b The argument for the proof of Theorem 3c does *not* carry over to Theorem 3a–b, since the non-atomic SO flow of the resulting scaled game $\Gamma^{[g]}$ could be of $o(1)$, and then the convergence rate of Colini-Baldeschi et al. [10] does not apply, when we still use the same scaling factor g as above, and when the condition, that $\frac{d_k}{T} \geq \xi_k$ for all $k \in \mathcal{K}$ and some constants $\xi_k > 0$ independent of T, does not hold. Interested readers may refer to Wu et al. [47] for a detailed explanation.

To prove Theorem 3a–b, we now employ the technique of *asymptotic decomposition* developed by Wu et al. [47], and show that Theorem 3a–b hold for an arbitrary infinite sequence of growing total demand, which then directly implies the convergence in Theorem 3a–b.

To that end, we now consider an arbitrary sequence $(\mathcal{S}_n)_{n\in\mathbb{N}}$ s.t. each component $\mathcal{S}_n$ is a tuple $\big(\mathcal{U}^{(n)}, d^{(n)}, \tilde{f}_{ran}^{(n)}, \tilde{\Pi}^{(n)}, \tilde{f}_{nat}^{(n)}, f_{nat}^{*(n)}\big)$ satisfying properties (S1)–(S3) below:

- (S1) $\mathcal{U}^{(n)} = \cup_{k\in\mathcal{K}} \mathcal{U}_k^{(n)}$ is an agent set of the game Γ, and $d^{(n)} = (d_{k,i}^{(n)})_{i\in\mathcal{U}_k^{(n)},\ k\in\mathcal{K}}$ is a vector of demands for the agents in $\mathcal{U}^{(n)}$. Here, $\mathcal{U}_k^{(n)}$ is an agent set of O/D pair $k \in \mathcal{K}$, $d_{k,i}^{(n)} \in (0,\ \upsilon]$ is the demand of agent $i \in \mathcal{U}_k$ of O/D pair $k \in \mathcal{K}$, and $\upsilon > 0$ is a finite *constant* upper bound of the maximum individual demand $d_{max}^{(n)} := \max_{k\in\mathcal{K}, i\in\mathcal{U}_k^{(n)}} d_{k,i}^{(n)}$, which is independent of the sequence $(\mathcal{S}_n)_{n\in\mathbb{N}}$. To facilitate our discussion, we denote the resulting game Γ equipped with $\mathcal{U}^{(n)}$ and $d^{(n)}$ by $\Gamma_n := (\tau, \mathcal{U}^{(n)}, d^{(n)})$ for each $n \in \mathbb{N}$.
- (S2) $\tilde{f}_{ran}^{(n)} = (\tilde{f}_{ran,p}^{(n)})_{p\in\mathcal{P}}$, $\tilde{f}_{nat}^{(n)} = (\tilde{f}_{nat,p}^{(n)})_{p\in\mathcal{P}}$ and $f_{nat}^{*(n)} = (f_{nat,p}^{*(n)})_{p\in\mathcal{P}}$ are an *arbitrary* mixed NE flow, an arbitrary non-atomic NE flow, and an arbitrary non-atomic SO flow of the game Γ_n, respectively. Moreover, $\tilde{\Pi}^{(n)} = (\tilde{\Pi}_{i,p}^{(n)})_{i\in\mathcal{U}_k^{(n)}, p\in\mathcal{P}_k, k\in\mathcal{K}}$ is the mixed profile of $\tilde{f}_{ran}^{(n)}$.
- (S3) $\lim_{n\to\infty} T(\mathcal{U}^{(n)}, d^{(n)}) = \infty$, where $T(\mathcal{U}^{(n)}, d^{(n)}) = \sum_{k\in\mathcal{K}} d_k^{(n)}$ is the total demand of Γ_n, and $d_k^{(n)} = \sum_{i\in\mathcal{U}_k^{(n)}} d_{k,i}^{(n)}$ is the demand of O/D pair $k \in \mathcal{K}$. To simplify notation, we write $T_n := T(\mathcal{U}^{(n)}, d^{(n)})$ in this proof.

Due to the arbitrary choice of $(\mathcal{S}_n)_{n\in\mathbb{N}}$, Theorem 3a–b hold if and only if

$$\lim_{n\to\infty} \frac{C\left(\mathbb{E}_{\tilde{\Pi}^{(n)}}\big(\tilde{f}_{ran}^{(n)}\big), \Gamma_n\right)}{C(f_{nat}^{*(n)}, \Gamma_n)} = 1 \ \text{ and } \ \lim_{n\to\infty} \frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}\left[C(\tilde{f}_{ran}^{(n)}, \Gamma_n)\right]}{C(f_{at}^{*(n)}, \Gamma_n)} = 1. \qquad \text{(A.28)}$$

Here, $f_{at}^{*(n)}$ is an arbitrary atomic SO flow of Γ_n. Note that Wu et al. [47] have proved that $\lim_{n\to\infty} \rho_{nat}(\Gamma_n) = \lim_{n\to\infty} \frac{C(\tilde{f}_{nat}^{(n)}, \Gamma_n)}{C(f_{nat}^{*(n)}, \Gamma_n)} = 1$ as $n \to \infty$ (i.e., $T_n \to \infty$). Hence, we can obtain (A.28) with Lemma 1, if (A.29) below holds.

$$\lim_{n\to\infty} \frac{C\big(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}_{ran}^{(n)}), \Gamma_n\big)}{C(\tilde{f}_{nat}^{(n)}, \Gamma_n)} = \lim_{n\to\infty} \frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}\big[C(\tilde{f}_{ran}^{(n)}, \Gamma_n)\big]}{C(\tilde{f}_{nat}^{(n)}, \Gamma_n)} = 1 \qquad \text{(A.29)}$$

Equation (A.29) means that the expected flow $\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}_{ran}^{(n)})$ is *asymptotically* as efficient as $\tilde{f}_{nat}^{(n)}$, and thus almost as efficient as $f_{nat}^{*(n)}$ when n is large enough. Moreover, the mixed NE flow $\tilde{f}_{ran}^{(n)}$ is also asymptotically as efficient as $f_{nat}^{*(n)}$ w.r.t. its expected total cost. Hence, all the flows, $\tilde{f}_{at}^{(n)}$, $f_{at}^{*(n)}$, $\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}_{ran}^{(n)})$, $\tilde{f}_{ran}^{(n)}$, $f_{ran}^{*(n)}$, $\tilde{f}_{nat}^{(n)}$, and $f_{nat}^{*(n)}$, are almost equally efficient, when T_n gets large and (A.29) holds.

To prove (A.29), we only need to consider NE flows. This avoids the difficulties of discussing the SO flows $\tilde{f}_{nat}^{*(n)}$ and $f_{at}^{*(n)}$. To facilitate our discussion, we assume, w.l.o.g., that

– (S4) $\lim_{n\to\infty} d_k^{(n)} \in [0,\infty]$ and $\lim_{n\to\infty} \frac{d_k^{(n)}}{d_{k'}^{(n)}} \in [0,\infty]$ exist for all $k, k' \in \mathcal{K}$.

Note that (A.29) holds for an arbitrary sequence $(\mathcal{S}_n)_{n\in\mathbb{N}}$ satisfying (S1)–(S3) if and only if (A.29) holds for an arbitrary sequence $(\mathcal{S}_n)_{n\in\mathbb{N}}$ satisfying (S1)–(S4). This follows since every infinite subsequence $(\mathcal{S}_{n_j})_{j\in\mathbb{N}}$ of a sequence $(\mathcal{S}_n)_{n\in\mathbb{N}}$ satisfying (S1)–(S3) has an infinite subsequence $(\mathcal{S}_{n_{j_l}})_{l\in\mathbb{N}}$ fulfilling (S1)–(S4). We will use similar subsequence arguments implicitly and repeatedly in this proof.

We now show (A.29) for an arbitrary sequence $(\mathcal{S}_n)_{n\in\mathbb{N}}$ satisfying (S1)–(S4) with the technique of asymptotic decomposition of Wu et al. [47].

Step I: The asymptotic decomposition of Γ_n :

We put $\mathcal{K}_{reg} := \{k \in \mathcal{K} : \lim_{n\to\infty} d_k^{(n)} = \infty\}$ and $\mathcal{K}\setminus\mathcal{K}_{reg} =: \mathcal{K}_{irreg}$. We obtain by (S3)–(S4) that $\mathcal{K}_{reg} \neq \emptyset$. We call k in $\mathcal{K}_{reg}$ *regular*, and k' in $\mathcal{K}_{irreg}$ *irregular*. So $d_k^{(n)}$ is bounded for $k \in \mathcal{K}_{irreg}$, and unbounded for $k \in \mathcal{K}_{reg}$.

We collect these $k \in \mathcal{K}_{reg}$ with an *equal* demand growth rate into one class, which, by property (S4), then results in an *ordered partition* $\mathcal{K}_1 \prec \cdots \prec \mathcal{K}_m$ of $\mathcal{K}_{reg}$ satisfying conditions (AD1)–(AD2).

– (AD1) $\lim_{n\to\infty} \frac{d_k^{(n)}}{d_{k'}^{(n)}} \in (0,\infty)$, i.e., $d_k^{(n)} \in \Theta(d_{k'}^{(n)})$, for all $k, k' \in \mathcal{K}_u$ for each $u \in \mathcal{M} := \{1,\ldots,m\}$,
– (AD2) $\lim_{n\to\infty} \frac{d_k^{(n)}}{d_{k'}^{(n)}} = 0$, i.e., $d_k^{(n)} \in o(d_{k'}^{(n)})$, for all $k \in \mathcal{K}_u, k' \in \mathcal{K}_l$ for all $u, l \in \mathcal{M}$ with $l < u$.

Here, $m \geq 1$ is an integer, and $\mathcal{K}_l \prec \mathcal{K}_u$ means that these $k' \in \mathcal{K}_l$ have demands $d_{k'}^{(n)}$ converging to ∞ much *faster* than the demands $d_k^{(n)}$ of those $k \in \mathcal{K}_u$.

W.r.t. this partition, Γ_n is decomposed into "subgames" $\Gamma_{n|\mathcal{K}_1}, \ldots, \Gamma_{n|\mathcal{K}_m}, \Gamma_{n|\mathcal{K}_{irreg}}$. Here, we call $\Gamma_{n|\mathcal{K}'}$ a *subgame* of Γ_n if $\Gamma_{n|\mathcal{K}'}$ is a restriction of Γ_n to the subset $\mathcal{K}'$ of O/D pairs, i.e., $\Gamma_{n|\mathcal{K}'}$ is the game obtained by removing all O/D pairs $k \in \mathcal{K}\setminus\mathcal{K}'$, and all agents $i \in \cup_{k\in\mathcal{K}\setminus\mathcal{K}'}\mathcal{U}_k^{(n)}$ together with their demands $d_{k,i}^{(n)}$ from Γ_n. We thus ignore completely the influence of all O/D pairs $k \in \mathcal{K}\setminus\mathcal{K}'$ when we consider the subgame $\Gamma_{n|\mathcal{K}'}$.

Clearly, each *regular subgame* $\Gamma_{n|\mathcal{K}_u}$ has the agent set $\mathcal{U}_{|\mathcal{K}_u}^{(n)} := \cup_{k\in\mathcal{K}_u}\mathcal{U}_k^{(n)}$, the demand vector $d_{|\mathcal{K}_u}^{(n)} := (d_{k,i}^{(n)})_{i\in\mathcal{U}_k^{(n)},k\in\mathcal{K}_u}$ and the total demand $T_{n|\mathcal{K}_u} := \sum_{k\in\mathcal{K}_u} d_k^{(n)}$ that tends to ∞ as $n \to \infty$. The *irregular subgame* $\Gamma_{n|\mathcal{K}_{irreg}}$ has the agent set $\mathcal{U}_{|\mathcal{K}_{irreg}}^{(n)} := \cup_{k\in\mathcal{K}_{irreg}}\mathcal{U}_k^{(n)}$, the demand vector $d_{|\mathcal{K}_{irreg}}^{(n)} := (d_{k,i}^{(n)})_{i\in\mathcal{U}_k^{(n)},k\in\mathcal{K}_{irreg}}$ and the total demand $T_{n|\mathcal{K}_{irreg}} := \sum_{k\in\mathcal{K}_{irreg}} d_k^{(n)}$ that tends to a *bounded constant* as $n\to\infty$. Moreover, we obtain by condition (AD2) that

$$\lim_{n\to\infty} \frac{T_{n|\mathcal{K}_u}}{T_{n|\mathcal{K}_l}} = \lim_{n\to\infty} \frac{T_{n|\mathcal{K}\setminus\cup_{l'=1}^{l}\mathcal{K}_{l'}}}{T_{n|\mathcal{K}_l}} = 0 \quad \forall u, l \in \mathcal{M} \text{ with } l < u. \tag{A.30}$$

Here, we observe that $T_n = T_{n|\mathcal{K}_{irreg}} + \sum_{l=1}^{m} T_{n|\mathcal{K}_l}$ and $T_{n|\mathcal{K}\setminus\cup_{l'=1}^{l}\mathcal{K}_{l'}} = T_{n|\mathcal{K}_{irreg}} + \sum_{l'=l+1}^{m} T_{n|\mathcal{K}_{l'}}$.

Note that each flow $f^{(n)}$ of Γ_n induces a *joint* total cost

$$C_{\mathcal{K}'}(f^{(n)}, \Gamma_n) := \sum_{k\in\mathcal{K}'}\sum_{p\in\mathcal{P}_k} f_p^{(n)}\cdot\tau_p(f^{(n)})$$

and an *independent* total cost

$$C(f_{|\mathcal{K}'}^{(n)}, \Gamma_{n|\mathcal{K}'}) = \sum_{k\in\mathcal{K}'}\sum_{p\in\mathcal{P}_k} f_p^{(n)}\cdot\tau_p(f_{|\mathcal{K}'}^{(n)}) = \sum_{a\in A} f_{a|\mathcal{K}'}^{(n)}\cdot\tau_a(f_{a|\mathcal{K}'}^{(n)})$$

for an arbitrary subset $\mathcal{K}'$ of $\mathcal{K}$, where $f_{|\mathcal{K}'}^{(n)} = (f_p^{(n)})_{p\in\cup_{k\in\mathcal{K}'}\mathcal{P}_k}$ is the restriction of $f^{(n)}$ into the subgame $\Gamma_{n|\mathcal{K}'}$, $f_{a|\mathcal{K}'}^{(n)} = \sum_{k\in\mathcal{K}'}\sum_{p\in\mathcal{P}_k:a\in p} f_p^{(n)}$ is the arc flow induced *independently* by the "flow" $f_{|\mathcal{K}'}^{(n)}$ of $\Gamma_{n|\mathcal{K}'}$, and $\tau_p(f_{|\mathcal{K}'}^{(n)}) = \sum_{a\in p}\tau_a(f_{a|\mathcal{K}'}^{(n)})$ is the *independent* path cost under the flow $f_{|\mathcal{K}'}^{(n)}$. Here, we use that $f_{|\mathcal{K}'}^{(n)}$ is indeed a flow of $\Gamma_{n|\mathcal{K}'}$, and so the independent total cost of $f^{(n)}$ is exactly the total cost of the flow $f_{|\mathcal{K}'}^{(n)}$ in $\Gamma_{n|\mathcal{K}'}$.

W.r.t. the above asymptotic decomposition, we obtain for an arbitrary flow $f^{(n)}$ of Γ_n that

$$\begin{aligned} C\left(f^{(n)}, \Gamma_n\right) &= C_{\mathcal{K}_{irreg}}\left(f^{(n)}, \Gamma_n\right) + \sum_{u=1}^{m} C_{\mathcal{K}_u}\left(f^{(n)}, \Gamma_n\right) \\ &\geq C\left(f_{|\mathcal{K}_{irreg}}^{(n)}, \Gamma_{n|\mathcal{K}_{irreg}}\right) + \sum_{u=1}^{m} C\left(f_{|\mathcal{K}_u}^{(n)}, \Gamma_{n|\mathcal{K}_u}\right), \\ f_a^{(n)} &= f_{a|\mathcal{K}_{irreg}}^{(n)} + \sum_{u=1}^{m} f_{a|\mathcal{K}_u}^{(n)} \quad \forall a\in A. \end{aligned}$$

The above inequality follows since the *joint* path cost $\tau_p(f^{(n)})$ considers all subgames and the independent path cost $\tau_p(f_{|\mathcal{K}'}^{(n)})$ considers only flow induced by agents from $k\in\mathcal{K}'$, and so $\tau_p(f^{(n)}) \geq \tau_p(f_{|\mathcal{K}'}^{(n)})$ for each subset $\mathcal{K}'$ of $\mathcal{K}$.

Step II: An equivalent transformation in the limit

Wu et al. [47] have shown for this decomposition of non-atomic NE flows that

$$\begin{aligned} &\lim_{n\to\infty}\frac{C(\tilde{f}_{nat}^{(n)}, \Gamma_n)}{C(\tilde{f}_{nat}^{(n,-)}, \Gamma_{n|\mathcal{K}_{irreg}}) + \sum_{l=1}^{m} C(\tilde{f}_{nat}^{(n,l)}, \Gamma_{n|\mathcal{K}_l})} \\ &= \lim_{n\to\infty}\frac{C_{\mathcal{K}_{irreg}}(\tilde{f}_{nat}^{(n)}, \Gamma_n) + \sum_{l=1}^{m} C_{\mathcal{K}_l}(\tilde{f}_{nat}^{(n)}, \Gamma_n)}{C(\tilde{f}_{nat}^{(n,-)}, \Gamma_{n|\mathcal{K}_{irreg}}) + \sum_{l=1}^{m} C(\tilde{f}_{nat}^{(n,l)}, \Gamma_{n|\mathcal{K}_l})} \\ &= \lim_{n\to\infty}\frac{\sum_{l=1}^{m} C_{\mathcal{K}_l}(\tilde{f}_{nat}^{(n)}, \Gamma_n)}{\sum_{l=1}^{m} C(\tilde{f}_{nat}^{(n,l)}, \Gamma_{n|\mathcal{K}_l})} = 1, \end{aligned} \tag{A.31}$$

where $\tilde{f}_{nat}^{(n,l)}$ and $\tilde{f}_{nat}^{(n,-)}$ are non-atomic NE flows of $\Gamma_{n|\mathcal{K}_l}$ and $\Gamma_{n|\mathcal{K}_{irreg}}$, respectively, and where $C_{\mathcal{K}_l}(\tilde{f}_{nat}^{(n)}, \Gamma_n) = \sum_{k\in\mathcal{K}_l}\sum_{p\in\mathcal{P}_k}\tilde{f}_{nat,p}^{(n)}\cdot\tau_p(\tilde{f}_{nat}^{(n)})$ is the joint total cost of $\Gamma_{n|\mathcal{K}_l}$ in the non-atomic NE flow $\tilde{f}_{nat}^{(n)}$ of Γ_n. Note that the restriction $\tilde{f}_{nat|\mathcal{K}_l}^{(n)} = (\tilde{f}_{nat,p}^{(n)})_{p\in\mathcal{P}_k,k\in\mathcal{K}_l}$ of $\tilde{f}_{nat}^{(n)}$ is a non-atomic flow of $\Gamma_{n|\mathcal{K}_l}$, but need not be a non-atomic NE flow of $\Gamma_{n|\mathcal{K}_l}$, and so has a total cost that may differ from $\tilde{f}_{nat}^{(n,l)}$.

The irregular subgame vanishes in the limit of (A.31), since it has a bounded total demand and thus a negligible influence on the limit, see [47] for details.

For each $n\in\mathbb{N}$, let $\tilde{f}_{nat}^{(n,l)}$, $l\in\mathcal{M}=\{1,\ldots,m\}$, be arbitrary non-atomic NE flows of subgames $\Gamma_{n|\mathcal{K}_l}$, and let $\tilde{f}_{nat}^{(n,-)}$ be an arbitrary non-atomic NE flow of $\Gamma_{n|\mathcal{K}_{irreg}}$.

Then (A.29) follows from (A.31) if and only if

$$\lim_{n\to\infty}\frac{C\left(\mathbb{E}_{\tilde{\Pi}^{(n)}}\left(\tilde{f}_{ran}^{(n)}\right),\Gamma_n\right)}{C(\tilde{f}_{nat}^{(n,-)},\Gamma_{n|\mathcal{K}_{irreg}})+\sum_{l=1}^m C(\tilde{f}_{nat}^{(n,l)},\Gamma_{n|\mathcal{K}_l})}=1, \tag{A.32}$$

$$\lim_{n\to\infty}\frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}\left[C(\tilde{f}_{ran}^{(n)},\Gamma_n)\right]}{C(\tilde{f}_{nat}^{(n,-)},\Gamma_{n|\mathcal{K}_{irreg}})+\sum_{l=1}^m C(\tilde{f}_{nat}^{(n,l)},\Gamma_{n|\mathcal{K}_l})}=1. \tag{A.33}$$

Step III: Further subsequence arguments

We will prove (A.32)–(A.33) by scaling each of the above regular subgames $\Gamma_{n|\mathcal{K}_u}$ independently. We define a scaling factor $g_n^{(u)} := T_{n|\mathcal{K}_u}^{\lambda_u}$ for each $u\in\mathcal{M}$, where $\lambda_u := \max_{k\in\mathcal{K}_u}\min_{p\in\mathcal{P}_k}\max_{a\in p}\beta_a\geq 0$. To facilitate the discussion, we also call a path $p\in\cup_{k\in\mathcal{K}_u}\mathcal{P}_k$ for $u\in\mathcal{M}$ *tight* if $\max_{a\in p}\beta_a\leq\lambda_u$, and *non-tight* if $\max_{a\in p}\beta_a>\lambda_u$. Clearly, every $k\in\mathcal{K}_u$ has *at least* one tight path. Moreover, each tight path $p\in\cup_{k\in\mathcal{K}_u}\mathcal{P}_k$ contains only arcs $a\in A$ with $\beta_a\leq\lambda_u$, while a non-tight path $p'\in\cup_{k\in\mathcal{K}_u}\mathcal{P}_k$ contains *at least* one arc $a\in A$ with $\beta_a>\lambda_u$, for each $u\in\mathcal{M}$. These simple facts will be very helpful in the further discussion.

To simplify the proof, we assume further that the sequence $(\mathcal{S}_n)_{n\in\mathbb{N}}$ satisfies properties (S5)–(S8) below.

(S5) $\lim_{n\to\infty}\frac{g_n^{(u)}}{g_n^{(l)}}\in[0,\infty]$ exists for $u,l\in\mathcal{M}$. We call $g_n^{(u)}$ and $g_n^{(l)}$ *mutually comparable*.

(S6) $\lim_{n\to\infty}\frac{\tilde{f}_{nat}^{(n,u)}}{T_{n|\mathcal{K}_u}} = \lim_{n\to\infty}\frac{\left(\tilde{f}_{nat,p}^{(n,u)}\right)_{p\in\mathcal{P}_k,k\in\mathcal{K}_u}}{T_{n|\mathcal{K}_u}} =: \tilde{f}_{nat}^{(\infty,u)} = (\tilde{f}_{nat,p}^{(\infty,u)})_{p\in\mathcal{P}_k,k\in\mathcal{K}_u}$ for $u\in\mathcal{M}$.

(S7) For $u\in\mathcal{M}$, $\lim_{n\to\infty}\frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}_{ran}^{(n)})_{|\mathcal{K}_u}}{T_{n|\mathcal{K}_u}} = \lim_{n\to\infty}\frac{\left(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}_{ran,p}^{(n)})\right)_{p\in\mathcal{P}_k,k\in\mathcal{K}_u}}{T_{n|\mathcal{K}_u}} =: \tilde{f}_{exp}^{(\infty,u)} = (\tilde{f}_{exp,p}^{(\infty,u)})_{p\in\mathcal{P}_k,k\in\mathcal{K}_u}$. Here, $\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}_{ran}^{(n)})_{|\mathcal{K}_u} = \mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}_{ran|\mathcal{K}_u}^{(n)})$ is the restriction of the expected flow $\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}_{ran}^{(n)}) = (\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}_{ran,p}^{(n)}))_{p\in\mathcal{P}}$ of the mixed NE flow $\tilde{f}_{ran}^{(n)}$ of Γ_n to the subgame $\Gamma_{n|\mathcal{K}_u}$, which is a non-atomic flow of $\Gamma_{n|\mathcal{K}_u}$.

(S8) $\lim_{n\to\infty}\frac{d_k^{(n)}}{T_{n|\mathcal{K}_u}} =: d_k^{(\infty,u)}\in(0,1]$ for each $k\in\mathcal{K}_u$ and each $u\in\mathcal{M}$. This actually follows directly from property (S4) and decomposition condition (AD1).

Note that (A.32)–(A.33) hold for an arbitrary sequence $(\mathcal{S}_n)_{n\in\mathbb{N}}$ fulfilling (S1)–(S4) if and only if they hold for an arbitrary sequence $(\mathcal{S}_n)_{n\in\mathbb{N}}$ satisfying (S1)–(S8). This follows again since every infinite subsequence $(\mathcal{S}_{n_j})_{j\in\mathbb{N}}$ of an sequence $(\mathcal{S}_n)_{n\in\mathbb{N}}$ fulfilling (S1)–(S4) contains an infinite subsequence $(\mathcal{S}_{n_{j_l}})_{l\in\mathbb{N}}$ fulfilling (S1)–(S8).

Step IV: The inductive assumptions

We will prove (A.32)–(A.33) by showing that the statements IA1–IA7 below hold for each $u \in \mathcal{M}$, using an induction on u over the set $\{0,\ldots,m\} = \{0\}\bigcup\mathcal{M}$. Here, we put $\mathcal{K}_0 := \emptyset$, $g_n^{(0)} := 0$ and identify $\Gamma_{n|\mathcal{K}_0}$ as the empty subgame and employ the convention that IA1–IA7 hold for $u = 0$.

IA1 $\max_{p\in\mathcal{P}_k:\ \mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p})>0} \tau_p(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran})) \in O\big(\max_{l=0}^{u}\ g_n^{(l)}\big)$ for $k \in \cup_{l=0}^{u}\mathcal{K}_l$, i.e., the *most costly* path used by agents of the subgame $\Gamma_{n|\cup_{l=0}^{u}\mathcal{K}_l}$ has a cost of *at most* $O\big(\max_{l=0}^{u}\ g_n^{(l)}\big)$ in the expected flow $\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran})$.

IA2 The joint total cost of $\Gamma_{n|\cup_{l=0}^{u}\mathcal{K}_l}$ in $\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran})$ is $\Theta\big(\max_{l=0}^{u}\ g_n^{(l)}\cdot T_{n|\mathcal{K}_l}\big)$, i.e.,

$$\begin{aligned}
&\sum_{l=0}^{u} C_{\mathcal{K}_l}(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran}),\Gamma_n)\\
&\quad=\sum_{l=0}^{u}\sum_{k\in\mathcal{K}_l}\sum_{p\in\mathcal{P}_k}\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p})\cdot\tau_p(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran}))\\
&\quad\in\Theta\big(\max_{l=0}^{u}\ g_n^{(l)}\cdot T_{n|\mathcal{K}_l}\big).
\end{aligned}$$

IA3 $\lim_{n\to\infty}\frac{\sum_{l=0}^{u}C_{\mathcal{K}_l}(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran}),\Gamma_n)}{\sum_{l=0}^{u}C(\tilde{f}^{(n,l)}_{nat},\Gamma_{n|\mathcal{K}_l})}=1.$

IA4 $\max_{p\in\mathcal{P}_k:\ \mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p})>0}\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tau_p(\tilde{f}^{(n)}_{ran}))\in O\big(\max_{l=0}^{u}\ g_n^{(l)}\big)$ for $k\in\cup_{l=0}^{u}\mathcal{K}_l$.

IA5 The expected joint total cost of $\Gamma_{n|\cup_{l=0}^{u}\mathcal{K}_l}$ in $\tilde{f}^{(n)}_{ran}$ is also $\Theta\big(\max_{l=0}^{u}\ g_n^{(l)}\cdot T_{n|\mathcal{K}_l}\big)$, i.e.,

$$\begin{aligned}
\mathbb{E}_{\tilde{\Pi}^{(n)}}\left[\sum_{l=0}^{u}C_{\mathcal{K}_l}\left(\tilde{f}^{(n)},\Gamma_n\right)\right]&=\sum_{l=0}^{u}\sum_{k\in\mathcal{K}_l}\sum_{p\in\mathcal{P}_k}\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p}\cdot\tau_p(\tilde{f}^{(n)}_{ran}))\\
&\in\Theta\left(\max_{l=0}^{u}\ g_n^{(l)}\cdot T_{n|\mathcal{K}_l}\right).
\end{aligned}$$

IA6 $\lim_{n\to\infty}\frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}[\sum_{l=0}^{u}C_{\mathcal{K}_l}(\tilde{f}^{(n)},\Gamma_n)]}{\sum_{l=0}^{u}C(\tilde{f}^{(n,l)}_{nat},\Gamma_{n|\mathcal{K}_l})}=1.$

IA7 For each $k\in\mathcal{K}_l$ and each $l=0,\ldots,u$,

$$\max_{p\in\mathcal{P}_k:\ \mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p})>0}\mathbb{E}_{\tilde{\Pi}^{(n)}}[\tilde{f}^{(n)}_{ran,p}\cdot\tau_p(\tilde{f}^{(n)}_{ran})]\in O\left(T_{n|\mathcal{K}_l}\cdot\max_{l'=0}^{l}\ g_n^{(l')}\right).$$

Among these inductive assumptions, IA3 and IA6 are the most crucial. We obtain trivially that

$$\lim_{n\to\infty} \frac{\sum_{l=0}^{m}\sum_{k\in\mathcal{K}_l}\sum_{p\in\mathcal{P}_k}\mathbb{E}_{\tilde{\Pi}^{(n)}}\left(\tilde{f}^{(n)}_{ran,p}\right)\cdot\tau_p\left(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran})\right)}{\sum_{l=0}^{m} C(\tilde{f}^{(n,l)}_{nat},\, \Gamma_{n|\mathcal{K}_l})} = 1, \tag{A.34}$$

$$\lim_{n\to\infty} \frac{\sum_{l=0}^{m}\sum_{k\in\mathcal{K}_l}\sum_{p\in\mathcal{P}_k}\mathbb{E}_{\tilde{\Pi}^{(n)}}\left(\tilde{f}^{(n)}_{ran,p}\cdot\tau_p(\tilde{f}^{(n)}_{ran})\right)}{\sum_{l=0}^{m} C(\tilde{f}^{(n,l)}_{nat},\, \Gamma_{n|\mathcal{K}_l})} = 1, \tag{A.35}$$

when IA3 and IA6 hold for all $u \in \mathcal{M} = \{1, \ldots, m\}$. Then (A.32)–(A.33) follow immediately from (A.34)–(A.35), since the subgame $\Gamma_{n|\mathcal{K}_{irreg}}$ has a bounded total demand and thus can be neglected in the limits by an argument similar to that in the proof of Fact A2 below.

Moreover, IA4 implies IA7. This follows since the random event "$\tilde{f}^{(n)}_{ran,p} \le T_{n|\mathcal{K}_l}$" occurs almost surely for each $p \in \cup_{k\in\mathcal{K}_l}\mathcal{P}_k$ and each $l \in \mathcal{M}$. In fact, IA4 also implies IA1, which we will claim later in Fact A1.

Now, we consider an arbitrary $u \in \{0, \ldots, m-1\}$ such that IA1–IA7 hold for each non-negative integer $l \le u$. We will prove IA1–IA7 for $u+1$, which then implies (A.34)–(A.35) by induction, and so completes the proof of Theorem 3a–b.

Step V: Validating IA1–IA7 for $u+1$

For each $k \in \mathcal{K}_{u+1}$ and each $p \in \mathcal{P}_k$, $\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p}) = 0$ implies that $\tilde{\Pi}^{(n)}_{i,p} = 0$ for every $i \in \mathcal{U}^{(n)}_k$ because of (2.2). So, for each $p \in \mathcal{P}_k$, $\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p}) = 0$ is equivalent to the fact that the random event "$\tilde{f}^{(n)}_{ran,p} = 0$" occurs *almost surely*, i.e., $\mathbb{P}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p} = 0) = 1$. Similarly, for each $a \in A$, $\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\cup_{l=0}^{u}\mathcal{K}_l}) = 0$ is equivalent to the fact that the random event "$\tilde{f}^{(n)}_{ran,a|\cup_{l=0}^{u}\mathcal{K}_l} = 0$" occurs almost surely. Therefore, we can directly remove $\tilde{f}^{(n)}_{ran,a|\cup_{l=0}^{u}\mathcal{K}_l}$ from the respective expectations of the random variables $\tau_a(\tilde{f}^{(n)}_{ran,a})$ and $\tilde{f}^{(n)}_{ran,a}\cdot\tau_a(\tilde{f}^{(n)}_{ran,a})$ when $\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\cup_{l=0}^{u}\mathcal{K}_l}) = 0$.

With the above observations and the inductive assumptions IA1 and IA4 of step u, we obtain (A.36)–(A.38) for every arc $a \in A$ and every path $p \in \cup_{k\in\mathcal{K}_{u+1}}\mathcal{P}_k$.

$$\begin{aligned}
&\tau_a\big[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a})\big]\\
&\quad= \tau_a\big[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\cup_{l=0}^{u}\mathcal{K}_l}) + \mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l})\big]\\
&\quad= \begin{cases} O\big(\max_{l=0}^{u}\, g_n^{(l)}\big) & \text{if } \mathbb{E}_{\tilde{\Pi}^{(n)}}[\tilde{f}^{(n)}_{ran,a|\cup_{l=0}^{u}\mathcal{K}_l}] > 0,\\ \tau_a\big[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l})\big] & \text{if } \mathbb{E}_{\tilde{\Pi}^{(n)}}[\tilde{f}^{(n)}_{ran,a|\cup_{l=0}^{u}\mathcal{K}_l}] = 0,\end{cases}
\end{aligned} \tag{A.36}$$

$$\begin{aligned}
&\mathbb{E}_{\tilde{\Pi}^{(n)}}\big[\tau_a(\tilde{f}^{(n)}_{ran,a})\big]\\
&\quad= \mathbb{E}_{\tilde{\Pi}^{(n)}}\big[\tau_a(\tilde{f}^{(n)}_{ran,a|\cup_{l=0}^{u}\mathcal{K}_l} + \tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\bigcup_{l=0}^{u}\mathcal{K}_l})\big]\\
&\quad= \begin{cases} O(\max_{l=0}^{u}\, g_n^{(l)}) & \text{if } \mathbb{E}_{\tilde{\Pi}^{(n)}}\big[\tilde{f}^{(n)}_{ran,a|\cup_{l=0}^{u}\mathcal{K}_l}\big] > 0,\\ \mathbb{E}_{\tilde{\Pi}^{(n)}}\big[\tau_a(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=1}^{u}\mathcal{K}_l})\big] & \text{if } \mathbb{E}_{\tilde{\Pi}^{(n)}}\big[\tilde{f}^{(n)}_{ran,a|\cup_{l=0}^{u}\mathcal{K}_l}\big] = 0,\end{cases}
\end{aligned} \tag{A.37}$$

$$\begin{aligned}
&\mathbb{E}_{\tilde{\Pi}^{(n)}}\big[\tilde{f}^{(n)}_{ran,p}\cdot\tau_a(\tilde{f}^{(n)}_{ran,a})\big]\\
&\quad=\mathbb{E}_{\tilde{\Pi}^{(n)}}\big[\tilde{f}^{(n)}_{ran,p}\cdot\tau_a(\tilde{f}^{(n)}_{ran,a|\cup_{l=0}^{u}\mathcal{K}_l}+\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l})\big]\\
&\quad=\begin{cases} O\big(T_{n|\mathcal{K}_{u+1}}\cdot\max_{l=0}^{u}\ g_n^{(l)}\big) & \text{if } \mathbb{E}_{\tilde{\Pi}^{(n)}}[\tilde{f}^{(n)}_{ran,a|\cup_{l=0}^{u}\mathcal{K}_l}]>0,\\ \mathbb{E}_{\tilde{\Pi}^{(n)}}\big[\tilde{f}^{(n)}_{ran,p}\cdot\tau_a(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l})\big] & \text{if } \mathbb{E}_{\tilde{\Pi}^{(n)}}[\tilde{f}^{(n)}_{ran,a|\cup_{l=0}^{u}\mathcal{K}_l}]=0.\end{cases}
\end{aligned}\tag{A.38}$$

(A.36) and (A.37) follow since IA1 and IA4 hold in steps $l\le u$, and since the expected arc flow $\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\cup_{l=0}^{u}\mathcal{K}_l})>0$ implies that arc a belongs to some path $p\in\cup_{l=0}^{u}\cup_{k\in\mathcal{K}_l}\mathcal{P}_k$ with $\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p})>0$. (A.38) follows immediately from (A.37) and the fact that $\tilde{f}^{(n)}_{ran,p}\le T_{n|\mathcal{K}_{u+1}}$ for every path $p\in\cup_{k\in\mathcal{K}_{u+1}}\mathcal{P}_k$. Here, we observe that (A.36)–(A.38) hold trivially when $u=0$, i.e., when $\cup_{l=0}^{u}\mathcal{K}_l=\emptyset$. □

With (A.36)–(A.37), we now show IA1, IA4 and IA7 for step $u+1$.

Fact A1 *IA1, IA4 and IA7 hold for step* $u+1$.

Proof of Fact A1 We only need to show IA4 and IA1, as IA4 implies IA7.

Proof of IA4 We obtain by (A.37) that $\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tau_p(\tilde{f}^{(n)}_{ran}))\in O(\max_{l=0}^{u+1}\ g_n^{(l)})$ for every *tight* path $p\in\cup_{k\in\mathcal{K}_{u+1}}\mathcal{P}_k$. This follows since a tight path consists of arcs a with degrees $\beta_a\le\lambda_{u+1}$, $g_n^{(u+1)}=T^{\lambda_{u+1}}_{n|\mathcal{K}_{u+1}}$, and $T_{n|\mathcal{K}_{u+1}}\in\Theta(T_{n|\mathcal{K}_l\setminus\cup_{l=0}^{u}\mathcal{K}_l})$, see (A.30). Then IA4 of step $u+1$ follows immediately from the facts that every $k\in\mathcal{K}_{u+1}$ has *at least* one tight path, that $\tilde{f}^{(n)}_{ran}$ is a mixed NE flow, and that $d^{(n)}_{k,i}\le\upsilon$ for all k and i. Here, we use that the choice of a single agent has a *negligible* influence on the expected cost of a path (compared to $\max_{l=0}^{u+1}\ g_n^{(l)}$) when n is large enough, since his demand is bounded from above by the constant υ and $T_{n|\mathcal{K}_{u+1}}\to\infty$ as $n\to\infty$. In fact, we can even think of $\tilde{f}^{(n)}_{ran}$ as a mixed WE flow (see Remark 1) in this proof.

Proof of IA1

We show for each $a\in A$ that

$$\tau_a[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a})]\le\mathbb{E}_{\tilde{\Pi}^{(n)}}[\tau_a(\tilde{f}^{(n)}_{ran,a})]+O(1),\tag{A.39}$$

which, combined with IA4 of step $u+1$, implies IA1 in step $u+1$, since $\max_{l=0}^{u+1}\ g_n^{(l)}\in\Omega(1)$ for every $u=0,\ldots,m-1$. Note that $\tau_a(\cdot)$ is convex on $[M_a,\infty)$ for some constant $M_a>0$, since $\tau_a(\cdot)$ is a non-decreasing polynomial with an integer degree

$\beta_a \geq 0$. Jensen's inequality from Lemma 8c then yields that

$$\begin{aligned}\tau_a\left[\mathbb{E}_{\tilde{\Pi}^{(n)}}\left(\tilde{f}^{(n)}_{ran,a}\right)\right] &= \tau_a\left[\mathbb{E}_{\tilde{\Pi}^{(n)}}\left(\tilde{f}^{(n)}_{ran,a}|\tilde{f}^{(n)}_{ran,a} \geq M_a\right)\right]\cdot\mathbb{P}_{\tilde{\Pi}^{(n)}}\left[\tilde{f}^{(n)}_{ran,a} \geq M_a\right]\\ &\quad + \tau_a\left[\mathbb{E}_{\tilde{\Pi}^{(n)}}\left(\tilde{f}^{(n)}_{ran,a}|\tilde{f}^{(n)}_{ran,a} < M_a\right)\right]\cdot\mathbb{P}_{\tilde{\Pi}^{(n)}}\left[\tilde{f}^{(n)}_{ran,a} < M_a\right]\\ &\leq \mathbb{E}_{\tilde{\Pi}^{(n)}}\left[\tau_a\left(\tilde{f}^{(n)}_{ran,a}\right)|\tilde{f}^{(n)}_{ran,a} \geq M_a\right]\cdot\mathbb{P}_{\tilde{\Pi}^{(n)}}\left[\tilde{f}^{(n)}_{ran,a} \geq M_a\right]\\ &\quad + \tau_a\left[\mathbb{E}_{\tilde{\Pi}^{(n)}}\left(\tilde{f}^{(n)}_{ran,a}|\tilde{f}^{(n)}_{ran,a} < M_a\right)\right]\cdot\mathbb{P}_{\tilde{\Pi}^{(n)}}\left[\tilde{f}^{(n)}_{ran,a} < M_a\right]\\ &\leq \mathbb{E}_{\tilde{\Pi}^{(n)}}\left[\tau_a\left(\tilde{f}^{(n)}_{ran,a}\right)|\tilde{f}^{(n)}_{ran,a} \geq M_a\right]\cdot\mathbb{P}_{\tilde{\Pi}^{(n)}}\left[\tilde{f}^{(n)}_{ran,a} \geq M_a\right]\\ &\quad + \mathbb{E}_{\tilde{\Pi}^{(n)}}\left[\tau_a\left(\tilde{f}^{(n)}_{ran,a}\right)|\tilde{f}^{(n)}_{ran,a} < M_a\right]\cdot\mathbb{P}_{\tilde{\Pi}^{(n)}}\left[\tilde{f}^{(n)}_{ran,a} < M_a\right]\\ &\quad + \tau_a\left(M_a\right) = \mathbb{E}_{\tilde{\Pi}^{(n)}}\left[\tau_a\left(\tilde{f}^{(n)}_{ran,a}\right)\right] + O\left(1\right).\end{aligned}$$

This proves IA1 for step $u+1$, and completes the proof of Fact A1[1]. □

Note that either $g_n^{(u+1)} \in O(\max_{l=0}^{u}\ g_n^{(l)})$ or $g_n^{(u+1)} \in \omega(\max_{l=0}^{u}\ g_n^{(l)})$, since the scaling factors are mutually comparable, i.e., the sequence $(\mathcal{S}_n)_{n\in\mathbb{N}}$ satisfies property (S5). To validate IA2–IA3 and IA5–IA6, we thus distinguish *two* subcases.

Subcases I: $g_n^{(u+1)} \in O(\max_{l=0}^{u}\ g_n^{(l)})$

Fact A2 shows IA2–IA3, and IA5–IA6 for step $u+1$ when $g_n^{(u+1)} \in O(\max_{l=0}^{u}\ g_n^{(l)})$. Then Fact A1–Fact A2 together imply IA1–IA7 for step $u+1$ when $g_n^{(u+1)} \in O(\max_{l=0}^{u}\ g_n^{(l)})$. Here, we observe that $g_n^{(u+1)} \in O(\max_{l=0}^{u}\ g_n^{(l)})$ happens only when $u > 0$, since $g_n^{(0)} = 0$ and $g_n^{(u+1)} \in \Omega(1)$ for each $u \in \{0,\ldots,m-1\}$.

Fact A2 *If* $g_n^{(u+1)} \in O(\max_{l=0}^{u}\ g_n^{(l)})$, *then IA2–IA3, and IA5–IA6 hold at step* $u+1$.

Proof of Fact A2 IA1 of step $u+1$ yields

$$\begin{aligned}\mathbb{E}_{\tilde{\Pi}^{(n)}}[\tilde{f}^{(n)}_{ran,p}]\cdot\tau_p\big[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran})\big] &\in O\big(T_{n|\mathcal{K}_{u+1}}\cdot\max_{l=0}^{u+1}\ g_n^{(l)}\big)\\ &= O\big(T_{n|\mathcal{K}_{u+1}}\cdot\max_{l=0}^{u}\ g_n^{(l)}\big)\end{aligned}$$

[1] There is an alternative proof that does not need the convexity of the polynomial cost functions. The random variable $X_n := \tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l}$ has a variance of at most $\upsilon\cdot\mathbb{E}_{\tilde{\Pi}^{(n)}}[X_n]$, and so the random event

Footnote 1 continued

"$X_n \leq \mathbb{E}_{\tilde{\Pi}^{(n)}}[X_n]-\sqrt{2\cdot\upsilon\cdot\mathbb{E}_{\tilde{\Pi}^{(n)}}[X_n]}$" occurs with a probability of at most $\frac{1}{2}$ by Chebyshev's inequality from Lemma 8b. This then implies $\tau_a(\mathbb{E}_{\Pi^{(n)}}[X_n]) \in O(\mathbb{E}_{\tilde{\Pi}^{(n)}}[\tau_a(X_n)])$, and so IA1 in step $u+1$ holds by IA4 of step $u+1$ and (A.36). Hence, Theorem 3 carries also over to non-decreasing polynomial cost functions with arbitrary non-negative real-valued degrees, since only the above proof of IA1 for step $u+1$ involves the convexity of the cost functions.

for every $p \in \cup_{k\in\mathcal{K}_{u+1}}\mathcal{P}_k$ with $\mathbb{E}_{\tilde{\Pi}^{(n)}}[\tilde{f}^{(n)}_{ran,p}] > 0$ when $g_n^{(u+1)} \in O(\max_{l=0}^{u} g_n^{(l)})$. This in turn implies with (A.30) that

$$C_{\mathcal{K}_{u+1}}\big[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran}), \Gamma_n\big] = \sum_{k\in\mathcal{K}_{u+1}}\sum_{p\in\mathcal{P}_k}\mathbb{E}_{\tilde{\Pi}^{(n)}}[\tilde{f}^{(n)}_{ran,p}]\cdot\tau_p\big[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran})\big]$$
$$\in O\big(T_{n|\mathcal{K}_{u+1}}\cdot\max_{l=0}^{u} g_n^{(l)}\big)\subseteq o\big(\max_{l=0}^{u} T_{n|\mathcal{K}_l}\cdot g_n^{(l)}\big). \tag{A.40}$$

Then IA2 of step $u+1$ follows from (A.30), $g_n^{(u+1)} \in O(\max_{l=0}^{u} g_n^{(l)})$, and IA2 of step u.

IA3 of step $u+1$ then follows from (A.40), IA3 of step u, and (A.41),

$$C(\tilde{f}^{(n,u+1)}_{nat}, \Gamma_{n|\mathcal{K}_{u+1}}) \in \Theta(T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)})\subseteq o\big(\max_{l=0}^{u} T_{n|\mathcal{K}_l}\cdot g_n^{(l)}\big), \tag{A.41}$$

see (A.43) of Fact A3 below.

(A.40)–(A.41) show that $\Gamma_{n|\mathcal{K}_{u+1}}$ is negligible when we compute its respective total cost in the expected flow of $\tilde{f}^{(n)}_{ran}$ and $\tilde{f}^{(n,u+1)}_{nat}$, and when $g_n^{(u+1)} \in O(\max_{l=0}^{u} g_n^{(l)})$. Similarly, we can obtain IA5–IA6 of step $u+1$ by showing that $\Gamma_{n|\mathcal{K}_{u+1}}$ is again negligible when we compute its joint expected total cost in $\tilde{f}^{(n)}_{ran}$ and when $g_n^{(u+1)} \in O(\max_{l=0}^{u} g_n^{(l)})$, where we use IA4 and IA7 of step $u+1$.

This completes the proof of Fact A2. □

Subcase II: $g_n^{(u+1)} \in \omega(\max_{l=1}^{u} g_n^{(l)})$

We now show IA2–IA3 and IA5–IA6 for step $u+1$ when $g_n^{(u+1)} \in \omega(\max_{l=1}^{u} g_n^{(l)})$. This, together with Fact A1 and Fact A2, completes the proof of IA1–IA7 for step $u+1$.

Fact A3 below states a helpful result from Wu et al. [47], which shows that the limit $\tilde{f}^{(\infty,u+1)}_{nat} = \lim_{n\to\infty}\frac{\tilde{f}^{(n,u+1)}_{nat}}{T_{n|\mathcal{K}_{u+1}}}$ in (S6) is a non-atomic NE flow of a limit game $\Gamma^{(\infty)}_{|\mathcal{K}_{u+1}}$, and the scaled non-atomic NE cost $\frac{C(\tilde{f}^{(n,u+1)}_{nat},\Gamma_{n|\mathcal{K}_{u+1}})}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}}$ of subgame $\Gamma_{n|\mathcal{K}_{u+1}}$ converges to the total cost of the non-atomic NE flow $\tilde{f}^{(\infty,u+1)}_{nat}$ of $\Gamma^{(\infty)}_{|\mathcal{K}_{u+1}}$. Here, $\Gamma^{(\infty)}_{|\mathcal{K}_{u+1}}$ is a (non-atomic) congestion game with (O/D pair) demand vector $d^{(\infty,u+1)} = (d_k^{(\infty,u+1)})_{k\in\mathcal{K}_{u+1}} = \lim_{n\to\infty}\frac{d^{(n)}_{|\mathcal{K}_{u+1}}}{T_{n|\mathcal{K}_{u+1}}}\lim_{n\to\infty}\frac{(d_k^{(n,u+1)})_{k\in\mathcal{K}_{u+1}}}{T_{n|\mathcal{K}_{u+1}}}$ and cost function

$$\tau_a^{(\infty,u+1)}(x) = \lim_{y\to x^+}\lim_{n\to\infty}\frac{\tau_a\big(T_{n|\mathcal{K}_{u+1}}\cdot y\big)}{g_n^{(u+1)}} = \begin{cases}\infty & \text{if } \beta_a > \lambda_{u+1},\\ \eta_a\cdot x^{\beta_a} & \text{if } \beta_a = \lambda_{u+1},\\ 0 & \text{if } \beta_a < \lambda_{u+1},\end{cases} \tag{A.42}$$

for every $x \in [0, 1]$ and every arc $a \in A$.

Fact A3 (See [47]) *For each* $u = \{0, \ldots, m-1\} = \{0\} \cup (\mathcal{M} \setminus \{m\})$,

$$\begin{aligned}
&\lim_{n\to\infty} \frac{C(\tilde{f}_{nat}^{(n,u+1)}, \Gamma_{n|\mathcal{K}_{u+1}})}{T_{n|\mathcal{K}_{u+1}} \cdot g_n^{(u+1)}} \\
&\quad = \lim_{n\to\infty} \sum_{k\in\mathcal{K}_{u+1}} \sum_{p\in\mathcal{P}_k} \frac{\tilde{f}_{nat,p}^{(n,u+1)}}{T_{n|\mathcal{K}_{u+1}}} \cdot \frac{\tau_p(\tilde{f}_{nat}^{(n,u+1)})}{g_n^{(u+1)}} \\
&\quad = \sum_{a\in A} f_{nat,a}^{(\infty,u+1)} \cdot \tau_a^{(\infty,u+1)}(f_{nat,a}^{(\infty,u+1)}) \in (0,\infty)
\end{aligned} \tag{A.43}$$

and $f_{nat}^{(\infty,u+1)}$ *is a non-atomic NE flow of* $\Gamma_{|\mathcal{K}_{u+1}}^{(\infty)}$ *s.t.* $f_{nat.p}^{(\infty,u+1)} = 0$ *for each non-tight* $p \in \cup_{k\in\mathcal{K}_{u+1}} \mathcal{P}_k$*. Here, we employ the convention that* $0 \cdot \infty = 0$.

Properties similar to Fact A3 actually carry over to the expected flow $\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}_{ran|\mathcal{K}_{u+1}}^{(n)})$ when $g_n^{(u+1)} \in \omega(\max_{l=0}^{u} g_n^{(l)})$. Here, we recall that $\lim_{n\to\infty} \frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}_{ran|\mathcal{K}_{u+1}}^{(n)})}{T_{n|\mathcal{K}_{u+1}}} = \tilde{f}_{exp}^{(\infty,u+1)}$, see (S7).

Consider an arbitrary arc $a \in A$ with $\beta_a \le \lambda_{u+1}$. Then (A.36) yields

$$\begin{aligned}
&\lim_{n\to\infty} \frac{\tau_a\big(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}_{ran,a}^{(n)})\big)}{g_n^{(u+1)}} \\
&= \lim_{n\to\infty} \frac{\tau_a\big(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}_{ran,a}^{(n)})\big)}{g_n^{(u+1)}} \cdot \mathbb{1}_{(0,\infty)}\Big(\mathbb{E}_{\tilde{\Pi}^{(n)}}[\tilde{f}_{ran,a|\cup_{l=0}^{u}\mathcal{K}_l}^{(n)}]\Big) \\
&\quad + \lim_{n\to\infty} \frac{\tau_a\Big(\mathbb{E}_{\tilde{\Pi}^{(n)}}\Big(\tilde{f}_{ran,a|\mathcal{K}_{u+1}}^{(n)} + \tilde{f}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u+1}\mathcal{K}_l}^{(n)}\Big)\Big)}{g_n^{(u+1)}} \cdot \mathbb{1}_{\{0\}}\Big(\mathbb{E}_{\tilde{\Pi}^{(n)}}[\tilde{f}_{ran,a|\cup_{l=0}^{u}\mathcal{K}_l}^{(n)}]\Big) \\
&= \lim_{n\to\infty} \frac{\tau_a\left(T_{n|\mathcal{K}_{u+1}} \cdot \mathbb{E}_{\tilde{\Pi}^{(n)}}\left(\frac{\tilde{f}_{ran,a|\mathcal{K}_{u+1}}^{(n)}}{T_{n|\mathcal{K}_{u+1}}} + \frac{\tilde{f}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u+1}\mathcal{K}_l}^{(n)}}{T_{n|\mathcal{K}_{u+1}}}\right)\right)}{T_{n|\mathcal{K}_{u+1}}^{\lambda_{u+1}}} \\
&= \tau_a^{(\infty,u+1)}\Big(\sum_{k\in\mathcal{K}_{u+1}} \sum_{p\in\mathcal{P}_k: a\in p} \tilde{f}_{exp,p}^{(\infty,u+1)}\Big) = \tau_a^{(\infty,u+1)}(\tilde{f}_{exp,a}^{(\infty,u+1)}).
\end{aligned} \tag{A.44}$$

Here, we use (A.30) to remove the influence of subgame $\Gamma_{n|\mathcal{K}\setminus\cup_{l=0}^{u+1}\mathcal{K}_l}$, and use (A.42) to obtain the limit. The subgame $\Gamma_{n|\cup_{l=0}^{u}\mathcal{K}_l}$ vanishes in the limit since $g_n^{(u+1)} \in \omega(\max_{l=0}^{u} g_n^{(l)})$ and (A.36).

Hence, we obtain for each *tight* path $p \in \cup_{k\in\mathcal{K}_{u+1}} \mathcal{P}_k$ that

$$\begin{aligned}
\lim_{n\to\infty} \frac{\tau_p(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}_{ran}^{(n)}))}{g_n^{(u+1)}} &= \lim_{n\to\infty} \frac{\tau_p(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}_{ran|\mathcal{K}_{u+1}}^{(n)}))}{g_n^{(u+1)}} \\
&= \tau_p^{(\infty,u+1)}(\tilde{f}_{exp}^{(\infty,u+1)}) \in [0,\infty),
\end{aligned} \tag{A.45}$$

since a tight path $p \in \cup_{k\in\mathcal{K}_{u+1}}\mathcal{P}_k$ contains *only* arcs $a \in A$ with $\beta_a \le \lambda_{u+1}$.

Lemma 9 shows another helpful result when we justify IA2–IA3 and IA5–(Sto–IA6) for the case that $g_n^{(u+1)} \in \omega(\max_{l=0}^{u} g_n^{(l)})$. We move the long proof of Lemma 9 to Appendix A.7.

Lemma 9 *Consider an arbitrary* $a \in A$, *an arbitrary* $u \in \{1,\ldots,m\}$, *an arbitrary polynomial function* $h(\cdot)$ *with degree* $\beta \ge 0$ *and a constant* $g_n := T_{n|\mathcal{K}_u}^{\lambda}$ *with an arbitrary constant exponent* $\lambda > 0$. *Assume that* $h(x)$ *is non-decreasing on* $[0,\infty)$. *Then*

$$\lim_{n\to\infty}\frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}(h(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u-1}\mathcal{K}_l}))}{g_n}=\lim_{n\to\infty}\frac{h(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u-1}\mathcal{K}_l}))}{g_n}\in[0,\infty]$$

if either of the two limits exist.

With Lemma 9, Fact A4 confirms IA2–IA3 and IA5–IA6 for the case that $g_n^{(u+1)} \in \omega(\max_{l=0}^{u}\ g_n^{(l)})$.

Fact A4 *IA2–IA3, IA5–IA6 hold at step* $u+1$ *when* $g_n^{(u+1)} \in \omega(\max_{l=0}^{u}\ g_n^{(l)})$.

Proof of Fact A4 We obtain by IA1 of step $u+1$ that $\mathbb{E}_{\tilde{\Pi}^{(n)}}\big(\tilde{f}^{(n)}_{ran,p}\big) \in o(T_{n|\mathcal{K}_{u+1}})$ for an arbitrary *non-tight* path $p \in \cup_{k\in\mathcal{K}_{u+1}}\mathcal{P}_k$. Otherwise, there is a non-tight path $p \in \cup_{k\in\mathcal{K}_{u+1}}\mathcal{P}_k$ with

$$\tau_p(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran}))=\sum_{a\in p}\tau_a(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a}))\ge\sum_{a\in p}\tau_a(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p}))\in\omega(g_n^{(u+1)})$$

and $\mathbb{E}_{\tilde{\Pi}^{(n)}}\big(\tilde{f}^{(n)}_{ran,p}\big) \in \Omega(T_{n|\mathcal{K}_{u+1}})$. This contradicts IA1 of step $u+1$, i.e., $\tau_{p'}\big(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran})\big) \in O(g_n^{(u+1)})$ for every $p' \in \cup_{k\in\mathcal{K}_{u+1}}\mathcal{P}_k$ with $\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p'}) > 0$ when $g_n^{(u+1)} \in \omega(\max_{l=0}^{u}\ g_n^{(l)})$. Here, we recall again that every non-tight path $p \in \cup_{k\in\mathcal{K}_{u+1}}\mathcal{P}_k$ contains at least one arc $a \in A$ whose cost function has a degree $\beta_a > \lambda_{u+1}$.

Consequently, we obtain for each *non-tight* path $p \in \cup_{k\in\mathcal{K}_{u+1}}\mathcal{P}_k$ that

$$\begin{aligned}
&\tilde{f}^{(\infty,u+1)}_{exp,p}=\lim_{n\to\infty}\frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p})}{T_{n|\mathcal{K}_{u+1}}}=0,\\
&\lim_{n\to\infty}\frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p})}{T_{n|\mathcal{K}_{u+1}}}\cdot\frac{\tau_p\big(\mathbb{E}_{\tilde{\Pi}^{(n)}}\big(\tilde{f}^{(n)}_{ran}\big)\big)}{g_n^{(u+1)}}\\
&\quad=\lim_{n\to\infty}\frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p})\cdot\mathbb{1}_{(0,\infty)}(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p}))}{T_{n|\mathcal{K}_{u+1}}}\cdot\frac{\tau_p\big(\mathbb{E}_{\tilde{\Pi}^{(n)}}\big(\tilde{f}^{(n)}_{ran}\big)\big)}{g_n^{(u+1)}}\\
&\quad=0.
\end{aligned}\tag{A.46}$$

Here, we used again IA1 of step $u+1$.

(A.45) and (A.46) together imply that

$$
\begin{aligned}
&\lim_{n\to\infty} \frac{C_{\mathcal{K}_{u+1}}\big[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran}),\, \Gamma_n\big]}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}} \\
&\quad= \lim_{n\to\infty} \frac{\sum_{k\in\mathcal{K}_{u+1}}\sum_{p\in\mathcal{P}_k}\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p})\cdot\tau_p(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran}))}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}} \\
&\quad= \lim_{n\to\infty} \frac{\sum_{k\in\mathcal{K}_{u+1}}\sum_{p\in\mathcal{P}_k:p\text{ is tight}}\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p})\cdot\tau_p(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran|\mathcal{K}_{u+1}}))}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}} \\
&\quad= \sum_{a\in A} f^{(\infty,u+1)}_{exp,a}\cdot\tau_a^{(\infty,u+1)}(f^{(\infty,u+1)}_{exp,a}),
\end{aligned}
\tag{A.47}
$$

where we again use the convention that $0\cdot\infty=0$. So IA2 of step $u+1$ holds.

When $\lambda_{u+1}>0$, then we obtain by Lemma 9 that $\tilde{f}^{(\infty,u+1)}_{exp}$ is a non-atomic NE flow of $\Gamma^{(\infty)}_{|\mathcal{K}_{u+1}}$. This follows since $g_n^{(u+1)}\in\omega(\max_{l=0}^u g_n^{(l)})$ and

$$
\begin{aligned}
\tau_p^{(\infty,u+1)}(\tilde{f}^{(\infty,u+1)}_{exp}) &= \lim_{n\to\infty}\frac{\tau_p(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}_{ran}))}{g_n^{(u+1)}} \\
&= \lim_{n\to\infty}\frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tau_p(\tilde{f}^{(n)}_{ran}))}{g_n^{(u+1)}} \le \lim_{n\to\infty}\frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tau_{p'}(\tilde{f}^{(n)}_{ran}))}{g_n^{(u+1)}} \\
&= \lim_{n\to\infty}\frac{\tau_{p'}(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}_{ran}))}{g_n^{(u+1)}} = \tau_{p'}^{(\infty,u+1)}(\tilde{f}^{(\infty,u+1)}_{exp})
\end{aligned}
\tag{A.48}
$$

for an arbitrary $k\in\mathcal{K}_{u+1}$ and two arbitrary *tight* paths $p,\ p'\in\mathcal{P}_k$ with $\tilde{f}^{(\infty,u+1)}_{exp,p}>0$. We used Lemma 9 to exchange the expectation and the function $\tau_p(\cdot)$ in (A.48), used (A.45) to obtain the limits on both sides, and used (A.36)–(A.37) to remove the influence of subgame $\Gamma_{|\cup_{l=0}^u\mathcal{K}_l}$ in the limits when $g_n^{(u+1)}\in\omega(\max_{l=0}^u g_n^{(l)})$ and the paths p and p' are tight. Moreover, the inequality in (A.48) follows since $\tilde{f}^{(n)}_{ran}$ is a mixed NE flow, which has a similar behavior with a mixed WE flow when we scale the path cost with $g^{(u+1)}$ and the maximum individual demand is bounded from above by υ.

When $\lambda_{u+1}=0$, then every tight path has constant cost. So (A.48) holds trivially and $\tilde{f}^{(\infty,u+1)}_{exp}$ is also a non-atomic NE flow of $\Gamma^{(\infty)}_{|\mathcal{K}_{u+1}}$. Here, we recall (A.46), i.e., $\tilde{f}^{(\infty,u+1)}_{exp,p}>0$ only if $p\in\cup_{k\in\mathcal{K}_{u+1}}\mathcal{P}_k$ is tight.

The above arguments together with Fact A3 and IA3 of step u imply IA3 for step $u+1$.

Below we show IA5–IA6 for step $u+1$ when $g_n^{(u+1)}\in\omega(\max_{l=0}^u g_n^{(l)})$.

Lemma 9 implies for each $\theta_n \in o(T_{n|\mathcal{K}_{u+1}})$ and each $a \in A$ with $\beta_a \leq \lambda_{u+1}$ that

$$\begin{aligned}
&\lim_{n\to\infty} \frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}\big[(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l} \pm \theta_n)\cdot \tau_a(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l})\big]}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}} \\
&= \lim_{n\to\infty} \frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}\big[\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l}\cdot \tau_a(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l})\big]}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}} \\
&= \lim_{n\to\infty} \frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}\big[\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l}\big]\cdot \tau_a\big[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l})\big]}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}} \\
&= \lim_{n\to\infty} \frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}\big[\tilde{f}^{(n)}_{ran,a|\mathcal{K}_{u+1}}\big]\cdot \tau_a\big[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\mathcal{K}_{u+1}})\big]}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}} \\
&= \tilde{f}^{(\infty,u+1)}_{exp,a}\cdot \tau_a^{(\infty,u+1)}(\tilde{f}^{(\infty,u+1)}_{exp,a}).
\end{aligned} \tag{A.49}$$

Here, $\mathbb{E}_{\tilde{\Pi}^{(n)}}[\theta_n\cdot\tau_a(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=1}^{u}\mathcal{K}_l})] \in o(T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)})$, as $\tau_a(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=1}^{u}\mathcal{K}_l}) \in O\big(g_n^{(u+1)}\big)$ holds almost surely when $\beta_a \leq \lambda_{u+1}$.

Lemma 9, (A.37)–(A.38), $g_n^{(u+1)} \in \omega(\max_{l=0}^{u} g_n^{(l)})$ and (A.46) together imply for each *non-tight* path $p \in \cup_{k\in\mathcal{K}_{u+1}}\mathcal{P}_k$ that

$$\begin{aligned}
&\lim_{n\to\infty} \frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}[\tilde{f}^{(n)}_{ran,p}\cdot\tau_p(\tilde{f}^{(n)}_{ran})]}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}} \\
&= \lim_{n\to\infty} \frac{\mathbb{1}_{(0,\infty)}[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p})]\cdot\mathbb{E}_{\tilde{\Pi}^{(n)}}[\tilde{f}^{(n)}_{ran,p}\cdot\tau_p(\tilde{f}^{(n)}_{ran})]}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}} \\
&= \lim_{n\to\infty} \frac{\sum_{a\in p:\beta_a>\lambda_{u+1}}\mathbb{1}_{(0,\infty)}[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p})]\cdot\mathbb{E}_{\tilde{\Pi}^{(n)}}\big[\tilde{f}^{(n)}_{ran,p}\cdot\tau_a(\tilde{f}^{(n)}_{ran,a})\big]}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}} \\
&= \lim_{n\to\infty} \frac{\sum_{a\in p:\beta_a>\lambda_{u+1}}\mathbb{1}_{(0,\infty)}[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p})]\cdot\mathbb{E}_{\tilde{\Pi}^{(n)}}\big[\tilde{f}^{(n)}_{ran,p}\cdot\tau_a(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l})\big]}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}} \\
&\leq \lim_{n\to\infty} \frac{\sum_{a\in p:\beta_a>\lambda_{u+1}}\mathbb{1}_{(0,\infty)}[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p})]\cdot\mathbb{E}_{\tilde{\Pi}^{(n)}}\big[\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l}\cdot\tau_a(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l})\big]}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}} \\
&= \lim_{n\to\infty} \frac{\sum_{a\in p:\beta_a>\lambda_{u+1}}\mathbb{1}_{(0,\infty)}[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p})]\cdot\mathbb{E}_{\tilde{\Pi}^{(n)}}\big[\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l}\big]\cdot\tau_a\big[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l})\big]}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}} \\
&\leq \lim_{n\to\infty} \frac{\sum_{a\in p:\beta_a>\lambda_{u+1}}\mathbb{1}_{(0,\infty)}[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p})]\cdot\mathbb{E}_{\tilde{\Pi}^{(n)}}\big[\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l}\big]\cdot\tau_p\big[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran})\big]}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}} \\
&= \lim_{n\to\infty} \frac{\sum_{a\in p:\beta_a>\lambda_{u+1}}\mathbb{1}_{(0,\infty)}[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p})]\cdot\mathbb{E}_{\tilde{\Pi}^{(n)}}\big[\tilde{f}^{(n)}_{ran,a|\mathcal{K}_{u+1}}\big]}{T_{n|\mathcal{K}_{u+1}}}\cdot O(1)
\end{aligned}$$

$$= \lim_{n\to\infty} \frac{\sum_{a\in p:\beta_a>\lambda_{u+1}} \sum_{p'\in\cup_{k\in\mathcal{K}_{u+1}}\mathcal{P}_k:a\in p'} \mathbb{1}_{(0,\infty)}[\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p})]\cdot\mathbb{E}_{\tilde{\Pi}^{(n)}}[\tilde{f}^{(n)}_{ran,p'}]}{T_{n|\mathcal{K}_{u+1}}} \cdot O(1)$$
$$= 0. \tag{A.50}$$

Here, we used that $T_{n|\mathcal{K}\setminus\cup_{l=1}^{u+1}\mathcal{K}_l} \in o\big(T_{n|\mathcal{K}_{u+1}}\big)$, that $\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p}) = 0$ implies $\mathbb{P}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p} \cdot \tau_p(\tilde{f}^{(n)}_{ran,a}) = 0) = 1$ for every $a \in A$, that $\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p}) > 0$ implies $\tau_p(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran})) \in O(g_n^{(u+1)})$, that p' is non-tight if p' contains an arc a with $\beta_a > \lambda_{u+1}$, and that

$$\lim_{n\to\infty} \frac{\sum_{a\in p:\beta_a\le\lambda_{u+1}} \mathbb{1}_{(0,\infty)}(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p}))\cdot\mathbb{E}_{\tilde{\Pi}^{(n)}}\big(\tilde{f}^{(n)}_{ran,p}\cdot\tau_a(\tilde{f}^{(n)}_{ran,a})\big)}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}}$$
$$= \lim_{n\to\infty} \frac{\sum_{a\in p:\beta_a\le\lambda_{u+1}} \mathbb{1}_{(0,\infty)}(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p}))\cdot\mathbb{E}_{\tilde{\Pi}^{(n)}}\big(\tilde{f}^{(n)}_{ran,p}\cdot\tau_a(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l})\big)}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}}$$
$$\le \lim_{n\to\infty} \frac{\sum_{a\in p:\beta_a\le\lambda_{u+1}} \mathbb{1}_{(0,\infty)}(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p}))\cdot\mathbb{E}_{\tilde{\Pi}^{(n)}}\big(\tilde{f}^{(n)}_{ran,p}\big)\cdot\tau_a(T_{n|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l})}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}}$$
$$= \lim_{n\to\infty} \frac{\sum_{a\in p:\beta_a\le\lambda_{u+1}} \mathbb{1}_{(0,\infty)}(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p}))\cdot\mathbb{E}_{\tilde{\Pi}^{(n)}}\big(\tilde{f}^{(n)}_{ran,p}\big)}{T_{n|\mathcal{K}_{u+1}}} \cdot O(1) = 0$$

when $p \in \cup_{k\in\mathcal{K}_{u+1}}\mathcal{P}_k$ is non-tight.

(A.50) means that non-tight paths are also negligible in the limit when we scale the joint (expected) total cost of the subgame $\Gamma_{n|\mathcal{K}_{u+1}}$ in the mixed NE flow $\tilde{f}^{(n)}_{ran}$ with the factor $T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}$.

(A.49)–(A.50), (A.38) and $g_n^{(u+1)} \in \omega(\max_{l=0}^{u}\ g_n^{(l)})$ together imply that

$$\lim_{n\to\infty} \frac{\sum_{k\in\mathcal{K}_{u+1}}\sum_{p\in\mathcal{P}_k} \mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p}\cdot\tau_p(\tilde{f}^{(n)}_{ran}))}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}}$$
$$= \lim_{n\to\infty} \frac{\sum_{k\in\mathcal{K}_{u+1}}\sum_{p\in\mathcal{P}_k:p\text{ is tight}} \mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p}\cdot\tau_p(\tilde{f}^{(n)}_{ran|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l}))}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}}$$
$$= \lim_{n\to\infty} \frac{\sum_{a\in A:\beta_a\le\lambda_{u+1}} \mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\text{ tight }p}\cdot\tau_a(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l}))}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}}$$
$$= \lim_{n\to\infty} \frac{\sum_{a\in A:\beta_a\le\lambda_{u+1}} \mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l}\cdot\tau_a(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l}))}{T_{n|\mathcal{K}_{u+1}}\cdot g_n^{(u+1)}} \tag{A.51}$$
$$= \sum_{a\in A:\beta_a\le\lambda_{u+1}} \tilde{f}^{(\infty,u+1)}_{exp,a}\cdot\tau_a^{(\infty,u+1)}(\tilde{f}^{(\infty,u+1)}_{exp,a})$$

$$= \sum_{a\in A} \tilde{f}_{exp,a}^{(\infty,u+1)} \cdot \tau_a^{(\infty,u+1)}(\tilde{f}_{exp,a}^{(\infty,u+1)}),$$

where we put $\tilde{f}^{(n)}_{ran,a|\text{tight }p} := \sum_{p'\in\cup_{k\in\mathcal{K}_{u+1}}\mathcal{P}_k : a\in p',\, p' \text{ is tight}} \tilde{f}^{(n)}_{ran,p'}$ for each $a \in A$ with $\beta_a \leq \lambda_{u+1}$. We also used that

$$\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=1}^{u}\mathcal{K}_l} - \tilde{f}^{(n)}_{ran,a|\text{ tight }p} \leq \sum_{k\in\mathcal{K}_{u+1}} \sum_{p'\in\mathcal{P}_k:\ p' \text{ is non-tight}} \tilde{f}^{(n)}_{ran,p'} + T_{n|\mathcal{K}\setminus\cup_{l=1}^{u+1}\mathcal{K}_l}$$
$$\in o(T_{n|\mathcal{K}_{u+1}}),$$

and that

$$\lim_{n\to\infty} \frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p'} \cdot \tau_a(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l}))}{T_{n|\mathcal{K}_{u+1}} \cdot g_n^{(u+1)}}$$
$$\leq \lim_{n\to\infty} \frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,p'})}{T_{n|\mathcal{K}_{u+1}}} \cdot O(1) = 0$$

when $\beta_a \leq \lambda_{u+1}$ and $p' \in \cup_{k\in\mathcal{K}_{u+1}}\mathcal{P}_k$ is non-tight. Here, we observe that the random event "$\tau_a(\tilde{f}^{(n)}_{ran,a|\mathcal{K}\setminus\cup_{l=0}^{u}\mathcal{K}_l}) \in O(g_n^{(u+1)})$" occurs almost surely when $\beta_a \leq \lambda_{u+1}$.

(A.51) together with Fact A3 proves IA5–IA6 for step $u+1$. Note that we have already shown that $\tilde{f}_{exp}^{(\infty,u+1)}$ is a non-atomic NE flow of $\Gamma^{(\infty)}_{|\mathcal{K}_{u+1}}$. This completes the proof of Fact A4. □

Therefore, IA1–IA7 hold for all $u \in \mathcal{M}$. This completes the whole proof by induction. □

A.7 Proof of Lemma 9

Consider an arbitrary arc $a \in A$, and an arbitrary $u \in \mathcal{M} = \{1,\ldots,m\}$. Let $g_n = T^{\lambda}_{n|\mathcal{K}_u}$ be a factor with an arbitrary exponent $\lambda > 0$, and let $h : [0,\infty) \to [0,\infty)$ be an arbitrary non-decreasing polynomial function with degree $\beta \geq 0$. To simplify notation, we assume that $\mathcal{K}_u = \mathcal{K}\setminus\cup_{l=0}^{u-1}\mathcal{K}_l$. The proof still holds when $\mathcal{K}_u$ is replaced by $\mathcal{K}\setminus\cup_{l=0}^{u-1}\mathcal{K}_l$, since (A.30) holds and $\lim_{n\to\infty}\frac{g_n}{T^{\lambda}_{n|\mathcal{K}\setminus\cup_{l=0}^{u-1}\mathcal{K}_l}} = \lim_{n\to\infty}\frac{g_n}{T^{\lambda}_{n|\mathcal{K}_u}} = 1.$

We assume, w.o.l.g., that the limit $\lim_{n\to\infty}\frac{h(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\mathcal{K}_u}))}{g_n} \in [0,\infty]$ exists.

To prove Lemma 9, we need tight probability lower and upper bounds for the random event $|\tilde{f}^{(n)}_{ran,a|\mathcal{K}_u} - \mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\mathcal{K}_u})| \in O(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\mathcal{K}_u}))$, for which we will need *Markov's inequality* from Lemma 8a.

Note that $\tilde{f}^{(n)}_{ran,a|\mathcal{K}_u} = \sum_{k\in\mathcal{K}_u}\sum_{i\in\mathcal{U}_k^{(n)}} d^{(n)}_{k,i} \cdot \mathbb{1}_{p_{k,i}(\tilde{\Pi}_i^{(n)})}(a)$ is a *weighted sum* of mutually independent *Bernoulli* random variables $\mathbb{1}_{p_{k,i}(\tilde{\Pi}_i^{(n)})}(a)$, $i \in \mathcal{U}^{(n)}_{|\mathcal{K}_u} = \cup_{k\in\mathcal{K}_u}\mathcal{U}_k^{(n)}$. Recall that $p_{k,i}(\tilde{\Pi}_i^{(n)})$ is the random path sampled by agent i using the probability distribution $\tilde{\Pi}_i^{(n)} = (\tilde{\Pi}_{i,p}^{(n)})_{p\in\mathcal{P}_k}$ for each $k \in \mathcal{K}_u$ and $i \in \mathcal{U}_k^{(n)}$, and that $\mathbb{1}_B(b)$ is the indicator function of the membership relation "$b \in B$" for an arbitrary set B and an arbitrary element b.

Fact A5a–d below show useful lower and upper probability bounds for a weighted sum of *arbitrary* Bernoulli random variables, and thus apply to the weighted sum $\tilde{f}_{ran,a|\mathcal{K}_u}^{(n)}$.

Fact A5 *Consider n mutually independent Bernoulli random variables $X_1, \ldots, X_n$ with success probabilities $q_1, \ldots, q_n \in [0, 1]$, respectively. Let $v_1, \ldots, v_n$ be non-negative weights with sum $V_n := \sum_{i=1}^n v_i$, and let $Y_n = \sum_{i=1}^n v_i \cdot X_i$ be the weighted sum of these n random variables. If $v_i \le \upsilon$ for a constant $\upsilon > 0$, then the following probability bounds hold.*

a) $\mathbb{P}\big(Y_n \ge (1+\delta)\cdot E(Y_n)\big) \le e^{-\frac{(\delta+1)\cdot E(Y_n)}{\upsilon}\cdot\left(\ln(\delta+1)-\frac{\delta}{\delta+1}\right)}$ *for all $\delta > 0$.*

b) $\mathbb{P}\big(Y_n \le (1-\delta)\cdot E(Y_n)\big) \le e^{-\frac{V_n-(1-\delta)\cdot\mathbb{E}(Y_n)}{\upsilon}\cdot\left(\ln\frac{V_n-(1-\delta)\cdot\mathbb{E}(Y_n)}{V_n-\mathbb{E}(Y_n)}-\frac{\delta\cdot\mathbb{E}(Y_n)}{V_n-(1-\delta)\cdot\mathbb{E}(Y_n)}\right)}$ *for all $\delta \in (0, 1)$.*

c) If $\lim_{n\to\infty}\mathbb{E}(Y_n) = 0$ and $\underline{\lim}_{n\to\infty} V_n > 1$, then there is an integer $N \in \mathbb{N}$ such that $\mathbb{P}\big(Y_n \ge 1+\delta\big) \le e^{-\frac{\delta+1}{\upsilon}\cdot\left(\ln(\delta+1)-\frac{\delta}{\delta+1}\right)}$ for all $\delta > 0$ and all $n \ge N$.

d) If $\lim_{n\to\infty}\frac{V_n}{\mathbb{E}(Y_n)} = 1$ and $\lim_{n\to\infty} V_n = \infty$, then there is an integer $N \in \mathbb{N}$ s.t. $\mathbb{P}\big(Y_n \le (1-\delta)\cdot(\mathbb{E}(Y_n)-c)\big) \le e^{-\frac{V_n-(1-\delta)\cdot(\mathbb{E}(Y_n)-c)}{\upsilon}\cdot\left(\ln\frac{V_n-(1-\delta)\cdot(\mathbb{E}(Y_n)-c)}{V_n-\mathbb{E}(Y_n)+c}-\frac{\delta\cdot\mathbb{E}(Y_n)-\delta\cdot c}{V_n-(1-\delta)\cdot(\mathbb{E}(Y_n)-c)}\right)}$ for all $\delta \in (0, 1)$, all $c \in (0,\ \mathbb{E}(Y_n))$, and all $n \ge N$.

Proof of Fact A5a Our proof is similar to that for the usual Chernoff bound in, e.g., [27, 33]. Using Markov's inequality and the fact that $X_1, \ldots, X_n$ are mutually independent Bernoulli random variables with success probabilities $q_1, \ldots, q_n$, we obtain for an arbitrary $t > 0$ and an arbitrary $\delta > 0$ that

$$\begin{aligned}
&\mathbb{P}\big(Y_n \ge (1+\delta)\cdot E(Y_n)\big)\\
&= \mathbb{P}\big(e^{t\cdot Y_n} \ge e^{t\cdot(1+\delta)\cdot E(Y_n)}\big) \le \frac{\prod_{i=1}^n \mathbb{E}(e^{t\cdot X_i\cdot v_i})}{e^{t\cdot(1+\delta)\cdot E(Y_n)}}\\
&= \frac{\prod_{i=1}^n \big(q_i\cdot e^{t\cdot v_i} + (1-q_i)\big)}{e^{t\cdot(1+\delta)\cdot E(Y_n)}} = \frac{\prod_{i=1}^n \big(q_i\cdot v_i\cdot t\cdot\frac{e^{t\cdot v_i}-1}{t\cdot v_i}+1\big)}{e^{t\cdot(1+\delta)\cdot E(Y_n)}}.
\end{aligned} \tag{A.52}$$

The function $\frac{e^x-1}{x}$ is non-decreasing on $(0, \infty)$ and $1+x \le e^x$ holds for all $x \in [0, \infty)$. So we obtain by (A.52) that

$$\begin{aligned}
\mathbb{P}\big(Y_n \ge (1+\delta)\cdot E(Y_n)\big) &\le \frac{\prod_{i=1}^n \big(q_i\cdot v_i\cdot t\cdot\frac{e^{t\cdot\upsilon}-1}{t\cdot\upsilon}+1\big)}{e^{t\cdot(1+\delta)\cdot E(Y_n)}}\\
&\le \frac{e^{\sum_{i=1}^n q_i\cdot v_i\cdot\frac{e^{t\cdot\upsilon}-1}{\upsilon}}}{e^{t\cdot(1+\delta)\cdot E(Y_n)}} = e^{E(Y_n)\cdot\left(\frac{e^{t\cdot\upsilon}-1}{\upsilon}-t\cdot(1+\delta)\right)}
\end{aligned} \tag{A.53}$$

for all $t > 0$. (A.53) implies that $\mathbb{P}\big(Y_n \geq (1+\delta)\cdot E(Y_n)\big) \leq e^{-\frac{(\delta+1)\cdot E(Y_n)}{v}\cdot\left(\ln(\delta+1)-\frac{\delta}{\delta+1}\right)}$ when we put $t = \frac{\ln(\delta+1)}{v}$ and observe that $\ln(\delta+1) - \frac{\delta}{\delta+1} > 0$ for all $\delta > 0$.

Proof of Fact A5b Let $Z_n := \sum_{i=1}^{n} v_i \cdot (1 - X_i) = V_n - Y_n$. Then $Z_n + Y_n = V_n$ and $\mathbb{E}(Z_n) + \mathbb{E}(Y_n) = V_n$. Fact A5a) implies for every $\delta \in (0, 1)$ that

$$\mathbb{P}\Big(Z_n \geq \mathbb{E}(Z_n) + \delta\cdot\mathbb{E}(Y_n) = \Big(1 + \frac{\delta\cdot\mathbb{E}(Y_n)}{\mathbb{E}(Z_n)}\Big)\cdot\mathbb{E}(Z_n)\Big) \\ \leq e^{-\frac{\mathbb{E}(Z_n)+\delta\cdot\mathbb{E}(Y_n)}{v}\cdot\left(\ln\frac{\mathbb{E}(Z_n)+\delta\cdot\mathbb{E}(Y_n)}{\mathbb{E}(Z_n)}-\frac{\delta\cdot\mathbb{E}(Y_n)}{\mathbb{E}(Z_n)+\delta\cdot\mathbb{E}(Y_n)}\right)}.$$

Since the random event $Z_n \geq \mathbb{E}(Z_n) + \delta\cdot\mathbb{E}(Y_n)$ is equivalent to the random event $Y_n \leq (1-\delta)\cdot\mathbb{E}(Y_n)$, we obtain that

$$\mathbb{P}(Y_n \leq (1-\delta)\cdot E(Y_n)) \leq e^{-\frac{\mathbb{E}(Z_n)+\delta\cdot\mathbb{E}(Y_n)}{v}\cdot\left(\ln\frac{\mathbb{E}(Z_n)+\delta\cdot\mathbb{E}(Y_n)}{\mathbb{E}(Z_n)}-\frac{\delta\cdot\mathbb{E}(Y_n)}{\mathbb{E}(Z_n)+\delta\cdot\mathbb{E}(Y_n)}\right)} \\ = e^{-\frac{V_n-(1-\delta)\cdot\mathbb{E}(Y_n)}{v}\cdot\left(\ln\frac{V_n-(1-\delta)\cdot\mathbb{E}(Y_n)}{V_n-\mathbb{E}(Y_n)}-\frac{\delta\cdot\mathbb{E}(Y_n)}{V_n-(1-\delta)\cdot\mathbb{E}(Y_n)}\right)} \tag{A.54}$$

for all $\delta \in (0, 1)$. (A.54) proves Fact A5b.

Proof of Fact A5c We say that n mutually independent Bernoulli random variables $X_1', \ldots, X_n'$ with success probabilities $q_1', \ldots, q_n'$ are *stochastically larger than* $X_1, \ldots, X_n$ if $q_i' \geq q_i$ for each $i = 1, \ldots, n$. Clearly, there are n mutually independent Bernoulli random variables $X_1', \ldots, X_n'$ that are stochastically larger than $X_1, \ldots, X_n$ and satisfy $\mathbb{E}(Y_n') = \mathbb{E}(\sum_{i=1}^{n} v_i \cdot X_i') = \sum_{i=1}^{n} v_i \cdot q_i' = 1$ for large enough n. This follows since $\mathbb{E}(Y_n) = \sum_{i=1}^{n} v_i \cdot q_i \to 0$ as $n \to \infty$, $\underline{\lim}_{n\to\infty} V_n = \underline{\lim}_{n\to\infty} \sum_{i=1} v_i > 1$, and the continuous multi-variate function $\alpha(x_1, \ldots, x_n) := \sum_{i=1}^{n} v_i \cdot (q_i + x_i)$ has $[\mathbb{E}(Y_n),\ V_n]$ as its range on the compact domain $\prod_{i=1}^{n}[0, 1 - q_i]$ for all $n \in \mathbb{N}$.

Fact A5c then follows from Fact A5a, if $\mathbb{P}(Y_n \geq c) \leq \mathbb{P}(Y_n' = \sum_{i=1}^{n} v_i \cdot X_i' \geq c)$ for an arbitrary constant $c \geq \mathbb{E}(Y_n)$, (since we can then obtain Fact A5c by applying Fact A5a to Y_n' with $c = 1 + \delta$ for large enough n).

Consider now an arbitrary constant $c \geq \mathbb{E}(Y_n)$. We prove below that $\mathbb{P}(Y_n \geq c) \leq \mathbb{P}(Y_n' \geq c)$ only for the particular case that $q_1' \geq q_1$ and $q_i' = q_i$ for all $i = 2, \ldots, n$. One can obtain a complete proof for the general case with a simple induction over $\{2, \ldots, n\}$.

Note that

$$\mathbb{P}(Y_n' \geq c) = \mathbb{P}\left(\sum_{i=2}^{n} v_i \cdot X_i' \geq c - v_1\right)\cdot\mathbb{P}(X_1' = 1) + \mathbb{P}\left(\sum_{i=2}^{n} v_i \cdot X_i' \geq c\right)\cdot\mathbb{P}(X_1' = 0) \\ = \mathbb{P}\left(\sum_{i=2}^{n} v_i \cdot X_i' \geq c - v_1\right)\cdot(q_1 + q_1' - q_1) + \mathbb{P}\left(\sum_{i=2}^{n} v_i \cdot X_i' \geq c\right) \\ \cdot(1 - q_1 + q_1 - q_1')$$

$$= \mathbb{P}(Y_n \geq c) + (q_1' - q_1) \cdot \left(\mathbb{P}\left(\sum_{i=2}^{n} v_i \cdot X_i \geq c - v_1\right) - \mathbb{P}\left(\sum_{i=2}^{n} v_i \cdot X_i \geq c\right)\right)$$

$$= \mathbb{P}(Y_n \geq c) + (q_1' - q_1) \cdot \mathbb{P}\left(c > \sum_{i=2}^{n} v_i \cdot X_i \geq c - v_1\right) \geq \mathbb{P}(Y_n \geq c).$$

This follows since the Bernoulli random variables X_i and X_i' can be identified for each $i = 2, \ldots, n$, as they have the same success probability q_i. □

Proof of Fact A5d It follows immediately from Fact A5b and the fact that there are n mutually independent Bernoulli random variables $X_1', \ldots, X_n'$ such that $X_1, \ldots, X_n$ are stochastically larger than $X_1', \ldots, X_n'$ and $\mathbb{E}(Y_n') = \mathbb{E}(Y_n) - c$ for a constant $c \in (0,\ \mathbb{E}(Y_n))$. Note that such Bernoulli random variables exist since $\lim_{n\to\infty} \frac{V_n}{\mathbb{E}(Y_n)} = 1$ and $\lim_{n\to\infty} V_n = \infty$.

This completes the proof of Fact A5. □

The two probability bounds in Fact A5a–b are similar to *Chernoff's bounds* and *Hoeffding's bounds*, see, e.g., [20, 27, 33]. However, a direct application of these known bounds to $\tilde{f}^{(n)}_{ran,a|\mathcal{K}_u}$ involves either the number $|\mathcal{U}^{(n)}_{|\mathcal{K}_u}|$ of agents in subgame $\Gamma_{n|\mathcal{K}_u}$, or the *minimum* individual demand $\min_{k\in\mathcal{K}_u, i\in\mathcal{U}^{(n)}_k} d^{(n)}_{k,i}$. Note that this minimum individual demand may vanish quickly as $n \to \infty$ and so the number $|\mathcal{U}^{(n)}_{|\mathcal{K}_u}|$ of agents need not be in $\Theta(T_{n|\mathcal{K}_u})$. Therefore, we include a proof tailored to our needs.

Note also that Fact A5a does not apply when $\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\mathcal{K}_u}) \in o(1)$, and Fact A5b does not apply when $\lim_{n\to\infty} \frac{\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\mathcal{K}_u})}{T_{n|\mathcal{K}_u}} = \tilde{f}^{(\infty,u)}_{exp,a} = 1$. We will instead use Fact A5c–d, respectively, in the proof of Lemma 9 in these two cases.

With all these preparations, we are now ready to prove Lemma 9.

The two limits in Lemma 9 are equal to 0 when $\lambda > \beta$, since both $h(\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\mathcal{K}_u}))$ and $\mathbb{E}_{\tilde{\Pi}^{(n)}}(h(\tilde{f}^{(n)}_{ran,a|\mathcal{K}_u}))$ are in $o(g_n)$ when $\lambda > \beta$.

We assume, w.l.o.g., that $\beta \geq \lambda > 0$. Moreover, we assume that $\lim_{n\to\infty}\mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\mathcal{K}_u}) \in [0, \infty]$ exists. Otherwise, we take an arbitrary infinite subsequence $(n_j)_{j\in\mathbb{N}}$ satisfying this condition. To simplify notation, we write $Y_n := \tilde{f}^{(n)}_{ran,a|\mathcal{K}_u}$, $E_n := \mathbb{E}_{\tilde{\Pi}^{(n)}}(\tilde{f}^{(n)}_{ran,a|\mathcal{K}_u})$, $\mathbb{P}_{\tilde{\Pi}^{(n)}}(\cdot) = \mathbb{P}(\cdot)$, and $\mathbb{E}_{\tilde{\Pi}^{(n)}}(\cdot) = \mathbb{E}(\cdot)$.

We distinguish four cases.

Case I: $E_n \in \Theta(1)$, i.e., $\lim_{n\to\infty} E_n \in (0, \infty)$. Let $\xi := \frac{\lambda}{2\cdot\beta} \in (0, 1)$. We obtain by Fact A5a with $\delta := T^{\xi}_{n|\mathcal{K}_u}$ that $\mathbb{P}[Y_n \geq (1 + \delta) \cdot E_n] \leq e^{-\frac{(1+\delta)\cdot E_n}{\upsilon}\cdot\left(\ln(\delta+1) - \frac{\delta}{\delta+1}\right)} = e^{-\omega(T^{\xi}_{n|\mathcal{K}_u}/\upsilon)}$.

This in turn implies that $\mathbb{E}(h(Y_n)) \le e^{-\omega\left(T^{\xi}_{n|\mathcal{K}_u}/\upsilon\right)} \cdot h(T_{n|\mathcal{K}_u}) + h\left((1+T^{\xi}_{n|\mathcal{K}_u}) \cdot E_n\right) \in o(g_n)$. So $\lim_{n\to\infty} \frac{h(E_n)}{g_n} = 0 = \lim_{n\to\infty} \frac{\mathbb{E}(h(Y_n))}{g_n}$.

Case II: $E_n \in o(1)$, i.e., $\lim_{n\to\infty} E_n = 0$. We obtain by Fact A5c that $\mathbb{P}\left[Y_n \ge 1 + T^{\xi}_{n|\mathcal{K}_u}\right] \le e^{-\omega\left(T^{\xi}_{n|\mathcal{K}_u}/\upsilon\right)}$. Then, $\lim_{n\to\infty} \frac{h(E_n)}{g_n} = 0 = \lim_{n\to\infty} \frac{\mathbb{E}(h(Y_n))}{g_n}$.

Case III: $\tilde{f}^{(\infty,u)}_{exp,a} = \lim_{n\to\infty} \frac{E_n}{T_{n|\mathcal{K}_u}} = 1$. We obtain by Fact A5d that

$$\begin{aligned}\mathbb{P}\left[Y_n \le (1-\delta)\cdot(E_n - c)\right] &\le e^{-\frac{T_{n|\mathcal{K}_u}-(1-\delta)\cdot(E_n-c)}{\upsilon}\cdot\left(\ln\frac{T_{n|\mathcal{K}_u}-(1-\delta)\cdot(E_n)-c)}{T_{n|\mathcal{K}_u}-E_n+c} - \frac{\delta\cdot E_n - \delta\cdot c}{T_{n|\mathcal{K}_u}-(1-\delta)\cdot(E_n-c)}\right)} \\ &= e^{-\Omega\left(\delta\cdot T_{n|\mathcal{K}_u}\right)},\end{aligned}$$

where $\delta \in (0, 1)$ is an arbitrary constant and $c := \sqrt{T_{n|\mathcal{K}_u}}$. Therefore,

$$\frac{\mathbb{E}(h(Y_n))}{h(E_n)} \ge (1 - e^{-\Omega(\delta\cdot T_{n|\mathcal{K}_u})}) \cdot \frac{h((1-\delta)\cdot(E_n - \sqrt{T_{n|\mathcal{K}_u}}))}{h(E_n)}.$$

This implies that $\underline{\lim}_{n\to\infty} \frac{\mathbb{E}(h(Y_n))}{h(E_n)} \ge (1-\delta)^{\beta}$ by letting $n \to \infty$ on both sides of the above inequality. So $\underline{\lim}_{n\to\infty} \frac{\mathbb{E}(h(Y_n))}{h(E_n)} \ge 1$ due to the arbitrary choice of $\delta \in (0, 1)$. However, on the other hand, $\overline{\lim}_{n\to\infty} \frac{\mathbb{E}(h(Y_n))}{h(E_n)} = \overline{\lim}_{n\to\infty} \frac{\mathbb{E}(h(Y_n))}{h(T_{n|\mathcal{K}_u})} \cdot \lim_{n\to\infty} \frac{h(T_{n|\mathcal{K}_u})}{h(E_n)} \le 1$. Hence, we have $\lim_{n\to\infty} \frac{\mathbb{E}(h(Y_n))}{h(E_n)} = 1$ when $\tilde{f}^{(\infty,u)}_{exp,a} = 1$.

Case IV: $\tilde{f}^{(\infty,u)}_{exp,a} < 1$ and $E_n \in \omega(1)$, i.e., $\lim_{n\to\infty} E_n = \infty$ and $T_{n|\mathcal{K}_u} - E_n \in \Theta(T_{n|\mathcal{K}_u})$. Clearly, Fact A5a–b apply in this case. We further distinguish two subcases.

(Subcase IV-I: $h(E_n) \in o(g_n)$**)** Then $E_n \in o(T^{\lambda/\beta}_{n|\mathcal{K}_u})$. We obtain further by Fact A5a that $\mathbb{E}(h(Y_n)) \in o(g_n)$. This follows since $\mathbb{P}(Y_n > \delta \cdot T^{\lambda/\beta}_{n|\mathcal{K}_u}) \le e^{-\Omega(\delta\cdot T^{\lambda/\beta}_{n|\mathcal{K}_u})}$ for all $\delta > 0$ when $E_n \in o(T^{\lambda/\beta}_{n|\mathcal{K}_u})$, and so $\overline{\lim}_{n\to\infty} \frac{\mathbb{E}(h(Y_n))}{g_n} \le \delta^{\beta} \cdot O(1)$ for all $\delta > 0$.

(Subcase IV-II: $h(E_n) \in \Omega(g_n)$**)** Then $E_n \in \Omega(T^{\lambda/\beta}_{n|\mathcal{K}_u})$. Fact A5a yields that $\mathbb{P}\left[Y_n \ge E_n + E_n^{2/3}\right] = e^{-\frac{E_n+E_n^{2/3}}{\upsilon}\cdot\left(\ln(1+E_n^{-1/3}) - \frac{E_n^{-1/3}}{1+E_n^{-1/3}}\right)} \le e^{-\Omega(E_n^{1/3})}$. Hence,

$$\overline{\lim_{n\to\infty}} \frac{\mathbb{E}(h(Y_n))}{h(E_n)} \le \overline{\lim_{n\to\infty}}\, e^{-\Omega(E_n^{1/3})} \cdot \frac{h(T_{n|\mathcal{K}_u})}{h(E_n)} + \overline{\lim_{n\to\infty}} \frac{h(E_n + E_n^{2/3})}{h(E_n)} = 1.$$

Moreover, $\underline{\lim}_{n\to\infty} \frac{\mathbb{E}(h(Y_n))}{h(E_n)} \ge 1$ follows from Fact A5b, since $\mathbb{P}[Y_n \le (1-\delta)\cdot E_n] \le e^{-\Omega(T_{n|\mathcal{K}_u})}$ for each $\delta \in (0, 1)$, when $T_{n|\mathcal{K}_u} - E_n \in \Theta(T_{n|\mathcal{K}_u})$.

All the above together prove Lemma 9. □

References

1. Beckmann, M., McGuire, C., Winsten, C.: Studies in the economics of transportation. Yale Univ. Press, New Haven, CT (1956)
2. Bingham, N., Goldie, C., Teugels, J.: Regular variation. Cambridge University Press, Cambridge (1987)
3. Brouwer, L.E.J.: Uber eineindeutige. stetige transformationen von flächen in sich. Math. Ann. **67**, 176–180 (1910)
4. Budny, K.: A generalization of chebyshev's inequality for hilbert-space-valued random elements. Statist. Probab. Lett. **88**, 62–65 (2014)
5. Bureau of Public Roads: Traffic assignment manual. USA, U.S, Department of Commerce, Urban Planning Division, Washington, D.C. (1964)
6. Chau, C., Sim, K.: The price of anarchy for non-atomic congestion games with symmetric cost maps and elastic demands. Operations Research Letter **31**, 327–334 (2003)
7. Christodoulou, G., Koutsoupias, E.: The price of anarchy in finite congestion games. In: Proceedings of the 37th Annual ACM Symposium on Theory of Computing–STOC'05, ACM, Baltimore, MD, 1–7 (2005)
8. Colini-Baldeschi, R., Cominetti, R., Scarsini, M.: On the price of anarchy of highly congested nonatomic network games. In: International Symposium on Algorithmic Game Theory, Springer, Lecture Notes in Computer Science 9928, Berlin Heidelberg, 117–128 (2016)
9. Colini-Baldeschi, R., Cominetti, R., Mertikopoulos, P., Scarsini, M.: The asymptotic behavior of the price of anarchy. In: WINE 2017. Lecture Notes in Computer Science, vol. 10674, pp. 133–145. Springer, Berlin Heidelberg (2017)
10. Colini-Baldeschi, R., Cominetti, R., Mertikopoulos, P., Scarsini, M.: When is selfish routing bad? the price of anarchy in light and heavy traffic. Oper. Res. **68**(2), 411–434 (2020). https://doi.org/10.1287/opre.2019.1894
11. Cominetti, R., Scarsini, M., Schröder, M., Stier-Moses, N.: Approximation and convergence of large atomic congestion games. Tech. rep., arXiv:2001.02797v6 [cs.GT] (2021)
12. Correa, J., Schulz, A., Stier-Moses, N.: Selfish routing in capacitated networks. Math. Oper. Res. **29**(4), 961–976 (2004)
13. Correa, J., Schulz, A., Stier-Moses, N.: On the inefficiency of equilibria in congestion games, extended abstract. In: Proceedings of Integer Programming and Combinatorial Optimization, Berlin, Germany, June 8-10, Lecture Notes in Computer Science 3509, Berlin Heidelberg, 167–181 (2005)
14. Dafermos, S.: Traffic equilibrium and variational inequalities. Transp. Sci. **14**(1), 42–54 (1980)
15. Dafermos, S., Sparrow, F.: The traffic assignment problem for a general network. Journal of Research of the US National Bureau of Standards **73B**, 91–118 (1969)
16. Fotakis, D., Kontogiannis, S., Spirakis, P.: Selfish unsplittable flows. Theoret. Comput. Sci. **348**(2), 226–239 (2005)
17. Harks, T., Klimm, M.: On the existence of pure nash equilibria in weighted congestion games. Math. Oper. Res. **37**(3), 419–436 (2012)
18. Harks, T., Klimm, M., Möhring, R.: Characterizing the existence of potential functions in weighted congestion games. Theory Computer Systems **49**(1), 46–70 (2011)
19. Haurie, A., Marcotte, P.: On the relationship between nash-cournot and wardrop equilibria. Networks **15**(3), 295–308 (1985)
20. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **58**(301), 13–30 (1963)
21. Jacquot, P., Wan, C.: Routing game on parallel networks: the convergence of atomic to nonatomic. Tech. rep., arXiv:1804.03081 [cs.GT] (2018)
22. Jacquot, P., Wan, C.: Nonatomic aggregative games with infinitely many types. Tech. rep., arXiv:1906.01986 [cs.GT] (2019)
23. Jahn, O., Möhring, R.H., Schulz, A.S., Stier-Moses, N.E.: System-optimal routing of traffic flows with user constraints in networks with congestion. Oper. Res. **53**(4), 600–616 (2005)
24. Koutsoupias, E., Papadimitriou, C.: Worst-case equilibria. In: Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science (STACS), Springer, Lecture Notes in Computer Science 1563, Berlin Heidelberg, 404–413 (1999)
25. McShane, E.: Jensen's inequality. Bull. Am. Math. Soc. **43**(8), 521–528 (1937)
26. Milchtaich, I.: Generic uniqueness of equilibrium in large crowding games. Math. Oper. Res. **25**(3), 349–364 (2000)

27. Mitrinovic, D.: Analytic Inequalities. Springer-Verlag, New York/Heidelberg/Berlin (1970)
28. Monderer, D., Shapley, L.: Potential games. Games Econom. Behav. **14**(1), 124–143 (1996)
29. Monnot, B., Benita, F., Piliouras, G.: How bad is selfish routing in practice? Tech. rep., arXiv:1703.01599v2 [cs.GT] (2017)
30. Nash, J.J.: Equilibrium points in n-person games. Proc. Nat. Acad. Sci, USA **36**(1), 48–49 (1950)
31. Neuts, M., Wolfson, D.: Convexity of the bounds induced by markov's inequality. Stochastic Processes and Their Applications **1**, 145–149 (1973)
32. Nisan, N., Roughgarden, T., Tardos, É., Vaz, V.: Algorithmic game theory. Cambridge University Press, Cambridge, UK (2007)
33. Nowak, R.: Chernoff's bound and hoeffding's inequality. Polyhedron **13**(1), 45–51 (1994)
34. O'Hare, S., Connors, R., Watling, D.: Mechanisms that govern how the price of anarchy varies with travel demand. Transportation Research Part B Methodological **84**, 55–80 (2016)
35. Papadimitriou, C.: Algorithms, games, and the internet. In: International Colloquium on Automata, Languages, and Programming, Springer, Lecture Notes in Computer Science 2076, Berlin Heidelberg, 1–3 (2001)
36. Perakis, G.: The price of anarchy under nonlinear and asymmetric costs. Math. Oper. Res. **32**(3), 614–628 (2007)
37. Rosenthal, R.: A class of games possessing pure-strategy nash equilibria. Internat. J. Game Theory **2**(1), 65–67 (1973)
38. Roughgarden, T.: Designing networks for selfish users is hard. Proceedings of Annual Symposium on Foundations of Computer Science **72**(72), 472–481 (2001)
39. Roughgarden, T.: The price of anarchy is independent of the network topology. Journal of Computer & System Sciences **67**(2), 341–364 (2003)
40. Roughgarden, T.: Selfish Routing and the Price of Anarchy. The MIT Press, Cambridge, MA (2005)
41. Roughgarden, T.: Intrinsic robustness of the price of anarchy. J. ACM **62**(32), 1–42 (2015)
42. Roughgarden, T., Tardos, É.: How bad is selfish routing? J. ACM **49**(2), 236–259 (2002)
43. Roughgarden, T., Tardos, É.: Bounding the inefficiency of equilibria in nonatomic congestion games. Games & Economic Behavior **47**(2), 389–403 (2004)
44. Roughgarden, T., Tardos, É.: Routing games. In: Nisan, N., Roughgarden, T., Tardos, É., Vazirani, V.V. (eds.) Algorithmic game theory, pp. 461–486. Cambridge University Press, Cambridge, MA (2007)
45. Wardrop, J.: Some theoretical aspects of road traffic research. Proc. Inst. Civ. Eng. **1**(2), 325–362 (1952)
46. Wu, Z., Möhring, R.: A sensitivity analysis of the price of anarchy in non-atomic congestion games. Tech. rep., arXiv:2007.13979v3 [cs.GT] (2021)
47. Wu, Z., Möhring, R., Chen, Y., Xu, D.: Selfishness need not be bad. Oper. Res. **69**(2), 410–435 (2021). https://doi.org/10.1287/opre.2020.2036
48. Youn, H., Gastner, M.T., Jeong, H.: Erratum: Price of anarchy in transportation networks: efficiency and optimality control. Phys. Rev. Lett. **101**, 128701 (2008)