Approximate Equilibria and Ball Fusion*

Elias Koutsoupias[†] *Marios Mavronicolas*[‡] *Paul Spirakis*[§]

Abstract

We consider selfish routing over a network consisting of m parallel links through which n selfish users route their traffic trying to minimize their own expected latency. We study the class of mixed strategies in which the expected latency through each link is at most a constant multiple of the optimum maximum latency had global regulation been available. For the case of uniform links it is known that all Nash equilibria belong to this class of strategies. We are interested in bounding the coordination ratio (or price of anarchy) of these strategies defined as the worst-case ratio of the maximum (over all links) expected latency over the optimum maximum latency. The load balancing aspect of the problem immediately implies a lower bound $\Omega\left(\frac{\ln m}{\ln \ln m}\right)$ of the coordination ratio. We give a tight (up to a multiplicative constant)

*A preliminary version of this work appears in the *Proceedings of the 9th International Colloquium on Structural Information and Communication Complexity (SIROCCO 2002)*, Andros, Greece, June 2002. This work has been partially supported by the IST Program of the European Union under contract numbers IST-1999-14186 (ALCOM-FT) and IST-2001-33116 (FLAGS), by the Joint Program of Scientific and Technological Collaboration between Greece and Cyprus under contract number 40/2000 (EPDS), by funds for the promotion of research at University of Cyprus, and by NSF.

[†]Department of Computer Science, University of California at Los Angeles. & Department of Informatics, University of Athens, 15771 Athens, Greece. Email: elias@di.uoa.gr

[‡]**Contact Author.** Department of Computer Science, University of Cyprus, Nicosia CY-1678, Cyprus. Email: mavronic@ucy.ac.cy

[§]Department of Computer Engineering and Informatics, University of Patras, 265 00 Patras, Greece, & Computer Technology Institute, 261 10 Patras, Greece. Part of the work of this author was performed while visiting Max-Planck Institut für Informatik, 66123 Saarbrücken, Germany. Email: spirakis@cti.gr

upper bound. To show the upper bound, we analyze a variant of the classical balls and bins problem, in which balls with arbitrary weights are placed into bins according to arbitrary probability distributions. At the heart of our approach is a new probabilistic tool that we call ball fusion; this tool is used to reduce the variant of the problem where balls bear weights to the classical version (with no weights). Ball fusion applies to more general settings such as links with arbitrary capacities and other latency functions.

1 Introduction

1.1 Motivation-Framework

We consider a *routing* problem in communication networks, where a collection of *noncooperating* entities called *users* want to select paths from sources to destinations. Different users may wish to optimize completely different objectives of performance and demand. Such networks are henceforth called *noncooperative*. Such noncooperative and antagonistic scenaria apply to various modern networking environments, where a *single* performance objective, to be achieved via cooperation among the users, is no longer an appropriate assumption. The objective of our work is to study the inherent costs due to the lack of a central authority to monitor and regulate network operation according to global objectives.

A natural framework in which to study such multiobjective problems with selfish objectives is (noncooperative) *Game Theory* [17]. Several notions of equilibria for noncooperative games have been defined and studied in the classical literature of Game Theory; the most famous of them is Nash equilibrium, originally defined in John Nash's seminal paper [15]. Roughly speaking, the strategies chosen by the players in a game constitute a *Nash equilibrium* if no player can do better by unilaterally adopting some other strategy.

Several variants and refinements of Nash equilibria have been studied in the literature of Game Theory in order to best model the appropriate solution concept for specific games (or classes of them). Examples include the *Stackelberg equilibrium* [19], where a distinguished *leader* among players in the game holds a powerful position, while the rest of the players, called *followers,* act rationally; a *perfect equilibrium* [14, 18] is obtained as a result of a limiting procedure that converges to suitable *mixed strategies* of the players; finally, a *saddle-point equilibrium* [16] is specially tailored for two-person, zero-sum finite games. In our work, we introduce and study a very general class of equilibria, which we call *approximate equilibria*; the definition is tailored to the specific selfish routing game we consider in this paper.

Unlike all previously known classes of equilibria, our approximate equilibria satisfy a very weak law: the expected latency induced on any link is at most a constant times the *optimal* (i.e., the least possible) maximum latency that could be achieved by a global algorithm. This law represents a very natural requirement to impose on any reasonable solution to the (specific) selfish routing problem we study. We shall see that Nash equilibria are a special class of approximate equilibria (for the specific game we consider).

In a recent paper Koutsoupias and Papadimitriou [10] introduced the notion of *coordination ratio* or *price of anarchy* to measure the performance loss due to lack of coordination. We extend the definition to approximate equilibria and define the *approximate coordination ratio,* which is the ratio between the *social cost* (specifically, the *expected maximum latency* in the setting we consider) in the *worst* possible approximate equilibrium, over the *social optimum,* which is the best "offline" global cost (specifically, the *least possible* maximum (over all links) latency in our setting) had a central network authority regulated traffic.

In this work, we follow Koutsoupias and Papadimitriou [10] (and its follow-up work [11]) to restrict attention on the simplest case of a network consisting of m parallel *links.* Each of n users fixes a *mixed strategy,* which is a probability distribution over links. We mainly focus on the case of *uniform* links in which the latency over each link is equal to the total traffic assigned to it. The problem for this simple network is a selfish *resource allocation* problem. Our ball fusion technique applies to more general settings such as links with arbitrary capacities or more general delays. It essentially reduces the problem to the case of (almost) equal traffic for each user; it is not however by itself sufficient to establish upper bounds on the coordination ratio for such cases.

1.2 Contribution

We view the selfish routing problem as an instance of the random experiment of independently placing n balls into m bins at random. This is the classical *balls and bins* problem (cf. [8]), which has been studied extensively both in its original form and with algorithmic extensions; (see, for example, the recent survey by Mitzenmacher *et al.* [12]). A central result in the theory of the balls and bins problem is that if one places *uniformly* and *independently* m *identical* balls into m bins, the expected maximum number of balls in a bin is $\Theta\left(\frac{\log m}{\log \log m}\right)$. To benefit from the well developed theory of the balls and bins problem, we represent users and links as balls and bins, respectively. However, since user traffics are not identical in the problem we study, balls are not identical either; instead, they come with *weights* representing the users' traffics, while the probabilities used in the random experiment are now arbitrary (in order to account for any set of mixed strategies for the users). The only property we retain in the revised version of the balls and bins problem we consider is that balls are still placed independently into the bins. Thus, the bounds shown by our analysis immediately translate to bounds on approximate coordination ratio.

Our analysis consists of two major steps. In the first step, called *ball fusion* (Section 3), we reduce the case of arbitrary weights to the case of almost equal weights (that is, where all weights are within a factor of 2 to each other). To do so, we "fuse" the two (currently) smallest balls together to form a new larger ball with weight equal to the sum of the two. Moreover, we assign new probabilities to the resulting ball in a way that the expected weight assigned to each link is preserved. We proceed to show (Lemma 3.1) that, roughly speaking, the social cost of the resulting game is no less than the social cost of the original game (before applying ball fusion). We repeat this fusion procedure until all remaining balls have weights within a factor of 2 to each other.

In the second step of our analysis (Section 4), we consider the special case where all balls have *identical* weights; the social cost for this case is no more than a multiplicative factor of 2 times the social cost for the case where all balls have weights within a factor of 2, resulting at the end of the previous step. We there apply standard techniques for estimating tails and Chernoff bounds [3] to show that (roughly speaking), for this case, the social cost is at most $O\left(\frac{\ln m}{\ln \ln m}\right)$ times the maximum expected number of balls that is placed into any link (Lemma 4.3). Putting together the two steps yields that, for balls with arbitrary weights and for any arbitrary probability distribution, the expected maximum is at most $O\left(\frac{\ln m}{\ln \ln m}\right)$ times the maximum expected number of balls that is placed into any link (Theorem 4.2). The remaining link needed for completing the proof of our main result (Theorem 4.4), namely that the approximate coordination ratio is $O(\ln m / \ln \ln m)$, is just the property we require in the definition of approximate equilibria: the maximum expected latency be at most a constant multiple of the optimum.

Recently and independently, Czumaj and Vöcking [4] obtained the same upper bound for the coordination ratio of uniform links and they greatly expanded the result to the case of links with arbitrary capacities. To show the result on uniform links they employ a lemma by Hoeffding [6] which bounds the tail probability of the sum of independent random variables. This lemma is stronger than the Chernoff bound we use in our approach and applies directly to balls of arbitrary weights thus bypassing the need to fuse balls. Their proof using the Hoeffding bound is shorter and more direct that ours. However our approach has its own merits. It is simpler and more intuitive and the ball fusion technique applies not only to the case of identical links but to a much wider class of latency functions —if it holds for pure equilibria it holds for mixed equilibria too.

2 Definitions, Background and Preliminaries

Throughout, for an integer m let $[m] = \{1, \ldots, m\}$. For an event $\mathcal{E}$ in a sample space, denote $\mathbf{Pr}(\mathcal{E})$ the probability of event $\mathcal{E}$ occurring. For a random variable X, denote $E(X)$ the *expectation* of X.

Following [10, 11], we consider a *network* consisting of a set of m parallel *links* $1, 2, \ldots, m$ from a *source* node to a *destination* node. Each of n *users* $1, 2, \ldots, n$ wishes to select a link from source to destination to route a particular amount of traffic along it; denote w_i the *traffic* of user i, $i \in [n]$. Define the $n \times 1$ *traffic vector* $\mathbf{w}$ in the natural way.

A *pure strategy* for user i is some specific link; a *mixed strategy* for user i is a probability distribution on the set of pure strategies. We often use subscripts for users and superscripts for links. A *set of pure strategies,* one per user, is represented by an n-tuple $\langle \ell_1, \ell_2, \ldots, \ell_n \rangle \in [m]^n$; a *set of mixed strategies,* one per user, is represented by an $n \times m$ *probability matrix* $\mathbf{P}$ of mn probabilities p_i^ℓ, $i \in [n]$ and $\ell \in [m]$, where p_i^ℓ is the probability that user i selects link ℓ. Clearly, $\mathbf{P} \cdot \mathbf{1} = \mathbf{1}$. For a probability matrix $\mathbf{P}$, define *indicator variables* $I_i^\ell \in \{0, 1\}$, where $i \in [n]$ and $\ell \in [m]$, such that $I_i^\ell = 1$ if and only if $p_i^\ell > 0$. This setting is reminiscent of classical *random allocation* problems, where n *balls* (with *weights*) are thrown into m *bins* at random (see, e.g., [8]). Thus, we will interchangeably use the terms users and balls, and links and bins, respectively, in our discussion.

Most of the time we assume that the *latency* for traffic w through link $\ell \in [m]$ is w. This corresponds to the model of *uniform speeds* or *capacities,* introduced in [10] and studied in [10, 11].

For any set of pure strategies $\langle \ell_1, \ell_2, \ldots, \ell_n \rangle$, the *latency cost* for user $i \in [n]$, denoted λ_i, is $\sum_{k:\ell_k=\ell_i} w_k$; that is, the latency cost for user i is the latency of the link it chooses. For any set of mixed strategies, the *expected latency cost* for user $i \in [n]$ on link $\ell \in [m]$, denoted λ_i^ℓ, is the expectation, over all random choices of the remaining users, of the latency cost for user i had its traffic been assigned to link ℓ. For any set of mixed strategies $\mathbf{P}$, denote θ^ℓ the *expected traffic* on link ℓ; clearly, $\theta^\ell = \sum_{i=1}^n p_i^\ell w_i$.

We introduce a special class of mixed strategies called *approximate equilibria.* In such equilibria, the probability matrix $\mathbf{P}$ has the property that the expected latency of any link is at most a constant times the *optimum* latency, i.e., the *least possible* maximum (over all links) latency that could be achieved if a centralized scheduler were scheduling all traffics to links.

Associated with a traffic vector $\mathbf{w}$ and a set of mixed strategies $\mathbf{P}$ is the *social cost* [10, Section 2], denoted $\mathsf{C}(\mathbf{w}, \mathbf{P})$, which is the expectation, over all random choices of the users, of the maximum (over all links) latency of traffic through a link; thus,

$$\begin{aligned}
\mathsf{C}(\mathbf{w}, \mathbf{P}) &= E\left(\max_{\ell \in [m]} \sum_{k:\ell_k=\ell} w_k\right) \\
&= \sum_{\langle \ell_1, \ell_2, \ldots, \ell_n \rangle \in [m]^n} \left(\prod_{k=1}^n p_k^{\ell_k} \cdot \max_{\ell \in [m]} \sum_{k:\ell_k=\ell} w_k\right).
\end{aligned}$$

On the other hand, the *social optimum* [10, Section 2] associated with a traffic vector $\mathbf{w}$, denoted OPT, is the *least possible* maximum (over all links) latency of traffic through a link; thus,

$$\mathsf{OPT} = \min_{\langle \ell_1, \ell_2, \ldots, \ell_n \rangle \in [m]^n} \max_{\ell \in [m]} \sum_{k:\ell_k=\ell} w_k .$$

The *approximate coordination ratio* is the maximum value, over all traffic vectors $\mathbf{w}$ and approximate equilibria $\mathbf{P}$, of the ratio $\frac{\mathsf{C}(\mathbf{w},\mathbf{P})}{\mathsf{OPT}(\mathbf{w})}$.

The above definitions can be generalized directly to the model of *arbitrary capacities* [10] where the latency for traffic w through link ℓ equals w/c^ℓ, where $c^\ell > 0$ is the *capacity* of link ℓ. More generally we consider delay functions $\mathbf{\Delta} = \langle \Delta^1, \Delta^2, \ldots, \Delta^m \rangle$ such that the latency of traffic w through link ℓ

is $\Delta^\ell(w)$. The above definitions can be extended in a straightforward way to arbitrary delay functions (i.e., use $\Delta^\ell(\sum_{k:\ell_k=\ell} w_k)$ instead of $\sum_{k:\ell_k=\ell} w_k$). In particular, let $\mathsf{C}_{\mathbf{\Delta}}(\mathbf{w}, \mathbf{P})$ and $\mathsf{OPT}_{\mathbf{\Delta}}(\mathbf{w})$ denote the social cost and social optimum for delay functions $\mathbf{\Delta}$. Notice that when we omit the delay functions as a subscript we refer to the uniform case.

3 Ball fusion

To bound the coordination ratio we consider the random experiment of placing randomly and independently n balls into m bins, and we derive general bounds on the expected maximum number of balls that are placed into any single bin.

In the most studied occupancy problem, n identical balls are placed uniformly into the m bins (see, e.g., the excellent research monograph [8]). Although many variants of the problem have been studied (see [12] for an excellent survey), the general case in which the balls have arbitrary weights and the probabilities are arbitrary has not been considered before.

In this section we develop a general technique which we call *ball fusion* to reduce the problem to the special case where all users have almost equal traffic. In the next section we will use it to bound the coordination ratio for the uniform case. However, the ball fusion technique is more general and in this section we treat general delay functions.

We start with an informal outline. Suppose we replace two balls with their sum and assign a probability to the sum so that the expected traffic for each bin remains the same. What will happen to $\mathsf{C}_{\mathbf{\Delta}}(\mathbf{w}, \mathbf{P})$? Naturally, given the positive correlation between the two balls, we expect $\mathsf{C}_{\mathbf{\Delta}}(\mathbf{w}, \mathbf{P})$ to either increase or remain the same (for natural increasing delay functions). We will show that this is indeed the case. With this in mind, we repeat the process of replacing the two smaller balls with their sum until all balls have weights within a factor of 2 to each other. Thus, we reduce the problem to the instance of identical balls (within a factor of 2). We now continue with the details of the formal proof.

Consider a mixed equilibrium (Nash or approximate). Then the strategies of the users are described by a $n \times m$ probability matrix $\mathbf{P}$. If the strategies are pure then the matrix is a 0-1 probability matrix.

Given the $n \times m$ probability matrix $\mathbf{P}$, define the $(n-1) \times m$ probability matrix $\widehat{\mathbf{P}}$ such that $\widehat{p}_i^\ell = p_i^\ell$ if $i < n-1$, and

$$\widehat{p}_{n-1}^\ell \quad = \quad \frac{p_{n-1}^\ell w_{n-1} + p_n^\ell w_n}{w_{n-1} + w_n} \, .$$

Thus, all rows of $\widehat{\mathbf{P}}$ are identical to the corresponding ones of $\mathbf{P}$, except for the last one, each entry of which is now a linear combination of the two corresponding entries in the last two rows of $\mathbf{P}$. Similarly, define $\widehat{\mathbf{w}}$ to be identical to $\mathbf{w}$ in its first $n-2$ entries, while its last entry is the sum of the last two entries of $\mathbf{w}$, i.e., $\widehat{w}_i = w_i$ for $i < n-1$ and $\widehat{w}_{n-1} = w_{n-1} + w_n$. We prove a crucial property of the resulting system:

Lemma 3.1 *For every traffic vector* $\mathbf{w}$*, if*

$$\mathsf{C}_{\mathbf{\Delta}}\left(\mathbf{w}, \mathbf{P}\right) \;\; \leq \;\; \mathsf{C}_{\mathbf{\Delta}}\left(\widehat{\mathbf{w}}, \widehat{\mathbf{P}}\right)$$

holds for probability matrices $\mathbf{P}$ *whose last two rows are 0-1, then it also holds for all probability matrices.*

To prove the lemma, consider a sequential placement of the balls into the bins. The processes for $\mathsf{C}_{\mathbf{\Delta}}\left(\mathbf{w}, \mathbf{P}\right)$ and $\mathsf{C}_{\mathbf{\Delta}}\left(\widehat{\mathbf{w}}, \widehat{\mathbf{P}}\right)$ are identical for the first $n-2$ balls. This suggests that we should study the case when the bins are not initially empty. Thus, we will prove a (slight) generalization of the lemma in which the bins have some initial weight. Let $\mathbf{L} = \langle L^1, \ldots, L^m \rangle$ be the initial weights of the bins, and let $\mathsf{C}_{\mathbf{\Delta}}\left(\mathbf{w}, \mathbf{P}, \mathbf{L}\right)$ be the expected maximum delay when we place balls with weights $\mathbf{w}$ according to probabilities $\mathbf{P}$ when links have initial weight $\mathbf{L}$. It is straightforward how to extend the definition of social cost to the case where links have initial loads $\mathbf{L}$. We prove

Lemma 3.2 *For every traffic vector* $\mathbf{w}$ *and initial load vector* $\mathbf{L}$*, if*

$$\mathsf{C}_{\mathbf{\Delta}}\left(\mathbf{w}, \mathbf{P}, \mathbf{L}\right) \;\; \leq \;\; \mathsf{C}_{\mathbf{\Delta}}\left(\widehat{\mathbf{w}}, \widehat{\mathbf{P}}, \mathbf{L}\right)$$

holds for probability matrices $\mathbf{P}$ *whose last two rows are 0-1, then it also holds for all probability matrices.*

Proof: We first show the lemma for the case where $n = 2$, and the general case will follow immediately from it.

Fix some traffic vector $\mathbf{w}$ and initial load vector $\mathbf{L}$ and assume that for 0-1 probability matrices $\mathbf{P}$ the inequality $\mathsf{C}_{\mathbf{\Delta}}\left(\mathbf{w}, \mathbf{P}, \mathbf{L}\right) \leq \mathsf{C}_{\mathbf{\Delta}}\left(\widehat{\mathbf{w}}, \widehat{\mathbf{P}}, \mathbf{L}\right)$

holds. We need to show that, for every probability matrix $\mathbf{P}$, $\mathsf{C}_{\mathbf{\Delta}}\left(\widehat{\mathbf{w}}, \widehat{\mathbf{P}}, \mathbf{L}\right) - \mathsf{C}_{\mathbf{\Delta}}\left(\mathbf{w}, \mathbf{P}, \mathbf{L}\right)$ is nonnegative. If we fix $\mathbf{L}$, $\mathbf{w}$ and the probabilities p_1^j, $j = 1, \ldots, m$, the difference $\mathsf{C}_{\mathbf{\Delta}}\left(\widehat{\mathbf{w}}, \widehat{\mathbf{P}}, \mathbf{L}\right) - \mathsf{C}_{\mathbf{\Delta}}\left(\mathbf{w}, \mathbf{P}, \mathbf{L}\right)$ is a linear function of the probabilities $\widehat{p}_2^j$, $j = 1, \ldots, m$, as it can be seen from the definition of social cost. Thus, subject to the condition that $\sum_j p_2^j = 1$, the difference $\mathsf{C}_{\mathbf{\Delta}}\left(\widehat{\mathbf{w}}, \widehat{\mathbf{P}}, \mathbf{L}\right) - \mathsf{C}_{\mathbf{\Delta}}\left(\mathbf{w}, \mathbf{P}, \mathbf{L}\right)$ is minimized when one of the probabilities $\widehat{p}_2^j$ is 1 and the rest are 0. Here is an alternative way to see this: when we fix $\mathbf{w}$, $\mathbf{L}$, and the probabilities of the first ball, the strategy that minimizes the difference $\mathsf{C}\left(\widehat{\mathbf{w}}, \widehat{\mathbf{P}}, \mathbf{L}\right) - \mathsf{C}\left(\mathbf{w}, \mathbf{P}, \mathbf{L}\right)$ is to place the second ball in the bin with the expected minimum value. Similarly, for the probabilities $\widehat{p}_1^j$, $1 \leq j \leq m$. In other words, for every pair of a traffic vector $\mathbf{w}$ and an initial load vector $\mathbf{L}$, the difference $\mathsf{C}_{\mathbf{\Delta}}\left(\widehat{\mathbf{w}}, \widehat{\mathbf{P}}, \mathbf{L}\right) - \mathsf{C}_{\mathbf{\Delta}}\left(\mathbf{w}, \mathbf{P}, \mathbf{L}\right)$ is minimized by some 0-1 probability matrix $\mathbf{P}$.

This settles the lemma when $n = 2$. The case $n > 2$, falls immediately if we consider placing the balls in order. To formalize it, we need to define $\mathbf{L}'$ the vector of expected weights after placing the first $n-2$ balls $w_1, \ldots, w_{n-2}$. The probability distribution for $\mathbf{L}'$ is the same for both $\mathsf{C}_{\mathbf{\Delta}}(\mathbf{w}, \mathbf{P}, \mathbf{L})$ and $\mathsf{C}_{\mathbf{\Delta}}(\widehat{\mathbf{w}}, \widehat{\mathbf{P}}, \mathbf{L})$. Let also $\mathbf{w}_{n-1}$ and $\mathbf{P}_{n-1}$ be the weights and probabilities of the last 2 balls. Then $\mathsf{C}_{\mathbf{\Delta}}(\widehat{\mathbf{w}}, \widehat{\mathbf{P}}, \mathbf{L}) = \sum_{\mathbf{L}'} \mathbf{Pr}(\mathbf{L}') \mathsf{C}_{\mathbf{\Delta}}(\widehat{\mathbf{w}}_{n-1}, \widehat{\mathbf{P}}_{n-1}, \mathbf{L}') \geq \sum_{\mathbf{L}'} \mathbf{Pr}(\mathbf{L}') \mathsf{C}_{\mathbf{\Delta}}(\mathbf{w}_{n-1}, \mathbf{P}_{n-1}, \mathbf{L}') = \mathsf{C}(\mathbf{w}, \mathbf{P}, \mathbf{L})$ where the inequality comes from the special case of 2 balls. ∎

Using Lemma 3.1, we can transform a set of non-identical balls into a set of almost identical ones as follows. Let $h = 2 \max_i w_i$. Keep replacing the two lighter balls with their sum while the resulting ball is no heavier than h. It is clear that the resulting weights $\widehat{w}_1, \ldots, \widehat{w}_k$ are all within a factor of 2 and in particular in the interval $[h/2, h]$—if a ball has weight less than $h/2$ it can be fused together with the (original) ball of maximum weight $h/2$.

We now show that ball fusion can be applied to pure strategies for the uniform case and the case of links with arbitrary capacities. By Lemma 3.1, we need only to show the following lemma:

Lemma 3.3 *For every traffic vector* $\mathbf{w}$ *and probability matrix* $\mathbf{P}$ *whose last two rows are 0-1*

$$\mathsf{C}\left(\mathbf{w}, \mathbf{P}\right) \quad \leq \quad \mathsf{C}\left(\widehat{\mathbf{w}}, \widehat{\mathbf{P}}\right).$$

Therefore the inequality holds for all probability matrices.

Proof: Assume that according to $\mathbf{P}$, the ball w_{n-1} is placed in bin j_{n-1} and the ball w_n is placed in bin j_n. If $j_{n-1} = j_n$ then clearly $\mathsf{C}(\widehat{\mathbf{w}}, \widehat{\mathbf{P}}) = \mathsf{C}(\mathbf{w}, \mathbf{P})$. Otherwise, let $\mathbf{L} = \langle L^1, \ldots, L^m \rangle$ be the expected weights of the bins resulting from the placement of the first $n-2$ balls. Then $\mathsf{C}(\mathbf{w}, \mathbf{P}) = \max(\max_j L^j, L^{j_{n-1}} + w_{n-1}, L^{j_n} + w_n)$. On the other hand, according to $\widehat{\mathbf{P}}$, a ball with weight $w_{n-1} + w_n$ is placed into bin j_{n-1} with probability $w_{n-1}/(w_{n-1} + w_n)$ and to bin j_n with the remaining probability. Then $\mathsf{C}(\widehat{\mathbf{w}}, \widehat{\mathbf{P}}) = \frac{w_{n-1}}{w_{n-1}+w_n} \max(\max_j L_j, L_{j_{n-1}} + w_{n-1} + w_n) + \frac{w_n}{w_{n-1}+w_n} \max(\max_j L_j, L_{j_n} + w_{n-1} + w_n)$. From this, it can be directly verified that $\mathsf{C}(\widehat{\mathbf{w}}, \widehat{\mathbf{P}}, \mathbf{L}) - \mathsf{C}(\mathbf{w}, \mathbf{P}, \mathbf{L})$ is nonnegative. A simple way to see this is to consider that $\mathsf{C}(\widehat{\mathbf{w}}, \widehat{\mathbf{P}}, \mathbf{L})$ is the *expected maximum* of the experiment with probabilities $\widehat{\mathbf{P}}$ but $\mathsf{C}(\mathbf{w}, \mathbf{P}, \mathbf{L})$ is the *maximum expected* of the same experiment. ■

An identical argument applies to the case of links with arbitrary capacities.

From the above lemma, it becomes clear that for the case of approximate equilibria, the worst-case coordination ratio occurs when all users have almost equal traffic. In fact, we can restrict our attention to exactly equal traffic; we loose at most a factor of 2 in the approximate coordination ratio. Indeed, if at the end of the fusion process when all weights are in the interval $[h/2, h]$ we keep the probabilities of the balls the same and increase their weight to h the value of $\mathsf{C}(\mathbf{w}, \mathbf{P})$ at most doubles.

4 Bounding the coordination ratio

In this section we use the technique of ball fusion to bound the coordination ratio for the uniform case. In a Nash equilibrium, each user assigns its traffic with positive probability only on links (possibly more than one of them) for which its expected latency cost is minimized; this implies that there is no incentive for a user to unilaterally deviate from its mixed strategy in order to avoid links on which its expected latency cost is higher than necessary. The following result from [10] relates the expected traffic of any link and the social optimum in a Nash equilibrium for the uniform case.

Proposition 4.1 (Koutsoupias and Papadimitriou [10]) *Take any Nash equilibrium* $\mathbf{P}$. *Then, for any link* $\ell \in [m]$,

$$\theta^\ell \quad \leq \quad \left(2 - \frac{1}{m}\right) \cdot \mathsf{OPT}\left(\mathbf{w}\right) .$$

Proposition 4.1 shows that Nash equilibria are a special case of the approximate equilibria introduced in this work. However, it is not difficult to construct approximate equilibria that are *not* Nash equilibria.

To bound the coordination ratio we need to bound $\mathsf{C}(\mathbf{w}, \mathbf{P})$. We argue that we cannot say much about the expected maximum $\mathsf{C}(\mathbf{w}, \mathbf{P})$ in the general case. For example, when all balls fall with probability one into the first bin, or when the weight of one ball dwarfs the rest, then $\mathsf{C}(\mathbf{w}, \mathbf{P})$ can be as high as $\sum_{i\in[n]} w_i$. But we will show that, if we exclude these two pathological cases, that is, when there are no very large balls and the expected weights in the bins are balanced, then, up to a constant factor, the worst case occurs when all balls are identical. Formally, we show:

Theorem 4.2 *For every traffic vector* $\mathbf{w}$ *and probability matrix* $\mathbf{P}$*,*

$$\mathsf{C}(\mathbf{w}, \mathbf{P}) \quad \in \quad \max\left\{\max_{i\in[n]} w_i, \max_{\ell\in[m]} \theta^{\ell}\right\} \cdot O\left(\frac{\ln m}{\ln\ln m}\right) .$$

Ball fusion allows as to consider only the case where all n balls are identical and have weight 1. Roughly speaking, we will establish that in this case the social cost is no more than $O(\ln m/\ln\ln m)$ times the maximum (over all links) of the expected number of balls in any particular bin. Note that this may not hold when the number of balls is small; for example, if there is only one ball which is placed uniformly into the m bins, then, the social cost (expectation of the maximum number of balls in a bin) is 1, while the maximum expected number of balls in any particular bin is $1/m$, which is substantially smaller. Thus, to fix this problem, we will bound the social cost (expected maximum) by either the maximum expected number of balls in a link or by the weight of each ball. Formally, we show:

Lemma 4.3 *For any arbitrary probability matrix* $\mathbf{P}$*,*

$$\mathsf{C}(\mathbf{1}, \mathbf{P}) \quad \le \quad \left(\frac{2e\ln m}{\ln\ln m} + 1\right) \cdot \max\left\{\max_{\ell\in[m]} \theta^{\ell}, 1\right\} .$$

Proof: The proof is not substantially different from the well-studied version of uniform probabilities. Roughly speaking, we focus on a fixed bin $\ell \in [m]$, and we show that the probability of the weight exceeding $\max\left\{1, \theta^{\ell}\right\}$ by a factor greater than $\Theta(\ln m/\ln\ln m)$ is small (specifically, $1/m^2$). The claim then follows by averaging according to the definition of social cost.

For each pair of a user $i \in [n]$ and link $\ell \in [m]$, define the (binary) random variable X_i^ℓ by $\mathbf{Pr}\left(X_i^\ell = 1\right) = p_i^\ell$ and $\mathbf{Pr}\left(X_i^\ell = 0\right) = 1 - p_i^\ell$. Let $X^\ell = \sum_{i=1}^n X_i^\ell$ and $\theta^\ell = E(X^\ell)$. For any parameter $\alpha > 0$, we upper bound the probability $\mathbf{Pr}(X^\ell \geq \alpha \max\left\{\theta^\ell, 1\right\})$, using the fact that X^ℓ is the sum of independent Bernoulli variables and apply Chernoff bounds to obtain that $\mathbf{Pr}\left(X^\ell \geq \alpha\theta^\ell\right) \leq \left(\frac{e^{\alpha-1}}{\alpha^\alpha}\right)^{\theta^\ell} \leq \left(\frac{e}{\alpha}\right)^{\alpha\theta^\ell}$. We now choose appropriately the parameter α so that this probability does not exceed $1/m^2$. We proceed by case analysis.

- Assume first that $\theta^\ell \geq 1$. Then, choose $\alpha = \frac{2e \ln m}{\ln\ln m}$. Thus, $\mathbf{Pr}\left(X^\ell \geq \alpha \max\left\{\theta^\ell, 1\right\}\right) = \mathbf{Pr}\left(X^\ell \geq \alpha\theta^\ell\right) \leq \left(\frac{e}{\alpha}\right)^{\alpha\theta^\ell} \leq \left(\frac{e}{\alpha}\right)^{\alpha} \leq \frac{1}{m^2}$.
- Assume now that $\theta^\ell < 1$. Then, choose $\alpha = \frac{1}{\theta^\ell}\frac{2e \ln m}{\ln\ln m}$. Thus, $\mathbf{Pr}\left(X^\ell \geq \alpha \max\left\{\theta^\ell, 1\right\}\right) = \mathbf{Pr}\left(X^\ell \geq \alpha\right) \leq \mathbf{Pr}\left(X^\ell \geq \alpha\theta^\ell\right) \leq \left(\frac{e}{\alpha}\right)^{\alpha\theta^\ell} \leq \left(\frac{e}{\alpha\theta^\ell}\right)^{\alpha\theta^\ell} \leq \frac{1}{m^2}$.

So, in both cases, $\mathbf{Pr}\left(X^\ell \geq \frac{2e\ln m}{\ln\ln m} \max\left\{\theta^\ell, 1\right\}\right) \leq \frac{1}{m^2}$. If we consider now all m bins, the probability that there exists a bin $\ell \in [m]$ such that X^ℓ exceeds $\frac{2e\ln m}{\ln\ln m} \max\left\{\theta^\ell, 1\right\}$ is at most $1/m$. We can now bound the expected maximum $\mathsf{C}\left(\mathbf{1}, \mathbf{P}\right)$: with probability at most $1 - 1/m$, the expected maximum is at most $\frac{2e\ln m}{\ln\ln m} \max\left\{\theta^\ell, 1\right\}$, and with the remaining probability, the expected maximum is at most n. Taking into account that $n/m \leq \max_\ell \theta^\ell$, we get that

$$\mathsf{C}\left(\mathbf{1}, \mathbf{P}\right) \;\leq\; \left(\frac{2e\ln m}{\ln\ln m} + 1\right) \cdot \max\left\{\max_{\ell\in[m]} \theta^\ell, 1\right\},$$

as needed. ■

The proof of Theorem 4.2 is now complete: we showed that

$$\mathsf{C}(\mathbf{w}, \mathbf{P}) \leq \mathsf{C}(\widehat{\mathbf{w}}, \widehat{\mathbf{P}}) \leq \mathsf{C}((2\max_i w_i)\cdot\mathbf{1}, \widehat{P}) = O(\frac{\ln m}{\ln\ln m}) \max(\max_i w_i, \max_\ell \theta^\ell),$$

where $\widehat{\mathbf{w}}$ and $\widehat{\mathbf{P}}$ are the weights and probabilities of the balls resulting from fusing the original balls.

We can now combine Theorem 4.2 with the definition of approximate equilibria to derive the desired upper bound on the approximate coordination ratio. By this definition on each link $\ell \in [m]$ the expected traffic on each link is at most $k\mathsf{OPT}$ where OPT is the social optimum and k is the constant involved in the definition of approximate equilibria.

Furthermore, since OPT is bounded below by the maximum weight, it follows that

$$\max\left\{\max_{i\in[n]} w_i, \max_{\ell\in[m]} \theta^\ell\right\} \quad \leq \quad \max\left\{\mathsf{OPT}, k\mathsf{OPT}\right\} ,$$

from which we conclude that

$$\mathsf{C}\left(\mathbf{w}, \mathbf{P}\right) \quad \in \quad O\left(\frac{\ln m}{\ln \ln m}\right) \mathsf{OPT} .$$

Thus, it follows:

Theorem 4.4 *The approximate coordination ratio (and therefore the standard coordination ratio) of selfish routing over m uniform parallel links and for any number of agents is $O\left(\frac{\ln m}{\ln \ln m}\right)$.*

5 Discussion and Directions for Further Research

We have given a method (i.e., ball fusion) which leads to almost tight bounds for the social cost of any arbitrary set of mixed strategies of n agents. We have shown also that only a very weak property of such strategies (namely, the approximate equilibria property) is needed to get tight bounds for the approximate coordination ratio of such selfish allocations.

An interesting research direction is to examine cases where the property of approximate equilibria is relaxed even more so that the expected latency per link is just a function of OPT. Finally, an obvious direction is to try to apply the technique of ball fusion to bound the coordination ratio for more general delay functions. Since ball fusion reduces essentially the problem to users with identical traffic, the remaining steps seem to require more game-theoretic arguments and less probabilistic ones. In a similar direction, recently Czumaj, Krysta, and Vöcking [5] gave bounds for some general delay functions.

References

[1] Y. Azar, A. Z. Broder, A. R. Karlin and E. Upfal, "Balanced Allocations," *SIAM Journal on Computing,* Vol. 29, No. 1, pp. 180–200, September 1999.

[2] P. Berenbrink, F. Meyer auf der Heide and K. Schröder, "Allocating Weighted Jobs in Parallel," *Proceedings of the 9th Annual ACM Symposium on Parallel Algorithms and Architectures,* pp. 302–310, June 1997.

[3] H. Chernoff, "A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations," *Annals of Mathematical Statistics,* Vol. 23, No. 4, pp. 493–509, December 1952.

[4] A. Czumaj and B. Vöcking, "Tight Bounds for Worst-case Equilibria," *Procedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms,* pp. 413–420, January 2002.

[5] A. Czumaj, P. Krysta, and B. Vöcking, "Selfish Traffic Allocation for Server Farms," Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC'02), pages 287 - 296, Montreal, Quebec, Canada, May 19 - 21, 2002.

[6] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *American Statistical Association Journal,* 58(301):13-30,1963.

[7] E. Koutsoupias, M. Mavronicolas and P. Spirakis, "A Tight Bound on Coordination Ratio," Technical Report 010029, Department of Computer Science, University of California at Los Angeles, 2001.

[8] V. F. Kolchin, V. P. Chistiakov and B. A. Sevastianov, *Random Allocations,* V. H. Winston, New York, 1978.

[9] Y. Korilis, A. Lazar and A. Orda, "Architecting Noncooperative Networks," *IEEE Journal on Selected Areas in Communications,* Vol. 13, No. 7, pp. 1241–1251, September 1995.

[10] E. Koutsoupias and C. H. Papadimitriou, "Worst-case Equilibria," *Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science,* G. Meinel and S. Tison eds., pp. 404–413, Vol. 1563, Lecture Notes in Computer Science, Springer-Verlag, March 1999.

[11] M. Mavronicolas and P. Spirakis, “The Price of Selfish Routing,” *Proceedings of the 33rd Annual ACM Symposium on Theory of Computing,* pp. 510–519, July 2001.

[12] M. Mitzenmacher, A. W. Richa, and R. Sitaraman, “The Power of Two Random Choices: A Survey of Techniques and Results,” *Handbook of Randomized Computing,* Kluwer Academic Publishers, 2001.

[13] R. Motwani and P. Raghavan, *Randomized Algorithms,* Cambridge University Press, 1995.

[14] R. B. Myerson, "Refinement of the Nash Equilibrium Concept,” *Journal of Game Theory,* Vol. 7, pp. 73–80, 1978.

[15] J. F. Nash, “Non-cooperative Games,” *Annals of Mathematics,* Vol. 54, No. 2, pp. 286–295, 1951.

[16] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior* (second edition), Princeton University Press, Princeton, NJ, 1947.

[17] M. J. Osborne and A. Rubinstein, *A Course in Game Theory,* The MIT Press, 1994.

[18] R. Selten, "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games,” *International Journal of Game Theory,* Vol. 4, pp. 25–55, 1975.

[19] H. von Stackelberg, *The Theory of the Market Economy,* Oxford University Press, Oxford, England, 1934.