Algorithmic Operations Research Vol.4 (2009) 133–143

From Balls and Bins to Points and Vertices

Ralf Klasing

LaBRI - Université Bordeaux 1 - CNRS, 351 cours de la Liberation, 33405 Talence cedex, France.

Zvi Lotker

Communication Systems Department, Ben Gurion University, Beer Sheva, Israel.

Alfredo Navarra

Dipartimento di Matematica e Informatica, Università degli Studi di Perugia, Via Vanvitelli 1, 06123 Perugia, Italy.

Stéphane Pérennes

MASCOTTE project, I3S-CNRS/INRIA/University of Nice, Sophia Antipolis, France.

Abstract

Given a graph $G = (V, E)$ with $|V| = n$, we consider the following problem. Place $m = n$ points on the vertices of G independently and uniformly at random. Once the points are placed, relocate them using a bijection from the points to the vertices that minimizes the maximum distance between the random place of the points and their target vertices.We look for an upper bound on this maximum relocation distance that holds with high probability (over the initial placements of the points). For general graphs and in the case $m \leq n$, we prove the $\#P$-hardness of the problem and that the maximum relocation distance is $O(\sqrt{n})$ with high probability. We present a Fully Polynomial Randomized Approximation Scheme when the input graph admits a polynomial-size family of witness cuts while for trees we provide a 2-approximation algorithm. Many applications concern the variation in which $m = (1 - \epsilon)n$ for some $0 < \epsilon < 1$. We provide several bounds for the maximum relocation distance according to different graph topologies.

Key words: uniform random points, grid, mapping, #P-hardness, approximation algorithm

1. Introduction

Given a set of n uniform random points inside a given square $D \subseteq \mathbb{R}^d$ and n points of a square grid covering D, an interesting question is the "cost" of ordering the random points P on the grid vertices. A natural cost function is the measure of the distance that the random points have to move in order to achieve the grid order (see for instance Figure 1).

Among all the possible bijections $f : P \to Grid$, we are interested in minimizing the maximum distance between P and $f(P)$, i.e. $\min_f \max_{1 \leq i \leq n} ||p_i - f(p_i)||_1$ with $p_i \in P$. In [17,25,26], the relation between two basic, fundamental structures like Uniform Random points and d-dimensional Grid points was studied. Those papers show that the expected minimax grid matching distance is $\Theta(\log(n)^{3/4})$ for $d = 2$ and $\Theta(\log(n)^{1/d})$ for $d > 2$. In a more general setting, we are interested in the *Points and Vertices* problem for arbitrary graphs $G = (V, E)$ with $|V| = n$ which can be described as follows:

(1) Throw n points independently and randomly onto the n vertices of G.
(2) Remap the points on G such that the load of each vertex is exactly 1, minimizing the maximal distance that any point has to move (on G).

⋆ The research was partially funded by the project "ALPAGE" of the ANR "Masse de données: Modélisation, Simulation, Applications", the project "CEPAGE" of INRIA, the European FET project AEOLUS, the European projects COST Action 293, "Graphs and Algorithms in Communication Networks" (GRAAL), and COST Action 295, "Dynamic Communication Networks" (DYNAMO). Preliminary results concerning this paper appeared in [16,19].

Email: Ralf Klasing [Ralf.Klasing@labri.fr], Zvi Lotker [zvilo@cse.bgu.ac.il], Alfredo Navarra [navarra@dmi.unipg.it], Stéphane Pérennes [Stephane.Perennes@sophia.inria.fr].

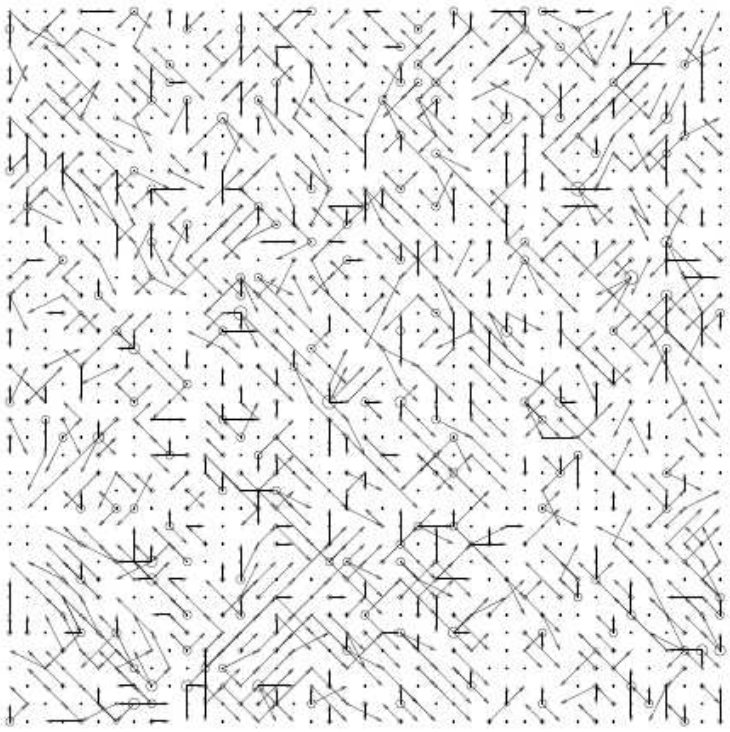

Figure 1. An example of a matching between points randomly thrown onto grid locations and the grid vertices. The radius of each circle associated to the grid vertices is proportional to the number of random points initially thrown on them.

In the following we distinguish such a problem from its variation called *Unbalanced Points and Vertices* in which the thrown points are $m = (1-\epsilon)n$ for some $0 < \epsilon < 1$ and the remapping function is slightly different. The final setting, in fact, is given by *at most* one point per vertex.

The Points and Vertices problem and the Unbalanced version may be viewed as an extension of the classical *Balls into Bins* problem, where m balls are thrown (independently and uniformly at random) into n bins, by adding graph-structural properties to the bins. The bins become vertices and there is an edge between two vertices if they are "close" enough (see e.g., [4,6,11,24] for a formal definition of the Balls into Bins problem and some of its variations). Usually, in the Balls into Bins problem the aim is to find out the distribution of the most loaded bin. In the Points and Vertices problem, instead, we are interested in the accumulation of several vertices, not only one.

The interest in the Points and Vertices problem arises from the fact that it captures in a natural way the "distance" between the *randomness* of throwing points (independently and uniformly at random) onto the vertices of G, and the *order* of the points being evenly balanced on G. In fact, our problem can be considered as the opposite of the "Discrepancy" (see for instance [3]).

Besides the pure theoretical interest, the *Points and Vertices* problem has applications in several fields. E.g., in the field of robot deployment as well as in sensor networks, one of the main problems is how to organize a huge number of randomly spread devices. The goal is usually to obtain a nearly equidistant formation so as to maximize the coverage of interesting areas [7,8]. In the field of computer graphics, the mapping of points onto cells (pixels) of a regular grid is a well-studied topic [1]. Another application in which our study can be applied concerns Geometric Pattern Matching problems [10]. In fact, we can derive good bounds on the number of edges of the bipartite graph. For more general topologies, instead, we can consider the token distribution [21,23] and load balancing problems [11,22]. The general case is constituted by a set of m tokens that must be assigned to n processors connected by a general graph. Our problem appears when the tokens are

arriving randomly uniformly and when the cost is the maximum distance that some token has to travel.

From the hardness point of view, the remapping distance in the Unbalanced version turns out to be much easier to capture than the perfectly balanced case. This is due to the fact that it becomes a much more local property. From the applications point of view this can be translated into solving the problem locally in a distributed way. This is very important for instance in the field of sensors networks where global information is rather hard to be collected.

As noted before, in Sensor Networks and Robotics, one of the main problems is to organize devices in order to maximize the coverage of the area of interest [7,8,9,20,27]. One of the main applications is given by wireless communications. Given an area of interest that can be partitioned into a set S of subareas such that $|S| = n$, and a set D of devices such that $|D| = (1-\epsilon)n$, the goal is to minimize the spent transmission capacity and/or energy needed for movements of the devices in order to monitor as many subareas as possible.

For general topologies of the underlying graph, another way to view the problem is that of having n sensors distributed over a grid (or a general graph) of n vertices but since, as in reality, a base station (a graph vertex) can handle many sensors, we can assume some capacity for each grid vertex hence again obtaining an unbalanced version of the Points and Vertices. [1]

Due to its locality another very interesting issue in which our Unbalanced Points and Vertices problem can be applied concerns hierarchy. In the field of sensors networks as well as for the internet, hierarchy turns out to be the easiest way to build a communication network while minimizing the needed resources. In fact, a network in which all the vertices have the same capabilities could be too expensive and sometimes useless. It is sufficient to consider problems like localization in sensor networks where some of the sensors are powered by GPS devices called anchors (see for instance [2]). It is easy to understand how expensive a solution could be where all the sensors have this capability. A natural question is then of finding suitable ratios between the desired level of hierarchy. An interesting issue where the unbalanced points and vertices can be applied is in exploring hierarchy problems by means of recursively solving the problem. Starting from the bottom level of the network, that is, spreading simple sensors, one question could be to understand which should be the right number of smarter (or more powerful) sensors in order to cover the whole network. Clearly the higher is the hierarchy level the less will be the thrown sensors, hence, since the underlaying network does not change, this is translated into solving the Unbalanced Points an Vertices problem with ϵ bigger and bigger.

[1] Indeed this corresponds to making the grid thicker.

1.1. *Our Results*

We formalize the *Points and Vertices* problem by defining a random variable $\rho(G)$ for the remapping distance on G. This distance turns out to be somewhat difficult to capture since it is related to global phenomena on G. We study $\rho(G)$ for general graphs and trees. (Note that results for classical topologies like paths and grids can be found in [17,25,26].) More specifically, we obtain:

(1) $\#P$-hardness for the general case.
(2) A Fully Polynomial Randomized Approximation Scheme (FPRAS) when the graph admits a polynomial-size family of witness cuts.
(3) $\rho(G) = O(\sqrt{n})$ with high probability for any connected graph G.
(4) A greedy algorithm $\mathcal{A}$ that remaps the points on any tree T with remapping distance $\rho_{\mathcal{A}}(T) \leq 2\rho(T)$.

We then concentrate our attention to the formalization of the *Unbalanced Points and Vertices* problem. While the approach is somehow quite similar to its original problem, the obtained results are very interesting due to the described locality issues. In particular, again we define a random variable $\rho'(G)$ for the remapping distance on G where we allow some imperfect (i.e. not totally balanced) remapping of the points on G. We study the behavior of $\rho'(G)$ for classical topologies like paths, trees and d-dimensional grids but also for general graphs. We derive the following results, all of them achieved with high probability:

(1) $\rho'(G_d) = \Theta(\sqrt[d]{\log n})$ for the d-dimensional square grid G_d, $d \geq 2$.
(2) $\rho'(H_d) = \Theta(1)$ for the d-dimensional hypercubes H_d.
(3) $\rho'(P_n) = \Theta(\log n)$ for the path P_n of n vertices, and $\rho'(G) = O(\log n)$ for any graph G.

The paper is organized as follows. Section 2. provides a formal definition of the *Points and Vertices* problem. Section 3. contains some general observations from which we derive the related computational hardness results. In Section 4., the greedy algorithm on arbitrary trees that computes the remapping distance up to

a factor of 2 is presented. Section 5. provides a formal definition of the *Unbalanced Points and Vertices* problem. Section 6. presents the achieved results for various topologies such as d-dimensional grids, hypercubes, paths, trees and general graphs. Finally, Section 7. gives some conclusive remarks and points out possible directions for further investigations.

2. Formalizing the Points and Vertices Problem

We study how far is a random structure from a fixed one assuming that two structures are close if there exists a "short" bijection from one to the other.

Actually, we are interested in bounding the maximum distance performed by the movement of the points randomly and uniformly distributed over the vertices of a graph $G = (V, E)$ to the vertices V by moving the points over the edges E in such a way that the final setting is given by one point per each vertex.

Definition 1 *Given a metric space with metric d and $\varrho \in \mathbb{R}^+$, a one-to-one mapping $f : A \to B$ is called* mapping with stretch ϱ from a set A to a set B *if $d(x, f(x)) \leq \varrho$ for all $x \in A$.*

Definition 2 *Given a metric space with metric d and two sets A and B, we define $\delta(A, B)$ as the minimum $\varrho \in \mathbb{R}^+$ such that there exists a one-to-one mapping with stretch ϱ from A to B.*

Let $G = (V, E)$ be a graph with $n = |V(G)|$ vertices. In what follows, $\Omega = V(G)^n$ is the probabilistic space associated to uniform independent choices of n points over the nodes $V(G)$. The events will either be considered as (indexed) sets or as positive integral weight functions on the ground set $V(G)$ with the adequate measure.

On graphs, we use the usual distance metric assuming edges of unitary length.

Problem 1 *Given a graph G with $n = |V(G)|$* vertices *and a random set $P(G, \omega)$, $\omega \in \Omega$ of n* points *lying on the vertices of G, the aim is to study the random variable $\rho(G, \omega) = \delta(P(\omega), V(G))$. In particular, what is the minimum $\varrho \in \mathbb{R}^+$ such that there exists a one-to-one mapping with stretch ϱ from $P(\omega)$ to $V(G)$?*

Problem 1 can be generalized as follows:

Problem 2 *Given a graph G, a set of locations $L \subseteq V(G)$ and a random set $P(L, \omega)$ of* points *chosen according to a distribution*[2] *F, with $\omega \in \Omega$ and $|P| = |L|$, the aim is to study the random variable $\rho(L, \omega) = \delta(P(L, \omega), L)$. In particular, what is the minimum $\varrho \in \mathbb{R}^+$ such that there exists a one-to-one mapping with stretch ϱ from $P(L, \omega)$ to L?*

In what follows, for any graph G and any $\varrho \in \mathbb{R}^+$, we will denote by $\mu(G, \varrho)$ the probability that there exists a stretch ϱ one-to-one mapping from $P(G, \omega)$ to $V(G)$, and we define $\rho(G, \omega) = \min\{\varrho \in \mathbb{R}^+ | \mu(G, \varrho) = 1 - o(1)\}$. For instance, $\rho(G, \omega) = \sqrt{n}$ means that there exists a function $f \in o(1)$ such that $\mu(G, \sqrt{n}) > 1 - f$. Whenever it will not be ambiguous, we omit the parameters G and ω.

3. Hardness Results

We will often replace our process by a Poisson process with intensity 1, since the points and vertices process is simply the Poisson process conditioned by the fact that the total number of points is $|V|$.[3] Note that the Poisson process will always fail when the number of points is not $|V|$. It follows that, denoting by $\mu_{Poisson}(G, \varrho)$ the probability of finding a stretch ϱ one-to-one mapping for the Poisson model, we have $\mu(G, \varrho) \sim \mu_{Poisson}(G, \varrho)\sqrt{2\pi|V|}$.

3.1. *Perfect matching and Duality*

The Points and Vertices problem can also be stated in terms of perfect matchings. Given a set of random points P, we build the following auxiliary bipartite graph. On one side of the graph we take as vertices the random points and on the other side the original vertices. We then connect any random point to the vertices at distance at most ϱ. A stretch ϱ mapping from P to $V(G)$ is exactly a perfect matching in the auxiliary graph. It follows that for any fixed event ω, $\rho(G, \omega)$ can be computed in polynomial time, moreover duality can be used to prove bounds on $\rho(G, \omega)$. In order to apply the König-Hall theorem (see for instance [5], Theorem 2.1.2) to the associated bipartite graph, we need the following notation. For any set $X \subseteq V(G)$ and any event $\omega \in \Omega$, we denote by $\eta(X, \omega)$ the number of random points that lie inside X. For any set $X \subseteq V(G)$, let $\Gamma^{\varrho}(X) = \{v \in V | d(X, v) \leq \varrho\}$ and $\partial^{\varrho}(X) = \Gamma^{\varrho}(X) \setminus X$. The König-Hall theorem can then be expressed as follows:

[2] In the rest of the paper, we will assume such a distribution to be the Uniform one unless differently specified.

[3] The intensity of a Poisson process represents the mean of the number of events occurring per time unit.

Lemma 1 $\rho(G,\omega) = \min\{\varrho \in \mathbb{R}^+ \mid \forall X \subseteq V(G) : |\eta(X,\omega) - |X|| \leq |\partial^\varrho(X)|\}$.

For a given ω, we will say that X is a *bad ϱ-cut* whenever $|\eta(X,\omega) - |X|| > |\partial^\varrho(X)|$. The lemma implies that the graph expansion properties are strongly related to the distribution of $\rho(G,\omega)$. The random variable $\eta(X,\omega)$ will "usually" be distributed almost like the sum of $|X|$ independent Poisson variables with intensity 1 [4]. So, $\eta(X,\omega)$ will be concentrated around its mean $|X|$ in a normal way, $\Pr(|\eta(X,\omega) - |X|| \geq t\sqrt{|X|}) \sim \frac{e^{-t^2}}{\sqrt{2t}}$.

It follows that, given a fixed $t > 0$, whenever there exists a set X such that $|X| \leq \frac{n}{2}, |\partial^\varrho(X)| \leq t\sqrt{|X|}$, the probability for X to be a bad ϱ-cut will be non-vanishing (around e^{-t^2}).

Isoperimetric properties may also lead to some upper bounds, but these will usually not be tight, indeed by the first moment method it follows:

$$\begin{aligned}\mu(G,\varrho) &= \Pr(\forall X \subseteq V, X \text{ is not a bad } \varrho\text{-cut}) \\ &\leq \sum_{X \subseteq V(G)} \Pr(X \text{ is not a bad } \varrho\text{-cut}).\end{aligned}$$

Note that such a bound is usually weak since when there exists a bad cut it is likely to happen that the event induces a very high number of bad cuts. Moreover, the bound is not easy to estimate since among the $2^{|V(G)|}$ cuts some are much more likely to be bad cuts than others (e.g., in the 2-dimensional grid a disk is much more likely to be bad than a random set of vertices).

3.2. *Computational Issues*

Our problem consists in computing the number of points in a polytope defined by an exponential number of constraints but that admits a polynomial time separation oracle (namely the perfect matching algorithm). Let the vector $(x_1, x_2, \ldots, x_n)$ with $\sum_{i=1}^n x_i = n$ represent the event with x_i points at vertex v_i, then the polytope F of feasible events for $\varrho = 1$ is the set satisfying the linear constraints:

$$\{(x_1, ..., x_n) : \ \forall X \subseteq V(G), \ |\sum_{v_i \in X} x_i - |X|| \leq |\partial(X)|\}$$

and we wish to compute $\sum_{x \in F} \ell(x)$ where $\ell(x)$ is a discrete measure derived from Ω (e.g., $\Pr(x_i = k) \sim \frac{1}{k!}$).

[4] This is not true when $|X|$ is too small or too close to n.

This suggests connections with $\#P$ counting problems or volume estimation and with $\#P$ problems for which the decision problem is in P: matchings, Eulerian cycles and in particular reliability estimation problems. With the next theorem we prove that Problem 2 is $\#P$-hard by means of a reduction from the problem of counting the number of matchings in a graph.

Theorem 1 *Problem 2 is $\#P$-hard.*

Proof. Let us assume that it is possible to compute $\mu(G,1)$ for any graph $G = (V,E)$ in polynomial time. Let $G' = (V',E')$ be the graph obtained from G by replacing each edge $e \in E$ by a path of length two (note that $|V'| = |V| + |E|$ and $|E'| = 2|E|$). We set as locations L the nodes corresponding to the original vertices of G, that is, $|L| = |V|$. Let F be a distribution of random points obtained by choosing $\frac{|V|}{2}$ vertices of G' and placing 2 points in each one. In order to obtain the number of matchings in G, it is then sufficient to multiply the probability to have a bijection between the thrown points and L with $\binom{|V'|}{\frac{|V|}{2}}$. □

Our sample space is extremely simple, and we can check if $\rho(G,\omega) \leq \varrho$ in polynomial time. So, for any fixed graph G, it is "usually" easy to compute a $(1+\varepsilon)$-approximation of $\mu(G,\varrho)$ (resp. $1-\mu(G,\varrho)$) using the Monte Carlo method. It is efficient only as long as successful (resp. failing) events can be observed. Indeed, as noted by Karp and Luby [15] if an event has probability p, a Monte Carlo estimation with $O(\frac{\log n}{\epsilon^2 p})$ samples is a $(1+\varepsilon)$-approximation of p with probability $\frac{1}{n}$. Since our goal is not to approximate $\mu(G,\varrho)$ when it is close to zero (since then we would consider $\varrho' > \varrho$), we are left with the problem of computing an approximation of $1 - \mu(G,\varrho)$ when $\mu(G,\varrho)$ is close to 1.

3.3. *FPRAS to estimate $1 - \mu(G,\varrho)$ when there is a small set of witness cuts*

We say that a family F of cuts is a family of *Witness Cuts*, whenever the probability that some cut $C \in F$ is a bad ϱ-cut, conditioned on the fact that some bad ϱ-cut exists is almost 1.

In the case we have a polynomial-size family of witness cuts, following [14], we can evaluate the probability that an event violates a cut of the family, conditioned on the fact that the event is bad. Then, we can estimate the probability of a conjunction of "simple" events like in the case of DNF formulas [15]. We refer to Vazirani [28] for a detailed comprehensive presentation.

Let $C_w, w \in W$ be a set of witness cuts, A_w be

the event *Cut C_w fails* (i.e. A_w is true when the cut fails), and p_w be the probability that this happens. By hypothesis, we have that $(1-\mu(\varrho)) \sim \Pr(A_1 \vee A_2 \vee \ldots \vee A_w)$.

Let $c(\omega)$ denote the number of cuts violated by an event ω. We have $E[c(\omega)] = E[c(\omega) \mid \omega \text{ fails}]Pr(\omega \text{ fails})$, and $E[c(\omega)]$ is simply $\sum_{w\in W} p_w$. It follows that computing $(1-\mu(\varrho))$ reduces to computing $E[c(\omega) \mid \omega \text{ fails}]$.

Consider the following sampling process:

(1) Choose ω with probability $\frac{p_w}{\sum_{v\in W} p_v}$.
(2) Pick uniformly an event ω failing for A_w.
(3) Output ω with weight $\frac{1}{c(\omega)}$.

This process samples the space of true events, moreover each true event is sampled with uniform probability. In order to get a sample space with measure m we need in the worst case $|W|m$ steps. If now we want to estimate, using T Monte Carlo trials, the value of $c(\omega)$ under the failed condition, we simply need to count $\sum_{1,2,\ldots,T} c(\omega)\frac{1}{c(\omega)} = T$ and to divide by the sample measure $\sum_{t=1,2,\ldots,T} \frac{1}{c(\omega)}$.

So, using a sample with T elements we have

$$(1-\mu(\varrho)) \sim \sum_{w\in W} p_w \frac{\sum_{t=1,2,\ldots,T} \frac{1}{c(\omega)}}{T}$$

In order to show that our algorithm is polynomial, we simply need to check that p_w can be estimated and that the space $\Omega \mid C_w$ fails can be sampled.

In the case of $i.i.d.$ points, this is straightforward, p_w is obtained via a closed formula and sampling $\Omega \mid C_w$ fails simply means conditioning on the event $\eta(C_w)$ whose distribution is also known.

3.4. *Graphs with no polynomial set of witness cuts*

Unfortunately, there exist graphs on which in order to solve the points and vertices problem, we have to consider an exponential number of cuts. In the example below, for any polynomial family of cuts, most of the events will satisfy all the cut inequalities while still violating some random [5] cut inequality.

Let us consider the following graph G. We start from a clique with k vertices and add ℓ "leaves" that are connected to all the clique nodes. The diameter of G is 2, so $\mu(2) = 1$. Let us study $\mu(1)$ with the Poisson paradigm. For any set X of leaves, we have $|\Gamma(X)| = |X|+k$, and a cut fails if $\eta(X,\omega) > |X|+k$ or $\eta(X,\omega) < |X|-k$. So, only two cuts induce the failure, but they are random, that is, the set of leaves with at least 1 point or the set of leaves with 0 points.

Since the probability for a vertex to receive p points is $\frac{1}{p!e}$, we find about $\frac{\ell}{e}$ leaves with 0 points and about $\frac{\ell}{e}$ extra points in the set of leaves with 1 or more points. So the set of vertices with 1 or more points is a bad cut with high probability as soon as $k < \frac{\ell}{e}(1-\epsilon)$.

Taking for instance $k = n^{1-\epsilon}$ and $\ell = n-k$ it follows that $\mu(1)$ is exponentially small. If we consider now a fixed cut X, it follows that $|\partial(X)| \geq k = n^{1-\epsilon}$ and the probability that $\eta(X,\omega)$ deviates from $|X|$ by more than $t\sqrt{n}$ is exponentially small. Consequently, in this graph, the probability that no matching exists is exponentially larger than the probability for a cut to fail. This means that cuts are not correlated and that the failure probability is induced by an exponential number of cuts. Note that to get an example with $\mu(1) \sim 1$, we can choose an appropriate $k \sim \frac{\ell}{e}$.

3.5. *Consequences for general graphs*

Let P_n be a path with n vertices. It is well-known that $\rho(P_n,\omega) = \sqrt{n}$ (see for instance [26]). From this example, we derive a general result for arbitrary graphs.

Intuitively, paths look like the graphs with the worst possible ρ. We can motivate this intuition as follows. Since for any graph G, G^3 contains a Hamiltonian path [13], we conclude that for any graph $\Pr(\rho(G,\omega) \geq 3k\sqrt{n}) \leq e^{-k^2}$ and so $\mu(G, \sqrt{n\log n}) \sim 1$.

Theorem 2 *For any graph G, $\rho(G,\omega) = O(\sqrt{n})$.*

The following example shows that for some graph G_0, $\rho(G_0,\omega) >> \sqrt{D} >> 1$ (where D is the diameter of G_0), i.e. $\Pr(\rho(G_0,\omega) \geq \sqrt{D})$ is small. Consider two complete graphs with n nodes connected with a path of length ℓ, with $\ell \leq \sqrt{n/4}$. If the number of points in one of the complete graphs deviates by more than ℓ (this happen with finite probability), $\rho(G_0,\omega)$ is larger than ℓ, so we have $D = \sqrt{n}$ and $\rho(G_0,\omega) = \Theta(D) = \Theta(\sqrt{n})$ with large probability. Note that we can replace the complete graphs by binary trees to get a bounded-degree example.

4. Trees

Previous results for paths and grids can be found in [17,25,26]. In this section, we consider tree topologies and we show that $\mu(\varrho)$ is quite well described by a

[5] In the Kolmogorov acceptation.

few cut inequalities. Hence, we describe a greedy algorithm that for a given tree T and a set of points $P(T,\omega)$ evaluates up to a factor of 2 the value $\rho(T,\omega)$.

4.1. *A greedy approximation algorithm*

We use a labeling process, each node v receives a family of labels. Label $+\ell$ (resp. $-\ell$) means that one point (resp. vertex) at distance ℓ from v in the subtree rooted at v need to be assigned an image (resp. a pre-image) outside the subtree.

To each leaf we associate a label -1 if there are no points inside it, 0 if there is 1 point, $p-1$ times $+1$ if there are p points. Then for each subtree whose vertices are already labeled except for the root, we compute the number of positive labels minus the number of negative labels. Let us call s such a number. If $s>0$ we label the root with the smallest $s-1$ positive numbers contained in the previous labels increased by 1 and a $+1$ for each point contained in it. If $s<0$, let s' be the number of points contained in the root. If $s'>|s|$ then we label the root with $s'+s-1$ times $+1$ (hence with 0 if $s'+s-1=0$); if $s'<|s|$ then with the biggest $|s|-s'$ negative numbers contained in the previous labels decreased by 1 and a -1; if $s'=|s|$ just with a -1. We can then continue the process until the whole tree is labeled. Since we are considering a number of points equal to the number of vertices, the last vertex will be labeled by just a 0, see Figure 2.

Let $m(v)$ be the biggest absolute value appearing as a label for a node v, it is possible to prove by induction that any matching will have to use a path with length at least $m(v)$ going through v. This property is due to the fact that the algorithm always pushes up the smallest possible set of "ordered" labels (according to the positive cone order $u>v$ when $u-v$ is a positive vector). It follows that if M is the biggest absolute value of a label, then $\rho \geq M$.

Now, remark that we can easily find a matching with stretch $2M$ by associating positive labels with negative labels.

4.2. *Analysis of the algorithm*

In order to compute the probability of finding a matching between random points and the tree vertices, we would normally apply the Hall theorem to every vertex-subset of the tree. The greedy algorithm tells us that we can actually reduce our attention to specific subsets obtained that correspond to edge-cuts. There are $2(n-1)$ such subsets, reducing the number of witness cuts from an exponential to a linear number.

Definition 3 *For a given tree T, $T'<T$ if T' is one of the two subtrees obtained by removing one edge of T.*

Lemma 2 *Given a tree $T=(V,E)$, $T'<T$ and stretch ϱ, it is possible to compute in polynomial time the probability that T' induces a bad cut for ϱ.*

Proof. Using standard binomial coefficient evaluation, we can compute

$$\Pr[\eta(T',\omega)>|\Gamma^{\varrho}(T')|] = \sum_{i=|\Gamma^{\varrho}(T')|+1}^{|V|} \binom{|V|}{i} \left(\frac{|T'|}{|V|}\right)^{i} \left(1-\frac{|T'|}{|V|}\right)^{|V|-i};$$

and do the same for $\Pr[\eta(T',\omega)<|\Gamma^{\varrho}(T')|]$. □

Theorem 3 *Given a tree $T=(V,E)$ and any stretch ϱ, it is possible to approximate $1-\mu(T,\varrho)$ within $2|V|$.*

Proof. From the previously described labeling scheme, $1-\mu(\varrho) \leq \Pr(\exists T'<T$ such that $\eta(T',\omega) \geq |\Gamma^{2\varrho}(T')|)$ and $\sum_{T'<T} Pr(\eta(T',\omega)\geq|\Gamma^{2\varrho}(T')|) \leq 2(|V|-1)\max_{T'<T} Pr(\eta(T',\omega)\geq|\Gamma^{2\varrho}(T')|)$.

Moreover, $\max_{T'<T}\Pr(\eta(T',\omega)>|\Gamma^{2\varrho}(T')|) \leq 1-\mu(2\varrho) \leq 1-\mu(\varrho)$. It follows that $1-\mu(\varrho) \leq 2(|V|-1)\max_{T'<T}\Pr(\eta(T',\omega)\geq|\Gamma^{2\varrho}(T')|) \leq (1-\mu(\varrho))2(|V|-1)$. Since by Lemma 2 such probabilities can be computed in polynomial time, the claim holds. □

The previous proof can be interpreted as follows. If there exists a bad cut then there exists a bad cut that is defined by a subtree obtained by removing one edge. This means that we can use n witness cuts and get a good estimation of the probability of failing using simple cut considerations. Since on the line there is only one witness cut (the half line) we wonder if the same happens for trees. Consider a subdivided star with k branches of length k, we see that for $\varrho=\sqrt{k}$ no cut is likely to be bad. Now, consider an event, with high probability any branch will contain $k+\Theta(\sqrt{k\log k})$ points, and then each branch is an independent Poisson process conditioned on its number of points. Let $C_i, i=1,2,\ldots,k$ be the set containing the $k/2$ points of branch i that are at distance at least $k/2$ from the central node. Since we have k branches, one of them will deviate by about $\sqrt{m\log k}$ where m is its mean. So for some C_i, $|\eta(C_i,\omega)-|C_i|| = \Theta(\sqrt{k\log k})$ and we need $\varrho=\Theta(\sqrt{k\log k})$ to get $\mu(\varrho)\sim\frac{1}{2}$, and $\mu(t\sqrt{k}) \leq e^{-t^2}k$. On this graph we find about $k=\sqrt{n}$ "independent"

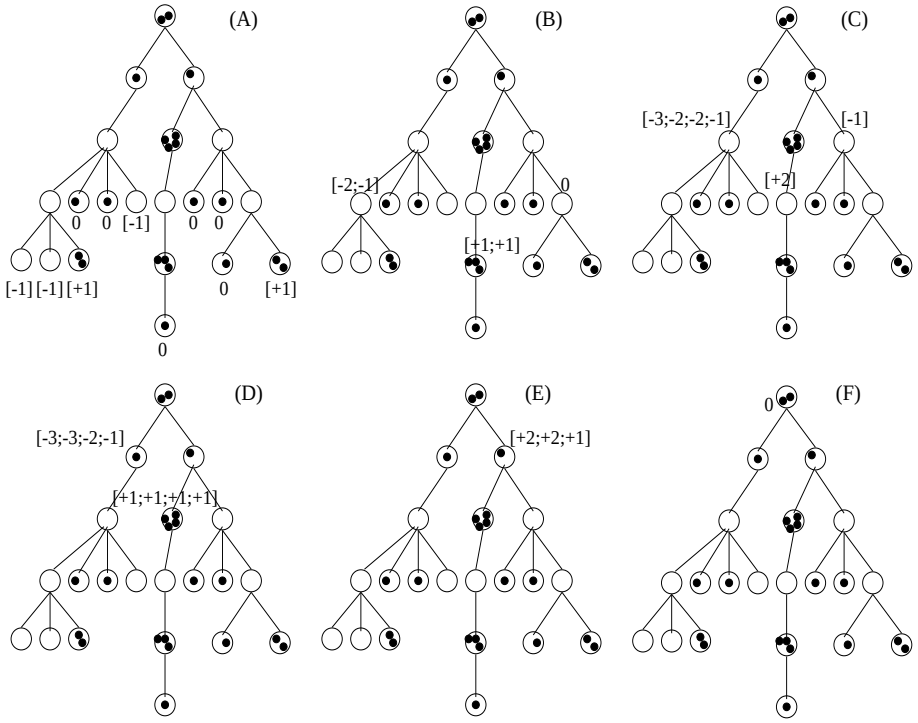

Figure 2. An example of the labeling process. The maximum absolute value obtained is $M = 3$.

cuts and ϱ is chosen such that the probability of each of these cuts to be bad is less than $\frac{1}{\sqrt{n}}$.

5. Formalizing the Unbalanced Points and Vertices Problem

In this section we slightly modify the previous problem by assuming less points than vertices. We then study how far is a random structure from an almost fixed one, in the sense that now points have some degree of liberty in order to be distributed. Again, we assume that two structures are close if there exists a "short" mapping from one to the other. We make use of a positive number ϵ and we consider a function that has to map $(1-\epsilon)n$ objects into n by following predetermined roots (the edges of a given graph).

Actually, we are interested in bounding the maximum distance performed by the movement of the points to the closest available vertices in such a way that in the final setting each vertex has associated at most one point.

Definition 4 *Given a metric space with metric d and $\varrho' \in \mathbb{R}^+$, a mapping $f : A \to B$ is called* mapping with stretch ϱ' from a set A to a set B *if $d(x, f(x)) \leq \varrho'$ for all $x \in A$.*

Definition 5 *Given a metric space with metric d and two sets A and B, we define $\delta'(A, B)$ as the minimum $\varrho' \in \mathbb{R}^+$ such that there exists a mapping with stretch ϱ' from A to B.*

The above definitions differ from Definition 1, 2, respectively, by the fact that the mapping function f is not one-to-one anymore.

Problem 3 *Given $0 < \epsilon < 1$, a graph G with $n = |V(G)|$* vertices *and a random set of* points *$P(\omega)$, with $\omega \in \Omega$ and $|P| = (1-\epsilon)n$, lying on $V(G)$, the aim is to study the random variable $\rho'(G, \omega) = \delta'(P(\omega), V(G))$.*

In what follows, for any graph G and any $\varrho' \in \mathbb{R}^+$, we will denote by $\mu'(G, \varrho')$ the probability that there exists a stretch at most ϱ' mapping from $P(\omega)$ to $V(G)$, and we define $\rho'(G) = \min\{\varrho' \in \mathbb{R}^+ | \mu'(G, \varrho') = 1 - o(1)\}$.

6. Achieved Bounds

In this section we consider the Unbalanced Points and Vertices problem on various topologies for the underlying graph. As defined we will use $\mu'(\varrho')$ to denote the probability that a stretch ϱ' mapping from the points to the vertices exists and ϱ'_0 to denote the minimum ϱ'

such that $\mu'(\varrho') \sim 1$.

6.1. *Case of the d-dimensional grids*

Theorem 4 *For d-dimensional grids $\mu'(c_{\epsilon,d}\sqrt[d]{\log n}) = 1 - \frac{1}{\log n}$ and $\mu'(d_{\epsilon,d}\sqrt[d]{\log n})) \leq \frac{1}{2}$ for some constants $c_{\epsilon,d} \leq d_{\epsilon,d}$.*

Proof. Note that for technical reasons we consider $\epsilon < \frac{1}{2}$. Anyway, from a practical point of view, the problem becomes much easier as ϵ increases since we have less points to remap.

For some constant c_ϵ to be specified, we partition regularly G_d into subgrids (boxes) containing on average $c_\epsilon \log n$ points, and hence $c_\epsilon \frac{\log n}{1-\epsilon}$ vertices. We prove that the points can be rearranged inside each box of the partition. Let X_i be the number of random points belonging to the i-th box, the points can be rearranged in this box whenever $X_i \leq c_\epsilon \frac{\log n}{1-\epsilon}$. From Chernoff (see [12]) and from the assumption that $\epsilon < \frac{1}{2}$, we have

$$\Pr(X_i \geq (1 + \frac{\epsilon}{1-\epsilon})c_\epsilon \log n) \leq e^{-((\frac{\epsilon}{1-\epsilon})^2 c_\epsilon \log n)/3}$$

So, for $c_\epsilon \geq 3\left(\frac{1-\epsilon}{\epsilon}\right)^2$, $\Pr(X_i$ cannot be rearranged locally $) \leq \frac{1}{n}$. Since there are $\Theta(\frac{n}{\log n})$ boxes, from the *union bound*, none will fail with probability $1 - \frac{1}{\log n}$.

Now, the diameter of each box is $d\left(\frac{c_\epsilon}{1-\epsilon}\right)^{1/d} \times (\log n)^{1/d}$. The result follows by setting $c_{\epsilon,d} = d\left(\frac{c_\epsilon}{1-\epsilon}\right)^{1/d}$.

In order to estimate a lower bound it is possible to apply [24] when the number of bins is $(1+\epsilon)n$ and the number of balls is n in order to show, with high probability, that one bin contains $O(\frac{\log n}{\log\log n})$ balls. Therefore, one of those balls has to move at least $O(\sqrt[d]{\frac{\log n}{\log\log n}})$ vertices in order to achieve the desired matching.

We now give a matching lower bound for Theorem 4. Let $L = \Theta(\log n)$ to be specified exactly later. We consider a partition of G into $\frac{n}{L}$ boxes containing L vertices and average number of points $\mu' = L(1-\epsilon)$.

Let $\delta' > 0$. For a given box B, the probability that $n(B)$ deviates by $\delta'\mu'$ is as follows:

$$\left(\frac{e^{\delta'}}{(1+\delta')^{1+\delta'}}\right)^{\mu'} = f(\delta')^{\mu'} = f(\delta')^{(1-\epsilon)L}$$

Note that $f(\delta') < 1$.

We say that a box B with L vertices fails whenever $n(B) \geq (1+\delta')\mu' = (\epsilon - \epsilon^2)L$. Let $\delta' = 2\epsilon$ and $L = \frac{\log n}{-3\log f(\delta')(1-\epsilon)}$. The probability that a box fails is about $n^{\frac{1}{3}}$, and the number of failing boxes is bigger than $n^{\frac{2}{3}}\sqrt{L}$. Since the number of failing boxes is a martingale[6] [29], we conclude that with high probability the number of failing boxes is $n^{\frac{2}{3}}\sqrt{L} + \Theta(\sqrt{n})$.

It follows that a box with $L = \frac{\log n}{-3\log f(\delta')(1-\epsilon)} = g(\epsilon)\log n$ fails, hence such a box misses $h(\epsilon)(\epsilon - \epsilon^2)L$ vertices.

Each box B has side length $L^{1/d}$. Let $\Gamma^{\varrho'}(B) = \{v \in V | d(B,v) \leq \varrho'\}$, in order to find $h(\epsilon)L$ points in $\Gamma^{\varrho'}(B)$, we need $\varrho' L^{d-1} d \geq h(\epsilon)L$. So, $\varrho' \geq L^{1/d}h(\epsilon) = f(\epsilon)^{1/d}h(\epsilon)L^{1/d} = f(\epsilon)^{1/d}h(\epsilon)g(\epsilon)^{1/d}(\log n)^{1/d}$. □

6.2. *Extreme case of the d-cube*

The graph $H_d = P_2^d$ is a limit case of grids or tori, since we cannot divide at all the sides that are too short. To get a tight result we need to partition the cube into small radius balls, that is exactly what error correcting codes are doing (see for instance [18]). For instance, using a 1 bit perfect correcting code when it exists, determines an exact partition into $\frac{2^d}{d}$ balls with radius 1. Those balls will be large enough (when ϵ is close enough to 1), and $\mu'(2) \sim 1$ for ϵ close to 1. For larger density of points we need larger boxes, balls of radius 2 with size $\sum_{i=0}^{2}\binom{d}{i} > d^2/2$ will be large enough for any fixed ϵ (and d large enough). Since almost perfect error correcting codes for distance 2 exist we have:

Lemma 3 *For any ϵ, $\mu'(H_d, 4) \sim 1$ when $d \to \infty$.*

6.3. *Paths and General Graphs*

As it was for the balanced version of the problem, paths look like the graphs with the worst possible ρ'. Again, since for any generic graph G, G^3 contains a Hamiltonian path [13], we conclude that for any graph $\Pr(\rho'(\omega) \geq 3k\sqrt{n}) \leq e^{-k^2}$ and so $\mu'(\sqrt{n\log n}) \sim 1$.

Theorem 5 *For any graph G, $\rho'(G) = O(\log n)$.*

Using the same idea as for grids we can try to partition the graph into boxes of small radius containing at least $f(\epsilon)\log n$ vertices for some function f depending just on ϵ.

[6] Indeed boxes are almost independent and we could use $L = \frac{\log n}{-3\log f(\delta')(1-\epsilon)}$.

Theorem 6 *$\mu'(\varrho'_0) \sim 1$ for $\rho'_0 = min\{\varrho' \mid \forall x \in V(G), |\partial^{\varrho'}(\{x\})| \geq f(\epsilon)\log n\}$*

Proof. Let $\rho'_0 = min\{\varrho' \mid \forall x \in V(G), |\partial^{\varrho'}(\{x\})| \geq f(\epsilon)\log n\}$. Using the usual covering and packing argument, we can greedily prune disjoint radius ϱ'_0 disks until any vertex is at distance less than ϱ'_0 from the pruned set. The remaining vertices are then spread arbitrarily into the existing disks. It follows that we can partition the graph into sets with size at least $f(\epsilon)\log n$ that have eccentricity $2\varrho'_0$, and hence diameter at most $4\varrho'_0$.

Using the same analysis as in the case of grids, with high probability we can find a matching in each subset. □

For tree topologies it is enough to note that an almost identical proof to the one presented in Section 4. still holds, hence obtaining a constant approximation factor.

Theorem 7 *Given a tree $T = (V, E)$ and any stretch ϱ', it is possible to approximate $1 - \mu'(T, \varrho')$ within $2|V|$.*

7. Conclusion and Future Work

We have introduced the *Points and Vertices* problem for a graph G, which captures in a natural way the "distance" between the *randomness* of throwing points (independently and uniformly at random) onto the vertices of G, and the *order* of the points being evenly balanced on G. We have derived several results on the problem with exact balancing of the points. Besides the pure theoretical interest, the *Points and Vertices* problem turns out to be of relevant interest in several fields motivating further investigation.

As a variation of the *Points and Vertices* problem, we have also introduced the *Unbalanced* version in which the thrown points are less than the vertices of the underlying graphs. After showing that this new problem has many applications in many interesting field such as robotics and wireless networking, we have derived bounds according to different topologies of the underlying graph. The obtained results confirm the intuition that this new mapping distance is much more local with respect to the original one, hence much more suitable for distributed and parallel environments.

As future work, further investigation concerning other topologies can be approached as well as experimental results concerning derived solutions for basic problems such as clustering or leader election in sensor networks.

The model suggests to investigate a similar problem which is very important in the field of sensor networks. Considering in fact $(1 - \epsilon)n$ points and n vertices, a natural problem could be to distribute the points in such a way that the final setting not only has at most one point per vertex but also it should guarantee the maximum coverage of the area of interest. For maximum coverage we intend to spread the points (sensors) as much as possible according to their power transmission range in such a way that the induced graph according to the neighborhood of each point is connected.

Also the variation of the Points and Vertices problem for which more points with respect to the number of vertices are thrown is of interest. From the applicability point of view, in the field of sensor networks, for instance, this can be translated into having some more sensors that can use less energy in order to achieve the desired communications. This would suggest to determine a suitable trade-off between the minimum number of sensor that must be thrown and the maximum transmission range that each sensor has to guarantee.

References

[1] BANEZ, J. M. D., HURTADO, F., LOPEZ, M. A., AND SELLARES, J. A. Optimal point set projections onto regular grids. In *Proceedings of the 14^{th} Annual International Symposium on Algorithms and Computation (ISAAC)* (2003), vol. 2906 of *LNCS*, pp. 270–279.

[2] CHAN, H., LUK, M., AND PERRIG, A. Using clustering information for sensor network localization. In *Proceedings of the 1st International Conference on Distributed Computing in Sensor Systems (DCOSS)* (2005), pp. 109–125.

[3] CHAZELLE, B. *The Discrepancy Method: Randomness and Complexity*. Cambridge University Press, 2002.

[4] COLE, R., FRIEZE, A. M., MAGGS, B. M., MITZENMACHER, M., RICHA, A. W., SITARAMAN, R. K., AND UPFAL, E. On balls and bins with deletions. In *Proceedings of the 2^{nd} International Workshop on Randomization and Approximation Techniques in Computer Science (RANDOM)* (1998), pp. 145–158.

[5] DIESTEL, R. *Graph Theory*. Springer-Verlag, New York, 2^{nd} edition, 2000.

[6] DRINEA, E., FRIEZE, A., AND MITZENMACHER, M. Balls and bins models with feedback. In *Proceedings of the 13^{th} Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)* (2002), pp. 308–315.

[7] DUDENHOEFFER, D. D., AND JONES, M. P. A formation behavior for large-scale micro-robot force deployment. In *Proceedings of the 32^{nd} Conference on Winter Simulation* (2000), pp. 972–982.

[8] GOBRIEL, S., MELHEM, R., AND MOSSE, D. A unified interference/collision analysis for power-aware adhoc networks. In *Proceedings of the 23rd Conference of the IEEE Communications Society* (2004).

[9] HSIANG, T.-R., ARKIN, E., BENDER, M. A., FEKETE, S., AND MITCHELL, J. Algorithms for rapidly dispersing robot swarms in unknown environments. In *Proceedings of the 5th Workshop on Algorithmic Foundations of Robotics (WAFR)* (2002), pp. 77–94.

[10] INDYK, P., MOTWANI, R., AND VENKATASUBRAMANIAN, S. Geometric matching under noise: combinatorial bounds and algorithms. In *Proceedings of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms* (1999), pp. 457–465.

[11] IWAMA, K., AND KAWACHI, A. Approximated two choices in randomized load balancing. In *Proceedings of the* 15th *Annual International Symposium on Algorithms and Computation (ISAAC)* (2004), vol. 3341 of *LNCS*, pp. 545–557.

[12] JUKNA, S. *Extremal Combinatorics with Applications in Computer Science*. Springer-Verlag, 2001.

[13] KARAGANIS, J. J. On the cube of a graph. *Canadian Mathematical Bulletin 11* (1969), 295–296.

[14] KARGER, D. R. A randomized fully polynomial time approximation scheme for the all-terminal network reliability problem. *SIAM Journal on Computing 29*, 2 (2000), 492–514.

[15] KARP, R. M., AND LUBY, M. G. Monte carlo algorithms for planar multiterminal network reliability problems. *Journal of Complexity 1* (1985), 45–64.

[16] KLASING, R., LOTKER, Z., NAVARRA, A., AND PERENNES, S. From balls and bins to points and vertices. In *Proceedings of the* 16th *Annual International Symposium on Algorithms and Computation (ISAAC)* (2005), vol. 3827 of *LNCS*, pp. 757–766.

[17] LEIGHTON, F. T., AND SHOR, P. W. Tight bounds for minimax grid matching with applications to the average case analysis of algorithms. *Combinatorica 9*, 2 (1989), 161–187.

Received 24-09-2007; revised 15-09-2008; accepted 30-10-2008

[18] LIN, S., AND COSTELLO, D. J. *Error Control Coding: Fundamentals and Applications*. Prentice Hall: Englewood Cliffs, NJ, 1983.

[19] LOTKER, Z., AND NAVARRA, A. Unbalanced points and vertices problem. In *Proceedings of the 1st IEEE PERCOM International Workshop on Foundations and Algorithms for Wireless Networking (FAWN)* (2006), pp. 96–100.

[20] MCCANN, J. A., NAVARRA, A., AND PAPADOPOULOS, A. A. Connectionless Probabilistic (CoP) routing: an efficient protocol for Mobile Wireless Ad-Hoc Sensor Networks. In *Proceedings of the 24th International Performance Computing and Communications Conference (IPCCC)* (2005), pp. 73–77.

[21] MEYER, F., OESTERDIEKHOFF, B., AND WANKA, R. Strongly adaptive token distribution. *Algorithmica 15* (1993), 413–427.

[22] MITZENMACHER, M., PRABHAKAR, B., AND SHAH., D. Load balancing with memory. In *Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science (FOCS)* (2002), pp. 799–808.

[23] PELEG, D., AND UPFAL, E. The token distribution problem. *SIAM J. Comput. 18*, 2 (1989), 229–243.

[24] RAAB, M., AND STEGER, A. "Balls into bins" - a simple and tight analysis. In *Proceedings of the 2nd International Workshop on Randomization and Approximation Techniques in Computer Science (RANDOM)* (1998), pp. 159–170.

[25] SHOR, P. W. The average-case analysis of some on-line algorithms for bin packing. *Combinatorica 6*, 2 (1986), 179–200.

[26] SHOR, P. W., AND YUKICH, J. E. Minimax grid matching and empirical measures. *The Annals of Probability 19*, 3 (1991), 1338–1348.

[27] SPEARS, W., HEIL, R., SPEARS, D., AND ZARZHITSKY, D. Physicomimetics for mobile robot formations. In *Proceedings of the 3rd International Conference on Autonomous Agents and Multi Agent Systems (AAMAS)* (2004).

[28] VAZIRANI, V. *Approximation Algorithms*. Springer, 2001.

[29] WILLIAMS, D. *Probability with Martingales*. Cambridge University Press, 1991.