Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

AMIN COJA-OGHLAN SVEN O. KRUMKE TILL NIERHOFF

Average Case Analysis of a Hard Dial-a-Ride Problem

ZIB-Report 03-12 (May 2003)

Average Case Analysis of a Hard Dial-a-Ride Problem

Amin Coja-Oghlan* Sven O. Krumke† Till Nierhoff‡

13th May 2003

Abstract

In the dial-a-ride-problem (DARP) objects have to be moved between given sources and destinations in a transportation network by means of a server. The goal is to find a shortest transportation for the server. We study the DARP when the underlying transportation network forms a caterpillar. This special case is strongly NP-hard in the worst case. We prove that in a probabilistic setting there exists a polynomial time algorithm which almost surely finds an optimal solution. Moreover, with high probability the optimality of the solution found can be certified efficiently. We also examine the complexity of the DARP in a semi-random setting and in the unweighted case.

Keywords: dial-a-ride-problem, average case analysis, MST-heuristic, Steiner trees

*Humboldt-Universität zu Berlin, Institut für Informatik, Unter den Linden 6, 10099 Berlin, Germany. Email: coja@informatik.hu-berlin.de. Research supported by the German Science Foundation (DFG, grant FOR 413/1-1)

†Konrad-Zuse-Zentrum für Informationstechnik Berlin, Department Optimization, Takustr. 7, D-14195 Berlin-Dahlem, Germany. Email: krumke@zib.de. Research supported by the German Science Foundation (DFG, grant Gr 883/10)

‡Humboldt-Universität zu Berlin, Institut für Informatik, Unter den Linden 6, 10099 Berlin, Germany. Email: nierhoff@informatik.hu-berlin.de. Research supported by the German Science Foundation (DFG, grant PR 296/6-3)

1 Introduction

In the dial-a-ride problem (DARP) we are given a number of transportation requests which have to be handled by means of a server. The server can handle at most one request at a time and moves within a specified transportation network. The aim is to find a shortest (closed) tour for the server which serves all requests. The DARP comprises many transportation and routing problems in combinatorial optimization such as the traveling salesman problem.

One of the applications that can be put within the DARP framework is elevator scheduling [2, 18, 21]. This corresponds to the special case of the DARP where the underlying transportation network forms a caterpillar (cf. Figure 1). Here, the vertices on the backbone correspond to the floors and the edges between vertices on the backbone and the feet can be used to model start- and stopping delays of the elevator. This special case is NP-hard, as has already been shown in [21].

In reality, the task of scheduling an elevator is in fact an *online problem*: transportation requests are unknown until their respective release times and an online algorithm must decide how to handle requests without knowledge of the future. Practice is even more demanding. An online algorithm is indeed required to deliver the next piece of the solution within a very tight time bound. Thus, one is interested in online algorithms which do not only deliver good so-

lutions but which also react in real-time.

A standard way to measure quality of online algorithms is via competitive analysis [7]. All known competitive algorithms for minimizing the total completion time (makespan) in online dial-a-ride problems have to solve instances of the (offline-) DARP during their run [2, 4, 12]. It is shown in [2] that an offline approximation algorithm for the DARP with approximation ratio ρ implies a $c(\rho)$-competitive algorithm for the online version, where $c(\rho) = \frac{1}{4}\left(4\rho + 1 + \sqrt{1+8\rho}\right)$. Moreover, as shown in [19, 20] even for the case of minimizing the maximum or average waiting time online, an offline algorithm for the DARP which optimizes the length of a tour proves to be helpful, since it can be used to derive online performance guarantees.

We conclude that there is a need to solve the (offline-) DARP in real-time, although it is an NP-hard problem.

Our Contribution In this paper we address the complexity of the DARP on caterpillars in a probabilistic setting. We show that the so-called MST-heuristic, a fast and simple algorithm (see Section 1.2), in most cases solves the problem exactly if the transportation requests are chosen uniformly at random. We expect this result to be of use in view of the real-time issue for online algorithms as mentioned above.

Note that our result is also interesting in the context of the algorithmic theory of random graphs [16]: the DARP constitutes another combinatorial optimization problem which is hard in the worst-case and easy on average. The proof that the problem is easy on average relies mainly on an analysis of the so-called "balancing operation". Although this operation has no effect in the worst case, it turns out that in the average case balancing glues all non-Eulerian connected components of requests together. Moreover, Eulerian components are rare. The analysis of balancing causes considerable technical challenges. The key is an appropriate description of the random model, namely as a direct product of a random walk and the choice of a random permutation.

We complement our algorithmic result with a hardness result about the solvability of the DARP in a semi-random setting, which, as a byproduct, implies NP-hardness in the unweighted case.

Related Work The DARP is also known as the Stacker-Crane-Problem. In [21] it is shown that the problem is NP-hard even on caterpillars (with appropriate edge lengths). An earlier NP-hardness result for the DARP on trees is contained in [14]. In [15] the authors present a $9/5$-approximation algorithm for the DARP on general graphs. An improved algorithm for trees with performance $5/4$ is given in [14]. On paths, the DARP can be solved in polynomial time [3]. The paper [21] considers the DARP when additional precedence constraints between the requests are specified.

Organisation of the Paper In the rest of this introduction we give a formal problem statement and a synopsis of the results of the paper. The synopsis has pointers to the proof sketches, which are contained in the other sections. After some concluding remarks and the bibliography, there is an appendix containing detailed proofs.

1.1 Problem Statement

In the dial-a-ride problem DARP we are given an edge-weighted undirected graph $G = (V, E)$ and a list of transportation requests L between the vertices of G. The goal is to find a shortest (closed) tour which serves all the requests in L. This task can be viewed as adding new arcs A (empty moves) to

the directed graph (V, L) such that the resulting directed multi-graph $(V, L \cup A)$ is Eulerian [3, 14, 21]. We thus state DARP formally in the graph theoretic framework as follows:

Definition 1 (Dial-a-Ride Problem DARP) *An instance of the dial-a-ride problem* DARP *consists of an undirected graph $G = (V, E)$ with edge-lengths $c\colon E \to \mathbb{R}_0^+$ and a list L of pairs of vertices, called* requests. *A solution is a multi-set A of pairs (u, v) where $\{u, v\} \in E$ such that the directed multi-graph $(V, L \cup A)$ is Eulerian. The cost of A is the total length of an Euler tour in $(V, L \cup A)$, where the length of arc (u, v) equals the length of a shortest path between u and v in G with respect to c.*

As mentioned before, in this paper we consider the situation where the undirected graph G in the DARP is a *caterpillar*. The caterpillar Cat_n (see Figure 1) consists of a path on n vertices $p_1, \ldots, p_n$ and n leaves $l_1, \ldots, l_n$ where l_i is attached to p_i, $i = 1, \ldots, n$. The edges $h_i := \{p_i, l_i\}$ are called *hairs*, the leaves l_i are called *feet*, the edges $b_i := \{p_i, p_{i+1}\}$ are called *backbone* edges. Obviously, Cat_n is a tree on $2n$ vertices.

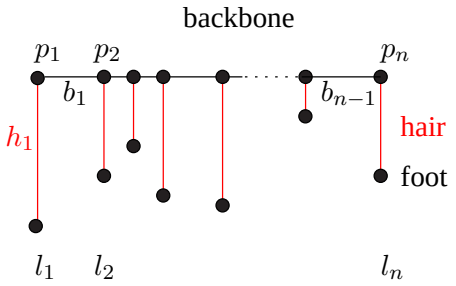

Figure 1: The caterpillar graph Cat_n.

We further assume that the requests extend between feet of the caterpillar. This is not a restriction of generality, as, instead of $G = \mathrm{Cat}_n$, we could consider $G' = \mathrm{Cat}_{2n}$, where every second backbone edge and every second hair had length 0. Then, introducing a request in G' for every request in G by replacing p_i with l'_{2i} and l_i with l_{2i-1}, results in an equivalent problem obeying the stated restriction. Note that the application background suggests that other requests than between feet in fact do not occur.

Any instance of this problem can be preprocessed by adding to the list L of requests a set B of "artificial" requests such that the value of an optimal solution does not increase. The set B is determined as follows. Removing an edge $e = \{u, v\}$ from G gives a graph $G - e$ that consists of two connected components $C_1 \ni u$ and $C_2 \ni v$. If the number of requests starting in C_2 and ending in C_1 exceeds the number of requests starting in C_1 and ending in C_2 by d, then add d copies of the request (u, v) to B. This operation is performed for all edges $e \in E$. After this procedure, which we call *balancing* in the sequel, the number of requests starting at any vertex v equals the number of requests ending at v. Moreover, every weakly connected component of (V, L) becomes a strongly connected component of $(V, L \cup B)$ [3, 21]. Therefore, the graph $(V, L \cup B)$ decomposes into several Eulerian components, and the remaining task is to connect these components at the least possible cost.

1.2 Results

Before we describe our results, we introduce the random model considered in this paper. Let $[n] = \{1, \ldots, n\}$.

Definition 2 *The* uniform model for the DARP, *is a list $L = L_{n,m}$ of requests $(i_k, j_k) \in [n] \times [n], k = 1, \ldots, m$. Each request (i_k, j_k) is chosen uniformly at random and independently of all others from $[n]^2$. This is obviously equivalent to choosing L from $[n]^{2m}$ uniformly at random.*

A list $L = L_{n,m}$ is interpreted as an instance of the DARP where $G = (V, E)$ is a caterpillar Cat_n, and L is the list of requests $\{(l_i, l_j) \mid (i, j) \in L\}$. Balancing the instance yields an additional set B of artificial requests. It is convenient to consider the directed multigraphs $D(L) = (V, L)$ and $D_B(L) = (V, L \cup B)$. To facilitate the analysis of the balancing operation, we give an equivalent formulation of the uniform model for the DARP in Section 2. We use the notion of connectedness in the digraphs $D(L)$ and $D_B(L)$ in a non-standard way: a component of $D(L)$ or $D_B(L)$ is a maximal connected subgraph *which contains at least one arc*. This extends to both weak and strong connectivity. The reason behind this concept is that for a solution of the problem isolated vertices need not be incorporated into the desired Euler tour.

The following lemma summarizes relevant statements about the components of $D(L)$ that follow from the theory of random graphs. See Section 4 for the nuts and bolts of the proof.

Lemma 3 *Let $L = L_{n,m}$ be chosen according to the uniform model.*

(i) If $m \geq 10n \ln n$, then $D(L)$ is weakly connected with probability $1 - o(1)$ as $m \to \infty$ as has no isolated vertices.

(ii) If $m \sim \gamma n$ for some fixed $\gamma > 0$, then the number of components of $D(L)$ that are directed cycles of length k has asymptotically Poisson distribution with parameter $\frac{1}{k}\left(\frac{\gamma}{\exp(2\gamma)}\right)^k$

(iii) If $m \ll n$ or $m \gg n$, then $D(L)$ has no Eulerian component with probability $1 - o(1)$ as $m \to \infty$.

If $D_B(L)$ has only one component, then any Euler tour in $D_B(L)$ already is an optimal solution. However, matters are not that simple in general: by Lemma 3, part (ii), $D(L)$ may contain several Eulerian components each of which remains a component of $D_B(L)$.

If $D_B(L)$ has more than one component then, as mentioned in the introduction, the DARP reduces to connecting the components at the least possible cost. The *MST-heuristic* for this task works as follows: first, the shortest distance of every pair of components is computed. According to the distances, a minimum spanning tree T on the components is determined. Finally, each edge of T connecting two components is replaced by a circuit of twice the edge length, connecting the same components. The MST-heuristic is a 2-approximation algorithm for the DARP on trees [14]. It can be shown that it is optimal in case of the DARP on paths [3].

The main result of this paper is the following theorem. The proof is sketched in Section 4.

Theorem 4 *Let $L = L_{n,m}$ be chosen according to the uniform model. The MST-heuristic finds an optimal solution with probability $1 - o(1)$ as $m \to \infty$. Moreover, optimality can be certified efficiently.*

The basis of this result is the following key technical lemma, which states essentially that it is unlikely for $D_B(L)$ to have more than one component, besides those that result from Eulerian components like in Lemma 3, part (ii):

Lemma 5 *Let $L = L_{n,m}$ be chosen according to the uniform model. With probability $1 - o(1)$ as $m \to \infty$, all non-Eulerian components of $D(L)$ are part of one single component of $D_B(L)$.*

Thus, in almost every case the balancing operation connects all components of $D(L)$ except the Eulerian ones. The proof of Lemma 5 is sketched in Section 3.

We complement our positive results about the solvability of the DARP in the average case by a hardness result in a semi-random setting, thereby improving upon the hardness results given in [13, 14, 21]. The following semi-random model for constructing instances of the DARP is inspired by a threshold result of Feige and Kilian on the complexity of the semi-random independent set problem [10]. First, a list $L = L_{n,m}$ of m requests is chosen according to the uniform model. Then, an adversary adds further requests, thereby producing a list $L' \supset L$. Note that the requests added by the adversary are *not* randomly chosen. We shall say that a polynomial time algorithm *solves the semirandom* (n, m)-DARP if with probability $1 - o(1)$ as $n \to \infty$ for any extension L' of the randomly chosen part $L = L_{n,m}$ on input L' the algorithm outputs an optimal solution of DARP; clearly, probability is taken over the choice of $L_{n,m}$. Obviously, if $m \gg n \ln n$, then the MST algorithm solves the semirandom (n, m)-DARP, because by Lemma 3 with high probability the graph $D(L)$ is connected and has no isolated vertices. Consequently $D(L') \supset D(L)$ is connected. We obtain the following theorem.

Theorem 6 *Let $L = L_{n,m}$ be chosen according to the uniform model. If $m \gg n \ln n$, then the MST-heuristic solves the semirandom (n, m)-DARP.*

Conversely, assume that the caterpillar Cat_n has uniform edge lengths. If $m \ll n \ln n$, then there is no polynomial time randomized algorithm that solves the semirandom (n, m)-DARP, unless $\mathsf{RP} = \mathsf{NP}$.

Note that the case $m = 0$ also gives a strong NP-hardness result for the plain worst case, as the edge lengths of the caterpillar are uniform. The proof of Theorem 6 is sketched in Section 5.

2 Random Walks and the Uniform Model

Given $L = L_{n,m}$, by $d_i(L)$ we denote the number of occurences of i in L. Clearly, $d_1(L) + \cdots + d_n(L) = 2m$. Now let $d_i \in \{0, 1, 2, 3, \ldots\}$ for $i = 1, \ldots, n$. Let $L_{n,m}(d_1, \ldots, d_n)$ denote the event that $d_i(L) = d_i$ for all i. In order to study the effect of the balancing operation on the directed multigraph $D(L_{n,m}(d_1, \ldots, d_n))$, we shall describe a simple random experiment that induces the same probability distribution as does the map

$$L_{n,m}(d_1, \ldots, d_n) \ni L \mapsto D(L). \tag{1}$$

Let

$$W_m = \{(x_1, \ldots, x_{2m}) | \sum_{i=1}^{2m} x_i = 0, x_i \in \{-1, +1\}\}$$

be the set of all ± 1-sequences of length $2m$ containing as many +1's as -1's. Note that the sequence $x_1, \ldots, x_{2m}$ is an unbiased Random Walk [11], conditional on $\sum_{i=1}^{2m} x_i = 0$. Then

$$\#W_m = \left(\begin{array}{c} 2m \\ m \end{array} \right).$$

Let $x = (x_1, \ldots, x_{2m}) \in W_m$ and $j \in \{1, \ldots, 2m\}$. We let

$$I_x(j) = \#\{i \leq j | \, x_i = x_j\},$$

that is, x_j is the $I_x(j)$th occurrence of the value of x_j.

Now let $\Omega = W_m \times \mathcal{S}_m$, where $\mathcal{S}_m$ is the symmetric group of order $m!$. We equip Ω with the uniform distribution. For each element $(x, \sigma) \in \Omega$, we construct a directed bipartite graph $H(x, \sigma)$ on the vertex set $\{a_1, \ldots, a_{2m}\}$ as follows. The arc (a_i, a_j) is present if and only if $x_i = 1$, $x_j = -1$, and $\sigma I_x(i) = I_x(j)$. Thus, the graph $H(x, \sigma)$ consists

of precisely m directed arcs. Finally, contracting the vertex sets

$$\{a_1, \ldots, a_{d_1}\}, \{a_{d_1+1}, \ldots, a_{d_1+d_2}\}, \\ \ldots, \{a_{d_1+\cdots+d_{n-1}+1}, \ldots, a_{2m}\}$$

gives a directed multigraph $D(x, \sigma)$ of order n (see Figure 2 for an illustration).

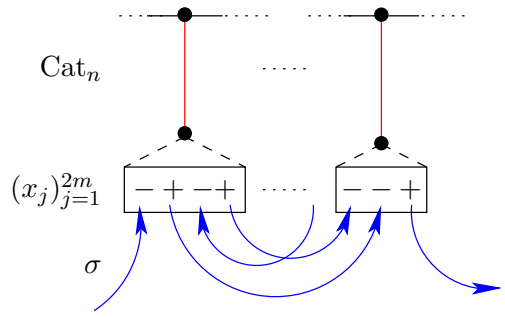

Figure 2: Illustration of the reformulated random model.

Lemma 7 *The distribution induced by (1) coincides with the distribution induced by the map* $\Omega \ni (x, \sigma) \mapsto D(x, \sigma)$. □

3 Non-Eulerian Components

Lemma 5 is equivalent to the statement that the artificial requests added by the balancing operation connect all non-Eulerian components of $D(L)$. Note that Lemma 5 vacuously follows from Lemma 3 if $m \geq 10n \ln n$. We therefore assume for the rest of the section that $m \ll n \ln n$.

To prove Lemma 5, the alternative formulation of the uniform model introduced in Section 2 turns out to be adequate. Let $d_1, \ldots, d_n$ be fixed. Then the set B of balancing requests depends only on the choice of $x \in W_m$. The first part of the proof is to bound the number of components of B. The second part is to show that with high probability the components of B and the non-Eulerian components of $D(L)$ glue together to form one large component. In fact, the probability that both parts hold turns out to be independent of the choice of $d_1, \ldots, d_n$. Thus, Lemma 5 follows from Lemma 7.

We begin collecting a bit of notation and some simple observations. The set B contains only requests along edges of $G = \mathrm{Cat}_n$. Let

$$E_B = \left\{ e \in E(G) : \begin{array}{l} \text{there exists a request} \\ \text{along } e \text{ in } B \end{array} \right\}.$$

Note that two components of $D(x, \sigma)$ are connected by B if and only if they are connected by E_B. Thus, the rest of analysis may be focused on the set E_B. Let us call a maximal set $S = \{b_i, b_{i+1}, \ldots, b_{i+l}\} \subset E_B$ of consecutive backbone edges in E_B a *backbone segment*.

Lemma 8 *With high probability, there are at most* $O(m^{3/4})$ *backbone segments.*

Sketch of proof. Let $i \in \{1, \ldots, n-1\}$. Observe that, unless

$$x_1 + \ldots + x_{d_1+\ldots+d_i} = x_{d_1+\ldots+d_i+1} + \ldots + x_{2m} = 0,$$

balancing yields a request along b_i and hence $b_i \in E_B$. Then, the number of gaps between backbone segments is bounded by the number Z_m of passages through zero of the random walk $x_1, \ldots, x_{2m}$, i.e., by the number of indices j where $x_1 + \cdots + x_j = 0$.

One can show that the expectation of z_m, divided by $\sqrt{m}$, converges to $\sqrt{\pi}$ as $m \to \infty$. Therefore, Lemma 8 follows by applying Markov's inequality. □

Consider the auxiliary directed bipartite graph $H(x, \sigma)$ from Section 2. A vertex a_j of H belongs to the vertex l_i of $D(x, \sigma)$, where $i = i(j)$ is chosen such that $d_1 + \ldots + d_{i-1} < j \leq d_1 + \ldots + d_i$.

It is called *active*, if $h_{i(j)} \in E_B$, and *inactive* otherwise. Observe that the active a_j are those that belong to vertices l_i of $D(x,\sigma)$ with different indegree and outdegree. As there are only requests between feet of G, $h_i \in E_B$ only if $b_i \in E_B$ or $b_{i-1} \in E_B$. Therefore, every hair h_i in E_B is incident to a backbone segment S, and thus each active a_j that belongs to l_i can be assigned to S. The purpose of this assignment is, that if $(a_j, a_{j'})$ is an edge of $H(x,\sigma)$ and both a_j and $a_{j'}$ are active, then the backbone segments that a_j and $a_{j'}$ are assigned to are connected through the hairs $h_{i(j)}$,$h_{i(j')}$ and the request $(l_{i(j)}, l_{i(j')})$. As every non-Eulerian component of $D(x,\sigma)$ contains a vertex with different indegree and outdegree, it is connected to a backbone segment. Thus, to complete the proof of Lemma 5 it suffices to show that all backbone segments are connected in the way just described. First we show that the number of active vertices of $H(x,\sigma)$ is large:

Lemma 9 *With probability* $1-o(1)$ *as* $m \to \infty$*, we have* $\#\{j \mid a_j \text{ is active}\} \geq m/2$.

Sketch of proof. The vertex a_j can only become inactive if $d_{i(j)}$ is even. The probability attains its maximum of $1/2 + o(1)$ if $d_{i(j)} = 2$. Therefore, the expected number of active vertices is at least $2m(1/2 - o(1))$. Based on the assumption that $m \ll n \ln n$, one can verify that the variance is dominated by the square of the expectation, and by Chebychev's inequality the lemma follows. □

Next, we reduce H to the active vertices in the following way: for each inactive a_j with $x_j = -1$ let τj be such that $x_{\tau j} = 1$, and $i(\tau j) = i(j)$. This is a perfect matching of the inactive vertices. If $(a_j, a_{j'})$ is an edge of H where a_j is active and $a_{j'}$ is inactive, then H contains another edge $(a_{\tau j'}, a_{j''})$. Replace these two edges with the edge $(a_j, a_{j''})$ and proceed until all edges are incident to active vertices. We call the resulting graph H'. Note that, by the construction, the backbone segments S and S' are connected through hairs in E_B and requests if $(a_j, a_{j'})$ is an edge of H', a_j is assigned to S, and $a_{j'}$ is assigned to S'.

For each backbone segment S let

$$m^+(S) = \{j \mid a_j \text{ assigned to } S \text{ and } x_j = +1\}$$

and

$$m^-(S) = \{j \mid a_j \text{ assigned to } S \text{ and } x_j = -1\}.$$

Clearly, for each backbone segment S, $\sum\{x_j \mid a_j \text{ assigned to } S\} = 0$ and hence $\#m^+(S) = \#m^-(S)$. As the choice of τ only depends on x, a simple counting argument shows that H' is a uniformly distributed matching of $m^+ = \bigcup_S m^+(S)$ with $m^- = \bigcup_S m^-(S)$.

Then the proof of the theorem is complete with the following

Lemma 10 *With probability* $1-o(1)$ *as* $m \to \infty$*, all backbone segments are in the same component of* $D_B(L)$.

Sketch of proof. Assume that there is a collection $S_1, \ldots, S_k$ of backbone components that are not in the same component of $D_B(L)$ as the others. We may assume that $l = \sum_{i=1}^{k} \#m^+(S_i) \leq m'/2$, where $m' \geq m/2$ is the number of active vertices, since otherwise we consider the collection of the remaining backbone segments. As $m^+(S) \geq 1$ for every backbone segment, $l \geq k$. The probability, taken over the distribution of H', that no edge of H' connects one of the S_i with a backbone component not in $S_1, \ldots, S_k$ is $\frac{l!(m'-l)!}{m'!} \leq 1/\binom{m'}{k}$. Thus, the expected number of such collections is at most $\sum_k \binom{m^{3/4}}{k}/\binom{m'}{k}$. As $\binom{m^{3/4}}{k}/\binom{m'}{k} = O((em^{-1/4})^k)$ the lemma follows. □

4 Eulerian Components

In this section, we first sketch the proof of Lemma 3. Then we estimate the number of vertices on Eulerian

components of $D(L_{n,m})$. Both results rely on results on the global structure of the random graph $G_{n,m}$; see [6] for a detailed exposition. Finally we sketch the proof of Theorem 4.

Given $L = L_{n,m}$, we obtain a simple graph $S(L)$ on $\{1, \ldots, n\}$ that consists of all edges $\{v, w\}$, $v \neq w$, such that $(v, w) \in L$ (or $(w, v) \in L$). Note that the connected components of $S(L)$ are in one-to-one correspondence with the connected components of the directed multigraph $D(L)$. Since the expected number of loops in $D(L)$ is m/n, and the expected number of multiple edges is $\leq m^3/n^4$, the number of edges of $D(L)$ is at least $m/2$, with high probability.

Suppose that $m \geq 10n \ln n$. Then with high probability $S(L)$ has at least $5n \ln n$ edges. Hence the first part of Lemma 3 follows from the fact that with high probability the random graph $G_{n,5n \ln n}$ is connected.

As for the proof of the second part of Lemma 3, denote by X_k the number of connected components of $D(L)$ that are directed k-cycles. Then a straightforward computation yields

$$E(X_k) \sim \frac{1}{k} \cdot \left(\frac{\gamma}{\exp(2\gamma)} \right)^k,$$

where $\gamma = m/n$. Moreover, for the rth factorial moment of X_k we have

$$\frac{E_r(X_k)}{E(X_k)^r} \sim 1.$$

Thus, [6, p. 25] entails that the distribution of X_k is asymptotically Poisson. Finally, a somewhat tedious computation proves part (iii).

Lemma 11 *Let $L = L_{n,m}$ be chosen according to the uniform model. Then with probability $1 - o(1)$ as $m \to \infty$ the number of vertices on Eulerian components of $D(L)$ is at most $m^{1/8}$.*

Sketch of proof. If $m \geq 10n \ln n$, then with high probability $D(L)$ is connected, by Lemma 3. Consequently, with high probability there are no Eulerian components at all. Now let us assume that $3n/4 \leq m \leq 10n \ln n$. Then results on the global structure of the random graph imply that with high probability $S(L)$ has no component of order at least $100 \ln n$ and at most $n^{2/3}$. Moreover, there is precisely one component of order $\geq n^{2/3}$, the so-called giant component. A simple counting argument proves that With high probability the component of $D(L)$ corresponding to the giant component of $S(L)$ is not Eulerian.

A lengthy computation shows that in the case $m \geq 3n/4$ the graph $D(L)$ has no Eulerian component of order at most $n^{1/4}$ that contains more edges than vertices. In addition, the number of vertices on components of $D(L)$ that are directed cycles is $O(1)$. Finally, in the case $m \leq 3n/4$ the expected number of vertices on Eulerian components is $O(1)$ with high probability. Thus, applying the Markov inequality completes the proof of Lemma 11. □

As for the proof of Theorem 4, note that by Lemma 5 with high probability there is only one component C_B in $D_B(L)$ in addition to the Eulerian components of $D(L)$. We may assume that $\sqrt{n} \leq m \leq 10n \ln n$, as otherwise, by Lemma 3, $D(L)$ has no Eulerian components. Hence by Lemma 11, the number of vertices in the Eulerian components is at most $m^{1/8}$. Thus with high probability $D_B(L)$ has the

Property 12 *Between any two feet l_i and l_j, $i < j$, that belong to Eulerian components of $D(L)$, there is a foot l_k, $i < k < j$, that belongs to C_B.*

As a consequence the distance graph on the components corresponds to a star metric where the center is C_B. Therefore, the MST heuristic finds an optimal tour. Observe that Property 12 can be checked in polynomial time. Hence it provides the desired certificate for the optimality of the solution produced by

the MST-heuristic, thereby proving Theorem 4.

5 A Hardness Result

As for the proof of Theorem 6, note that in the case $m \gg n \ln n$ the graph $D(L_{n,m})$ is connected with high probability. Consequently, the MST-heuristic finds an optimal solution of the semirandom problem.

Now suppose $m \ll n \ln n$. Consider an instance (S, T, E) of the bipartite Steiner tree problem, where S denotes the set of Steiner vertices, T the set of terminals, and E is the edge set. The bipartite Steiner tree problem is NP-hard even in the case $d(s) = 4$ for all $s \in S$ [1, 5]. We shall prove that a polynomial time algorithm that solves the semirandom dial-a-ride-problem optimally yields a randomized algorithm for the bipartite Steiner tree problem, implying that $NP = RP$.

Let $L = L_{n,m}$. We shall show how the adversary can include the instance (S, T, E) of the Steiner tree problem into the graph $D(L)$ such that an optimal solution of the dial-a-ride-problem gives an optimal Steiner tree. With high probability there are at least $n^{23/24}$ vertices in $\{l_1, \ldots, l_n\}$ that are not incident with any edge in $D(L)$. Partition the set $\{l_1, \ldots, l_n\}$ into $\sqrt{n}$ pieces

$$\{l_1, \ldots, l_{\sqrt{n}}\}, \{l_{\sqrt{n}+1}, \ldots, l_{2\sqrt{n}}\}, \ldots, \{l_{n-\sqrt{n}+1}, \ldots, l_n\}$$

With high probability there are at least $N = n^{1/8}$ pieces, which we denote by $B_1, \ldots, B_N$, starting with six vertices not incident with arcs in $D(L)$ each. Let I_j denote the set of the first six vertices of B_j, $j = 1, \ldots, N$ and let I be the union of the sets I_j. We may assume that $\#S = N/(4!)$.

First, for each request (u_i, v_i) in L, $i = 1, \ldots, m$, the adversary adds a request (v_i, u_i). Then, for each vertex $v \in \{l_1, \ldots, l_n\} \setminus I$ the adversary adds the requests (v, l_n), (l_n, v). Let L' denote the resulting list of requests. Then $D(L')$ has one large Eulerian component on the vertex set $\{l_1, \ldots, l_n\} \setminus I$.

For each Steiner vertex s and each permutation $\sigma \in \mathcal{S}_4$ the adversary picks a set $I_\sigma(s) \in \{I_1, \ldots, I_N\}$ such that each I_j is used precisely once. Assume that $I_\sigma(s)$ consists of the vertices $v_6, \ldots, v_1$, from left to right and let t_1, t_2, t_3, t_4 be the neighbors of s. The adversary labels v_2 with $t_{\sigma(1)}$, v_3 with $t_{\sigma(2)}$, v_4 with $t_{\sigma(3)}$, and v_5 with $t_{\sigma(4)}$. The vertices v_1 and v_6 are not labeled. Finally, for each $t \in T$ the adversary adds to L' a directed cycle connecting all vertices that are labeled with t. Let L'' denote the resulting list of requests.

In summary, the DARP instance constructed by the adversary consists of disjoint cycles, one for each terminal, and one giant component C_r containing all randomly chosen requests. Every Steiner vertex is represented by $4!$ gadgets, each consisting of six hairs where the first and the last foot of each gadget are isolated; each of the four feet in the middle lies on a cycle corresponding to the terminal the foot is labeled with.

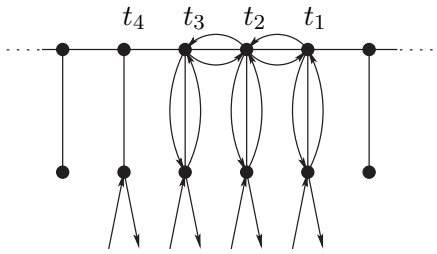

Figure 3: In M the Steiner vertex s is connected with the terminals t_1, t_2, and t_3.

Let M be a Steiner tree in (S, T, E) of cost c_M. The following tour in the DARP instance corresponding to $D(L'')$ has cost $2c_M + 2\#T + 4$ plus the total length of the requests in L'': For every $s \in S$

with $d_M(s) = k > 0$ let the neighbors of s in M be $t_1, \ldots, t_k$. Let $\sigma \in \mathcal{S}_4$ be such that v_{i+1} is labeled with t_i, $i = 1, \ldots, k$, in $I_\sigma(s) = \{v_1, \ldots, v_6\}$. Connect $v_2, \ldots, v_{k+1}$ with a total of $4k - 2$ requests of length 1 as indicated in Figure 3. Then the cycles corresponding to the terminals are connected since the terminals are connected by M. Finally, connect C_r with the cycles using 6 requests of length 1. The total length of the added requests is

$$\begin{aligned} & 6 + \sum_{\substack{s \in S \\ d_M(s) > 0}} 4d_M(s) - 2 \\ &= 6 + 4c_M - 2\#\{s \in S \mid d_M(s) > 0\} \\ &= 6 + 4c_M - 2c_M + 2\#T - 2 \\ &= 2c_M + 2\#T + 4, \end{aligned}$$

where we have made use of the fact that

$$c_M = \#T + \#\{s \in S \mid d_M(s) > 0\} - 1.$$

Conversely, given a solution of the DARP which is by c longer than the total length of the requests in L'', one can compute in polynomial time a Steiner tree in (S, T, E) of length at most $c/2 - 2 - \#T$. This Steiner tree can be found changing the given solution to a solution of at most the same length that results from the desired Steiner tree by applying the procedure described in the previous paragraph.

6 Conclusions and Open Problems

We have shown that the DARP on caterpillars, while NP-hard in the worst-case, is solvable efficiently on average. In view of the online setting described in the introduction, this can be regarded as a first step towards an investigation of the *online* dial-a-ride problem in the average case.

On the other hand, the (offline-) DARP is interesting in its own right. In this respect it would be interesting to investigate whether our methods carry over to more general transportation networks such as arbitrary trees.

Another potential extension concerns the distribution of the requests. This distribution might be biased according to given weights $0 \leq p_1, \ldots, p_n$, where $\sum_i p_i = 1$. Then the probability that a random request is (l_i, l_j) equals $p_i p_j$. Note that the uniform distribution corresponds to the case $p_i \equiv 1/n$. We expect that our methods extend to this biased random model, though the calculations seem to become considerably more complicated.

References

[1] P. ALIMONTI AND V. KANN, *Hardness of approximating problems on cubic graphs*, in Proceedings of the 3rd Italian Conference on Algorithms and Complexity, vol. 1203 of Lecture Notes in Computer Science, Springer, 1997, pp. 288–298.

[2] N. ASCHEUER, S. O. KRUMKE, AND J. RAMBAU, *Online dial-a-ride problems: Minimizing the completion time*, in Proceedings of the 17th International Symposium on Theoretical Aspects of Computer Science, vol. 1770 of Lecture Notes in Computer Science, Springer, 2000, pp. 639–650.

[3] M. J. ATALLAH AND S. R. KOSARAJU, *Efficient solutions to some transportation problems with applications to minimizing robot arm travel*, SIAM Journal on Computing, 17 (1988), pp. 849–869.

[4] G. AUSIELLO, E. FEUERSTEIN, S. LEONARDI, L. STOUGIE, AND M. TALAMO, *Algorithms for the on-line traveling salesman*, Algorithmica, 29 (2001), pp. 560–581.

[5] M. BERN AND P. PLASSMANN, *The Steiner problem with edge lengths 1 and 2*, Information Processing Letters, 32 (1989), pp. 171–176.

[6] B. BOLLOBÁS, *Random graphs*, Cambridge University Press, Cambridge, UK, 2 ed., 2001.

[7] A. BORODIN AND R. EL-YANIV, *Online Computation and Competitive Analysis*, Cambridge University Press, 1998.

[8] E. G. COFFMAN, C. A. COURCOUBETIS, M. R. GAREY, D. S. JOHNSON, L. A. MCGEOCH, P. W. SHOR, R. R. WEBER, AND M. YANNAKAKIS, *Fundamental discrepancies between average-case analyses under discrete and continuous distributions - a bin packing case study*, in Proceedings of the 23th Annual ACM Symposium on the Theory of Computing, 1991, pp. 230–240.

[9] E. G. COFFMAN AND G. S. LUEKER, *Probabilistic analysis of Packing and Partitioning Algorithms*, John Wiley, New York, 1991.

[10] U. FEIGE AND J. KILIAN, *Heuristics for semirandom graph problems*, Journal of Computer and System Sciences, 63 (2001), pp. 639–671.

[11] W. FELLER, *An Introduction to Probability Theory and Its Applications*, vol. 1, John Wiley & Sons, Inc., 3 ed., 1968.

[12] E. FEUERSTEIN AND L. STOUGIE, *On-line single server dial-a-ride problems*, Theoretical Computer Science, (2001). To appear.

[13] G. N. FREDERICKSON, *A note on the complexity of a simple transportation problem*, SIAM Journal on Computing, 22 (1993), pp. 57–61.

[14] G. N. FREDERICKSON AND D. J. GUAN, *Nonpreemptive ensemble motion planning on a tree*, Journal of Algorithms, 15 (1993), pp. 29–60.

[15] G. N. FREDERICKSON, M. S. HECHT, AND C. E. KIM, *Approximation algorithms for some routing problems*, SIAM Journal on Computing, 7 (1978), pp. 178–193.

[16] A. FRIEZE AND C. MCDIARMID, *Algorithmic theory of random graphs*, Random Structures and Algorithms, 10 (1997), pp. 5–42.

[17] G. R. GRIMMETT AND D. R. STIRZAKER, *Probability and Random Processes*, Oxford University Press, New York, 1992.

[18] M. GRÖTSCHEL, S. O. KRUMKE, AND J. RAMBAU, eds., *Online Optimization of Large Scale Systems*, Springer, Berlin Heidelberg New York, 2001.

[19] D. HAUPTMEIER, S. O. KRUMKE, AND J. RAMBAU, *The online dial-a-ride problem under reasonable load*, in Proceedings of the 4th Italian Conference on Algorithms and Complexity, vol. 1767 of Lecture Notes in Computer Science, Springer, 2000, pp. 125–136.

[20] ——, *The online dial-a-ride problem under reasonable load*, Theoretical Computer Science, (2002). A preliminary version appeared in the Proceedings of the 4th Italian Conference on Algorithms and Complexity, 2000, vol. 1767 of Lecture Notes in Computer Science.

[21] D. HAUPTMEIER, S. O. KRUMKE, J. RAMBAU, AND H.-C. WIRTH, *Euler is standing in line*, Discrete Applied Mathematics, 113 (2001), pp. 87–107. A preliminary version appeared in the Proceedings of the 25th International Workshop on Graph-Theoretic Concepts in Computer Science, 2000, vol. 1665 of Lecture Notes in Computer Science.

Appendix

A Proof of Lemma 7

Fix $d = (d_1, \ldots, d_n)$. Let $D_{n,m}$ denote the set of all directed multigraphs on the vertex set $\{1, \ldots, n\}$ with precisely m edges. Further, let $D_{n,m}(d_1, \ldots, d_n)$ denote the set of all $D \in D_{n.m}$ such that the vertex i is incident with precisely d_i arcs for all i (a loop contributes 2 incidences). Let $\varphi : L_{n,m}(d) \to D_{n,m}(d)$ denote the map (1). Finally, let $\psi : W_m \times \mathcal{S}_m \to D_{n,m}(d)$ denote the map $(x, \sigma) \mapsto D(x, \sigma)$. Then we are to prove that the distributions P_φ and P_ψ coincide, where both $L_{n,m}(d)$ and $W_m \times \mathcal{S}_m$ are equipped with the uniform distribution.

Thus, let $G \in D_{n,m}(d)$. We have to show that

$$\frac{\#\psi^{-1}(G)}{\#W_m \times \mathcal{S}_m} = \frac{\#\varphi^{-1}(G)}{\#L_{n,m}(d)}.$$

Suppose that G has precisely v_i arcs of multiplicity i, $i = 1, 2, \ldots$, and that v_0 is the number of loops of G. Indeed, let

$$E_i = \{e_1^{(i)}, \ldots, e_{v_i}^{(i)}\}$$

be the set of all arcs of multiplicity i. Then we can count the inverse images of G under the map φ as follows:

- Choose one of the m positions in the list m for $e_1^{(1)}$.
- Choose one of the remaining $m - 1$ positions for $e_2^{(1)}$.
- $\cdots$
- Choose a set of i remaining positions for $e_j^{(i)}$.
- $\cdots$

Thus,

$$\#\varphi^{-1}(G) = \prod_{i=1}^{\infty} \prod_{j=0}^{v_i - 1} \binom{m - ij - \sum_{k=1}^{i-1} k v_k}{i} = \frac{m!}{\prod_{i=1}^{\infty} i!^{v_i}}$$

In order to determine $\#\psi^{-1}(G)$, let $A(i)$ denote the set of all arcs that are incident with vertex i. Let us first count the number of maps

$$\sigma_i : \{1, \ldots, d_i\} \to E(i)$$

such that $\#\sigma_i^{-1}(e)$ equals the multiplicity $V(e)$ of e for all $e \in E(i)$. Let $E(i) = \{e_1, \ldots, e_{d_1}\}$. Obviously, the number of such maps is

$$\prod_{j=1}^{d_i} \binom{d_i - \sum_{k=1}^{j-1} V(e_k)}{V(e_j)} = \frac{d_i!}{\prod_{j=1}^{d_i} V(e_j)!}$$

Therefore, the number of tuples $\sigma = (\sigma_1, \ldots, \sigma_n)$ is

$$\frac{d_1! \cdots d_n!}{\prod_{e \in E(1)} V(e)! \cdots \prod_{e \in E(n)} V(e)!} = \frac{d_1! \cdots d_n!}{\prod_{i=1}^{\infty} i!^{2v_i}}.$$

Note that each tuple $\sigma = (\sigma_1, \ldots, \sigma_n)$ gives precisely $\prod_{i=1}^{\infty} i!^{v_i}$ inverse images of G under the map ψ, because for each arc of multiplicity i from vertex x to vertex y there are precisely $i!$ ways to map the corresponding $+1$s to the corresponding -1s. Moreover, each element of $\psi^{-1}(G)$ is counted precisely once.

Finally, observe that

$$\#W_m \times \mathcal{S}_m = \left(\begin{array}{c} 2m \\ m \end{array} \right) m! = \frac{(2m)!}{m!}$$

and that

$$\#L_{n,m}(d) = \left(\begin{array}{c} 2m \\ d_1 \end{array} \right) \left(\begin{array}{c} 2m - d1 \\ d_2 \end{array} \right) \cdots \left(\begin{array}{c} 2m - d_1 - \cdots - d_{n-1} \\ d_n \end{array} \right) = \frac{(2m)!}{d_1! \cdots d_n!}.$$

We conclude that

$$\frac{\#\psi^{-1}(G)}{\#W_m \times \mathcal{S}_m} = \frac{d_1! \cdots d_n! m!}{(2m)!} = \frac{\#\varphi^{-1}(G)}{\#L_{n,m}(G)},$$

thereby proving the lemma.

B Proofs for Section 3

In this section we give full proofs of Lemma 8 and of Lemma 9. Let $s_j := x_1 + \ldots + x_j$ and $D_i := d_1 + \ldots + d_i$.

Lemma 8 The missing detail in the sketch of the proof of Lemma 8 is the statement that $E[z_m] = \sqrt{\pi m}$. This can be deduced from the so-called *arc sine law* for the unbiased random walk. Let $X_1, X_2, \ldots$ be i.i.d. with $\mathrm{P}(X_i = 1) = \mathrm{P}(X_i = -1) = 1/2$ and let $S_i = \sum_{j \le i} X_j$. Define T_{2m} to be the largest index $2i \le 2m$ with $S_{2i} = 0$. The *arc sine law for last visit to origin* [17, p.80] states that

$$\mathrm{P}(T_{2m} = 2i) = \mathrm{P}(S_{2i} = 0)\, \mathrm{P}(S_{2m-2i} = 0).$$

Note that $\mathrm{P}(x_1 + \ldots + x_{2j} = 0) = \mathrm{P}(S_{2j} = 0 \mid S_{2m} = 0)$ and that

$$\mathrm{P}(S_{2m} = 0 \mid S_{2j} = 0) = \mathrm{P}(S_{2m-2j} = 0).$$

Therefore

$$\begin{aligned} E[z_m] &= \sum_{j\le m} \mathrm{P}\left(x_1 + \ldots + x_{2j} = 0\right) \\ &= \sum_{j\le m} \mathrm{P}\left(S_{2j} = 0 \mid S_{2m} = 0\right) \\ &= \sum_{j\le m} \frac{\mathrm{P}\left(S_{2j}=0\right)}{\mathrm{P}\left(S_{2m}=0\right)} \mathrm{P}\left(S_{2m} = 0 \mid S_{2j} = 0\right) \\ &= \sum_{j\le m} \frac{\mathrm{P}\left(S_{2j}=0\right)}{\mathrm{P}\left(S_{2m}=0\right)} \mathrm{P}\left(S_{2m-2j} = 0\right) \\ &= \frac{1}{\mathrm{P}\left(S_{2m}=0\right)} \sum_{j\le m} \mathrm{P}\left(T_{2m} = 2j\right) \\ &= \frac{1}{\mathrm{P}\left(S_{2m}=0\right)}, \end{aligned}$$

where $\mathrm{P}\left(S_{2m}=0\right) \sim 1/\sqrt{\pi m}$.

Lemma 9 The following lemma implies that, by the assumption that $m \ll n \ln n$, we can assume that $\max_i d_i \le n^{1/2}$.

Lemma 13 *Suppose that $m \le n^{5/4}$. Then with high probability there is no vertex of degree $\ge n^{1/2}$ in $D(L_{n,m})$.*

Proof. The probability that a fixed vertex v has degree d is

$$\left(\begin{array}{c} m \\ d \end{array}\right) (2n-1)^d (n-1)^{2(m-d)} n^{-2m} \le \left(\frac{2em}{dn}\right)^d.$$

Consequently, the expected number of vertices of degree $\ge d$ is

$$\le n \sum_{j=d}^{\infty} \left(\frac{2em}{jn}\right)^j \le n \left(\frac{2em}{dn}\right)^d \frac{dn}{dn-2em}.$$

By our assumption $m \le n^{5/4}$, in the case $d = n^{1/2}$ the right hand side is $o(1)$, whence with high probability there are no vertices of degree $\ge n^{1/2}$. □

Assume that $d_i = 2k$. A vertex a_j in $H(x,\sigma)$ where $i(j) = i$ is inactive iff the indegree of l_i in $D(x,\sigma)$ equals its outdegree, i.e. $s_{D_i} - s_{D_{i-1}} = 0$. The probability of this event is the same as the probability that $s_{d_i} = 0$, which is

$$f(k) := \frac{\binom{2k}{k}\binom{2(m-k)}{m-k}}{\binom{2m}{m}}$$

To bound this probability we may assume by symmetry that $k \leq m/2$. We first observe that

$$\frac{f(k+1)}{f(k)} = \frac{2j+1}{j+1}\frac{m-j}{2m-2j-1} = (2-\frac{1}{j+1})(\frac{1}{2}+\frac{1}{4m-4j-2}) \leq 1-\frac{1}{2j+2}+\frac{1}{2m-2j-1}.$$

The last term is less than 1 if $k \leq m/2-1$, thus $f(k)$ is maximal for $k=1$:

$$f(k) \leq f(1) = \frac{\binom{2}{1}\binom{2(m-1)}{m-1}}{\binom{2m}{m}} = \frac{m}{2m-1} = 1/2+o(1). \tag{2}$$

Let the random variable Y be the number of active vertices and $\bar{Y} = 2m - Y$ be the number of inactive vertices. Denote by v_i the event that $s_{D_i} - s_{D_{i-1}} \neq 0$. Then, by the previous computation, $\mathrm{E}(Y) = \sum_{i=1}^{n} \mathrm{P}(v_i)\, d_i \geq m - O(1)$. Therefore, by Chebychev's inequality, $\mathrm{P}(Y \leq m/2) \leq (4+o(1))\mathrm{Var}(Y)/\mathrm{E}(Y)^2$. To bound this further, let

$$\mathrm{Var}(Y) = \mathrm{Var}\left(\bar{Y}\right) = \mathrm{E}\left(\bar{Y}^2\right) - \mathrm{E}\left(\bar{Y}\right)^2 = A + B,$$

where

$$A = \sum_{i \neq j} d_i d_j (\mathrm{P}(\bar{v_i} \wedge \bar{v_j}) - \mathrm{P}(\bar{v_i})\,\mathrm{P}(\bar{v_j}))$$

and

$$B = \sum_{i \leq n} d_i^2 (\mathrm{P}(\bar{v_i}) - \mathrm{P}(\bar{v_i})^2).$$

We compute

$$\begin{aligned} A/\mathrm{E}\left(\bar{Y}\right)^2 &\leq \frac{\sum_{i\neq j} d_i d_j \binom{2d_i}{d_i}\binom{2d_j}{d_j}\binom{2m}{m}^{-2}\left[\binom{2m}{m}\binom{2(m-d_i-d_j)}{m-d_i-d_j} - \binom{2(m-d_i)}{m-d_i}\binom{2(m-d_j)}{m-d_j}\right]}{\sum_{i\neq j} d_i d_j \binom{2d_i}{d_i}\binom{2d_j}{d_j}\binom{2m}{m}^{-2}\binom{2(m-d_i)}{m-d_i}\binom{2(m-d_j)}{m-d_j}} \\ &\leq \max_{i\neq j} \frac{\binom{2m}{m}\binom{2(m-d_i-d_j)}{m-d_i-d_j}}{\binom{2(m-d_i)}{m-d_i}\binom{2(m-d_j)}{m-d_j}} - 1 \\ &= O\left(\frac{1/\sqrt{\pi^2 m(m-d_i-d_j)}}{1/\sqrt{\pi^2(m-d_i)(m-d_j)}} - 1\right), \end{aligned}$$

where the last equality follows from Stirling's formula. By the assumption that $\max_{i\leq n} d_i \leq m^{1/2}$, we get that

$$\max_{i\neq j}\sqrt{\frac{(m-d_i)(m-d_j)}{m(m-d_i-d_j)}} - 1 = \max_{i\neq j}\sqrt{1+\frac{d_i d_j}{m(m-d_i-d_j)}} - 1 \leq \sqrt{1+\frac{2}{m}} - 1 \leq 1/m$$

and, with (2), $A/\mathrm{E}\,(Y)^2 \le A/\mathrm{E}\left(\bar{Y}\right)^2 = O(1/m)$. For B, we simply note that

$$B/\mathrm{E}\,(Y)^2 = \frac{\sum_{i\le n} d_i^2 \mathrm{P}\,(v_i)\,\mathrm{P}\,(\bar{v}_i)}{\sum_{i\le n} d_i \mathrm{P}\,(v_i)\,\mathrm{E}\,(Y)} \le \max_{i\le n} \frac{d_i \mathrm{P}\,(\bar{v}_i)}{\mathrm{E}\,(Y)} \le 1/\sqrt{m}.$$

We conclude that $\mathrm{Var}\,(Y)\,/\mathrm{E}\,(Y)^2 = O(1/\sqrt{m})$ and thus $Y \ge m/2$ with high probability.

C Proofs for Section 4

In this section, we shall prove Lemma 3 and the assertions made in section 4. Though the proofs turn out to be quite technical and rather lengthy, we give all arguments in full detail. The first part of Lemma 3 follows from Lemma 16. The second part of Lemma 3 is Lemma 28 below. The third part follows from the proofs of Lemma 23 and Lemma 25. Lemma 11 summarizes the results of this section up to Lemma 22 (the case of so-called complex components) and Lemmas 23, 24, and 25 (the case of directed cycles). The fact that Property 13 is valid almost surely is a consequence of Corollary 27.

Let $L = L_{n,m}$. We call $I \subset \{1,\dots,m\}$ a *k-fold edge* if $\#I = k$ and the following conditions hold.

(i) For all $i,j \in I$ the ith entry (x_i, y_i) and the jth entry (x_j, y_j) of L coincide up to the direction, i.e. $\{x_i, y_i\} = \{x_j, y_j\}$.

(ii) There is no proper superset $J \supset I$ that satisfies 1.

Let $v_k(L)$ denote the number of k-fold edges of L.

A *loop* of L is an index i such that for the ith entry (x_i, y_i) of L we have $x_i = y_i$. By $v_0(L)$ we denote the number of loops of L. Furthermore, we put

$$v(L) = (v_0(L), v_1(L), v_2(L), \dots).$$

Conversely, if $v = (v_k)_{k=0,1,2,\dots}$, then L_v denotes the set of all $L \in L_{n,m}$ such that $v(L) = v$. Put

$$m(v) = \sum_{k=1}^{\infty} v(k).$$

Lemma 14 *Given a sequence v, the map $S|L_v : L_v \to G_{n,m(v)}$ maps the uniform distribution on L_v onto the uniform distribution on the space $G_{n,m(v)}$ of all simple graphs with n vertices and $m(v)$ edges.*

Proof. First observe that the map

$$S_1 : \begin{pmatrix} (a_1,b_1) \\ \vdots \\ (a_m,b_m) \end{pmatrix} \mapsto \begin{pmatrix} \{a_1,b_1\} \\ \vdots \\ \{a_m,b_m\} \end{pmatrix}$$

maps the uniform distribution on L_v onto the uniform distribution on its image $S_1(L_v)$ For any element of $S_1(L_v)$ has precisely 2^{m-v_0} inverse images.

Further, we claim that the map

$$S_2 : \begin{pmatrix} e_1 \\ \vdots \\ e_m \end{pmatrix} \mapsto \{e_1, \ldots, e_m\}$$

maps the uniform distribution on $S_1(L_v)$ onto the uniform distribution on $G_{n,m(v)}$. First observe that S_2 is well-defined, because $\#S_2(S_1(L)) = m(v)$ for all $L \in L_v$. Now let E be a set of $m(v)$ edges. Then there are $m(v)!$ tuples $(e_1, \ldots, e_{m(v)})$ such that $E = \{e_1, \ldots, e_{m(v)}\}$. Each such tuple gives an element of $S_2^{-1}(E)$ by writing down one copy of each of the first v_1 entries of $(e_1, \ldots, e_{m(v)})$, 2 copies of each of the following $v(2)$ entries, and so on. Finally, insert loops into the remaining s places. Thus, each tuple gives rise to

$$n^s \prod_{j=1}^{\infty} \prod_{i=0}^{v_j - 1} \begin{pmatrix} m - ij - \sum_{r=1}^{j-1} rv(r) \\ j \end{pmatrix} = \frac{n^s (m)_{m-s}}{\prod_{j=1}^{\infty} j!^{v_j}}$$

inverse images in the above manner. Since two tuples $(e_1, \ldots, e_{m(v)})$ and $(e'_1, \ldots, e'_{m(v)})$ give rise to the same inverse images if and only if the tuples coincide up to the order in that the edges that become k-fold edges occur, each set E has precisely

$$\frac{n^s (m)_{m-s} m(v)!}{\prod_{j=1}^{\infty} v(j)! j!^{v_j}} \tag{3}$$

inverse images. Because the quantity (3) does not depend on the particular choice of E, we have shown that S_2 maps the uniform distribution onto the uniform distribution. Finally, observe that the map $S|L_v$ is simply the composite of S_1 and S_2. □

Lemma 15 *The expected number of loops in $L_{n.m}$ is m/n. The expected number of k-fold edges, $k \geq 3$, is $\leq m^3/n^4$, provided $m \ll n^2$.*

Proof. The probability that the ith entry of $L = L_{n,m}$ is a loop is

$$\frac{n n^{2(m-1)}}{n^{2m}} = 1/n.$$

Thus, the expected number of loops is m/n.

Let $I \subset \{1, \ldots, m\}$ be a set of cardinality k. Then the probability that I is a k-fold edges is

$$\frac{n^2 n^{2(m-k)}}{n^{2m}} = n^{-2(k-1)}.$$

Thus, the expected number of k-fold edges is

$$\begin{pmatrix} m \\ k \end{pmatrix} n^{-2(k-1)} \leq \frac{m^k}{k! n^{2(k-1)}}.$$

Thus, the expected number of k-fold edges, $k \geq 3$, is

$$\leq \sum_{k=3}^{\infty} \frac{m^k}{k! n^{2(k-1)}} \leq \frac{m^3}{6n^4} \sum_{k=0}^{\infty} \frac{m^k}{n^{2k}} \leq \frac{m^3}{n^4}.$$

□

Lemma 16 *Suppose that $m \geq 10n \ln n$. Then with high probability the graph $D(L_{n,m})$ is connected.*

Proof. Let us first assume that $m = 10n \ln n$. Then, by the lemma before, with high probability the number of loops in $L_{n,m}$ is at most $m/100$. Moreover, with high probability $L_{n,m}$ has no ≥ 3-fold edges. The expected number of double edges is at most $m^2/(2n^2) \leq m/100$. Thus, with high probability we have

$$m(v(L_{n,m})) \geq m/2 \geq 5n \ln n.$$

Consequently, $S(L_{n,m})$ is connected with high probability, whence $D(L_{n,m})$ is connected with high probability.

If $m \geq 10n \ln n$, then decompose $L_{n,m}$ into pieces of size $10n \ln n$. With high probability at least one of these pieces connects all vertices of $D(L_{n,m})$. □

Lemma 17 *There exists a function $f(n) = o(1)$ such that in the case $m \geq 3n/4$ with probability $\geq 1 - f(n)$ the graph $D(L_{n,m})$ has no (weak) component of order $> 100 \ln n$ and $< n^{2/3}$.*

Proof. By the pervious lemma, we may assume that $m \leq 10n \ln n$. Then almost all $L = L_{n,m}$ satisfy $m(v(L)) \geq 5n/8$. Thus, the assertion follows from [6, p. 137]. □

Lemma 18 *There exists a function $f(n) = o(1)$ such that for all $m \geq 3n/4$ with probability $\geq 1 - f(n)$ the graph $D(L_{n,m})$ has precisely one component of order $\geq n^{2/3}$ (the so called "giant component").*

Proof. Again, we may assume that $m \leq 10n \ln n$. Then the assertion follows from [6, p. 142]. □

Lemma 19 *Suppose that $3n/4 \leq m \leq 10n \ln n$. There exists a function $f(n) = o(1)$ such that with probability $1 - f(n)$ the giant component of $D(L_{n,m})$ is not Eulerian.*

Proof. Let C be the giant component of $D(L)$, $L = L_{n,m}$. Then C is of order $\geq n^{2/3}$ almost surely. With high probability, the number of edges of multiplicity > 1 in $D(L_{n,m})$ is $\leq n^{1/2}$. Consequently, C has $\Omega(n^{2/3})$ edge of multiplicity 1. Assume that C is Eulerian. Then changing the direction of precisely one edge of multiplicity 1 in L that is mapped into C gives a new list $L' \in L_{n,m}$ such that the giant component of L' is not Eulerian. Conversely, given L', it is obvious how to recover L. Therefore, each L with Eulerian giant component C gives $\Omega(n^{2/3})$ elements of $L_{n,m}$ with non-Eulerian giant components. □

Lemma 20 *Suppose that $m \geq n$. Then with high probability the graph $D(L_{n,m})$ has no Eulerian component of order $\leq n^{1/4}$ that contains more arcs than vertices.*

Proof. If $m \geq 10n\ln n$, then with high probability $D(L_{n,m})$ is connected. Thus, let us assume that $m < 10n\ln n$. Put $\gamma = m/n$. Given $l \geq 1$, we can bound the expected number of Eulerian components of order $k \leq n^{1/4}$ and size $k+l$ as follows:

$$\begin{aligned}
&\sum_{k=1}^{n^{1/4}} \binom{n}{k} (m)_{k+l} k^{k+l} (n-k)^{2(m-l-l)} n^{-2m} \\
\leq\ & (2\pi)^{-1/2} \sum_{k=1}^{n^{1/4}} \left(\frac{n}{k}\right)^k \left(\frac{n}{n-k}\right)^{n-k} \\
&\qquad \left(\frac{n}{k(n-k)}\right)^{1/2} \gamma^{k+l} k^k \left(\frac{k}{n}\right)^l \left(\frac{n-k}{n}\right)^{2(m-k-l)} n^{-k} \\
\leq\ & (2\pi)^{-1/2} \sum_{k=1}^{n^{1/4}} \gamma^k \left(\frac{\gamma k}{n}\right)^l \left(\frac{n}{k(n-k)}\right)^{1/2} \left(\frac{n-k}{n}\right)^{2(m-k-l)} \left(\frac{n}{n-k}\right)^{n-k} \\
\leq\ & \frac{2}{(2\pi)^{1/2}} \sum_{k=1}^{n^{1/4}} \gamma^k \left(\frac{10k\ln n}{n}\right)^l \left(\frac{n}{n-k}\right)^{2l} \left(\frac{n-k}{n}\right)^{2(m-k)-(n-k)} \\
\leq\ & \sum_{k=1}^{n^{1/4}} \left(\frac{10kn\ln n}{(n-k)^2}\right)^l \left(\frac{k+m-k}{n}\right)^k \left(\frac{n-k}{n}\right)^{m-k} \\
\leq\ & \left(\frac{20n^{5/4}\ln n}{n^2}\right)^l \sum_{k=1}^{n^{1/4}} \exp(k(m-k)/n) \left(1-\frac{k}{n}\right)^{m-k} \\
\leq\ & \left(\frac{20\ln n}{n^{3/4}}\right)^l \sum_{k=1}^{n^{1/4}} \exp(k(m-k)/n - k(m-k)/n) \leq \left(\frac{20\ln n}{n^{3/4}}\right)^l n^{1/4}.
\end{aligned}$$

Summing over $l = 1, \ldots, m$ gives the estimate

$$\begin{aligned}
\sum_{l=1}^{m} \left(\frac{20\ln n}{n^{3/4}}\right) n^{1/4} &\leq n^{1/4} \sum_{l=1}^{\infty} \left(\frac{20\ln n}{n^{3/4}}\right)^l \\
&\leq \frac{20\ln n}{n^{1/4}} \sum_{l=0}^{\infty} \left(\frac{20\ln n}{n^{3/4}}\right)^l \leq \frac{40\ln n}{n^{1/4}} \ll 1,
\end{aligned}$$

which proves the lemma. □

Lemma 21 *Suppose that $3n/4 \leq m \leq n$. Then with high probability $D(L_{n,m})$ has no Eulerian component of order $\leq n^{1/4}$ that consists of more arcs than vertices.*

Proof. We use the same notations as in the pervious lemma. We have $1 \geq \gamma \geq 3/4$, whence $2m-n \geq n/2$. Thus,

$$\begin{aligned}
& \sum_{k=1}^{n^{1/4}} \left(\begin{array}{c} n \\ k \end{array} \right) (m)_{k+l} k^{k+l} (n-k)^{2(m-k-l)} n^{-2m} \\
\leq \; & (2\pi)^{-1/2} \sum_{k=1}^{n^{1/4}} \left(\frac{n}{k}\right)^k \left(\frac{n}{n-k}\right)^{n-k} \left(\frac{n}{k(n-k)}\right)^{1/2} \\
& \qquad \gamma^{k+l} k^{k+l} \left(\frac{n-k}{n}\right)^{2(m-k-l)} n^{-(k+l)} \\
\leq \; & (2\pi)^{-1/2} \sum_{k=1}^{n^{1/4}} 2 \left(\frac{n-k}{n}\right)^{(n/2)-k} \left(\frac{n}{n-k}\right)^{2l} \left(\frac{k}{n}\right)^l \\
\leq \; & \sum_{k=1}^{n^{1/4}} \left(\frac{kn}{(n-k)^2}\right)^l \leq n^{1/4} \left(\frac{2}{n^{3/4}}\right)^l .
\end{aligned}$$

Thus, summing over l, we can bound the expected number of Eulerian components as in the lemma by $4/n^{1/4}$. □

Lemma 22 *Suppose that $m \leq 3n/4$. Then the expected number of vertices in Eulerian components of $D(L_{n,m})$ is at most 3.*

Proof. Given k, there are $\leq n^k (m)_k / k$ possibilities for a Eulerian component containing precisely k arcs. If the component has order $l \leq k$, then there are $(n-l)^{2(m-k)}$ possibilities for the remaining graph. Thus, the expected number of vertices in Eulerian components is

$$\leq \sum_{k=1}^{m-1} \frac{k}{k} n^k (m)_k n^{-2k} \leq \sum_{k=1}^{m-1} \left(\frac{m}{n}\right)^k \leq 3,$$

as stated. □

In summary, we have shown the following: *There is a function $f(m) = o(1)$ such that the probability that $\geq m^{1/8}$ vertices belong to such Eulerian components of $D(L_{n,m})$ that contain more arcs than vertices is $\leq f(m)$.* The remaining task is to estimate the number of vertices that lie on isolated directed cycles.

Lemma 23 *Suppose that $m \geq n$. Then the expected number of vertices that lie on isolated directed cycles is $O(1)$.*

Proof. Put $\gamma = m/n$. Then the expected number of vertices on isolated directed cycles can be bounded as follows:

$$\begin{aligned}
& \sum_{k=1}^{n-1} k(n)_k (m)_k (n-k)^{2(m-k)} k^{-1} n^{-2m} \leq \sum_{k=1}^{n-1} \gamma^k \left(1 - \frac{k}{n}\right)^{2(m-k)} \\
= & \sum_{k=1}^{n-1} \left(\frac{k}{n} + \frac{m-k}{n}\right)^k \left(1 - \frac{k}{n}\right)^{2(m-k)} \\
\leq & \sum_{k=1}^{n-1} \left(1 + \frac{m-k}{n}\right)^k \exp(-2(m-k)k/n) \\
\leq & \sum_{k=1}^{n-1} \exp((m-k)k/n - 2(m-k)k/n) = \sum_{k=1}^{n-1} \exp(-k(m-k)/n) \\
\leq & 2 \sum_{k=1}^{\lfloor n/2 \rfloor} \exp(-k(m-k)/n) \leq 2 \sum_{k=1}^{\lfloor n/2 \rfloor} \exp(-k(n-k)/n) \\
\leq & 2 \sum_{k=1}^{\lfloor n/2 \rfloor} \exp(-k(n-n/2)/n) = 2 \sum_{k=1}^{\lfloor n/2 \rfloor} \exp(-k/2) = O(1).
\end{aligned}$$

Note that in the case $m \gg n$ it drops out that the above expectation is $\ll 1$. □

Lemma 24 *Suppose that $n/2 \leq m \leq n$. Then the expected number of vertices on isolated directed cycles is $O(1)$.*

Proof. The expectation is

$$\begin{aligned}
& \sum_{k=1}^{m-1} k(n)_k (m)_k (n-k)^{2(m-k)} k^{-1} n^{-2m} \\
\leq & \sum_{k=1}^{m-1} \left(\frac{m}{n}\right)^k \left(1 - \frac{k}{n}\right)^{2(m-k)} \leq \sum_{k=1}^{m-1} \exp(-2(m-k)k/n) \\
\leq & 2 \sum_{k=1}^{\lfloor m/2 \rfloor} \exp(-2k(m-k)/n) \leq 2 \sum_{k=1}^{m-1} \exp(-2km/(2n)) \\
\leq & 2 \sum_{k=1}^{\lfloor m/2 \rfloor} \exp(-k/4) = O(1).
\end{aligned}$$

□

Lemma 25 *Suppose that $m \le n/2$. Then the expected number of vertices on isolated directed cycles is $O(1)$.*

Proof. Put $\gamma = m/n \le 1/2$. Then the expectation is

$$\begin{aligned}\sum_{k=1}^{m-1} k(n)_k(m)_k(n-k)^{2(m-k)}k^{-1}n^{-2m} &\le \sum_{k=1}^{m-1}\gamma^k\left(1-\frac{k}{n}\right)^{2(m-k)}\\ &\le \sum_{k=1}^{m-1}\gamma^k = O(1).\end{aligned}$$

Note that in the case $\gamma \ll 1$ the last sum is $\ll 1$. □

If $L = L_{n,m}$ and $\sigma \in \mathcal{S}_n$ is a permutation, then we define $\sigma L \in L_{n,m}$ in the natural way. We equip the space $L_{n,m} \times \mathcal{S}_n$ with the uniform distribution. Obviously, the map $(L, \sigma) \mapsto \sigma L$ maps the uniform distribution onto the uniform distribution.

Lemma 26 *There exists a function $f(m) = o(1)$ such that in the space $L_{n,m} \times \mathcal{S}_n$ the following event has probability $\le f(m)$: There is $k \in \{1, \dots, n\}$ such that in $D(\sigma L)$ the vertices k and $k+1$ both belong to Eulerian components.*

Proof. Let X be the number of vertices in Eulerian components of $D(\sigma L)$. Let the random variable X_k take value 1 if k belongs to an Eulerian component of $D(\sigma L)$ and 0 otherwise. Let $Y = \sum_{k=1}^{n-1} X_k X_{k+1}$. Conditioning on $X \le x$, we obtain

$$P(X_k X_{k+1} = 1) \le \frac{x(x-1)(n-2)!}{n!} \le \frac{x^2}{n(n-1)}.$$

Consequently, $E(Y) \le x^2/n$. Therefore, if $x \le n^{1/4}$, then $Y = 0$ with high probability. However, the event $X \le n^{1/4}$ occurs with high probability, as seen above. □

Corollary 27 *There is a function $f(m) = o(1)$ such that the following event has probability $\le f(m)$: There is $k \in \{1, \dots, n-1\}$ such that the vertices k and $k+1$ both belong to Eulerian components of $D(L_{n,m})$.*

Let us examine the number of small Eulerian components in the case $m = \Theta(n)$ more closely.

Lemma 28 *Let $L = L_{n,m}$, where $m = \gamma/n$ for some fixed $\gamma > 0$. Then the number of isolated directed cycles of length k converges in law to the Poisson distribution $P(\lambda)$, $\lambda = \frac{1}{k}\left(\frac{\gamma}{e^{2\gamma}}\right)^k$.*

Proof. Let X be the number of directed isolated k-cycles in $D(L_{n,m})$. Then

$$E(X) = \frac{(n)_k(m)_k(n-k)^{2(m-k)}}{kn^{2m}}.$$

Furthermore, the rth factorial moment of X is

$$E_r(X) = \frac{(n)_{kr}(m)_{kr}(n-kr)^{2(m-kr)}}{k^r n^{2m}}.$$

Consequently,

$$\begin{aligned}\frac{E_r(X)}{E(X)^r} &= \frac{(n)_{kr}}{(n)_k^r}\cdot\frac{(m)_{kr}}{(m)_k^r}\cdot\frac{(n-kr)^{2(m-kr)}n^{2mr}}{(n-k)^{2(m-k)r}n^{2m}} \\ &= \frac{(n)_{kr}}{(n)_k^r}\cdot\frac{(m)_{kr}}{(m)_k^r}\cdot\frac{(n-kr)^{2m}}{n^{2m}}\cdot\frac{n^{2mr}}{(n-k)^{2mr}}\cdot\frac{(n-k)^{2kr}}{(n-kr)^{2kr}}.\end{aligned}$$

Since k, r are fixed, we have

$$1 \geq \frac{(n)_{kr}}{(n)_k^r} \geq \left(\frac{n-kr+1}{n}\right)^{kr} = \left(1-\frac{kr+1}{n}\right)^{kr} \to 1 \text{ as } n\to\infty.$$

Similarly, $(m)_{kr}/(m)_k^r \to 1$ and $\left(\frac{n-k}{n-kr}\right)^{2kr} \to 1$ as $n\to\infty$ Moreover, for any fixed $\varepsilon > 0$

$$\exp(-2(1+\varepsilon)kr\gamma) \leq \left(\frac{n-kr}{n}\right)^{2m} \leq \exp(-2kr\gamma)$$

for large n. Similarly, because

$$\frac{n}{n-k} = \frac{1}{1-\frac{k}{n}} \leq \frac{1}{\exp(-2(1+\varepsilon)k)} = \exp(2(1+\varepsilon)k),$$

we have

$$\exp(2kr\gamma) \leq \left(\frac{n-k}{n}\right)^{-2mr} = \left(\frac{n}{n-k}\right)^{2mr} \leq \exp(2(1+\varepsilon)kr\gamma).$$

Consequently,

$$\exp(-2kr\gamma\varepsilon) \leq \left(\frac{n-kr}{n}\right)^{2m}\left(\frac{n}{n-k}\right)^{2mr} \leq \exp(2kr\gamma\varepsilon),$$

whence $\lim_{n\to\infty} E_r(X)/E(X)^r = 1$. Finally, we note that

$$E(X) = \frac{1}{k}\cdot\frac{(n)_k}{n^k}\cdot\frac{(m)_k}{m^k}\cdot\left(\frac{n-k}{n}\right)^{2(m-k)},$$

where $\lim\frac{(n)_k}{n^k} = 1$, $\lim\frac{(m)_k}{m^k} = \gamma^k$, and

$$\left(\frac{n-k}{n}\right)^{2(m-k)} = \left(1-\frac{k}{n}\right)^{2\gamma n}\left(1-\frac{k}{n}\right)^{-2k} \to \exp(-2k\gamma) \text{ as } n\to\infty.$$

Thus,

$$E(X) \to \frac{1}{k}\cdot\left(\frac{\gamma}{\exp(2\gamma)}\right)^k = \lambda \text{ as } n\to\infty.$$

From [6, p. 23] our assertion follows. □

D Proofs for Section 5

In Section 5 we claimed that there should be $4! \cdot \#S$ pieces out of $\sqrt{n}$ of length $\sqrt{n}$ that start with six isolated feet and are otherwise biconnected with l_n. This follows from Lemma 30, assuming that $N = n^{1/8} \geq 4! \cdot \#S$, and connecting superfluous isolated vertices to l_n:

Lemma 29 *Suppose that $m \ll n \ln n$. Let $\varepsilon > 0$. Then with high probability there are at least $n^{1-\varepsilon}$ isolated feet in $D(L_{n,m})$.*

Proof. Let X denote the number of isolated vertices of $D(L_{n,m})$. The probability that a given foot v is isolated is

$$\left(\frac{n-1}{n}\right)^{2m} = \left(1-\frac{1}{n}\right)^{2m} \geq \exp\left(-\frac{2m}{n} - \frac{2m}{n^2}\right).$$

Similarly, the probability that two given feet $v \neq w$ are isolated is

$$\left(\frac{n-2}{n}\right)^{2m} = \left(1-\frac{2}{n}\right)^{2m} \leq \exp\left(-\frac{4m}{n}\right).$$

Thus,

$$\mathrm{E}(X) = n \cdot \left(1-\frac{1}{n}\right)^{2m} \geq n \cdot \exp\left(-\frac{2m}{n}(1+1/n)\right)$$

and

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{v \neq w} P(\text{both } v,w \text{ are isolated}) - P(v \text{ isolated})P(w \text{ isolated}) \\
&\qquad + \sum_v P(v \text{ isolated}) - P(v \text{ isolated})^2 \\
&\leq \sum_{v \neq w} \left\{ \exp\left(-\frac{4m}{n}\right) - \exp\left(-\frac{4m}{n}(1+1/n)\right) \right\} + \sum_v P(v \text{ isolated}) \\
&\leq n^2 \exp\left(-\frac{4m}{n}\right)\left(1-\exp\left(-\frac{4m}{n^2}\right)\right) + n\exp(-2m/n),
\end{aligned}$$

where v, w range over the feet of the caterpillar. Consequently,

$$\begin{aligned}
\frac{\mathrm{Var}(X)}{\mathrm{E}(X)^2} &\leq \frac{n^2 \exp\left(-\frac{4m}{n}\right)\left(1-\exp\left(-\frac{4m}{n^2}\right)\right) + n\exp\left(-\frac{2m}{n}\right)}{n^2 \exp\left(-\frac{4m}{n}\right)\exp\left(-\frac{4m}{n^2}\right)} \\
&= \exp\left(\frac{4m}{n^2}\right) - 1 + 2\exp(2m/n)/n \ll 1.
\end{aligned}$$

Thus, by Chebyshev's inequality with high probability we have

$$X \geq \mathrm{E}(X)/2 \geq n^{1-\varepsilon},$$

where $\varepsilon > 0$ is arbitrary. □

Lemma 30 *Suppose that $m \ll n \ln n$. Then with high probability $D(L_{n,m})$ contains at least $n^{1/8}$ pieces of length $n^{1/2}$ starting with 6 consecutive isolated feet.*

Proof. Split the backbone of the caterpillar into l pieces of equal length. Note that the operation of the symmetric group $\mathcal{S}_n$ leaves the distribution on $L_{n,m}$ invariant. Let us for a moment work under the condition that the number of isolated feet in $L = L_{n,m}$ is $N \geq n^{1-\varepsilon}$. Then the probability that L has isolated vertices at K given positions after applying $\sigma \in \mathcal{S}_n$ is precisely

$$\frac{(N)_K}{(n)_K} = \prod_{j=0}^{K-1} \frac{N-j}{n-j} \sim \left(\frac{N}{n}\right)^K,$$

provided K is constant. Thus, the expected number of pieces that start with K isolated vertices is

$$\sim l \left(\frac{N}{n}\right)^K.$$

In order to estimate the variance, we let X_i take value 1 if the ith piece starts with K isolated vertices and 0 otherwise, $i = 1, \ldots, l$. Let $X = \sum X_i$. Then

$$\mathrm{E}(X^2) = \frac{l(l-1)(N)_{2K}}{(n)_{2K}} + \mathrm{E}(X),$$

and

$$\mathrm{E}(X)^2 = \frac{l(l-1)(N)_K^2}{(n)_K^2} + \frac{l(N)_K^2}{(n)_K^2} \geq \frac{l(l-1)(N)_K^2}{(n)_K^2}.$$

Thus,

$$\begin{aligned} \frac{\mathrm{Var}(X)}{\mathrm{E}(X)^2} &\leq \frac{l(l-1)\left(\frac{(N)_{2K}}{(n)_{2K}} - \frac{(N)_K^2}{(n)_K^2}\right) + \mathrm{E}(X)}{\mathrm{E}(X)^2} \\ &\leq \frac{(N)_{2K}(n)_K^2}{(N)_K^2(n)_{2K}} - 1 + \frac{1}{\mathrm{E}(X)} = \frac{1}{\mathrm{E}(X)} + o(1). \end{aligned}$$

Now put $K = 6$ and $l = n^{1/2}$, $\varepsilon = 1/100$. Then $\mathrm{E}(X) \geq n^{2/5}$. Thus, with high probability $X \geq n^{1/8}$, provided that the number of isolated vertices is $\geq N$. Taking into account the previous lemma, our assertion follows. □

The last claim in Section 5 is that any solution to the DARP instance constructed in Section 5 can be transformed to a solution of at most the same cost which has certain properties (*normalized solution*).

Lemma 31 *Let L'' be the instance of the* DARP *on Cat_n that has been constructed in Section 5. Let D be the edges of a solution other than requests. Then D consists of Eulerian components $C_1, \ldots, C_i, \ldots$ induced on the feet and the following assumptions on D can be ensured without increasing the cost of the solution:*

(i) Each C_i visits only feet of a single I_j and, possibly, neighboring feet of $V \setminus I$

(ii) For each Steiner vertex s at most one of the $I_\sigma(s)$, $\sigma \in \mathcal{S}_n$, is visited by a C_i. If C_i visits a labeled foot $l_k \in I_j$ then it also visits the labeled feet $l_\ell \in I_j$ where $\ell > k$

(iii) There is only one C_i that visits a foot of $V \setminus I$

Proof. The fact that D consists of Eulerian components follows directly from the fact that L'' and $D \cup L''$ are Eulerian. Observe that balancing does not yield any additional request and that all arcs of D are between feet, because all of L'' are. Note that all vertices outside $I = \bigcup_j I_j$ make up one large Eulerian component M. As a consequence, if a C_i visits M twice, this can safely be short cut to one visit.

If an Euler tour of a C_i crosses all of $B_j \setminus I_j$ for some j, then this crossing must be both ways. It can therefore be split and short cut to the first and the last foot of $B_j \setminus I_j$. As the distance between these two feet is almost $\sqrt{n}$ this shortens the total length of D dramatically. This proves assertion (i) of the lemma.

Let $s \in S$ and assume that C_1 connects k feet of $I_\sigma(s)$ while C_2 connects ℓ feet of $I_{\sigma'}(s)$ ($\sigma, \sigma' \in \mathcal{S}_n$) and these sets of feet share a label. Then the length of C_1 is at least $4k-2$ and the length of C_2 is at least $4\ell-2$. On the other hand, there is $\tau \in \mathcal{S}_n$ such that the set of labels of the feet in C_1 and C_2 are contained in the labels of $v_2, \ldots, v_{\ell+k}$, where $I_\tau(s) = \{v_6, \ldots, v_1\}$. Thus C_1 and C_2 can be replaced with a component C' on I_τ with length $4\ell + 4k - 6$, short cutting D. Observe that this procedure also works if C_1 or C_2 visit additionally a foot of $V \setminus I$ and that it can be achieved that this foot lies to the right.

If C_1 and C_2 visit feet with disjoint labels then we may assume that a foot in C_1, say with label t_1, has a neighboring foot with label t_2 of C_2, possibly changing σ appropriately. Then, increasing the length of C_1 by 4, label t_2 can also be visited by C_1. On the other hand, in D there is already another path between labels t_1 and t_2. Say, the next label on this path is t_3 and the connection between label t_1 and t_3 is in component C_3. Again we may assume, that the labels t_1 and t_3 are neighboring in C_3, and t_1 is leftmost. Then the connection can be split, saving a length of 4 in C_3. This proves assertion (ii) of the lemma.

Assume that there are two components C_1 and C_2 that both visit a foot in $V \setminus I$, and let t_1 and t_2 be labels visited by C_1 and C_2, respectively. Split off C_1 from $V \setminus I$, saving a length of 6 (Note that the labeled feet are separated from $V \setminus I$ by an unlabeled foot). If this splits D into two connected components, then let T_1 and T_2 be the sets of labels reachable from t_1 and t_2, respectively. As (S, T, E) is connected there must be $t_1' \in T_1$ and $t_2' \in T_2$ that share a neighbor $s \in S$. With an appropriate σ, t_1' and t_2' are the two rightmost labels in $I_\sigma(s)$ and can thus be connected at a cost of 6. This proves assertion (iii) of the lemma. □

The solution as provided by Lemma 31 can be turned into a Steiner tree in (S, T, E) by connecting $t \in T$ to $s \in S$ iff label t is visited in a component on $I_\sigma(s)$ for some σ.