Concentration of the cost of a random matching problem

Martin Hessler and Johan Wästlund

Mathematics Subject Classification: Primary: 60C05, 90C27, 90C35

August 20, 2008

Abstract

Let M_n be the minimum cost of a perfect matching on a complete graph on n vertices whose edges are assigned independent exponential costs. It follows from work of D. Aldous that M_n converges in probability to $\pi^2/12$. This was earlier conjectured by M. Mézard and G. Parisi. We establish the more precise result that $E\left|M_n - \pi^2/12\right| = O(n^{-1/2})$.

1 Introduction

We consider the exponential matching problem on the complete graph K_n, where n is an even positive integer. The edges of K_n are assigned independent costs from exponential (mean 1) distribution. The *minimum matching* is the set of edges of minimum total cost subject to the constraint that each vertex must be incident to exactly one edge in the set. We let M_n denote the cost of the minimum matching.

Example 1.1. The vertices are labeled $v_1, \ldots, v_n$, and the edge between v_i and v_j has cost $X_{i,j}$. When $n = 2$, there is only one edge, and this edge has cost $X_{1,2}$. The unique matching consists of this single edge, and M_2 is therefore exponentially distributed with mean $EM_2 = 1$.

We can compute the expectation of M_4 as follows: There are three perfect matchings of costs $X_{1,2} + X_{3,4}$, $X_{1,3} + X_{2,4}$ and $X_{1,4} + X_{2,3}$ respectively. The probability that a sum of two independent exp(1) variables is greater than t

is equal to the probability that in a rate 1 Poisson process there is at most one event in the interval $[0, t]$. Hence

$$P(X_{1,2} + X_{3,4} > t) = (1 + t)e^{-t}.$$

Since the costs of the three matchings are independent, we can compute the expectation of M_4 as

$$EM_4 = \int_0^\infty P(M_4 > t)\, dt = \int_0^\infty P(\text{all three matchings have cost} > t)\, dt$$
$$= \int_0^\infty (1 + t)^3 e^{-3t}\, dt = \frac{26}{27}.$$

For larger n there seems to be no simple description of the distribution of M_n, but it has been known for some time that M_n converges in probability to $\pi^2/12$. Our main result is the following concentration inequality:

Theorem 1.2.

$$E\left|M_n - \frac{\pi^2}{12}\right| < \frac{1.706}{\sqrt{n}}. \tag{1}$$

We also show that except for the constant, (1) is best possible in the sense that the right hand side cannot be replaced by anything of smaller order. This is a consequence of the following theorem:

Theorem 1.3. *For every real number x,*

$$P\left(|M_n - x| \geq \frac{1.114}{\sqrt{n}}\right) \geq \frac{1}{8}. \tag{2}$$

2 Background and outline of our approach

In a series of papers in the 1980's [5, 6, 7, 8, 9] Marc Mézard and Giorgio Parisi conjectured that as $n \to \infty$,

$$M_n \xrightarrow{\text{p}} \frac{\pi^2}{12}. \tag{3}$$

The conjectured mode of convergence was not explicitly stated, but it is clear that at least convergence in probability was intended. In principle, (3) follows from the results of David Aldous [1, 2], and so does the fact that

$$EM_n \to \frac{\pi^2}{12}, \tag{4}$$

which strictly speaking does not follow from (3). It is known that the method, which in [1, 2] is applied to the assignment problem (bipartite matching), will apply also to matching on the complete graph provided certain technical lemmas are modified. The same is true of the bound on the variance established by Michel Talagrand in [10]. There, an upper bound of

$$O\left(\frac{(\log n)^4}{n \log\log n}\right)$$

was established for the variance in the bipartite matching problem on $K_{n,n}$.

A self-contained proof of (4) was given in [13], where it was shown that

$$\frac{\pi^2}{12} < EM_n < \frac{\pi^2}{12} + \frac{\log n}{n}. \tag{5}$$

With related methods we here establish a $O(1/n)$-bound on the variance of the cost C_n of a certain LP-relaxation of the matching problem. It is known from [14, 15] that the expected value of this relaxation is given by the explicit formula

$$EC_n = 1 - \frac{1}{4} + \frac{1}{9} - \cdots - \frac{1}{n^2} = \frac{\pi^2}{12} + O\left(\frac{1}{n^2}\right). \tag{6}$$

Hence the expected difference between M_n and C_n is $O(\log n/n)$, smaller than the standard deviation of either of them. Combining these results we establish (1), thereby obtaining an explicit proof of (3).

3 The relaxed matching problem

Following [14, 15], we consider a certain LP-relaxation of the matching problem on K_n. This is a special case of the *linear flow problem* defined in [15]. The relaxation is defined on a graph with random edge costs called the *friendly model* [15]. In this model there are also loops (edges connecting a vertex to itself). One of the features of the friendly model is that the feasible solutions corresponding to perfect matchings exist also for odd n.

The vertices are labeled $v_1, \ldots, v_n$. There are edges $e_{i,j}$ of cost $X_{i,j}$ for $1 \le i \le j \le n$. The edge costs are independent exponentially distributed, and $X_{i,j}$ has mean 1 if $i \neq j$ and mean 2 if $i = j$. In [15], there are also multiple edges between each pair of vertices, but these are irrelevant for the matching problem.

The relaxed k-matching problem asks for coefficients $\sigma(i,j) \in [0,1]$ minimizing

$$\sum_{1 \le i \le j \le n} \sigma(i,j) X_{i,j},$$

subject to the constraints

$$\sum_{1 \le i \le j \le n} \sigma(i,j) = k, \tag{7}$$

and for $1 \le i \le n$,

$$\sum_{j \le i} \sigma(j,i) + \sum_{j \ge i} \sigma(i,j) \le 1, \tag{8}$$

Notice that in (8), $\sigma(i,i)$ is counted twice. Also notice that if we require the coefficients $\sigma(i,j)$ to be either 0 or 1, a feasible solution is a matching. The expression (8) is called the *degree* of v_i. The degree of v_i in the optimum solution is denoted $\delta_{k,n}(i)$.

A feasible solution (relaxed matching) is also called a *flow*. Let $C_{k,n}$ denote the cost of the minimum k-flow. One of the results of [15] is the following formula for the expectation of $C_{k,n}$.

Proposition 3.1. *Let k be such that $2k$ is an integer. Then*

$$EC_{k,n} = \sum_{\substack{0 \le i \le j \\ i+j<2k}} \frac{1}{(n-i)(n-j)}. \tag{9}$$

It can be verified that (9) specializes to (6) if $k = n/2$ (also for odd n).

4 The extended graph

An important idea of [15] is to introduce an *extended graph* that contains an extra vertex v_{n+1}. Here we let the extra vertex have weight γ and unrestricted capacity, and we shall explain what this means. The costs of the edges connecting v_{n+1} to the ordinary vertices $v_1, \ldots, v_n$ are exponential of rate γ. There is also a potentially infinite sequence of loops at v_{n+1}, and the costs of these loops are given by the times of the events in a Poisson process of rate $\gamma^2/2$. In particular the cost of the cheapest loop is exponential of rate $\gamma^2/2$. In the end, γ will tend to zero. Informally, it is natural to think of γ as infinitesimal.

The relaxed matching problem on the extended graph asks for the minimum, denoted $C_{k,n+1}$, of

$$\sum_e \sigma(e) X_e,$$

where e ranges over the edge set, and the coefficients $\sigma(e)$ are restricted to the interval $[0,1]$. Naturally the constraint (7) is replaced by

$$\sum_e \sigma(e) = k.$$

We still require (8) to hold for $1 \le i \le n$, but we put no constraint on the degree of v_{n+1}. We let σ_k denote the minimum k-flow, assuming that the edge costs are generic in the sense that no two flows have the same cost.

In [15], the formula (6) and its generalization (9) are established inductively by computing the expectation of the degree $\delta_{k,n+1}(n+1)$ of the extra vertex in the extended problem. Similarly, knowledge about the correlation of $\delta_{k,n+1}(n+1)$ with the cost $C_{k,n+1}$ of the minimum flow would allow us to inductively compute the variance of $C_{k,n}$.

We consider the extended problem, and condition on all edge costs except the cost $X_{n,n+1}$ of the edge $e_{n,n+1}$. Let $f(x)$ be the cost of the minimum k-flow given that $X_{n,n+1} = x$. In other words, $f(x) = (C_{k,n+1} | X_{n,n+1} = x)$.

Notice that f is continuous, and that $f'(x) = \sigma_{k,n+1}(n, n+1)$ except at a finite number of points where the right derivative is not equal to the left derivative. It follows by partial integration that

$$\begin{aligned} E\left(f^2(x)\right) &= \int_0^\infty f^2(x)\gamma \exp(-\gamma x)dx \\ &= f^2(0) + \frac{1}{\gamma}\int_0^\infty 2f'(x)f(x)\gamma\exp(-\gamma x)dx. \quad (10) \end{aligned}$$

We now want to let γ tend to zero, and for technical reasons we want to take this limit for a fixed point in the probability space. We therefore think of the costs of the edges to v_{n+1} as being generated from underlying exponential variables (and a Poisson process for the loops) of rate 1. Then the actual costs are obtained by dividing by the rate (that is, γ for the ordinary edges and $\gamma^2/2$ for the loops).

This means that as $\gamma \to 0$, the costs of the edges at v_{n+1} except $e_{n,n+1}$ tend to infinity. It follows that (for fixed costs of the ordinary edges, as $\gamma \to 0$)

$$f^2(0) \to C^2_{k-1,n-1}.$$

Similarly, by the principle of dominated convergence,

$$E(f^2(x)) \to C_{k,n}^2.$$

By the observation that $f'(x) = \sigma_{k,n+1}(n, n+1)$, we can rewrite (10) as

$$\lim_{\gamma \to 0} \frac{2}{\gamma} E(\sigma_{k,n+1}(n, n+1) \cdot C_{k,n+1}) = EC_{k,n}^2 - C_{k-1,n-1}^2.$$

We now take average over all edge costs,and note that by symmetry, $\sigma_{k,n+1}(n, n+1)$ can be replaced by $1/n \cdot \delta_{k,n+1}(n+1)$. We conclude that

$$\lim_{\gamma \to 0} \frac{2}{n\gamma} E(\delta_{k,n+1}(n+1) \cdot C_{k,n+1}) = E(C_{k,n}^2) - E(C_{k-1,n-1}^2). \tag{11}$$

By the same argument [13], it also follows that

$$\lim_{\gamma \to 0} \frac{1}{n\gamma} E(\delta_{k,n+1}(n+1)) = E(C_{k,n}) - E(C_{k-1,n-1}). \tag{12}$$

In order to inductively calculate the variance of $C_{k,n}$, it would be sufficient to calculate the left hand side of (11). Unfortunately we are still unable to calculate this exactly, but we will show that the correlation between $\delta_{k,n+1}(n+1)$ and $C_{k,n+1}$ is negative. This gives an upper bound on the variance of $C_{k,n}$ which turns out to be of the right order of magnitude (and better than what follows from the Talagrand inequality).

5 A correlation inequality

Let $X_1, \ldots, X_m$ be random variables (not necessarily independent), and let f and g be two real valued functions of $X_1, \ldots, X_m$. Let $f_i = E(f|X_1, \ldots, X_i)$, and similarly $g_i = E(g|X_1, \ldots, X_i)$, supposing that these expectations exist. Then $f_0 = E(f_j)$ for every j, $f_m = f$, and similarly for g. The following lemma is crucial for our approach. It is valid whenever the expectations under consideration exist. In our application, f will have finite expectation and g will be bounded, but the lemma is valid under more general conditions.

Lemma 5.1. *Suppose that for every i and every outcome of $X_1, \ldots, X_m$,*

$$(f_{i+1} - f_i)(g_{i+1} - g_i) \geq 0. \tag{13}$$

Then f and g are positively correlated, in other words,

$$E(fg) \geq E(f)E(g). \tag{14}$$

Proof. Equation (13) can be written

$$f_{i+1}g_{i+1} \geq (f_{i+1} - f_i)g_i + (g_{i+1} - g_i)f_i + f_i g_i.$$

Notice that $E((f_{i+1} - f_i)g_i|X_1, \ldots, X_i) = g_i E(f_{i+1} - f_i|X_1, \ldots, X_i) = g_i f_i - g_i f_i = 0$. It follows that $E\left((f_{i+1} - f_i)g_i\right) = 0$, and similarly for the other term. We conclude that $E(f_{i+1}g_{i+1}) \geq E(f_i g_i)$ and by induction that

$$E(fg) = E(f_m g_m) \geq E(f_0 g_0) = f_0 g_0 = E(f)E(g).$$

□

6 The oracle process

We study the so called *oracle process*, which has been described in [13, 14, 15]. This is a stochastic process governed by the edge costs in the extended graph. We think of an "oracle" who knows all the edge costs. We ask questions to the oracle in order to determine the degree $\delta_{k,n+1}(n+1)$ of v_{n+1} in the minimum relaxed k-matching in the extended graph. Here and in the next section, γ will be a fixed positive number. The questions are chosen in such a way that we can control the conditional distribution of the edge costs in the process.

In the process, we successively find the minimum relaxed r-matchings for $r = 1/2, 1, 3/2, \ldots, k$. The following three lemmas are proved in [15].

Lemma 6.1. *If $2k$ is an integer, then there is a minimum relaxed matching in which every edge has coefficient 0, $1/2$ or 1.*

Lemma 6.2. *If the edge costs are fixed and generic, then the degree $\delta_{k,n+1}(i)$ of a given vertex v_i in the minimum k-flow is a nondecreasing function of k. Moreover, supposing $2k$ is an integer, if $\delta_{k,n+1}(i)$ is an integer for every i, then $\delta_{k+1/2,n+1}$ is obtained from $\delta_{k,n+1}$ by either increasing the value by $1/2$ at two vertices, or increasing by 1 at one vertex. If $\delta_{k,n+1}(i)$ is not an integer for every i, then there are precisely two vertices for which it is $1/2$ plus an integer. In this case $\delta_{k+1/2,n+1}$ is obtained from $\delta_{k,n+1}$ by adding $1/2$ at these two vertices.*

Notice that $\delta_{k,n+1}(i) \in \{0, 1/2, 1\}$ except possibly when $i = n + 1$.

Definition 6.3. We say that a flow (in the extended graph) is *stable* if all edges of coefficient other than 0 and 1 go between ordinary vertices that have degree 1.

Hence the minimum flow is stable if the edges of coefficient 1/2 form closed cycles that do not contain v_{n+1}. The other possibility is that they form a cycle including v_{n+1} or a path (necessarily of odd length) with two distinct endpoints.

Lemma 6.4. *Suppose that the edge costs are generic. If $2k$ is an integer and the minimum k-flow σ_k is not stable, then*

$$\sigma_{k+1/2} = 2\sigma_k - \sigma_{k-1/2}. \tag{15}$$

In a generic step of the oracle process, we have found the minimum r-flow σ_r for a certain r such that $0 \leq r < k$ and $2r$ is an ineger. We assume that σ_r is stable. Let Γ be the set of ordinary vertices of degree 1 in σ_r. The following is known:

1. The costs of all edges for which both endpoints belong to Γ.
2. The cost of all other edges of nonzero coefficient in σ_r.
3. For each vertex $v \in \Gamma$, the minimum cost of the remaining edges that connect v to a vertex not in Γ (but not the location of this edge).
4. The minimum cost of the remaining edges between vertices not in Γ (but again not the location of the edge that has this minimum cost).

Using this information only, we can essentially compute the minimum $(r + 1/2)$-flow. By Lemma 6.2, $\sigma_{r+1/2}$ is obtained from σ_r by "switching" an alternating path that connects two vertices of degree 0, that is, the coefficients of the edges in the path are alternatingly increased and decreased by 1/2. The path can be degenerate in a number of ways, and in particular the two endpoints need not be distinct. The information in 1–4 allows us to compute everything except the location of the endpoints of this path.

By the memorylessness property of the Poisson process, the unknown endpoints (whether one or two) are chosen independently among the vertices outside Γ with probabilities proportional to the total rates of the competing edge costs. Notice that this holds also if the path consists of only one edge (and that this edge can in principle turn out to be a loop at v_{n+1}).

After computing $\sigma_{r+1/2}$ (except for the unknown endpoints), we ask the oracle for the information that will be needed according to (1–4) in the next round of the process. We begin by asking for the locations of the endpoints of the alternating path. There are essentially three possibilities.

1. If there are two distinct endpoints v_i and v_j, then their degrees increase by 1/2. In this case $\sigma_{r+1/2}$ is not stable, and by Lemma 6.4, σ_{r+1} is determined once $\sigma_{r+1/2}$ is known. The same two vertices will again increase their degrees by 1/2, so that $\delta_{r+1,n+1}(i) = \delta_{r,n+1}(i) + 1$ and $\delta_{r+1,n+1}(j) = \delta_{r,n+1}(j) + 1$. Since $\sigma_{r+1}(e) - \sigma_r(e)$ is an integer for every e, the flow σ_{r+1} is stable.

2. Even if the alternating path starts and ends in two distinct edges, these edges can turn out to go to the same vertex v_i. Then $\delta_{r+1/2,n+1}(i) = \delta_{r+1/2,n+1}(i) + 1$. In this case $\sigma_{r+1/2}$ is stable.

3. The third possibility is that the alternating path starts and ends with the same edge. It is then clear that the endpoints of the path will coincide. This endpoint v_i is again chosen among the vertices not in Γ, with probabilities proportional to the weights. In this case too, $\delta_{r+1/2,n+1}(i) = \delta_{r+1/2,n+1}(i) + 1$, and $\sigma_{r+1/2}$ is stable.

7 Negative correlation

Lemma 7.1. *For every $\gamma > 0$, we have*

$$E(C_{k,n+1} \cdot \delta_{k,n+1}(n+1)) \leq E(C_{k,n+1}) \cdot E(\delta_{k,n+1}(n+1)).$$

Proof. When we apply Lemma 5.1 to the relaxed matching problem, we take the variables $X_1, \ldots, X_m$ to be the information driving the oracle process. We let f be the cost of the minimum k-flow, and let $g = -\delta_{k,n+1}(n+1)$.

There are two types of information we get from the oracle. One is about the endpoints of certain edges whose cost is alread known. The other is about the minimum cost of certain edge sets. The process is designed in such a way that the latter, information about the minimum cost of certain edge sets, does never change the conditional distribution of the $\delta_{k,n+1}(n+1)$ (that is, conditioning on the information we have received so far in the oracle process).

We therefore consider what happens when we get information about an endpoint of an edge (whose cost is known). By the stability assumption, all ordinary vertices that are potential endpoints are vertices that have degree zero in σ_r, and have no known edge to them. By symmetry, the conditional expectation of f will change in the same way regardless of which of them is chosen.

Moreover it is clear, and can actually be verified by exact formulas, that the conditional expectation of $\delta_{k,n+1}(n+1)$ will increase if the unknown endpoint turns out to be v_{n+1}, and therefore decrease if it turns out to be another vertex.

To clarify exactly what we are asking for in case we are asking for the location of the minimum cost edge of all edges not in Γ, suppose that every edge is randomly given an orientation (uniformly and independently). The orientation is immaterial for the optimization problem, but it allows us to ask the oracle about the two endpoints one at a time in a well-defined way.

We need only consider the case that the oracle tells us that an endpoint goes to v_{n+1}, and to show that in this case, the conditional expectation of $C_{k,n+1}$ decreases. It then follows that the conditional expectation increases in the other case. We first consider the case that the edge we are asking for is the minimum cost edge from a given vertex v in Γ to the vertices not in Γ. It does not matter whether this is the first or second edge we are asking for in this round of the oracle process. Let E' be the set of edges that we are comparing, that is, those that go from v to a vertex not in Γ. Now we condition on the costs of all edges except those in E'. We are going to use a coupling argument to show that, given all other edge costs, if the minimum cost edge in E' goes to v_{n+1}, then $C_{k,n+1}$ is smaller than otherwise.

Now think of the edge costs in E as generated by a joint Poisson process, so that we first see the edge costs, and then determine for each event in this process to which vertex outside Γ the corresponding edge goes. Then it is clear that if we want to find a cheap flow, it is advantageous if the first edge goes to v_{n+1}, since there is no capacity constraint on this vertex.

The same argument applies if we are asking for the cheapest edge connecting vertices not in Γ. Suppose we are asking for the first (in the sense of the arbitrarily chosen orientation of the edges) endpoint of the cheapest edge outside Γ. Then we condition on the costs of all edges, and the second endpoint of every edge (although in the oracle process, this information would only be given to us later). Knowing the second endpoint of the cheapest edge, the situation is now similar to that of the previous case.

Hence if at a certain point we ask the oracle about the endpoint of a certain augmenting path, and we are informed that this endpoint goes to the extra vertex v_{n+1}, then the conditional expectation of $C_{k,n+1}$ decreases. By Lemma 5.1, this shows that $C_{k,n+1}$ is negatively correlated with the degree $\delta_{k,n+1}(n+1)$ of v_{n+1}. □

Corollary 7.2.

$$\mathrm{var}(C_{k,n}) \leq \mathrm{var}(C_{k-1,n-1}) + (E(C_{k,n}) - E(C_{k-1,n-1}))^2.$$

Proof. We have previously proved (11), [13] the two equations

$$E((C_{k,n})^2) - E((C_{k-1,n-1})^2) = \lim_{\gamma \to 0} \frac{2}{n\gamma} E(\delta_{k,n+1}(n+1) \cdot C_{k,n+1}), \tag{16}$$

and

$$E(C_{k,n}) - E(C_{k-1,n-1}) = \lim_{\gamma \to 0} \frac{1}{n\gamma} E(\delta_{k,n+1}(n+1)). \tag{17}$$

Using (16) in Lemma 7.1 and taking the limit $\gamma \to 0$ we find that

$$\lim_{\gamma \to 0} \frac{2}{n\gamma} E(\delta_{k,n+1}(n+1) \cdot C_{k,n+1}) \leq \lim_{\gamma \to 0} E(C_{k,n+1}) \cdot \lim_{\gamma \to 0} \frac{2}{n\gamma} E(\delta_{k,n+1}(n+1)).$$

By the principle of dominated convergence we know that

$$\lim_{\gamma \to 0} E(C_{k,n+1}) = E(C_{k,n}),$$

which implies that

$$\begin{aligned} \mathrm{var}(C_{k,n}) + (E(C_{k,n}))^2 \leq \mathrm{var}(C_{k-1,n-1}) + \\ + (E(C_{k-1,n-1}))^2 + 2E(C_{k,n}) \lim_{\gamma \to 0} \frac{1}{n\gamma} E(\delta_{k,n+1}(n+1)). \end{aligned}$$

By (17), this is equivalent to

$$\mathrm{var}(C_{k,n}) \leq \mathrm{var}(C_{k-1,n-1}) + (E(C_{k,n}) - E(C_{k-1,n-1}))^2.$$

□

8 An explicit bound on the variance of C_n

By Proposition 3.1, we know the expected values of incomplete LP-relaxed matchings. Hence using Corollary 7.2 we recursively obtain an upper bound

on the variance of $C_{n/2,n} = C_n$. For even n,

$$\begin{aligned}\mathrm{var}C_n \leq \frac{1}{n^2}\left(1+\frac{1}{2}+\cdots+\frac{1}{n}\right)^2 \\ + \frac{1}{(n-1)^2}\left(\frac{1}{2}+\cdots+\frac{1}{n-1}\right)^2 \\ \vdots \\ + \frac{1}{(n/2+1)^2}\left(\frac{1}{n/2}+\frac{1}{n/2+1}\right)^2. \quad (18)\end{aligned}$$

for odd n, a similar formula holds, but this is not relevant for our application to the (non-relaxed) matching problem. It is straightforward to verify that (18) implies that

$$\mathrm{var}(C_n) = O\left(\frac{1}{n}\right), \quad (19)$$

but we shall be a little more careful.

Theorem 8.1.

$$\mathrm{var}(C_n) \leq \frac{2(n+3)^2}{(n+1)(n+2)^2}. \quad (20)$$

Proof. It follows from (18) that

$$\begin{aligned}\mathrm{var}C_n \leq \frac{1}{n^2}\left(\log(n+1/2)-\log(1/2)\right)^2 \\ + \frac{1}{(n-1)^2}\left(\log(n-1/2)-\log(3/2)\right)^2 \\ \vdots \\ + \frac{1}{(n/2+1)^2}\left(\log(n/2+3/2)-\log(n/2-1/2)\right)^2. \quad (21)\end{aligned}$$

Further simplifying, we obtain

$$\begin{aligned}\frac{(n/2+1)^2}{(n/2+3/2)^2}\cdot\mathrm{var}(C_n) \\ \leq \frac{\left(\log(n+1/2)-\log(1/2)\right)^2}{(n+1/2)^2} + \frac{\left(\log(n-1/2)-\log(3/2)\right)^2}{(n-1/2)^2} + \cdots \\ \cdots + \frac{\left(\log(n/2+3/2)-\log(n/2-1/2)\right)^2}{(n/2+3/2)^2}. \quad (22)\end{aligned}$$

Now we put

$$f(x) = \frac{(\log x - \log(n+1-x))^2}{x^2}.$$

We verify that f is a convex function on $0 < x < n+1$ by calculating the second derivative of the function $g(x) = (n+1)^2 f(x(n+1))$:

$$\frac{d^2g}{dx^2} = \big(6((1-x)\log(x) - (1-x)\log(1-x))^2 + \\ + 2(5-6x)(\log(1-x) - \log(x)) + 2\big)\frac{1}{x^4(1-x)^2}. \quad (23)$$

The only term in the parenthesis which is not positive is the second one. But it is only negative from $x_0 = 1/2$ to $x_4 = 5/6$. Let $x_1 = 1/(1+\exp(-1/2))$, $x_2 = 1/(1+\exp(-5/6))$ and $x_3 = 1/(1+\exp(-4/3))$. The linear factor and the logarithmic factor are both decreasing on the interval (x_0, x_4). Hence on (x_0, x_1) the negative term is bounded from below by:

$$2(5-6x_0)(\log(1-x_1) - \log(x_1)) = -2.$$

The positive terms in the parenthesis is bounded from below on (x_1, x_4) by:

$$6((1-x_1)\log(x_1) - (1-x_1)\log(1-x_1))^2 + 2 > 2.2.$$

It is then simple to bound the negative term as above for the given points and by this method confirming the convexity of the function:

$$2(5-6x_1)(\log(1-x_2) - \log(x_2)) > -2.11,$$

$$2(5-6x_2)(\log(1-x_3) - \log(x_3)) > -2.19,$$

$$2(5-6x_3)(\log(1-x_4) - \log(x_4)) > -0.82.$$

The convexity of f implies that the right hand-side of (22) can be estimated as:

$$f(n/2+3/2) + \cdots + f(n+1/2) \le \int_{n/2+1}^{n+1} f(x)\,dx \le \int_{n/2+1/2}^{n+1} f(x)\,dx \\ = \frac{1}{n+1}\int_{1/2}^{1} \frac{(\log y - \log(1-y))^2}{y^2}\,dy = \frac{2}{n+1}. \quad (24)$$

This establishes (20).

□

The integrand

$$\frac{(\log y - \log(1-y))^2}{y^2}$$

in (24) actually has the primitive

$$\frac{-(1-y)\left[(\log(1-y) - \log y)^2 - 2(\log(1-y) - \log y)\right] - 2}{y}.$$

We believe that the bound $\mathrm{var}(C_n) \le 2/n$ can be established from (21), but we have not been able to prove this.

9 Proof of Theorem 1.2

We can now prove Theorem 1.2. By equation (9) of [13, p10],

$$EM_n \le \frac{1}{2}\left(1 + \frac{1}{4} + \cdots + \frac{1}{(n/2)^2}\right) + \frac{1}{n+1}\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right).$$

Moreover, the alternating formula (6) for EC_n can be rewritten as

$$EC_n = \frac{1}{2}\left(1 + \frac{1}{4} + \cdots + \frac{1}{(n/2)^2}\right) + \frac{1}{(n/2+1)^2} + \cdots + \frac{1}{n^2}.$$

Hence

$$E(M_n - C_n) \le \frac{1}{n+1}\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right) - \left(\frac{1}{(n/2+1)^2} + \cdots + \frac{1}{n^2}\right).$$

By an integral estimate we further have

$$\begin{aligned}\frac{1}{(n/2+1)^2} + \cdots + \frac{1}{n^2} &\ge \int_{n/2+1}^{n+1} \frac{dx}{x^2} + \frac{1}{2}\left(\frac{1}{(n/2+1)^2} - \frac{1}{(n+1)^2}\right) \\ &= \frac{n(2n^2+9n+8)}{2(n+1)^2(n+2)^2} \ge \frac{1}{n+2}, \quad (25)\end{aligned}$$

provided $n^2 \ge 2n+4$, which holds when $n \ge 4$. Hence

$$E(M_n - C_n) \le \frac{1}{n+1}\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right) - \frac{1}{n+2}.$$

This inequality holds (actually with equality) also when $n = 2$.

Now we have all the necessary ingredients. It follows that

$$E\left|M_n - \frac{\pi^2}{12}\right| \le E\left|M_n - C_n\right| + E\left|C_n - \frac{\pi^2}{12}\right|$$

$$= (EM_n - EC_n) + E\left|C_n - \frac{\pi^2}{12}\right| \le (EM_n - EC_n) + \left|EC_n - \frac{\pi^2}{12}\right| + E\left|C_n - EC_n\right|$$

$$\le \frac{1}{n+1}\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right) - \frac{1}{n+2} + \frac{1}{(n+1)^2} + \sqrt{\mathrm{var}C_n}$$

$$\le \frac{1}{n+1}\left(1 + \frac{1}{2} + \cdots + \frac{1}{n+1}\right) - \frac{1}{n+2} + \frac{n+3}{n+2}\sqrt{\frac{2}{n+1}}. \quad (26)$$

The last expression of (26) is

$$\le \frac{1}{n+1}\left(1 + \frac{1}{2} + \cdots + \frac{1}{n+1}\right) + \sqrt{\frac{2}{n+1}}$$

$$\le \frac{1 + \log(n+1)}{n+1} + \sqrt{\frac{2}{n+1}} \le \frac{1}{\sqrt{n}}\left(\frac{1 + \log(n+1)}{\sqrt{n}} + \sqrt{2}\right). \quad (27)$$

We verify that for large n,

$$\frac{1 + \log(n+1)}{\sqrt{n}} + \sqrt{2} \le 1.706.$$

We have

$$\frac{1 + \log(n+1)}{\sqrt{n}} + \sqrt{2} \le \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n}} + \sqrt{2} + \frac{\log n}{n} < 1.48 + \frac{\log n}{\sqrt{n}}, \quad (28)$$

for $n \ge 1000$. By differentiating, we see that $\log n/\sqrt{n}$ is decreasing for $n \ge e^2$. Hence for $n > 1000$, (28) is smaller than

$$1.48 + \frac{\log 1000}{\sqrt{1000}} < 1.7.$$

For $n = 2, 4, \ldots, 1000$, we proceed by verifying the inequality (1) directly, using the strongest bounds available. We therefore quote an even stronger bound on EM_n from [13, p9]:

$$EM_n \le \frac{2}{n}\left(1+\frac{1}{3}+\frac{1}{5}+\cdots+\frac{1}{n-1}\right) + \frac{2}{n-1}\left(\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{n-2}\right)+ \vdots + \frac{2}{n/2+1}\cdot\frac{1}{n/2}. \quad (29)$$

Using the best bounds thus available, together with (26), we obtain

$$E\left|M_n-\frac{\pi^2}{12}\right| \le 8\cdot\sum_{k=1}^{n/2}\frac{1}{n+2k}\sum_{i=1}^{k}\frac{1}{n-2k+4i-2}-2\sum_{k=1}^{n}\frac{(-1)^{k-1}}{k^2}+\frac{\pi^2}{12} + 4\sqrt{\sum_{k=1}^{n/2}\frac{1}{(n+2k)^2}\cdot\left(\sum_{i=1}^{2k}\frac{1}{n-2k+2i}\right)^2}. \quad (30)$$

It is straightforward to check by computer that this is smaller than $1.706/\sqrt{n}$ for $n = 2, 4, \ldots, 1000$ (the bound on the constant is worst for $n = 16$). This establishes (1).

In view of what we have established so far, it seems more or less clear that the variance of M_n must be of order $1/n$. However, we have been unable to find a proof of this. Since the variance in the bipartite matching problem on $K_{n/2,n/2}$ is of order $1/n$ [11], it follows that the contribution to the variance of M_n from the "tail" of values larger than say 10 (any constant larger than $\pi^2/6$) is $O(1/n)$. From this we can establish that $\mathrm{var}(M_n) = O(\log n/n)$. The scenario that we cannot so far exclude is that with probability $\log n/n$, $M_n - C_n$ is of order 1, while in the remaining cases it is very small.

10 Proof of Theorem 1.3

In this section we justify our claim that the bound in equation (1) is essentially sharp, by proving Theorem 1.3. In the bipartite case the corresponding

statement is easily proved. By conditioning on the cost of the minimum edge from each vertex on one side of the graph, we see that the cost of the minimum assignment in $K_{n,n}$ is the sum of n independent exponentials of rate n plus a remainder which is independent of them all. Since the sum of exponentials has deviations of order $n^{-1/2}$ (see Lemma 10.1 below), we are done.

For the complete graph, things are not quite that simple. The problem is that we cannot partition the edge set into a large number of subsets for which a prescribed number of edges will participate in the minimum matching. Our idea is based on the same approach as in the bipartite case, but the remainder will not be independent of the sum of exponentials. Fortunately the correlation is positive and monotone, so that by applying the Harris inequality we can prove that the remainder cannot cancel the deviations of the exponential sum.

We prove Theorem 1.3 by establishing the inequality (valid for every x)

$$P\left(|M_n - x| \geq \frac{4}{9}\sqrt{\frac{2\pi}{n}}\right) \geq \frac{1}{8}. \tag{31}$$

Here $4\sqrt{2\pi}/9 \approx 1.114$, which leads to Theoem 1.3.

10.1 The distribution of a sum of independent exponentials

Lemma 10.1. *Let $X = X_1 + \cdots + X_{k+1}$ be the sum of $k+1$ independent* exp(1)*-variables. Let ζ be the median of X. Then*

$$P\left(X - \zeta \geq \frac{\sqrt{2\pi k}}{4}\right) \geq \frac{1}{4}, \tag{32}$$

and

$$P\left(X - \zeta \leq -\frac{\sqrt{2\pi k}}{4}\right) \geq \frac{1}{4}. \tag{33}$$

Proof. The density function of X is

$$\frac{x^k}{k!}e^{-x},$$

and the derivative of this is

$$\frac{x^{k-1}}{k!}e^{-x} \cdot (k-x),$$

which is zero when $x = k$. Hence the maximum value of the density is

$$\frac{k^k e^{-k}}{k!}.$$

By Stirling's formula,

$$k! \geq \sqrt{2\pi k} \cdot k^k e^{-k},$$

which implies that the density of X is at most

$$\frac{1}{\sqrt{2\pi k}}.$$

Hence

$$P\left(\zeta < X < \zeta + \frac{\sqrt{2\pi k}}{4}\right) \leq \frac{1}{4}.$$

The inequality (32) follows, and (33) is established in the same way as above. □

10.2 An operation on the cost matrix

Consider a random matching problem on K_n where the edge costs are non-negative linear combinations of a set of independent exponential random variables $X_1, \ldots, X_m$. We start with each edge cost given by a single exponential random variable. Let $k < n/2$ be a positive integer that will be chosen later as a function of n, and suppose that there is a set S of $n-k$ vertices such that every edge between two vertices of S has a variable X_i occurring with coefficient 1 in its cost, such that X_i does not occur (in other words has coefficient zero) in any other edge cost. Hence we can in the first step pick any set S with $n-k$ vertices. The following operation on the edge costs will be referred to as *reduction*: Choose the set S, and for every edge within S a variable X_i. Condition on the location of the minimum among the chosen variables, and subtract this minimum from all of them. Then add the same amount to all edges between vertices in $V - S$.

We first show that after reduction, the costs are still given by non-negative linear combinations of exponential variables. The variable being minimal

is replaced by a zero, and the minimum, which is itself exponentially distributed, is added to the edges in $V - S$. The variables that are not minimal are still independent and exponentially distributed, which means that they can be regarded as unchanged by the reduction.

Secondly, notice that if we start with independent exponential mean 1 edge costs, then reduction can be performed $k + 1$ times, since every round removes only one of the original variables. After k rounds, we can choose S by removing k vertices, one for each of the original variables that has disappeared.

Thirdly, notice how reduction affects the cost of the minimum matching. When the vertex set on n vertices is partitioned in two parts of sizes k and $n-k$, the larger part, with $n-k$ vertices, will always contain exactly $n/2-k$ more edges in any perfect matching than the smaller part on k vertices. Let $\xi_1, \ldots, \xi_{k+1}$ be the values of the minima obtained in the $k + 1$ reductions. Then

$$M_n = \left(\frac{n}{2} - k\right)(\xi_1 + \cdots + \xi_{k+1}) + R,$$

where R is the cost of the minimum matching given the edge costs after the $k+1$ rounds of reduction. The ξ_i's are exponential of rate $(n-k)(n-k-1)/2$.

We rescale by the factor

$$\left(\frac{n}{2} - k\right) \cdot \frac{2}{(n-k)(n-k-1)} = \frac{n-2k}{(n-k)(n-k-1)},$$

and apply Lemma 10.1 to conclude that if ζ is the median of $(n/2-k)(\xi_1 + \cdots + \xi_{k+1})$, then

$$P\left[\left(\frac{n}{2} - k\right)(\xi_1 + \cdots + \xi_{k+1}) - \zeta \geq \frac{\sqrt{2\pi k} \cdot (n-2k)}{(n-k)(n-k-1)}\right] \geq \frac{1}{4},$$

and similarly for deviations in the opposite direction.

Now we condition on the remaining random variables occurring in the edge costs, that is, those except the ξ_i's. Let ν be the median of R. Both $(n/2-k)(\xi_1 + \cdots + \xi_{k+1})$ and R are increasing as functions of $\xi_1, \ldots, \xi_{k+1}$. Therefore by the Harris inequality [3], the events

$$\left(\frac{n}{2} - k\right)(\xi_1 + \cdots + \xi_{k+1}) - \zeta \geq \frac{\sqrt{2\pi k} \cdot (n-2k)}{(n-k)(n-k-1)}$$

and

$$R \geq \nu$$

are positively correlated. We conclude that

$$P\left[M_n - (\zeta + \nu) \geq \frac{\sqrt{2\pi k} \cdot (n-2k)}{(n-k)(n-k-1)}\right] \geq \frac{1}{4} P(R \geq \nu) = \frac{1}{8},$$

and by the same argument that

$$P\left[M_n - (\zeta + \nu) \leq \frac{\sqrt{2\pi k} \cdot (n-2k)}{(n-k)(n-k-1)}\right] \geq \frac{1}{8}.$$

It follows that for *every* x,

$$P\left[|M_n - x| \geq \frac{\sqrt{2\pi k} \cdot (n-2k)}{(n-k)(n-k-1)}\right] \geq \frac{1}{8}.$$

Notice that this conclusion holds regardless of the values of the variables on which we have conditioned.

It remains to choose k as a function of n in order to maximize

$$\frac{\sqrt{k} \cdot (n-2k)}{(n-k)(n-k-1)}. \tag{34}$$

Fine tuning completely would give

$$\frac{k}{n} \approx \frac{\sqrt{17} - 3}{4},$$

but it suffices to plug in $k = n/4$ giving

$$\frac{\sqrt{k} \cdot (n-2k)}{(n-k)(n-k-1)} \geq \frac{\sqrt{k} \cdot (n-2k)}{(n-k)^2} = \frac{4}{9\sqrt{n}},$$

and $k = n/4 + 1/2$ giving

$$\frac{\sqrt{k} \cdot (n-2k)}{(n-k)(n-k-1)} = \frac{4\sqrt{n+2}}{3(3n-2)} \geq \frac{4}{9\sqrt{n}}$$

in order to conclude that we can always choose k such that (34) becomes at least $4/(9\sqrt{n})$. This establishes (31), except that we have to check the

case $n = 2$ separately, since this involved a division by zero in the algebraic simplification above. If $n = 2$ then $\frac{4}{9}\sqrt{\frac{2\pi}{n}}$ evaluates to

$$\frac{4\sqrt{\pi}}{9},$$

and checking the inequality (31) reduces to verifying that

$$\int_0^{8\sqrt{\pi}/9} e^{-x}\,dx \le \frac{7}{8}.$$

Indeed,

$$\int_0^{8\sqrt{\pi}/9} e^{-x}\,dx = 1 - e^{-8\sqrt{\pi}/9} \approx 0.793 \le \frac{7}{8}.$$

This shows that in (1), the right hand-side cannot be replaced by a function of smaller order, even if the number $\pi^2/12$ is replaced by a function of n.

11 A conjecture on asymptotic normal distribution

The inequalities (1) and (31) show that

$$0.139 \le \sqrt{n} \cdot E\left|M_n - \frac{\pi^2}{12}\right| \le 1.706. \tag{35}$$

Here the lower bound is $\sqrt{2\pi}/18 \approx 0.139$. We conjecture that

$$\sqrt{n}\left(M_n - \frac{\pi^2}{12}\right) \xrightarrow{d} N(0, 2\zeta(2) - 2\zeta(3)). \tag{36}$$

The evidence for this conjecture is quite strong. We believe that M_n is asymptotically normal, since the contributions to M_n from edges that are not connected by a short (order $1/n$) path are almost independent of each other. We also believe that the dependencies of the contributions of "nearby" edges will average out so that each edge that participates in the solution will contribute by a certain constant times $1/n^2$ to the variance. This constant should be the same in the complete graph K_n as in the complete bipartite

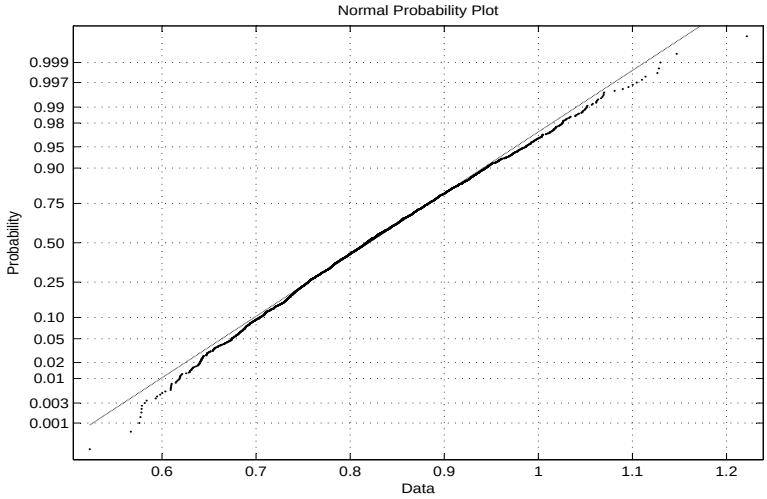
graph $K_{n,n}$. In the bipartite case, the variance is given by an explicit formula [11], and is asymptotically $n^{-1}(4\zeta(2) - 4\zeta(3))$. Since in K_n, a perfect matching has only $n/2$ edges, compared to n in the bipartite case, we expect the variance of M_n to be roughly half of the variance for perfect matching on $K_{n,n}$. This agrees very well with simulations up to $n = 100$. For larger n the computations needed to run the simulation makes it hard to get a sufficient number of samples. We have run simulations for n up to 400, but already at $n = 200$ we are forced to limit the sample size so much that the approximation of the variance of $C_{n/2,n}$ was still changing considerably when the simulation ended. But no simulation results contradict the conjecture (36).

Since for a $N(0, \sigma^2)$-variable X, the expectation of $|X|$ is $\sigma\sqrt{2/\pi}$, the conjecture (36) would suggest that:

$$\sqrt{n} \cdot E\left|M_n - \frac{\pi^2}{12}\right| \to \frac{\sqrt{2} \cdot \sqrt{2\zeta(2) - 2\zeta(3)}}{\sqrt{\pi}} = 2\sqrt{\frac{\pi}{6} - \frac{\zeta(3)}{\pi}} \approx 0.751,$$

well within the bounds given by (35). From our knowledge about the variance of the cost of a matching it is natural to guess that for large n, the cost will behave approximately as a random variable with normal distribution. But at the same time we can observe that the cost is strictly larger than zero, and therefore for any fixed n the cost cannot be a random variable with a normal distribution. It is therefore natural to ask how close the cost is to a normal random variable for a fixed n. By simulation we can observe that it is, apart from the expected truncation error, quite close.

We made a simulation with $n = 100$ and 2500 samples. By choosing a sufficiently large sample size we can make a very good approximation of the true distribution. Using Matlab we have made a so-called *normal probability plot* of the data. The 2500 numbers are sorted, and their values can be seen on the horizontal axis. The scaling of the vertical axis is such that a normal distribution would be plotted to a straight line.



The figure indicates that the sample of costs is taken from a distribution which is close to normal. The straight line represents a normal distribution with the same mean and variance as the plotted sample. In this case the data gives

$$\mathrm{var}(C_{50,100}) \approx 0.0088768.$$

This agrees well with the conjecture that the variance is asymptotically given by

$$\mathrm{var}(C_{n/2,n}) \approx \frac{2\zeta(2) - 2\zeta(3)}{n}.$$

This formula gives that

$$E(C_{50,100}) \approx 0.0088575.$$

Certainly it seems to be a hard problem how to investigate the explicit relation between the cost random variable and a normal distribution. But the scope of such an investigation might be far larger than just the matching problem described in this article.

References

[1] Aldous, David, *Asymptotics in the random assignment problem*, Probab. Theory Relat. Fields, **93** (1992) 507–534.

[2] Aldous, David, *The $\zeta(2)$ limit in the random assignment problem*, Random Structures & Algorithms **18** (2001), no 4. 381–418.

[3] Harris, T. E., *Lower bound for the critical probability in a certain percolation process*, Proc. Cambridge Phil. Soc. **56** (1960), 13–20.

[4] Häggström, O., *Problem solving is often a matter of cooking up an appropriate Markov chain*, manuscript 2007, to appear in Scandinavian Journal of Statistics.

[5] Mézard, Marc and Parisi, Giorgio, *Replicas and optimization*, Journal de Physique Lettres **46** (1985), 771–778.

[6] Mézard, Marc and Parisi, Giorgio, *Mean-field equations for the matching and the travelling salesman problems*, Europhys. Lett. **2** (1986) 913–918.

[7] Mézard, Marc and Parisi, Giorgio, *On the solution of the random link matching problems*, Journal de Physique Lettres **48** (1987),1451–1459.

[8] Mézard, M., Parisi, G. and Virasoro, M. A., Spin Glass Theory and Beyond, World Scientific, Singapore 1987.

[9] Parisi, Giorgio, *Spin glasses and optimization problems without replicas*, in Les Houches, Session 46, 1986 —Le hasard et la matière/Chance and matter (editors J. Souletie, J. Vannimenus and R. Stora), Elsevier Science Publishers B. V., Netherlands 1987.

[10] Talagrand, Michel, *Concentration of measure and isoperimetric inequalities in product spaces*, Inst. Hautes Études Sci. Publ. Math. **81** (1995), 73–205.

[11] Wästlund, J., *The variance and higher moments in the random assignment problem*, Linköping Studies in Mathematics No. 8, 2005.

[12] Wästlund, J,. *An easy proof of the zeta(2) limit in the random assignment problem*, submitted for publication, available at the author's webpage.

[13] Wästlund, J., *Random matching problems on the complete graph*, Electronic Communications in Probability **13** (2008), 258–265.

[14] Wästlund, J., *Optimization in mean field models*, extended abstract from the conference *Common Concepts in Statistical Physics and Computer Science*, ICTP, Trieste, Italy, July 2–6, 2007. Available through the ICTP webpage.

[15] Wästlund, J., *The mean field traveling salesman and related problems*, manuscript submitted for publication, available at the author's webpage.