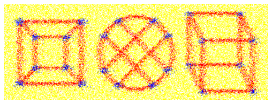
Journal of Graph Algorithms and Applications
http://jgaa.info/ vol. 20, no. 2, pp. 411–434 (2016)
DOI: 10.7155/jgaa.00400

A Binomial Distribution Model for the Traveling Salesman Problem Based on Frequency Quadrilaterals

Y. Wang[1] *J. B. Remmel*[2]

[1]School of Renewable Energy,
North China Electric Power University, Chang Ping, Beijing 102206. China
[2]Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112. USA

Abstract

We study the symmetric traveling salesman problem via frequency graphs. One computes the frequency of edges by computing how many times an edge occurs in an optimal path involving four vertices. The edges that are in the Optimal Hamiltonian Cycle (OHC) have a higher frequency than most edges that are not in the OHC and thus edges with a low frequency can safely be ignored when searching for the optimal solution. A binomial distribution model is introduced for the symmetric traveling salesman problem based on frequency quadrilaterals. When the frequency of each edge is computed with N frequency quadrilaterals, our model suggests that the minimum frequency of an edge in the OHC is $F_{\min} = (\epsilon_{\min}+1)N$ where $\frac{4}{3}+\frac{4}{3(n-2)} < \epsilon_{\min} < 4$. This suggests a heuristic to reduce the number of edges that need to be considered in the search for the OHC which is to keep only those edges whose frequencies are $\geq F_{\min}$. We explore this heuristic in several real-world examples.

Submitted:	Reviewed:	Revised:	Reviewed:	Revised:
December 2015	March 2016	April 2016	May 2016	May 2016
Reviewed:	Revised:	Accepted:	Final:	Published:
June 2016	June 2016	July 2016	July 2016	July 2016

Article type: Regular paper
Communicated by: D. Wagner

Research supported by Grant NSFC-51205129, Fundamental Research Funds for the Central Universities-2015ZD10

E-mail addresses: yongwang@ncepu.edu.cn (Y. Wang) jremmel@ucsd.edu (J. B. Remmel)

1 Introduction

We consider the symmetric traveling salesman problem (TSP). That is, we are given the complete graph K_n on the vertices $\{1, \ldots, n\}$ such that there is a distance function d such that for any $x, y \in \{1, \ldots, n\}$ and $x \neq y$, $d(x, y) = d(y, x)$ is the distance between x and y. The goal is to find the optimal Hamiltonian cycle (OHC) with respect to this distance function. That is, we want to find a permutation $\sigma = (\sigma_1 \ldots \sigma_n)$ of $1, \ldots, n$ such that $\sigma_1 = 1$ and $d(\sigma) := d(\sigma_n, 1) + \sum_{i=1}^{n-1} d(\sigma_i, \sigma_{i+1})$ is as small as possible. The TSP has been extensively studied to find special classes of graphs where polynomial-time algorithms exist for either finding an exact solution, that is, finding an OHC, or finding an approximate solution, that is, finding a permutation τ of the vertices which gives a Hamiltonian cycle such that $d(\tau) \leq cd(\sigma)$ where σ is the OHC and c is some fixed constant. We will call algorithms that find exact solutions *exact algorithms* and algorithms that find approximate solutions *approximation algorithms*. There are a number of special classes of graphs where one can find the OHC in a reasonable computation time, see [13].

Karp [17] has shown that the question of whether a graph has a Hamiltonian cycle is NP-complete which implies that TSP is NP-hard.

The computation time of exact algorithms is $O(a^n)$ for some $a > 1$ for the general TSP. For example, Held and Karp [14] and independently Bellman [3] gave a dynamic programming approach to solve the TSP that required $O(n^2 2^n)$ time. Integer programming techniques, such as either branch-and-bound [7, 10] or cutting-plane [18, 2], have been able to solve examples of the TSP with thousands of nodes. In 2006, a VLSI application (Euclidean TSP) with 85,900 nodes has been solved with an improved cutting-plane method on a computer system with 128 nodes [2].

On the other hand, the computation times of approximation algorithms and heuristics have been significantly improved [20]. For example, the MST approximation algorithm [8] and Christofides' approximation algorithm [16] are able to find the 2-approximation and $\frac{3}{2}$-approximation in time $O(n^2)$ and $O(n^3)$, respectively, for metric TSP. In 2011, Mömke and Svensson [22] gave a 1.461-approximation algorithm for metric graphs with respect to the Held-Karp lower bound. In most cases, the LKH heuristic [15] can generate "high quality" solutions within 5% of the optimum in nearly $O(n^{2.2})$ time. However, these approximation algorithms and heuristics are not guaranteed to find the OHC in polynomial time.

In recent years, researchers have developed polynomial-time algorithms to solve the TSP for sparse graphs. In sparse graphs, the number of Hamiltonian cycles (HC) is greatly reduced. For example, Sharir and Welzl [25] proved that in a sparse graph of average degree d, the number of HCs is less than $e(\frac{d}{2})^n$ where e is the base of the natural logarithm. Gebauer [12] gave a lower bound for the number of HCs roughly as $(\frac{d}{2})^n$ for a sparse graph of average degree d. In addition, Björklund [4] proved that the TSP in graphs with bounded degree could be solved in time $O((2-\epsilon)^n)$, where ϵ depends on the maximum degree of a vertex in the graph. For a cubic graph, Eppstein [11] introduced an

algorithm to solve the TSP with running time $O(1.260^n)$. This run time was improved by Liśewicz and Schuster [19] to $O(1.2553^n)$. Aggarwal, Garg and Gupta [1] and Boyd, Sitters, Van der Ster and Stougie [6] independently gave two approximation algorithms to solve the TSP with approximation factor $\frac{4}{3}$ for metric cubic graphs. Mömke and Svensson [22] also proved one $\frac{4}{3}$-approximation for degree three bounded and claw-free graphs with respect to the Held-Karp lower bound. For cubic connected graphs, Correa, Larré and Soto [9] proved that the approximation threshold of the TSP in cubic graphs was strictly below $\frac{4}{3}$. For the general bounded-genus graphs, Borradaile, Demaine and Tazari [5] gave a polynomial-time approximation scheme for TSP. In the case of the asymmetric version of the TSP, Gharan and Saberi [23] designed constant-factor approximation algorithms for the TSP for planar graphs with bounded genus where the constant-factor is $22.51(1+\frac{1}{n})$. Thus, whether one is trying to find exact solutions or approximate solutions to the TSP, one has a variety of more efficient algorithms available if one can reduce a given instance of TSP to finding the OHC in a sparse graph.

In this paper, we use a binomial distribution model based on frequency quadrilaterals to convert a complete graph (or dense graph) into a sparse graph for TSP. The sparse graphs generally have $O(|V|)$ or $O(|V|\ln(|V|))$ edges. In addition, if the resulting graph has bounded degree or genus or is planar or k-edge connected, then there are even more efficient algorithms available to find exact or approximate solutions to the TSP.

In previous work [26, 27, 30], the first author introduced frequency graphs as a way to reduce the number of edges that one has to consider to find the OHC. The basic idea of frequency graphs is the following. Suppose that we are given a sequence of k vertices $\vec{v} = (v_1, v_2 \ldots, v_{k-1}, v_k)$ in K_n where $k \geq 4$. Let $\vec{v}_\sigma = (v_{\sigma_1}, \ldots, v_{\sigma_k})$ be a permutation of $v_1, \ldots, v_k$ where $v_{\sigma_1} = v_1$ and $v_{\sigma_k} = v_k$ and let $d(\vec{v}_\sigma) = \sum_{i=1}^{k-1} d(v_{\sigma_i}, v_{\sigma_{i+1}})$. We assume that $d(\vec{v}_\sigma)$ all have different values, so there is an optimal path $\vec{v}_\sigma = (v_{\sigma_1}, \ldots, v_{\sigma_k})$ of the k vertices (or $k-1$ edges) connecting v_1 and v_k using the intermediate vertices $v_2, \ldots, v_{k-1}$ which makes $d(\vec{v}_\sigma)$ as small as possible. We will call such a path the *optimal k-vertex path* for $(v_1, \ldots, v_k)$. In general, if we are given a set of k vertices $\{v_1, \ldots, v_k\}$, we have $\binom{k}{2}$ ways to pick the end points of a k-vertex path using these vertices so that there are $\binom{k}{2}$ optimal k-vertex paths that arise from the set $\{v_1, \ldots, v_k\}$.

Let $\mathcal{OP}^k$ denote the set of all optimal k-vertex paths. Then the frequency $f(x, y)$ of an edge (x, y) in K_n with the distance function $d(x, y)$ is the number of optimal k-vertex paths which contain (x, y) as an edge. Our intuition is that if (x, y) is an edge in the OHC for K_n with the distance function $d(x, y)$, then its frequency is likely to be much higher than the average frequency. This has been born out by studying many real-world TSP instances, see [26, 27, 30]. This suggests that we can safely eliminate the edges of low frequency below the average frequency and still keep the OHC intact. The hope is that by eliminating the edges of low frequency, we can be left with a sparse graph which has $O(n\log(n))$ edges so that the techniques for either finding or approximating the OHC for sparse graphs can be applied. The questions then become what

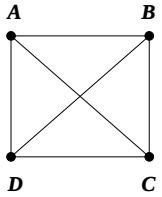

Figure 1: The quadrilateral $ABCD$

value of k should we use and what bound on the frequency should we use to eliminate edges.

The outline of this paper is as follows. First in Section 2, we shall introduce the concept of frequency quadrilaterals in the case where $k = 4$. In Section 3, we shall first discuss the combinatorics of frequency quadrilaterals. Then we shall introduce our binomial distribution model and study the combinatorics of frequency quadrilaterals for edges in the OHC. In Section 4, we shall discuss some heuristic estimates of various parameters of frequency graphs under our binomial distribution model. In Section 5, we will compare our heuristic estimates of such parameters to the actual values of those parameters that are computed for some graphs in the database [24].

2 The frequency quadrilateral

Suppose that we are given 4 vertices $\{A, B, C, D\}$ in K_n. Since we are assuming that the vertex set of $K_n = \{1, 2, \ldots, n\}$, there is a total order on the elements in $\{A, B, C, D\}$ induced by the natural ordering on $\{1, \ldots, n\}$, which we will assume to be $A < B < C < D$. K_n restricted to $\{A, B, C, D\}$ gives us the graph pictured in Figure 1. We will list the pairs of endpoints according to their lexicographic order to find the six optimal 4-vertex paths.

For any pair of vertices $U, V \in \{A, B, C, D\}$, we shall write UV for $d(U, V)$. There are $\binom{4}{2} = 6$ ways to pick end points of 4-vertex paths using $\{A, B, C, D\}$, namely, (I) A, B, (II) A, C, (III) A, D, (IV) B, C, (V) B, D, and (VI) C, D. Then, for example, there are two 4-vertex paths with end points A, B, namely, (A, C, D, B) and (A, D, C, B). To pick the optimal 4-vertex path with endpoints A, B, we must compare $AC + CD + DB$ and $AD + DC + CB$. Since we are assuming that we are in the symmetric TSP, we know that $CD = DC$. Thus, we must compare $AC + BD$ and $AD + BC$ to determine which one of the two 4-vertex paths (A, C, D, B) and (A, D, C, B) is the optimal 4-vertex path. Below we list the comparisons that we must make for each of the cases (I)-(VI).

Case	End points	Sum 1	Sum 2
I	A, B	$AC + BD$	$AD + BC$
II	A, C	$AB + CD$	$AD + BC$
III	A, D	$AB + CD$	$AC + BD$
IV	B, C	$AB + CD$	$AC + BD$
V	B, D	$AB + CD$	$AD + BC$
VI	C, D	$AC + BD$	$AD + BC$

It is easy to see that our comparisons involve only three sums of distances, namely, (1) $AB + CD$, (2) $AC + BD$, and (3) $AD + BC$. Thus, the relative frequency graph for the quadrilateral $ABCD$ depends only on the relative order of (1), (2), and (3). For example, if $AB+CD < AC+BD < AD+BC$, then the optimal 4-vertex paths for our six possible end points are given in the following table.

Case	End points	Inequality formula	Optimal 4-vertex path
I	A, B	$AC + BD < AD + BC$	(A, C, D, B)
II	A, C	$AB + CD < AD + BC$	(A, B, D, C)
III	A, D	$AB + CD < AC + BD$	(A, B, C, D)
IV	B, C	$AB + CD < AC + BD$	(B, A, D, C)
V	B, D	$AB + CD < AD + BC$	(B, A, C, D)
VI	C, D	$AC + BD < AD + BC$	(C, A, B, D)

These choices lead to the relative frequency graph for the quadrilateral $ABCD$ pictured in Figure 2 (a).

On the other hand, if $AB+CD < AD+BC < AC+BD$, then the optimal 4-vertex paths for our six possible end points are given in the following table.

Case	End points	Inequality formula	Optimal 4-vertex path
I	A, B	$AD + BC < AC + BD$	(A, D, C, B)
II	A, C	$AB + CD < AD + BC$	(A, B, D, C)
III	A, D	$AB + CD < AC + BD$	(A, B, C, D)
IV	B, C	$AB + CD < AC + BD$	(B, A, D, C)
V	B, D	$AB + CD < AD + BC$	(B, A, C, D)
VI	C, D	$AD + BC < AC + BD$	(C, B, A, D)

These choices lead to the relative frequency graph for the quadrilateral $ABCD$ pictured in Figure 2 (b). One can carry out similar computations for the other 4 possible orderings of $AB+CD$, $AC+BD$, and $AD+BC$. They are called four-vertex and three-line inequalities [29] to derive the optimal 4-vertex paths for any given four vertices A, B, C, D in K_n. We list the resulting frequency graphs in each case according to the corresponding four-vertex and three-line inequalities, see Figure 2 (c)-(f).

We will study the properties of frequency graphs on K_n. Let us denote by N_z the number of frequency graphs in K_n such that the relative frequency graph is of type (z), for $(z) \in \{(a), (b), (c), (d), (e), (f)\}$ in Figure 2 and note that

$$N_z = \frac{1}{6}\binom{n}{4}.$$

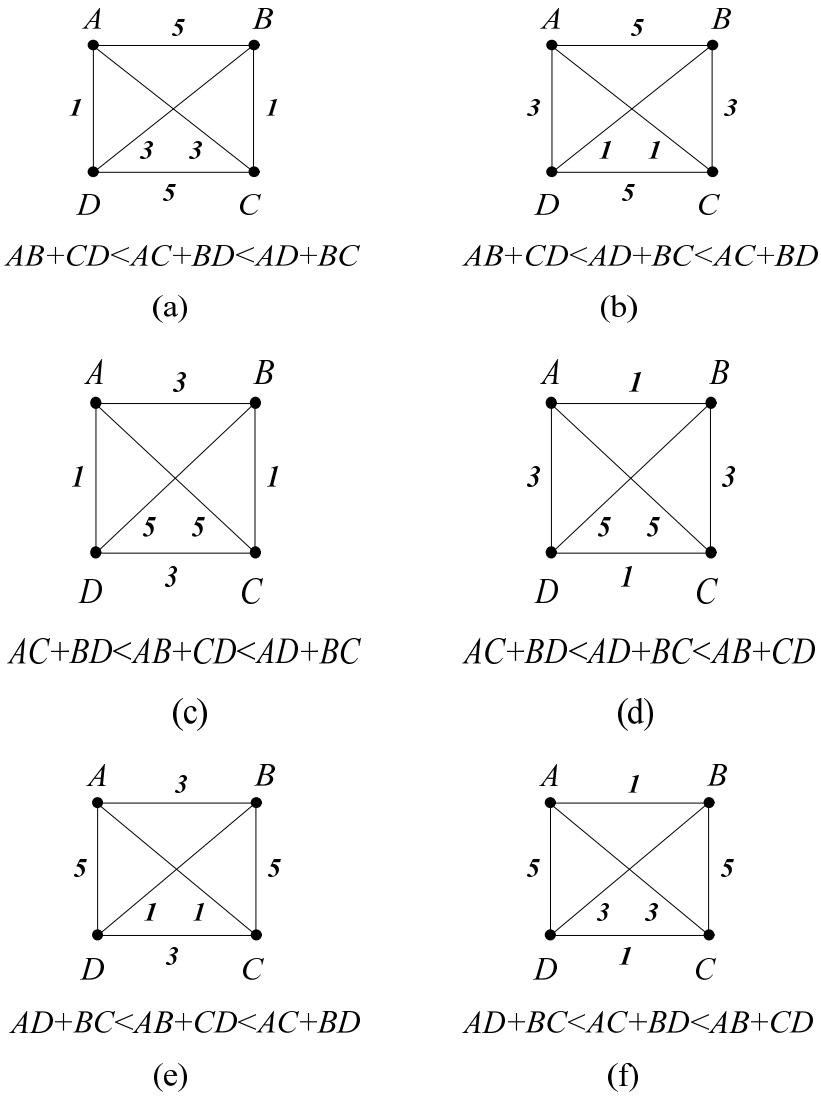

Figure 2: The six frequency quadrilaterals $ABCD$ in view of four-vertex and three-line inequality arrays in quadrilaterals $ABCD$.

We shall call this model the binomial distribution model. As we will see, this binomial distribution model suggests a heuristic on what bounds on the frequency graph should be used to eliminate edges according to their frequency.

3 The binomial distribution model

Our binomial distribution model for frequency graphs is to consider picking for each set of four vertices A, B, C, D in K_n a total order on the sums of the distances $AD + BC$, $AB + CD$, and $AC + BD$ at random. Then we want to study the properties of frequency graphs that arise from such random choices. For example, in cases (a)-(f) of Figure 2, we picture the six relative frequency graphs that arise from the six ways to put a total order on the $AD + BC$, $AB + CD$, and $AC + BD$. Note that of the six possible frequency quadrilaterals that are possible for $ABCD$, one sees that the possible frequency for any given edge e is 1, 3, or 5. Moreover, e is assigned frequency 1 in 2 cases, frequency 3 in 2 cases, and frequency 5 in 2 cases. Thus, the average frequency assigned to e over these six frequency graphs is 3. For $i \in \{1, 3, 5\}$, let $p_i(e)$ be the probability that edge e is assigned frequency i in a frequency quadrilateral $ABCD$ containing the edge e. Clearly, $p_1(e) = p_3(e) = p_5(e) = \frac{1}{3}$. More generally, for any subset $S \subseteq \{1, 3, 5\}$, we let $p_S(e)$ denote the probability that edge e is assigned any frequency i where $i \in S$ in a frequency quadrilateral $ABCD$ containing the edge e. Then $p_{\{1,3\}}(e) = p_{\{1,5\}}(e) = p_{\{3,5\}}(e) = \frac{2}{3}$ and $p_{\{1,3,5\}}(e) = 1$.

Next we want to study the expected frequency of edges e that are in the OHC under such a probability model.

First, suppose that A, B, C, D are consecutive vertices in the OHC. In that case, we know that path (A, B, C, D) must have smaller weight than path (A, C, B, D) which implies that $AB + BC + CD < AC + BC + BD$ and hence $AB + CD < AC + BD$. In the three frequency quadrilaterals in Figure 2 for the quadrilateral $ABCD$ where $AB + CD < AC + BD$, we see that the frequencies assigned to the edges (A, B), (B, C), and (C, D) are $5, 1, 5$, $5, 3, 5$, and $3, 5, 3$, respectively. Thus, the average frequency of (A, B) is $\frac{13}{3}$, the average frequency of (B, C) is 3, and the average frequency of (C, D) is $\frac{13}{3}$. For any given edge e in the OHC, e will be an edge in three optimal 4-vertex paths with consecutive edges in the OHC so that the total contribution to its frequency from the three optimal 4-vertex paths in OHC is $\frac{13}{3} + 3 + \frac{13}{3} = \frac{35}{3} = 11\frac{2}{3}$ as opposed the expected value of 9 for an edge that appears in 3 random quadrilaterals.

Second, suppose that (A, B) and (C, D) are two vertex-disjoint edges on the OHC. That is, we have the situation pictured in Figure 3. In this situation, we note that there is a Hamiltonian cycle which starts at vertex A and follows that OHC in clockwise direction to vertex D, then uses the edges (B, D), then follows the OHC in a counter-clockwise direction to vertex C, and then uses the edge (A, C). Since this Hamiltonian cycle is not the OHC, it must be the case that $AB + CD < AC + BD$.

If one looks at the three quadrilaterals $ABCD$ for which this inequality holds in Figure 2, one finds that the frequencies for the edge (A, B) are 5, 5

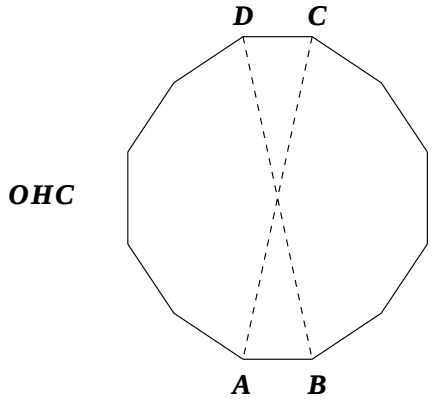

Figure 3: Two vertex-disjoint edges in the OHC.

and 3, respectively, so that the summed frequency for the 3 quadrilaterals of this form is 13 as opposed to the expected value of 9. Note that there are n edges in the OHC. Since we are assuming that the edges (A, B) and (C, D) have no vertices in common, then (C, D) cannot be one of the edges adjacent to (A, B) in the OHC. Hence we have $n - 3$ choices for (C, D). Thus, for an edge $(A, B) \in OHC$, it is contained in at least $n - 3$ such quadrilaterals which are composed of (A, B) and the other $n - 3$ non-adjacent edges in the OHC.

Note for any given edge (A, B), (A, B) is part of $\binom{n-2}{2}$ quadrilaterals in K_n. Using the OHC, we have found at least $n-3$ pairs (C, D) where the frequency of (A, B) relative to the quadrilateral $ABCD$ is either 3 or 5. Assuming that for the remaining choices of quadrilaterals, the probability $p_1(e)$ that $e = (A, B)$ has frequency 1 in each quadrilateral is $\frac{1}{3}$, the probability $p_3(e)$ that (A, B) has frequency 3 in each quadrilateral is $\frac{1}{3}$, and the probability $p_5(e)$ that (A, B) has frequency 5 in each quadrilateral is $\frac{1}{3}$, we see that

$$p_{\{3,5\}}(e) = \frac{\frac{2}{3}(\binom{n-2}{2} - (n-3)) + (n-3)}{\binom{n-2}{2}} = \frac{2}{3} + \frac{2}{3(n-2)}.$$

Note that this is a very conservative lower bound since we did not take into account the 3 possibilities where e is part of three consecutive edges in the OHC. Thus, we shall assume that the probabilities $p_{\{3,5\}}(e)$ and $p_{\{1\}}(e)$ for an edge e in the OHC are equal to the formula (1).

$$\begin{aligned} p_{\{3,5\}}(e) &= \frac{2}{3} + \frac{2}{3(n-2)} \text{ and} \\ p_1(e) &= \frac{1}{3} - \frac{2}{3(n-2)} \end{aligned} \tag{1}$$

We let X denote the random variable which gives the number of frequency quadrilaterals where the frequency of edge $e = (A, B)$ is either 5 or 3. Our

assumptions mean that if we select N quadrilaterals from the $\binom{n-2}{2}$ quadrilaterals which contain the pair (A, B), then X has the binomial distribution $X \backsim B_0(N, p_{\{3,5\}}(e))$. In such a situation, the probability $P(X = m)$ that $X = m$ is given by formula (2).

$$P(X = m) = \binom{N}{m}(p_{\{3,5\}})^m(1 - p_{\{3,5\}})^{N-m} \tag{2}$$

For a binomial distribution model, the function $P(X = m)$ is monotone increasing if $m < (N + 1)p_{\{3,5\}} - 1$ and is monotone decreasing if $m > (N + 1)p_{\{3,5\}}$. Thus, the maximum probability P_0 is achieved for an integer m when m equals

$$m_0 = \lfloor (N + 1)p_{\{3,5\}} \rfloor = \left\lfloor \left(\frac{2}{3} + \frac{2}{3(n-2)} \right)(N + 1) \right\rfloor$$

or

$$m_0 = \lceil (N + 1)p_{\{3,5\}} \rceil - 1 = \left\lceil \left(\frac{2}{3} + \frac{2}{3(n-2)} \right)(N + 1) \right\rceil - 1.$$

Thus, if we select N frequency quadrilaterals containing the edge (A, B) at random, we see that the case where there are m_0 frequency quadrilaterals with the frequency of edge (A, B) being greater than or equal to 3 has the maximum probability. In these m_0 frequency quadrilaterals, we assume that the number of frequency quadrilaterals with the frequency of (A, B) equal to 5 also has a binomial distribution $X \backsim B(m_0, \delta_0)$ where $0 \leq \delta_0 \leq 1$ is the ratio between the number of frequency quadrilaterals with the frequency of (A, B) equal to 5 and m_0. Thus, if $X = \lfloor \delta_0(m_0 + 1) \rfloor$ or $\lceil \delta_0(m_0 + 1) \rceil - 1$, the maximum probability will be obtained. For any edge e in the OHC, e is contained in $n - 3$ frequency quadrilaterals consisting of the vertices of e and the vertices of another edge f in the OHC and we assume that in those frequency quadrilaterals, e has equal probability of having frequency 5 or 3. Given the possible relative frequency graphs pictured in Figure 2, it is easy to see that $\delta_0 = \frac{1}{2}$ on average in our binomial distribution model.

If we use N random quadrilaterals to compute the frequency of $e = (A, B)$ in the OHC, its total frequency will be equal to formula (3) where $\epsilon_0 = \dfrac{4(1 + \delta_0)(n - 1)}{3(n - 2)}$.

$$\begin{aligned}
F_0 &= N(p_1(e) + 3(1 - \delta_0)p_{\{3,5\}} + 5\delta_0 p_{\{3,5\}}) \\
&= N\left(\frac{1}{3} - \frac{2}{3(n-2)} + 3(1 - \delta_0)\left(\frac{2}{3} + \frac{2}{3(n-2)}\right) + 5\delta_0\left(\frac{2}{3} + \frac{2}{3(n-2)}\right)\right) \\
&= N\left(\frac{1}{3} - \frac{2}{3(n-2)} + 3\left(\frac{2}{3} + \frac{2}{3(n-2)}\right) + 2\delta_0\left(\frac{2}{3} + \frac{2}{3(n-2)}\right)\right) \\
&= N\left(1 + 4(1 + \delta_0)\left(\frac{1}{3} + \frac{1}{3(n-2)}\right)\right) \\
&= (\epsilon_0 + 1)N
\end{aligned} \tag{3}$$

The minimum frequency $F_{\min}$ of an OHC edge is given by formula (4) where $\epsilon_{\min} = \dfrac{4(1+\delta_{\min})(n-1)}{3(n-2)}$.

$$F_{\min} = (\epsilon_{\min} + 1)N \tag{4}$$

In the worst case, $\delta_{\min} = 0$ which would mean that all m_0 frequency quadrilaterals would assign the frequency of (A, B) to be 3 which means that

$$\left(\frac{7}{3} + \frac{4}{3(n-2)}\right) N$$

is a lower bound for $F_{\min}$.

For edges (A, B) in the OHC, computational evidence suggests that $\delta_{\min}$ is approximately $\frac{1}{2}$ or larger. If $\delta_{\min} = \frac{1}{2}$, then $\epsilon_{\min} = 2 + \frac{2}{n-2}$ so that $F_{\min} = \left(3 + \frac{2}{n-2}\right) N$ which is bigger than the expected frequency $F_{\text{avg}} = 3N$. This is the intrinsic reason that the frequencies of the OHC edges computed with optimal 4-vertex paths for the examples in [30] are much bigger than those of most of the other edges. However, one can construct examples of TSP where

$$\left(\frac{7}{3} + \frac{4}{3(n-2)}\right) N \leq F_{\min} < \left(3 + \frac{2}{n-2}\right) N.$$

However, the probability that the minimum frequency of $(A, B) \in OHC$ equal to $(\frac{7}{3} + \frac{4}{3(n-2)})N$ approaches 0 as n approaches infinity.

For reasonable-sized graphs, such as the instances appearing in the database [24], one can compute the probabilities of the frequency of a given edge e using all $\binom{n-2}{2}$ quadrilaterals containing e. If N_i is the number of quadrilaterals containing e where the frequency is i for $i \in \{1, 3, 5\}$, then

$$\begin{aligned} p_5(e) &= \frac{2N_5}{(n-2)(n-3)}, \\ p_3(e) &= \frac{2N_3}{(n-2)(n-3)}, \text{ and} \\ p_1(e) &= \frac{2N_1}{(n-2)(n-3)}. \end{aligned}$$

Thus, when N frequency quadrilaterals are chosen at random, the total frequency F of e is given by formula (5).

$$\begin{aligned} F &= N(p_1(e) + 3p_3(e) + 5p_5(e)) \\ &= N(p_1(e) + p_3(e) + p_5(e) + 2p_3(e) + 4p_5(e)) \\ &= N(1 + 2p_3(e) + 4p_5(e)) \\ &= (\epsilon + 1)\, N \end{aligned} \tag{5}$$

Where

$$\begin{aligned}\epsilon &= 2p_3(e) + 4p_5(e)\\ &= 2(p_3(e) + p_5(e)) + 2p_5(e)\\ &= 2(1+\delta)p_{\{3,5\}}\end{aligned}$$

and $\delta = \frac{N_5}{N_3+N_5} = \frac{p_5(e)}{p_3(e)+p_5(e)}$. Here we assume $N_3+N_5 \neq 0$ and $p_3(e)+p_5(e) \neq 0$ for edge e. The number N_5 clearly plays a fundamental role in determining ϵ and F. As N_5 increases, both ϵ and F increase. In the extreme case where $N_5 = \binom{n-2}{2}$, $\epsilon = 4$. Thus, ϵ always lies in the interval $[0, 4]$. Based on the binomial distribution model (2) where $p_{\{3,5\}} = \frac{2}{3} + \frac{2}{3(n-2)}$, we know $P(\epsilon = 0) = P(X = 0)$ or $P(\epsilon = 4) = P(X = N_5 = \binom{n-2}{2})$ approaches zero for big n. This means that the number of edges with $\epsilon \approx 0$ or $\epsilon \approx 4$ have a very small probability. On average, when $p_3(e) = p_5(e) = \frac{1}{3} + \frac{1}{3(n-2)}$ for an edge e in the OHC, it follows that the ϵs of the OHC edges will be bigger than $2 + \frac{2}{n-2}$. Computational evidence from the graphs in [30] suggests that $\epsilon_{\min}$ is bigger than $2 + \frac{2}{n-2}$ due to the fact that N_5 is large.

This suggests the following criterion to determine whether a given edge e is likely to be in the OHC. That is, we should compare ϵ and $\epsilon_{\min}$. For any given edge e, if $\epsilon > \epsilon_{\min} = \frac{4(1+\delta_{\min})(n-1)}{3(n-2)}$, then $F > F_{\min}$ and e is more likely to be in the OHC. For example, if $\delta_{\min} \geq 0.5$, then the criterion becomes that $\epsilon > 2$. This suggests that we can safely trim the edges e with $\epsilon < 2$ and still keep the edges in the OHC. In Section 5, we shall give several examples to show what happens using this criterion for TSP instances in the database [24].

4 Some heuristic for the binomial distribution model

Recall that under our binomial distribution model, for each set of four vertices A, B, C, D of K_n, we are essentially picking one of the six relative frequency quadrilaterals pictured in Figure 2 (a)-(f) at random. If we compute many frequency graphs where each frequency graph is computed with N random frequency quadrilaterals with edge e, then the cumulative probability $P(X \leq m)$ of m frequency quadrilaterals where the frequency f associated with e is either 5 or 3 is computed with formula (6).

$$P(X \leq m) = \sum_{k=0}^{m} \binom{N}{k} (p_{\{3,5\}})^k (1 - p_{\{3,5\}})^{N-k} \tag{6}$$

If N is large, then $X \backsim B(N, p_{\{3,5\}})$ approximately follows a normal distribution $X \backsim \mathcal{N}(Np_{\{3,5\}}, Np_{\{3,5\}}(1 - p_{\{3,5\}}))$ because $Np_{\{3,5\}} > 5$ and $N(1 - p_{\{3,5\}}) > 5$. In this case, $P(X \leq m)$ will approach 1 when $m = Np_{\{3,5\}} + 3\sigma$, where $\sigma = \sqrt{Np_{\{3,5\}}(1 - p_{\{3,5\}})}$ is the standard deviation.

For the edges e with $p_{\{3,5\}} > \frac{2}{3} + \frac{2}{3(n-2)}$, the number of frequency quadrilaterals where the frequency of e is either 5 and 3 will be bigger than m_0. Their frequencies F computed with N frequency quadrilaterals will be bigger than $F_{\min}$. The cumulative probability $P(X \geq m_0)$ is computed as formula (7).

$$P(X \geq m_0) = \sum_{k=m_0}^{N} \binom{N}{k} (p_{\{3,5\}})^k (1 - p_{\{3,5\}})^{N-k} \tag{7}$$

The bigger the difference between $p_{\{3,5\}}$ and $\frac{2}{3} + \frac{2}{3(n-2)}$, the closer the probability $P(X \geq m_0)$ approaches 1. For the edges e with $p_{\{3,5\}}$ above $\frac{2}{3} + \frac{2}{3(n-2)}$, F has a high probability of being above $F_{\min}$ if it is computed with the same number of random frequency quadrilaterals. Meanwhile, these edges with big $p_{\{3,5\}}$ will have a small probability according to the binomial distribution (2).

We have seen that for edges e in OHC, their $p_{\{3,5\}}$s are on average bigger than the expected value of $p_{\{3,5\}}$ which is $\frac{2}{3}$. On the other hand, the edges e with $p_{\{3,5\}}$ below $\frac{2}{3} + \frac{2}{3(n-2)}$ have a small probability that their frequency F is above $F_{\min}$. The bigger the difference between $\frac{2}{3} + \frac{2}{3(n-2)}$ and $p_{\{3,5\}}$ in such cases, the closer the probability $P(X \geq m_0)$ approaches 0. For most of the edges not in the OHC, their $p_{\{3,5\}}$s are generally smaller than the average probability $\frac{2}{3}$.

Next, we consider the edges e with frequency above the average frequency. In view of the six frequency quadrilaterals, we know that the expected value of $p_{\{3,5\}}$ is $\frac{2}{3}$. In other words, every edge has the probability $\frac{2}{3}$ that its frequency is bigger than the average frequency 3 in a frequency quadrilateral in K_n. Consider the event that the total frequency F of e is greater than $3N$ where N represents the number of random frequency quadrilaterals with edge e which we denote by $P(F > 3N)$. The expected value of $P(F > 3N)$ is $\frac{2}{3}$ over all the $\binom{n}{2}$ edges. This suggests that we can throw away $\frac{1}{3}\binom{n}{2}$ edges with small frequency. As $n \to \infty$, the number of edges with F above $3N$ conforms to the normal distribution $\mathcal{N}(\frac{2}{3}\binom{n}{2}, \frac{2}{9}\binom{n}{2})$. This suggests that we should select only the $\frac{2}{3}\binom{n}{2}$ edges with top frequency in our search for a solution to TSP.

We also want to estimate the number of edges e such that their cumulative frequencies F_e satisfy $F_e > F_{\min}$ when we compute such frequencies with N random frequency quadrilaterals containing the edge e. If there are K such edges e where $F_e > F_{\min}$, the number of edges e with $F_e \leq F_{\min}$ will be $R = \binom{n}{2} - K$. Note that the total number of frequency quadrilaterals chosen is $\binom{n}{2}\frac{N}{6} = \frac{n(n-1)}{12}N$. Let $\overline{F_K}$ and $\overline{F_R}$ denote the average frequency of the K edges e with $F_e > F_{\min}$ and the average frequency of the R edges e with $F_e \leq F_{\min}$. Note that the six possible quadrilaterals containing vertices A, B, C, D give a cumulative frequency of 18 to each quadrilateral. It follows that the formula (8) holds.

$$\frac{18n(n-1)}{12}N = K\overline{F_K} + \frac{n}{2}(n - 1 - \frac{2K}{n})\overline{F_R} \tag{8}$$

If we let $\overline{F_K} = (1 + \overline{\epsilon_K})N$ and $\overline{F_R} = (1 + \overline{\epsilon_R})N$ as in equation (5), then we can substitute these expressions into (8) and solve for K in which case we see that the formula (9) is derived.

$$K = \frac{2 - \overline{\epsilon_R}}{\overline{\epsilon_K} - \overline{\epsilon_R}} \binom{n}{2} \tag{9}$$

For a given instance of the TSP with n vertices, we can fix some ordering of the edges e_k for $1 \leq k \leq \binom{n}{2}$ and compute $\epsilon_k = \frac{4\left(1+\frac{1}{\delta_k}\right)N_5}{(n-2)(n-3)} = 2(1 + \delta_k)p_{\{3,5\}}$ where the frequency of e_k is given by $F_{e_k} = (1 + \epsilon_k)N$. In fact, we will assume that we have fixed an ordering where the sequence $(\epsilon_1, \epsilon_2, \cdots, \epsilon_k, \cdots, \epsilon_{\binom{n}{2}})$ is weakly decreasing. In this case, $\sum_{k=1}^{\binom{n}{2}} \epsilon_k = n(n-1)$. Because we are assuming that $\overline{\epsilon_R} < 2 < \overline{\epsilon_K}$, it will follow that $K \leq \binom{n}{2}$. In addition, $K < \frac{n(n-1)}{4}$ if $\overline{\epsilon_K} + \overline{\epsilon_R} > 4$.

The sum of the monotone sequence $(\epsilon_1, \epsilon_2, \cdots, \epsilon_k, \cdots, \epsilon_{\binom{n}{2}})$ is equal to $n(n-1)$. The expected value $\mu(\epsilon)$ is equal to 2. When n is large, we expect that ϵ_k decreases relatively smoothly for $1 \leq k \leq \binom{n}{2}$. Indeed, if we form the graph of all points (k, ϵ_k) for $1 \leq k \leq \binom{n}{2}$, we should expect that the graph is almost flat in any given small interval. This suggests that we can roughly approximate ϵ_k with the linear function $\epsilon(k) = -\frac{8}{n(n-1)}k + 4$ so that $\epsilon_1 = 4$ and $\epsilon_{\binom{n}{2}} = 0$. The standard deviation is computed as $\sigma(\epsilon) = \frac{2}{\sqrt{3}}$. If ϵ_1 decreases more gradually to $\epsilon_{\binom{n}{2}}$, $\sigma(\epsilon)$ will be less than $\frac{2}{\sqrt{3}}$. Applying Chebyshev's inequality, we obtain the following formula (10).

$$P(|\epsilon - \mu(\epsilon)| \geq t\sigma(\epsilon)) \leq \frac{1}{t^2} \tag{10}$$

Thus, no more than $\frac{1}{t^2}\binom{n}{2}$ ϵ_ks can be more than $\frac{2}{\sqrt{3}}t$ away from the mean $\mu(\epsilon) = 2$. Note, however, the random variables in Chebyshev's inequality are assumed to have an infinite range. In our case, the maximum possible value of ϵ is $\epsilon_{\max} = 4$. In our situation, where $\epsilon_1, \ldots, \epsilon_{\binom{n}{2}}$ is a weakly decreasing sequence, it will be the case that if the ϵ_ks are distributed symmetrically around $\mu(\epsilon) = 2$, then the number of $\epsilon_k \in [2 + t\sigma(\epsilon), 4]$ will be less than $\frac{1}{2}(\frac{1}{t^2} - \frac{\sigma^2(\epsilon)}{4})\binom{n}{2}$.

If we compute the average ϵ of an edge e with N random frequency quadrilaterals, the average ϵ will conform to a normal distribution $\mathcal{N}(\mu(\epsilon), \sigma^2(\epsilon))$ as N becomes large according to the central limit theorem. We must confirm that every random ϵ_e of e in a frequency quadrilateral has a well-defined expected value $\mu(\epsilon_e)$ and variance $\sigma(\epsilon_e)$. The expected value will be $\mu(\epsilon_e) = 2$ in the six frequency quadrilaterals. This suggests that we can compute $\sigma(\epsilon_e)$ based on the six frequency quadrilaterals as follows. The ϵ_e corresponding to an edge e is equal to $2(1 + \delta_e)p_{\{3,5\}}$, where $p_{\{3,5\}} = \frac{2}{3}$ according to our assumption about the distribution of the frequency of e in the six frequency quadrilaterals in our binomial distribution model. We need to compute the $\delta_e = \frac{p_5(e)}{p_5(e)+p_3(e)}$ for edge e to determine the $\sigma(\epsilon_e)$.

For every edge, its frequency is 5, 3 or 1 in a frequency quadrilateral. Therefore, we draw the pairwise frequency from $\{5,\ 3\ ,1\}$ to form three frequency sets $\{5,\ 3\}$, $\{5,\ 1\}$ and $\{3,\ 1\}$. In the three frequency sets, the corresponding δ_e is either 0.5, 1.0, 0 and each occurs with probability $\frac{1}{3}$. Of course, the expected value $\mu(\delta_e) = 0.5$. For every edge e which corresponds to $\delta_e = 0.5$ $(1.0,\ 0)$, the corresponding ϵ_e is 2 $(\frac{8}{3},\ \frac{4}{3})$ and the expected value of ϵ_e, $\mu(\epsilon_e) = 2$. It follows that $\sigma^2(\epsilon_e) = (\frac{2}{3})^3 \approx 0.2963$ and $\sigma(\epsilon_e) \approx 0.5443$. One can compute that in the normal distribution $\epsilon_e \sim \mathcal{N}(2, (\frac{2}{3})^3)$. $P(\epsilon_e \geq 4) \leq 0.000119184$. However, we know that $P(\epsilon_e > 4) = 0$ for every edge e.

Note that there are in total $6\binom{n}{4}$ ϵ_es because a K_n has $\binom{n}{4}$ quadrilaterals and each quadrilateral contains 6 edges. Let ϵ_e denote the ϵ associated with edge e. If we draw N ϵ_es, i.e., $\{\epsilon_{e^1},\ \epsilon_{e^2},\ \cdots,\ \epsilon_{e^N}\}$ where ϵ_{e^k} means the k^{th} ϵ_e, at random, then we let $\epsilon = \frac{1}{N}\sum_{k=1}^{N}(\epsilon_{e^k})$ denote the associated mean value and $\sigma^2(\epsilon) = \sigma^2(\frac{1}{N}\sum_{k=1}^{N}(\epsilon_{e^k}))$ denote the associated variance. Obviously, $\sqrt{N}(\epsilon - \mu(\epsilon))$ conforms to a normal distribution based on the central limit theorem. Here $\sqrt{N}(\epsilon - \mu(\epsilon)) \sim \mathcal{N}(0, (\frac{2}{3})^3)$ or $\sqrt{N}\epsilon \sim \mathcal{N}(2\sqrt{N}, (\frac{2}{3})^3)$. The maximum and minimum ϵ are 4 and 0, respectively. As N becomes big, the $\sqrt{N}\epsilon$ also increases. However, the variance of all these ϵs remains unchanged. This means that the probability that ϵ is close to 4 becomes smaller as N becomes large.

The ϵ computed according to formula (5) for every edge e is just the mean value of the $\binom{n-2}{2}$ ϵ_es. The ϵ of every edge will conform to the normal distribution $\sqrt{N}(\epsilon - \mu(\epsilon)) \sim \mathcal{N}(0, (\frac{2}{3})^3)$ where $N = \binom{n-2}{2}$. This means that the probability that the ϵs deviate from their expected value $\mu(\epsilon) = 2$ and approach 4 tends to zero as $n \to \infty$. Thus, the number of ϵs close to 4 is very small. In the next section, we will see the ϵs of the OHC edges increase with the scale of TSP n until they approach the maximum value 4.

A linear transformation does not change the probability properties of random variables. Therefore, we can use the ϵs computed according to formula (5) to analyze their distribution for TSP. For the $\binom{n}{2}$ ϵs, the expected value $\mu(\epsilon)$ and variance $\sigma^2(\epsilon)$ are computed as follows. We assume $M = \binom{n}{2}$, $N = \binom{n-2}{2}$ and $\{\epsilon_1,\ \epsilon_2,\ \cdots,\ \epsilon_M\}$ are the ϵs of the M edges. For the j^{th} edge, $\epsilon_j = \frac{1}{N}\sum_{i=1}^{N}(\epsilon_{ij})$, where every $\epsilon_{ij} = \epsilon_e \in \{1, 3, 5\}$. In addition, we suppose all ϵ_{ij}s are independently and uniformly distributed. The expected value of ϵ_j is $\mu(\epsilon_j) = \frac{1}{N}\sum_{i=1}^{N}\mu(\epsilon_{ij}) = 2$. The variance of ϵ_j is $\sigma^2(\epsilon_j) = \frac{1}{N^2}\sum_{i=1}^{N}\sigma^2(\epsilon_{ij}) = \frac{1}{N}(\frac{2}{3})^3$. This holds in our binomial distribution model because we are assuming that the frequency of edge e_j being 1, 3, or 5 has the equal probability $\frac{1}{3}$. In real graphs, it is often the case that short edges have a high probability of having frequency 5 and 3 in their frequency quadrilaterals. On the other hand, it is often the case that for long edges, there is a small probability that the edge will have frequency 5 or 3 in their frequency quadrilaterals. Based on the $N \times M$ matrix of ϵ_{ij}s, we can derive the expected value and variance of them. The expected value of the ϵ_{ij}s is $\mu(\epsilon_{ij}) = \frac{1}{MN}\sum_{j=1}^{M}\sum_{i=1}^{N}\mu(\epsilon_{ij}) = 2$. Meanwhile, the variance of the ϵ_{ij}s is computed as $\sigma^2(\epsilon_{ij}) = \frac{1}{MN}\sum_{j=1}^{M}\sum_{i=1}^{N}\sigma^2(\epsilon_{ij}) = (\frac{2}{3})^3$.

In general, if we compute the ϵs of edges with random frequency quadri-

laterals, the expected value of ϵ, $\mu(\epsilon)$, will be 2 and the variance $\sigma^2(\epsilon)$ can also be determined. One would expect that ϵs of the $\binom{n}{2}$ edges will approximately conform to the normal distribution $\mathcal{N}(\mu(\epsilon), \sigma^2(\epsilon))$ according to the central limit theorem. The probability $P(\epsilon \geq \mu(\epsilon) + t\sigma(\epsilon))$ is equal to $1 - \Phi(t)$ where $\Phi(t) = \frac{1}{2}[1 + \text{erf}(\frac{t}{\sqrt{2}})]$ and $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{n!(2n+1)}$ is the Gauss error function. It follows that $P(\epsilon \geq \mu(\epsilon) + t\sigma(\epsilon))$ is given by the formula (11).

$$P(\epsilon \geq \mu(\epsilon) + t\sigma(\epsilon)) = \frac{1}{2} - \frac{1}{2}\text{erf}(\frac{t}{\sqrt{2}}) = \frac{1}{2} - \frac{1}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n (\frac{t}{\sqrt{2}})^{2n+1}}{n!(2n+1)} \quad (11)$$

Thus, $P(\epsilon \geq \mu(\epsilon) + t\sigma(\epsilon))$ is a function of the variable t which will reach a maximum at some value $t_{\max}$. In our frequency graphs, the maximum ϵ is 4. This means $P(\epsilon \geq 4)$ approaches 0 which is not consistent with a normal distribution. When t reaches the maximum value $t_{\max}$, $t_{\max}\ \sigma(\epsilon) = 2$ holds for a given $\sigma(\epsilon)$. Therefore, we can compute P for a distribution of ts to determine the $t_{\max}$ and then compute the corresponding $\sigma(\epsilon)$ later. For example, suppose one uses the first 14 terms of formula (11) to compute the probability. Then the change in this probability P according to $t_{\max}$ is shown in Figure 4. If we take a threshold at 0.0025 as a small probability (which is reasonable considering the 3σ rule for the normal distribution), then $t_{\max} = 2.819$ and $\sigma(\epsilon) \approx 0.7094$ which is bigger than the theoretical value 0.5443 (or $(\frac{2}{3})^{\frac{3}{2}}$) of the ideal case. If we want to compute a more accurate approximation $\sigma(\epsilon)$, then we must use more terms in the expansion of (11). We tried 22 terms of formula (11) to compute the other small $P(\epsilon \geq \mu(\epsilon) + t\sigma(\epsilon))$ and $t_{\max}$ and found that the corresponding graph did not differ much from the graph pictured in Figure 4. If we choose $\sigma(\epsilon) = 0.7094$, the probability density function (PDF) of the ϵs is approximated by formula (12).

$$f(\epsilon; \mu(\epsilon), \sigma^2(\epsilon)) = \frac{1}{0.7094\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\epsilon-2}{0.7094})^2} \quad (12)$$

Since we are assuming that the distribution of the ϵs nearly conforms to the normal distribution with the exception that $P(\epsilon > 4) = 0$, we can use some characteristics of the normal distribution to approximately analyze their distribution. For example, we can use the 3σ rule of the normal distribution with $t_{\max} = 3$ to compute the distribution of $P(\epsilon \geq \mu(\epsilon) + t\sigma(\epsilon))$ in which case we find that $\sigma(\epsilon) = \frac{2}{3}$.

The number of edges with ϵ above $\epsilon_{\min}$ decreases exponentially in proportion to the difference between $\epsilon_{\min}$ and $\mu(\epsilon) = 2$. For TSP of large size, our results suggest that $\epsilon_{\min}$ will be close to 4 and the number of edges with ϵs above $\epsilon_{\min}$ is close to n. For TSP of medium size, our computer experiments described in the next section suggest we will end up with a sparse graph if we keep only the edges with ϵ above $\mu(\epsilon)+2\sigma(\epsilon)$ or $\mu(\epsilon)+2.5\sigma(\epsilon)$. For TSP of small size, our computer experiments described in the next section suggest that we will end up with a

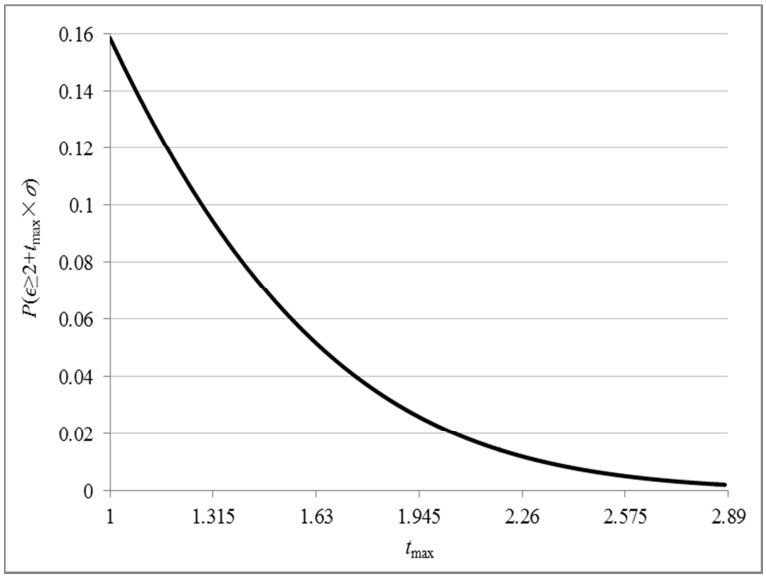

Figure 4: The change of $P(\epsilon \geq \mu(\epsilon) + t_{\max}\sigma(\epsilon))$ according to $t_{\max}$.

sparse graph if we keep only the edges with ϵ above $\mu(\epsilon)+\sigma(\epsilon)$ or $\mu(\epsilon)+1.5\sigma(\epsilon)$. The number of edges with $\epsilon \in [\mu(\epsilon) + t\sigma(\epsilon), 4]$ can be approximated according to formulas (11) and (12).

For the *OHC* edges, the distribution of their ϵs will conform to another normal distribution based on the central limit theorem. The expected value is $\lim_{n\to\infty} \mu_o(\epsilon) = 4$ and the standard deviation is $\lim_{n\to\infty} \sigma_o(\epsilon) = 0$. Thus, the probability density function becomes a Dirac delta function. That is, it is zero everywhere except at $\mu_o(\epsilon) = 4$, with an integral of one over the span $[0, 4]$.

5 Examples and analysis

The Concorde package on-line (NEOS Server for Concorde) [21] has computed the *OHC* for several families of TSP. In this section, we will report on some computer experiments where we used the *OHC* that had been computed for such TSP instances to compute the corresponding $\epsilon_{\min}$, $\sigma(\epsilon)$, $\mu_o(\epsilon)$, $\sigma_o(\epsilon)$, $\overline{\epsilon_K}$, $\overline{\epsilon_R}$. In each case, we keep only those edges whose corresponding ϵ is larger than $\epsilon_{\min}$. Let $r = \frac{2-\overline{\epsilon_R}}{\overline{\epsilon_K}-\overline{\epsilon_R}}$ be the ratio between K and $\binom{n}{2}$, which shows how sparse the graph is. The smaller the values r are, the sparser the graphs are. We also compute $c = \frac{K}{n\log_2(n)} = \frac{r(n-1)}{2\log_2(n)}$ for comparisons. If c is much smaller than the size of the number of vertices n of TSP, then we are reduced to considering graphs with only $O(n\log_2(n))$ edges, and we can use various efficient algorithms which work on sparse graphs to search for solutions to our given TSP.

The results are listed in Table 1 according to r as r ranges from big to small

values. Six digits after the decimal point are kept. In most cases, we found that $\epsilon_{\min}$ is bigger than $\mu(\epsilon) = 2$. As the number of vertices gets larger, $\epsilon_{\min}$ seems to approach or exceed 3 and $\sigma(\epsilon)$ is close to 0.7094. Similarly, as the number of vertices get larger, the corresponding values of ϵ related to the OHC edges which we call $\mu_{\rm o}(\epsilon)$ seem to approach 4 and and the corresponding variance $\sigma_{\rm o}(\epsilon)$ is much smaller than $\sigma(\epsilon)$. Similarly, we see that $\overline{\epsilon_K}$ is much bigger than $\overline{\epsilon_R}$. In general, we found that $\overline{\epsilon_K} + \overline{\epsilon_R} > 4$ and r is less than 0.5, except for the instance brg180. The deviation for brg180 from the other examples that we computed is probably due to the fact that brg180 has a lot of equal-weight edges so that the distribution of the computed frequency quadrilaterals does not conform to our binomial distribution model. As expected, our examples also show that r decreases quickly as $\epsilon_{\min}$ grows. Our results nearly conform to the results predicted by formula (11) and Figure 4. The number of edges with ϵ above $\epsilon_{\min}$ approximately conforms to the normal distribution formula (12). In the last column, we see that c is much smaller than n and they are smaller than 6.5 for all the *TSP* instances in Table 1. In such cases, the graph that remains after keeping only those edges with $\epsilon > \epsilon_{\min}$ are sparse enough to be resolved with the current exact algorithms that work only under the assumption that the underlying graph is sparse.

We note that one of the basic assumptions is that for all vertices A, B, C, D, the sum of the distances for the path (A, B, C, D) is always different from the sum of the distances for the path (A, C, B, D). This allows us to always pick the optimal 4-vertex paths for each of the 6 possible pairs paths in our frequency quadrilateral. Thus, a natural question arises of what should be done when there are lots of sets of four vertices, A, B, C, D, such that $AC + CB + BD = AB + BC + CD$. In such a situation, we have no criterion to determine which of (A, B, C, D) and (A, C, B, D) should be used as the optimal 4-vertex path. In our computer experiments, this issue is resolved by numbering the vertices from $1, \ldots, n$, and then making the choice between (A, B, C, D) and (A, C, B, D) by choosing the one that is smallest in lexicographic order based on our labeling of the vertices. For example for given four vertices $A < B < C < D$ and $AC + CB + BD = AB + BC + CD$, we choose the path (A, B, C, D) rather than (A, C, B, D) as an optimal 4-vertex path in our computer experiments. In such a situation, the frequency or ϵ of edges computed with them will deviate from our binomial distribution model. The problem instance brg180 is an example of a graph where there are many such sets of 4 vertices. In such a situation, some edges in the OHC may have small frequency in their frequency quadrilaterals due to our selection strategy for the optimal 4-vertex paths. This seems to produce a smaller $\epsilon_{\min}$ and bigger values of r and c.

In another computer experiment, we computed the number of edges e whose ϵs are greater than $\epsilon_{\min}$ as $\epsilon_{\min}$ varied from 2.0 to 3.4 for the instances pr144, brg180, kroA200, pr226, fl417 and gr431. We shall call the resulting graph in each case the residual graph. The number of edges in the original graphs equals $\binom{n}{2}$. The experimental results are shown in Table 2. One sees as the threshold $\epsilon_{\min}$ grows, the number of edges in the residual graph decreases sharply. The numbers in parenthesis in Table 2 represent the number of edges from the OHC

Table 1: The computational results of some *TSP* instances (n is the TSP scale)

TSP	n	$\epsilon_{\min}$	$\sigma(\epsilon)$	$\mu_o(\epsilon)$	$\sigma_o(\epsilon)$	$\overline{\epsilon_K}$	$\overline{\epsilon_R}$	r	c
brg180	180	2.077745	0.69068	3.302347	0.203426	2.584331	1.352981	0.535548	6.404693
pr144	144	2.077695	0.740290	3.699199	0.295564	2.741467	1.472082	0.415885	4.147292
pr226	226	2.355422	0.737526	3.742785	0.221221	2.948051	1.581856	0.306065	4.403008
kroA200	200	2.610097	0.722493	3.595294	0.310257	3.085069	1.700230	0.216465	2.817722
fl417	417	2.512650	0.714986	3.710468	0.217121	3.160457	1.697672	0.206680	4.939098
lin318	318	2.738794	0.739881	3.691154	0.248299	3.189280	1.728645	0.185779	3.542209
gr431	431	2.691909	0.710105	3.581473	0.270107	3.096364	1.751279	0.184911	4.542725
si175	175	3.024279	0.795722	3.866228	0.163097	3.447857	1.712433	0.165704	1.934752
rd400	400	2.866249	0.744384	3.694329	0.233032	3.271729	1.771145	0.152511	3.519950
d657	657	2.867387	0.739090	3.723399	0.208537	3.263634	1.776356	0.150371	5.269552
pcb442	442	2.863235	0.742620	3.695240	0.221822	3.276539	1.774525	0.150115	3.766582
pr439	439	2.911339	0.734087	3.707294	0.213165	3.292903	1.792572	0.138255	3.449257
rat575	575	2.884824	0.714930	3.633131	0.255639	3.262541	1.804531	0.134066	4.197140
d493	493	2.888322	0.722681	3.648368	0.245401	3.266274	1.805411	0.133201	3.663031
ail535	535	2.919370	0.734018	3.706185	0.205480	3.285298	1.822376	0.121417	3.576842
u724	724	2.975773	0.724267	3.708220	0.231430	3.345162	1.824725	0.115279	4.386740
att532	532	2.981047	0.722783	3.650877	0.241076	3.327321	1.828218	0.114590	3.359767
rat783	783	3.006706	0.715987	3.672295	0.236411	3.347090	1.839564	0.106423	4.328717
rl1323	1323	3.121597	0.729567	3.765055	0.162765	3.402678	1.850251	0.095126	6.063715

Table 2: The number of edges in the sparse graphs and the number of lost *OHC* edges in the parentheses according to $\epsilon_{\min}$

$\epsilon_{\min}$	$\binom{n}{2}$	2.1	2.4	2.5	2.6	2.7	2.8	2.9	3.0	3.4
pr144	10296	4195(1)	2981(1)	2582(2)	2242(2)	1948(3)	1642(4)	1406(4)	1192(5)	556(13)
brg180	16110	8252(1)	3100(1)	2868(1)	2868(1)	2862(6)	2856(12)	2856(12)	2216(12)	28(155)
kroA200	19900	7964(0)	5652(0)	5001(0)	4382(0)	3756(2)	3219(4)	2728(8)	2285(10)	845(51)
pr226	25425	10113(0)	7429(1)	6471(1)	5649(2)	4903(2)	4237(2)	3623(3)	3113(4)	1604(12)
fl417	86736	31611(0)	20355(0)	16407(0)	16392(1)	15172(4)	14227(5)	13047(8)	11513(11)	5256(29)
gr431	92665	38941(0)	26821(0)	23197(0)	19845(0)	16872(1)	14007(4)	11523(8)	9245(12)	3049(107)

whose corresponding ϵ is less than $\epsilon_{\min}$. For $\epsilon_{\min} \leq 2.7$, we see that the number of lost *OHC* edges do not change much. Indeed, when $\epsilon_{\min} = 2.7$ is taken as the threshold, only a few *OHC* edges are lost whereas the number of edges that we keep is sharply reduced. Based on our experimental results, we suggest that one should use $\epsilon_{\min} = 2.7$ as the frequency threshold to compute the residual graph for most small *TSP* instances. In most cases, the residual graph has less than 15.86% of total number of edges in the original graph. If the residual graph includes the *OHC*, then significant computation time will be saved to resolve *TSP*. Of course, in theory, the residual graph may not even have an *HC* if we use a big threshold, such as $\epsilon_{\min} > 2.7$. We used the improved genetic algorithm [28] to search the new *OHC* in the sparse graphs computed with $\epsilon_{\min} > 2.7$, but in nearly all of the cases we failed to find any *HC*. For the small instances of the *TSP*, $\mu(\epsilon) + 2\sigma(\epsilon)$ is too big to take as the $\epsilon_{\min}$. In such cases, many of the *OHC* edges are not included in the residual graph.

There is another possibility for dealing with graphs where there are many sets of vertices A, B, C, D where $AB + BC + CD = AC + CB + BD$ so that we can not choose between the paths (A, B, C, D) and (A, C, B, D) based on the sum of the distances of their edges. One way to resolve this problem is to add a small random distance $rd \in [0, 1]$ to the distance of every edge, i.e., the $d(A, B)$ of an edge (A, B) becomes $d(A, B) + rd(A, B)$. For symmetrical TSP, $rd(A, B) = rd(B, A) \in [0, 1]$ for an arbitrary edge (A, B). The random distance rd is so small that it does not change the OHC. However, the small random distance converts the "special" TSP into a general TSP so that our binomial distribution model can work well. In addition, rds are generated at random for every edge. Therefore, the random distance rd has the nearly equivalent impact on the probability $p_{\{3,5\}}$ for any edge e. Meanwhile, the $\epsilon_k (1 \leq k \leq \binom{n}{2})$ of every edge also complies with the binomial distribution model. We carried out experiments using this idea to generate the same kind of statistics as shown in Table 2. These results are illustrated in Table 3.

Our computer experiments focused on the instances brg180, pr144, pr226 and fl417 which have many equal-weight edges. We added a random distance $rd \in [0, 1]$ to the distance of every edge in order to compute the 6 optimal 4-vertex paths in each weighted quadrilateral. Because rds are random variables, the results may vary in different trials. That is, for any given quadrilateral, we may not compute the same six optimal 4-vertex paths. However, our experiments showed that the final result for the frequency graphs does not change very much. On average, the added random distances to the edges allowed us to generate 6 exact optimal 4-vertex paths for most of the weighted quadrilaterals. For some parameters of brg180, refer to Table 4. The number of edges in the sparse graphs is computed with $\epsilon_{\min}$ varying from 2.0 to 3.4. The number of the lost OHC edges is also recorded in the parentheses. Our results showed that the ϵs of edges have only small changes when we add a random distance rd to their distances as compared to the results in Table 1. However, the number of edges with ϵs above $\epsilon_{\min}$ is changed to some extent. For example, the $\epsilon_{\min}$ of brg180 becomes 2.337671 which is bigger than that in Table 1. The r is computed as 0.407138. It means that 2068 more edges are removed comparing with the results in Table 1.

Table 3: The experiments for the TSP with many equal-weight edges

$\epsilon_{\min}$	2.1	2.4	2.5	2.6	2.7	2.8	2.9	3.0	3.4
pr144	4193(1)	2981(1)	2581(2)	2243(2)	1947(3)	1637(4)	1398(4)	1191(5)	557(13)
brg180	8732(0)	5932(1)	4922(1)	3973(1)	2990(6)	2856(12)	2856(12)	2677(12)	28(155)
kroA200	7962(0)	5656(0)	5000(0)	4377(0)	3756(2)	3219(3)	2728(8)	2286(10)	843(51)
pr226	9988(0)	7346(1)	6407(1)	5600(2)	4850(2)	4173(2)	3561(3)	3050(4)	1664(12)
fl417	31624(0)	20351(0)	18143(0)	16402(1)	15187(4)	14215(5)	13051(8)	11574(11)	5277(29)
gr431	38945(0)	26817(0)	23201(0)	19845(0)	16877(1)	14000(4)	11532(8)	9244(12)	3051(106)

Finally, we carried out a few more experiments of this type for brg180 which includes a lot of equal-weight edges. In each experiment, we multiply the small random distance $rd \in [0, 1]$ with a different coefficient co, i.e., $co \times rd$ and

$rd \in [0,1]$. The number of equal-weight edges will be reduced greatly so that we can compute just 6 optimal 4-vertex paths for nearly every given quadrilateral. The $\epsilon_{\min}$ is recorded and r is computed according to the coefficients co. The results are given in Table 4. We note that the results in Table 3 for brg180 were computed according co=1.0 in Table 4.

The $\epsilon_{\min}$s are not equal for different coefficients co. The number of edges with ϵs above $\epsilon_{\min}$ changes according to $co \cdot rd$. In the worst experiment, the residual graph includes $0.565988 \times \frac{180\times179}{2} = 9118$ edges. In the best experiment, the residual graph includes $0.307886 \times \frac{180\times179}{2} = 4960$ edges. For the coefficients $co = 1.5,\ 2.0,\ 2.5$ and 2.8, the $\epsilon_{\min}$s are less than $2 + \frac{2}{(n-2)}$ which is probably due to the fact that we still have many quadrilaterals where we cannot compute the right optimal 4-vertex paths. One reason is that there are still many edges with equal weights. The other reason is that adding random distances leads to many inappropriate optimal 4-vertex paths for brg180. Although we compute 6 optimal 4-vertex paths for a given quadrilateral, the frequency of some OHC edges may not be as big as the the frequency of the other OHC edges in their frequency quadrilaterals.

With the other coefficients, the corresponding $\epsilon_{\min}$s are much bigger than $2 + \frac{2}{(n-2)}$. This suggests that these coefficients are able to change brg180 into a weighted graph to which our binomial distribution model applies. Thus, adding random increments to the distances of edges can still allow the binomial distribution model to work well. Because the rds are generated at random, we cannot expect to obtain the best results with just one experiment. We found that by using many experiments, we were able to acquire some results where $\epsilon_{\min}$s were much bigger than $2 + \frac{2}{(n-2)}$.

In each of the experiments represented in Table 4, we extracted the 180 ϵs of the OHC edges and ordered them from big to small values. In Table 4, one sees that $\epsilon_{\min}$s vary quite a bit. However, the second smallest value was approximately 2.674499 in the 11 experiments with different cos. In addition, the 3^{rd}, 4^{th}, 5^{th} and 6^{th} smallest values, which were approximately 2.680589, did not change much in the 11 experiments. Moreover, the seventh smallest ϵ was bigger than 2.7. Thus, when $\epsilon_{\min} = 2.7$ is taken as the frequency threshold, the number of lost OHC edges is at most 6 in each of the experiments. When one adds random distances to the edges for a TSP with a lot of equal-weight edges, the number of edges in the residual graph will vary from experiment to experiment. This suggests that one should do multiple experiments when adding random distances to the edges until one finds an $\epsilon_{\min}$ which is bigger than $2 + \frac{2}{(n-2)}$. In this way, we can compute a residual graph with a relatively small number of edges.

6 Conclusions

The main result of this paper is to give a heuristic to cut down the number of edges in the search for an OHC in a symmetric TSP based on computing frequency graphs. That is, first one adds a small increment of distance to

Table 4: The experiments with different $co \cdot rd$ for brg180

co	1.0	1.5	2.0	2.5	2.8	3.0
$\epsilon_{\min}$	2.337671	1.752619	1.827020	1.726974	1.893164	2.434364
r	0.407138	0.565988	0.560896	0.568031	0.556175	0.345190
co	3.2	3.5	4.0	4.5	5	
$\epsilon_{\min}$	2.497838	2.386143	2.051816	2.203885	2.281959	
r	0.307886	0.376662	0.545750	0.493540	0.438427	

each edge to ensure that one can distinguish between the distances of the path (A, B, C, D) versus the path (A, C, B, D) for all sets of 4 vertices A, B, C, D. Next, one computes the frequency graph based on randomly choosing N frequency quadrilaterals with each edge and then eliminates those edges e whose corresponding ϵ is less than a pre-specified $\epsilon_{\min}$. The analysis of our binomial distribution model for such randomly chosen frequency quadrilaterals and our computer experiments suggest $\epsilon_{\min} = 2.7$ is a good first choice. In this case, the residual graph has less than 15.86% of the total number of edges in the original graph. The cost of computing a frequency graph is $O(n^4)$.

One could ask whether we can produce a similar analysis by working with optimal 5-vertex paths and pentalaterals instead of optimal 4-vertex paths and quadrilaterals. In this case, there are 32 different frequency graphs that are possible using five vertices A, B, C, D, E and the distribution of frequencies is not uniform as it is in the case of frequency quadrilaterals using optimal 4-vertex paths. Thus, it is much harder to analyze. Another drawback is the cost of computing a frequency graph is $O(n^5)$ in this case.

When end with two questions for further research. The first question is what happens if we iterate the procedure of computing the residual graphs. In theory, we can throw away more and more edges. We shall pursue such an analysis in a subsequent paper. The second question is to estimate the complexity of the algorithms that we use an exact or approximation algorithm to resolve TSP based the residual graphs that we produce. This will be the next focus of our future research.

Acknowledgements

We acknowledge W. Cook, H. Mittelmann who created the Concorde and G. Reinelt, et al. who provide the TSP data to TSPLIB. We also thank the anonymous referees for their corrections and suggestions for improvements to presentation of the paper. The authors acknowledge the support provided by NSFC (No.51205129) and the Fundamental Research Funds for the Central Universities (No.2015ZD10).

References

[1] N. Aggarwal, N. Garg, and S. Gupta. A 4/3-approximation for TSP on cubic 3-edge-connected graphs. *arXiv 1101.5586*, pages 1–5, 2011. URL: `https://arxiv.org/abs/1101.5586`.

[2] D. Applegate, R. Bixby, V. Chvátal, W. Cook, D. Espinoza, M. Goycoolea, and K. Helsgaun. Certification of an optimal TSP tour through 85900 cities. *Operations Research Letters*, 37(1):11–15, 2009. `doi:10.1145/351827.384248`.

[3] R. Bellman. Dynamic programming treatment of the traveling salesman problem. *Journal of the ACM*, 9(1):61–63, 1962. `doi:10.1145/321105.321111`.

[4] A. Björklund, T. Husfeldt, P. Kaski, and M. Koivisto. The traveling salesman problem in bounded degree graphs. *ACM Transactions on Algorithms*, 8(2):1–12, 2012. `doi:10.1145/2151171.2151181`.

[5] G. Borradaile, E. Demaine, and S. Tazari. Polynomial-time approximation schemes for subset-connectivity problems in bounded-genus graphs. *Algorithmica*, 68(2):287–311, 2014. `doi:10.1007/s00453-012-9662-2`.

[6] S. Boyd, R. Sitters, S. van der Ster, and L. Stougie. The traveling salesman problem on cubic and subcubic graphs. *Mathematical Programming*, 144(1-2):227–245, 2014. `doi:10.1007/s10107-012-0620-1`.

[7] G. Carpaneto, M. Dell'Amico, and P. Toth. Exact solution of large-scale, asymmetric traveling salesman problems. *ACM Transactions on Mathematical Software*, 21(4):394–409, 1995. `doi:10.1145/212066.212081`.

[8] T. Cormen, C. Leiserson, R. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, 2009.

[9] J. Correa, O. Larré, and J. Soto. TSP tours in cubic graphs: beyond 4/3. *SIAM Journal on Discrete Mathematics*, 29(2):915–939, 2015. `doi:10.1137/140972925`.

[10] E. de Klerk and C. Dobre. A comparison of lower bounds for the symmetric circulant traveling salesman problem. *Discrete Applied Mathematics*, 159(16):1815–1826, 2011. `doi:10.1016/j.dam.2011.01.026`.

[11] D. Eppstein. The traveling salesman problem for cubic graphs. *Journal of Graph Algortihms and Applications*, 11(1):61–81, 2007. `doi:10.7155/jgaa.00137`.

[12] H. Gebauer. Enumerating all hamilton cycles and bounding the number of hamiltonian cycles in 3-regular graphs. *The Electric Journal of Combinatorics*, 18(1):1457–1480, 2011.

[13] G. Gutin and A. Punnen. *The traveling salesman problem and its variations*, volume 12 of *Combinatorial Optimization*. Springer, 2007.

[14] M. Held and R. Karp. A dynamic programming approach to sequencing problems. *Journal of the Society for Industrial and Applied Mathematics*, 10(1):196–210, 1962. doi:10.1137/0110015.

[15] K. Helsgaun. An effective implementation of the lin-kernighan traveling salesman heuristic. *European Journal of Operation Research*, 126(1):106–130, 2000. doi:10.7155/jgaa.0027910.1016/S0377-2217(99)00284-2.

[16] J. Hoogeveen. Analysis of Christofides' heuristic: Some paths are more difficult than cycles. *Operations Research Letters*, 10(5):291–295, 1991. doi:10.1016/0167-6377(91)90016-I.

[17] R. Karp. Reducibility among combinatorial problems. *Complexity of Computer Computations*, pages 85–103, 1972. doi:10.1007/978-3-540-68279-0-8.

[18] M. Levine. Finding the right cutting planes for the TSP. *Journal of Experimental Algorithmics*, 5(6):1–20, 2000. doi:10.1145/351827.384248.

[19] M. Liśkiewicz and M. Schuster. A new upper bound for the traveling salesman problem in cubic graphs. *Journal of Discrete Algorithms*, 7:1–20, 2014. doi:10.1016/j.jda.2014.02.001.

[20] Y. Liu. Diversified local search strategy under scatter search framework for the probabilistic traveling salesman problem. *European Journal of Operational Research*, 191(2):332–346, 2008. doi:10.1016/j.ejor.2007.08.023.

[21] H. Mittelmann. Neos server for concorde. URL: http://neos-server.org/neos/solvers/co:concorde/TSP.html.

[22] T. Mömke and O. Svensson. Approximating graphic TSP by matchings. pages 560–569, 2011. doi:10.1109/FOCS.2011.56.

[23] S. Oveis Gharan and A. Saberi. The asymmetric traveling salesman problem on graphs with bounded genus. pages 967–975, 2011.

[24] G. Reinelt. URL: http://comopt.ifi.uni-heidelberg.de/software/TSPLIB95/.

[25] M. Sharir and E. Welzl. On the number of crossing-free matchings, cycles, and partitions. *SIAM Journal on Computing*, 36(3):695–720, 2006. doi:10.1137/050636036.

[26] Y. Wang. The frequency graph for the traveling salesman problem. *International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering*, 6(10):1019–1022, 2012.

[27] Y. Wang. A representation model for TSP. pages 204–209, 2013. doi:10.1109/HPCC.and.EUC.2013.38.

[28] Y. Wang. The hybrid genetic algorithm with two local optimization strategies for traveling salesman problem. *Computers & Industrial Engineering*, 70(4):124–133, 2014. doi:10.1016/j.cie.2014.01.015.

[29] Y. Wang. An approximate algorithm for triangle TSP with a four-vertex-three-line inequality. *International Journal of Applied Metaheuristic Computing*, 6(1):35–46, 2015. doi:10.4018/ijamc.2015010103.

[30] Y. Wang. An approximate method to compute a sparse graph for traveling salesman problem. *Expert Systems with Applications*, 42(12):5150–5162, 2015. doi:10.1016/j.eswa.2015.02.037.