PROBABILISTIC ANALYSIS OF A RELAXATION FOR THE K-MEDIAN PROBLEM*†

SANG AHN,‡ COLIN COOPER,§ GÉRARD CORNUÉJOLS**
AND ALAN FRIEZE**

This paper provides a probabilistic analysis of the so-called “strong” linear programming relaxation of the k-median problem when the number of points, say n, goes to infinity. The analysis is performed under four classical models in location theory, the Euclidean, network, tree and uniform cost models. For example, we show that, for the Euclidean model and $k \to \infty$, $k = o(n/\log n)$, the value of the relaxation is almost surely within 0.3 percent of the optimum k-median value. For the uniform cost model and the same conditions on k, we show that the value of the relaxation is almost surely 50 percent of the optimum k-median value. We also show that, under various assumptions, branch and bound algorithms that use the strong linear programming relaxation as a bound must almost surely expand a nonpolynomial number of nodes to solve the k-median problem to optimality. Finally, we report extensive computational experiments. As predicted by the probabilistic analysis, the relaxation was not as tight for the problem instances drawn from the uniform cost model as for the other models.

1. Introduction. The k-median problem has been widely studied both from the theoretical point of view and for its applications. An interesting theoretical development was the successful probabilistic analysis of several heuristics for this problem (e.g. Fisher and Hochbaum [8] and Papadimitriou [22]). On the other hand, the literature on the k-median problem abounds in exact algorithms. Most are based on the solution of a certain relaxation to be defined later. The computational experience reported in the literature seems to indicate that this particular relaxation yields impressively tight bounds compared to what can usually be expected in integer programming. In this paper we analyze to what extent this relaxation is tight. We perform our analysis under various probabilistic assumptions and identify conditions under which the relaxation can be expected to be tight and others under which it can be expected to give a poor bound. For example, for a classical Euclidean model in the plane, we show that the relaxation can be expected to provide a bound within one third of one percent of the optimum value of the k-median problem. In addition to the probabilistic analysis, we also report extensive computational experiments, based on the solution of thousands of medium-size problems. Some of the results predicted for very large problems by our probabilistic analysis can already be observed on these test problems.

Consider a set $X = \{X_1, \ldots, X_n\}$ of n points, a positive integer $k \leq n$ and let $d_{ij} \geq 0$ be the distance between $X_i$ and $X_j$ for each $1 \leq i \leq n$ and $1 \leq j \leq n$. (Unless otherwise specified, it is assumed that $d_{ii} = 0$, $d_{ij} = d_{ji}$ and $d_{ij} \leq d_{ik} + d_{kj}$, for all

*Received July 18, 1985; revised January 14, 1986.

AMS 1980 subject classification. Primary: 90C05.

I AOR 1973 subject classification. Main: Programming/Linear; Cross References: Programming/Probabilistic.

OR/MS Index 1978 subject classification. Primary: 643 Mathematical Programming/Linear/Algorithms.

Key words: Probabilistic analysis, Location, k-median problem, Branch and bound algorithm.

†This work was supported in part by NSF grant ECS8503192.

‡Seoul National University.

§Polytechnic of North London.

**Carnegie-Mellon University.
The *k*-median problem consists of finding a set $S \subseteq X$, $|S| = k$, that minimizes

$$
\sum_{i=1}^{n} \sum_{j \in S} d_{ij} y_{ij},
$$

(1)

$$
\sum_{j=1}^{n} y_{ij} = 1 \quad \text{for } i = 1, \ldots, n,
$$

(2)

$$
\sum_{j=1}^{n} x_{j} = k,
$$

(3)

$$
0 \leq y_{ij} \leq x_{j} \leq 1 \quad \text{for } i, j = 1, \ldots, n,
$$

(4)

$$
x_{j} \in \{0, 1\} \quad \text{for } j = 1, \ldots, n.
$$

(5)

In this formulation $x_{j} = 1$ if $X_{j} \in S$, 0 otherwise and, for $1 \leq i \leq n$, we can set $y_{ij} = 1$ for an index $j$ that achieves $\min_{j \in S} d_{ij}$.

The formulation (1)–(4) is called the linear programming (LP) relaxation of the *k*-median problem. In other words, the LP relaxation is obtained by ignoring the integrality conditions on $x_{j}$, $1 \leq j \leq n$. The optimum value $z_{LP}$ of this relaxation clearly satisfies $z_{LP} \leq z_{IP}$. The bound $z_{LP}$ has been used extensively in exact algorithms for the *k*-median problem. (E.g. Marsten [15], Garfinkel, Neebe and Rao [10], ReVelle and Swain [23], Diehr [5], Schrage [24], Guignard and Spielberg [11], Narula, Ogbu and Samuelsson [20], Cornuejols, Fisher and Nemhauser [3], Erlenkotter [6], Galvão [9], Magnanti and Wong [14], Nemhauser and Wolsey [21], Mulvey and Crowder [19], Mavrides [16], Mirchandani, Oudjit and Wong [17], Christofides and Beasley [2], Beasley [1]).

Most of the computational experiments have been reported on test problems with $n \leq 100$. For many of these test problems, $z_{IP} = z_{LP}$. Recently, Beasley [1] solved 40 larger problems (with $100 \leq n \leq 900$) and found a small but positive gap $z_{IP} - z_{LP}$ for many of them. The average of $(z_{IP} - z_{LP})/z_{IP}$ over these problems was 0.0024.

In this paper we analyze the ratio $(z_{IP} - z_{LP})/z_{IP}$ from a probabilistic point of view as $n$ goes to infinity, under various assumptions on the probability distribution of problem instances. We do not address the worst-case analysis of this ratio except to note that this question was solved by Cornuejols, Fisher and Nemhauser [3], when $d_{ij} \leq 0$. The analysis of [3] does not carry over when the $d_{ij}$'s are nonnegative and satisfy the distance axioms. In fact, this worst-case analysis is an interesting open question. It would also be interesting to know the worst-case value of $(z_{IP} - z_{LP})/z_{IP}$ when the $d_{ij}$'s are further restricted to represent Euclidean distances. Once again, these questions are not addressed here as we focus on a probabilistic approach.

We will often write statements like $X_{n} \leq u_{n}$ almost surely (a.s.) for a sequence of random variables $(X_{n})$ and real sequence $(u_{n})$. This is a well-defined terminology of probability theory and details can be found in Stout [25] for example. We will invariably prove that $\sum_{n=1}^{\infty} \Pr(X_{n} > u_{n}) < \infty$ which implies the above statement. Nonprobabilists will be satisfied that we show $\Pr(X_{n} > u_{n}) \to 0$ as $n \to \infty$. If $X_{n} \leq u_{n}(1 + o(1))$ a.s. and $X_{n} \geq u_{n}(1 - o(1))$ a.s. then we write $X_{n} \sim u_{n}$ a.s.

First we study the *k*-median problem in the plane. When the points $X_{1}, \ldots, X_{n}$ are uniformly distributed in a unit square and $d_{ij}$ is the Euclidean distance between $X_{i}$ and $X_{j}$, $1 \leq i, j \leq n$, we show that $(z_{IP} - z_{LP})/z_{IP} \sim 0.00284$ almost surely, for any $k$
such that \( \omega \leq k \leq n / \omega \log n \) where \( \omega = \omega(n) \to \infty \). (In this paper we abbreviate \( f(n) \to a \) as \( n \to \infty \) by \( f(n) \to a \).)

In a second model, the points \( X_1, \ldots, X_n \) are the nodes of a random graph \( G_v(\rho) \) where \( \rho \) is the probability that an edge is in the graph, and \( d_{ij} \) is the minimum number of edges on a path from \( X_i \) to \( X_j \). We assume \( \rho \geq (\omega \log n) / n \) where \( \omega = \omega(n) \to \infty \) (this guarantees that \( G_v(\rho) \) is almost surely connected), and \( kp^2 \geq (\omega \log n) / n \). We prove that \( \frac{z_{LP} - z_{IP}}{z_{IP}} \leq 1 / (e + 1) \) almost surely, where \( e \) is the base of natural logarithms. More specifically, if \( \log b n \leq k \leq n \) where \( b = 1 / (1 - \rho) \), then \( z_{IP} = z_{LP} \) almost surely. If \( 2 \leq k \leq \log b n \), \( kp \to \alpha \) where \( 0 \leq \alpha \leq \infty \) and \( p \to \beta \) where \( 0 \leq \beta < 1 \), define \( \gamma = e \) if \( \beta = 0 \) and \( \frac{1 - \beta}{\beta} \) if \( \beta > 0 \); then \( \frac{z_{IP} - z_{LP}}{z_{IP}} - f(\alpha, \beta) \) almost surely where \( f(\alpha, \beta) = \frac{(1 - (1 - \alpha) + a^\alpha)}{(1 + a^\alpha)} \). (The maximum of this function is \( 1 / (e + 1) \) attained when \( \alpha = 1 \) and \( \beta = 0 \). When \( \alpha = 0 \) or \( \infty \) the function takes the value 0.)

We also analyze the \( k \)-median problem on random trees and on another model where it is assumed that the \( d_{ij} \)'s are independently and uniformly distributed on \([0, 1]\).

In \$6\$, we put our probabilistic results in perspective by presenting extensive computational experiments.

In \$7\$, we show how our results for the \( k \)-median problem relate to the simple plant location problem (SPLP). In the SPLP, the data comprise \( n \) points \( X_1, \ldots, X_n \), distances \( d_{ij} \) for \( 1 \leq i, j \leq n \), and fixed costs \( f_j \) associated with each point \( X_j \), \( 1 \leq j \leq n \). The SPLP consists of finding a nonempty set \( S \subseteq X \) that minimizes \( \sum_{j=1}^n \min_{s \in S} d_{ij} + \sum_{j \in S} f_j \). (Note that, in this problem, \(|S|\) is not restricted as in the \( k \)-median problem.) An integer programming formulation of SPLP is

\[
z_{IP} = \min \sum_{i=1}^n \sum_{j=1}^n d_{ij} y_{ij} + \sum_{j=1}^n f_j x_j
\]

subject to (2), (4) and (5). The LP relaxation is obtained by relaxing the integrality conditions (5).

In the remainder of this section we state some useful results from the literature. Our proofs use the following lemma (see Hoeffding [12]).

**Lemma 1.** If \( Y_1, \ldots, Y_n \) are independent random variables and \( 0 \leq Y_i \leq 1 \) for \( i = 1, \ldots, n \), then, for \( 0 < \epsilon < 1 \),

\[
\Pr(\bar{Y} \geq (1 + \epsilon) \mu) \leq e^{-\epsilon^2 n \mu / 3} \quad \text{and} \quad \Pr(\bar{Y} \leq (1 - \epsilon) \mu) \leq e^{-\epsilon^2 n \mu / 2},
\]

where \( \bar{Y} = (\sum_{i=1}^n Y_i) / n \) and \( \mu \) is the expected value of \( \bar{Y} \).

Given a vector \( x = (x_j; \quad j = 1, \ldots, n) \), define

\[
z_{LP}(x) = \min \sum_{i=1}^n \sum_{j=1}^n d_{ij} y_{ij},
\]

\[
\sum_{j=1}^n y_{ij} = 1 \quad \text{for} \quad i = 1, \ldots, n,
\]

\[
0 \leq y_{ij} \leq x_j \quad \text{for} \quad i, j = 1, \ldots, n.
\]

Note that \( z_{LP} = \min z_{LP}(x) \)

\[
\sum_j x_j = k, \quad 0 \leq x_j \leq 1 \quad \text{for} \quad j = 1, \ldots, n.
\]

The following lemma is well known in the \( k \)-median literature and is easy to prove.
Lemma 2. An optimal solution \( y = (y_{ij}; i, j = 1, \ldots, n) \) of \( z_{LP}(x) \) is obtained as follows. For each \( i \), sort the values \( d_{ij}, j = 1, \ldots, n \), so that \( d_{i,j_1(i)} \leq d_{i,j_2(i)} \leq \cdots \leq d_{i,j_n(i)} \), and let \( p \) be such that \( \sum_{h=1}^{p-1} x_h \leq 1 < \sum_{h=1}^{p} x_h \).

Then

\[
y_{ij} = \begin{cases} 
x_j & \text{for } j = j_1(i), \ldots, j_{p-1}(i), \\
1 - \sum_{h=j_p(i)}^{p-1} x_h & \text{for } j = j_p(i), \\
0 & \text{for } j = j_{p+1}(i), \ldots, j_n(i).
\end{cases}
\]

Proof. The program \( z_{LP}(x) \) separates for each \( i \) into a linear program with upper bounded variables and a single constraint.

Let \( d_i = \sum_{j=1}^{n} d_{ij} y_{ij} \) where the values of \( y_{ij} \) are those defined in Lemma 2 and \( x \) satisfies (3) and \( 0 \leq x \leq 1 \). Note that \( x_{LP} \leq \sum_{i=1}^{n} d_i \) since this bound is derived from a primal feasible solution. This bound will be used repeatedly in our proofs where it is computed for the vector \( x \) defined by \( x_j = k/n \) for \( j = 1, \ldots, n \).

The dual of the LP relaxation is

\[
z_{LP} = \max \sum_{i=1}^{n} u_i - \sum_{j=1}^{n} v_j - kw,
\]

\[
u_i - t_{ij} \leq d_{ij} \quad \text{for all } i, j,
\]

\[
\sum_{i=1}^{n} t_{ij} - v_j - w \leq 0 \quad \text{for all } j,
\]

\[
t_{ij}, v_j \geq 0 \quad \text{for all } i, j.
\]

For any given vector \( u = (u_i; i = 1, \ldots, n) \), define

\[
\rho_j(u) = \sum_{i=1}^{n} (u_i - d_{ij})^{+} \quad \text{for } j = 1, \ldots, n,
\]

where \( a^{+} \) denotes \( \max(0, a) \). Let \( z_D(u) = \sum_{i=1}^{n} u_i - k \max_{j=1, \ldots, n} \rho_j(u) \).

Lemma 3. \( z_{LP} \geq z_D(u) \) for any vector \( u \).

Proof. It can be checked that, for any given \( u \), a feasible solution of (6) is obtained by setting \( t_{ij} = (u_i - d_{ij})^{+} \), \( v_j = 0 \) and \( w = \max_{j=1, \ldots, n} \rho_j(u) \).

2. The Euclidean model in the plane. This section is concerned with the following Euclidean model: \( n \) points \( X_1, \ldots, X_n \) are chosen independently and uniformly at random in the unit square \( S_0 = [0, 1]^2 \). The distance matrix is given by \( d_{ij} = \|X_i - X_j\| \) for \( 1 \leq i, j \leq n \) where \( \| \cdot \| \) denotes the Euclidean norm. We assume that

\[
k \to \infty \quad \text{and } n/(k \log n) \to \infty.
\]

The following theorem was proved by Papadimitriou [22].

Theorem 1. Under the above conditions, \( z_{IP} \sim (0.3771967 \ldots)n/\sqrt{k} \) a.s. where

\[
\sqrt{2/(3\sqrt{3})} \left( \frac{1}{3} + \frac{1}{4} \log 3 \right) = 0.3771967 \ldots.
\]
This result was obtained by comparing $z_{\text{IP}}$ to the value $z_{\text{C}}$ of finding $k$ points in $X = \{ X_1, \ldots, X_n \}$ that minimize the sum of the distances to a continuum of points in the unit square. Papadimitriou showed that, when (7) holds, $z_{\text{IP}} \sim z_{\text{C}}$ almost surely. Actually, he used a weaker notion of probabilistic convergence, but Zemel [26] showed that almost sure convergence holds as well. It should be pointed out, however, that the continuous problem yielding $z_{\text{C}}$ is very different from the LP relaxation. In fact, for the LP relaxation, we prove

**THEOREM 2.** Under the above conditions,

$$z_{\text{LP}} = \frac{2}{3\sqrt{n}} n/\sqrt{k} \quad a.s.$$  

where $2/(3\sqrt{\pi}) = 0.3761264 \ldots$ .

Our method of proof consists of conjecturing a near-optimal solution to the LP relaxation and a near-optimal solution to its dual. Then we show that, almost surely, these lower and upper bounds on $z_{\text{LP}}$ are the same, up to small order terms. The probabilistic arguments are based on the estimates of the tails of the binomial distribution given in Lemma 1.

The proof of Theorem 2 will actually provide a constructive way of obtaining an upper bound $z_{\text{LP}}(x)$ and a lower bound $z_{\text{D}}(u)$ on the optimum value of the LP relaxation of the $k$-median problem.

**COROLLARY 1.** Let $x_j = k/n$ for $j = 1, \ldots, n$ and $u_i = \sqrt{k/\pi}$ for $i = 1, \ldots, n$. Then $z_{\text{D}}(u) \leq z_{\text{LP}} \leq z_{\text{LP}}(x)$ and, under condition (7),

$$z_{\text{D}}(u) \sim z_{\text{LP}} \quad \text{almost surely},$$

$$z_{\text{LP}}(x) \sim z_{\text{LP}} \quad \text{almost surely}.$$

In addition, in [22], Papadimitriou gives a heuristic which almost surely provides a solution with value $z_H \sim z_{\text{IP}}$. The complexity of the heuristic is $O(n \log n)$. Combining this result with the fact that $z_{\text{D}}(u)$ can be computed in linear time, we have a very fast procedure which will almost surely

(i) find a solution with a value close to the optimum,

(ii) prove that the value of this solution is within 0.3% of the optimum.

Finding the exact optimum is much more expensive as will be shown in Theorem 3. But first we give the proof of Theorem 2.

**PROOF OF THEOREM 2.** To obtain a probabilistic upper bound on $z_{\text{LP}}$, we are first going to consider the LP solution $x_j = k/n$ for $j = 1, \ldots, n$ and the values of $y_{ij}$ as defined in Lemma 2. Let $d_i = \sum_{j=1}^n d_{ij} y_{ij}$ for $i = 1, \ldots, n$. We must get a probabilistic estimate of $d_i$ for $i = 1, \ldots, n$. Let $\epsilon = ((k \log n)/n)^{1/3}$, $r = (1/k\pi(1-\epsilon))^{1/2}$ and let $S_r$ be the square $[r, 1-r]^2$. We show first

$$\Pr\left(d_i \geq \frac{2}{3\sqrt{k\pi}} (1 + o(1)) | X_i \in S_r\right) \leq 2e^{-\epsilon^2 n/9k}, \quad (8)$$

$$\Pr\left(d_i \geq \frac{4}{3\sqrt{k\pi}} (1 + o(1)) | X_i \notin S_r\right) \leq 2e^{-4\epsilon^2 n/9k}. \quad (9)$$

If $X_i \in S_r$, then a circle $C_r$ of radius $r$ centered at $X_i$ is entirely contained in $S_0$. The number $N$ of points lying in this circle stochastically dominates the binomial $B(n, \pi r^2)$.
(since \( X_i \in C_i \)). We define independent random variables \( W_j, \ j = 1, 2, \ldots, n \) as follows:

Let

\[
W_j = \begin{cases} 
    d_{ij} & \text{if } X_j \in C_i, \\
    0 & \text{otherwise.}
\end{cases}
\]

We note that \( E(W_j) = 2\pi r^3 / 3 \ (j \neq i) \). If \( N \geq \lfloor n/k \rfloor \) then

\[
d_i \leq \frac{k}{n} \sum_{j=1}^{n} W_j.
\]

Now, by Lemma 1,

\[
\Pr \left( N < \left\lfloor \frac{n}{k} \right\rfloor \right) = \Pr \left( N \leq (1 - \epsilon)n\pi r^2 \right) \leq e^{-\epsilon^2 n\pi r^2 / 2}.
\]

Furthermore, if \( \hat{W}_j = W_j / r \in [0, 1] \), then by Lemma 1,

\[
\Pr \left( \sum_{j=1}^{n} \hat{W}_j \geq (1 + \epsilon)(n - 1) \frac{2\pi r^2}{3} \right) \leq e^{-(\epsilon^2 / 3)(n-1)(2\pi r^3 / 3)}
\]

and (8) follows.

To prove (9), we note that if \( X_i \in S_0 - S_r \), we can at worst find a quadrant of a circle centered at \( X_i \) with radius 2r and contained entirely within \( S_0 \). The area of this quadrant is \( \pi(2r)^2 / 4 \) and we apply the same method as above with \( E(W) = 4\pi r^3 / 3 \).

We are now ready to bound \( z_{LP} \).

\[
z_{LP} \leq \sum_{i=1}^{n} d_i = \sum_{X_i \in S_r} d_i + \sum_{X_i \in S_0 - S_r} d_i.
\]

By Lemma 1,

\[
\Pr \left( \|X \cap S_r\| \leq n(1 - 2r)^2(1 - \epsilon) \right) \leq e^{-(\epsilon^2 / 2)(n(1-2r)^2)}
\]

and thus

\[
\Pr \left( z_{LP} \geq (1 + o(1)) \left((1 - 2r)^2 n \frac{2}{3\sqrt{k}\pi} + (1 - (1 - 2r)^2) n \frac{4}{3\sqrt{k}\pi} \right) \right)
\]

\[
\leq (2n + 1) e^{-(2\epsilon^2 / 9)(n/k)}
\]

giving

\[
z_{LP} \leq (1 + o(1)) \frac{2n}{3\sqrt{k}\pi} \text{ almost surely.} \tag{10}
\]

To obtain a probabilistic lower bound on \( z_{LP} \), we consider the dual problem (6). Let \( u_i = r \) for \( i = 1 \ldots n \). Then by Lemma 3

\[
z_{LP} \geq \sum_{i=1}^{n} u_i - k \max_j \left( \sum_{i=1}^{n} (u_i - d_{ij})^+ \right). \tag{11}
\]

For fixed \( j \), consider random variables \( U_i = (u_i - d_{ij})^+ \).
Setting \( u_i = r \) we find \( E(U_i) = \pi r^3 / 3 \) for \( i \neq j \) and \( X_j \in S_\varepsilon \), whereas these values decrease for points \( X_j \in S_0 - S_\varepsilon \). Rescaling \( U \) to \([0,1]\) and applying Lemma 1 to \( X_j \in S_\varepsilon \) we find

\[
\Pr \left( \sum_{i=1}^{n} U_i \geq (1 + \epsilon) \frac{n \pi r^3}{3} \right) \leq e^{- (\varepsilon^2 / 9)(n/k)}
\]

and thus for \( k = o(n / \log n) \) we have

\[
\max_j \left( \sum_{i=1}^{n} U_i \right) \leq (1 + \epsilon) \frac{n \pi r^3}{3} \quad \text{a.s.}
\]

giving

\[
z_{LP} \geq nr - (1 + \epsilon) k n \pi r^3/3 = (1 - o(1)) \frac{2n}{3\sqrt{k\pi}} \quad \text{a.s.}
\] (12)

Combining this with (10) yields the theorem. \( \blacksquare \)

One might expect then that an LP-based branch and bound procedure performs well, since \( z_{LP} \) provides a good bound. However, we can prove

**Theorem 3.** Assume \( k / \log n \to \infty \) and \( n / k^2 \log n \to \infty \).

Then there exists a constant \( \alpha > 0 \) such that a branch and bound procedure that branches by fixing a variable \( x_j \) to 0 or 1 at each node of the search tree which is not pruned and uses the LP bound to prune the search tree will almost surely explore at least \( n^{\alpha k} \) nodes.

**Proof.** Each node of the branch and bound tree is associated with two sets \( J_0 \) and \( J_1 \) where \( J_t = \{ j : x_j \) is fixed at \( t \) in the associated subproblem \} for \( t = 0, 1 \). Let \( z_{LP}(J_0, J_1) \) denote the LP bound computed at this node, i.e. the value of \( z_{LP} \) when we make the restriction \( x_j = t \) for \( j \in J_t, t = 0, 1 \). We prove the theorem by showing that for some constants \( \beta, \gamma > 0 \) (to be determined) the following holds almost surely:

For any \( J_0, J_1 \subset \{1, \ldots, n\} \) such that

\[
J_0 \cap J_1 = \emptyset, \quad |J_0| \leq \beta n / k \log n, \quad |J_1| \leq \gamma k, \quad \text{we have}
\] (13)

\[
z_{LP}(J_0, J_1) \leq 0.3769 \frac{n}{\sqrt{k}}.
\]

For then we almost surely have to branch at every node in which \( |J_0| \leq \beta n / k \log n \) and \( |J_1| \leq \gamma k \) even if we have an optimal solution of the integer program as our current best solution—by Theorem 1.

This implies that the algorithm must explore at least

\[
\left( \left\lfloor \beta n / k \log n \right\rfloor + \left\lfloor \gamma k \right\rfloor \right) = n^{\gamma (1 - o(1)) k} \text{ nodes.}
\] (14)

To verify (14) imagine that setting \( x_j = 0 \) means branching to the left and setting \( x_j = 1 \) means branching to the right. (13) implies that our tree contains a copy of all possible paths which make \( \left\lfloor \gamma k \right\rfloor \) right branches and \( \left\lfloor \beta n / k \log n \right\rfloor \) left branches. The number of such paths is precisely the left-hand side of (14).
Let $F$ denote the family of such pairs $J_0, J_1$. Thus let $J_0, J_1 \subseteq \{1, \ldots, n\}$ be disjoint, $\bar{J} = \{ j \notin J_0 \cup J_1 \}$, $\bar{n} = |\bar{J}|$, and $\bar{k} = k - |J_1|$. Consider the following solution to the associated linear program.

$$x_j = \begin{cases} 0 & \text{if } j \in J_0, \\ 1 & \text{if } j \in J_1, \\ \bar{k}/\bar{n} & \text{if } j \in \bar{J}. \end{cases}$$

The values of $y_{ij}$ are then defined as in Lemma 2, but only using $j \in \bar{J}$ to form the sequence $j_1(i), j_2(i), \ldots, j_\bar{n}(i)$. This choice of $y_{ij}$ is feasible although usually not optimum. However this is sufficient since we only need to compute an upper bound on $z_{LP}(J_0, J_1)$. We can assume w.l.o.g. that $|J_0| = \lceil \beta n/k \log n \rceil$ and $|J_1| = \lceil \gamma k \rceil$. Let $\epsilon > 0$ be small and $r = \sqrt{1/(1 - \epsilon)}$ and proceed as in the proof of Theorem 2, defining variables $W_1, W_2, \ldots, W_\bar{n}$ for each $i$. We find that for $\epsilon < 1/2$ and $n$ large

$$\Pr\left( z_{LP}(J_0, J_1) > \frac{2n}{3\sqrt{\pi k}} (1 + 3\epsilon) \right) \leq \frac{(2n + 1)e^{-2\epsilon^2 n/3k}}{\sqrt{k}}.$$ 

Since $|F| \leq n^{\beta n/k \log n + \gamma k}$ we find

$$\Pr\left( \exists (J_0, J_1) \epsilon F : z_{LP}(J_0, J_1) > \frac{2n}{3\sqrt{\pi k}} (1 + 3\epsilon) \right) \leq (2n + 1)n^{\beta n/k \log n + \gamma k}e^{-2\epsilon^2 n/3k}. $$

Taking $\beta = \epsilon^2/5$, $\gamma = \epsilon$ and $\epsilon$ sufficiently small that $2(1 + 3\epsilon)/3\sqrt{\pi(1 - \epsilon)} \leq 0.3769$ yields

$$\max\{ z_{LP}(J_0, J_1) : (J_0, J_1) \in F \} \leq 0.3769 \frac{n}{\sqrt{k}}$$ almost surely.

Any $\alpha < \gamma$ can be used to give the theorem. ■

3. A graphical model. This section is concerned with the following graphical model. Let $G$ be a random graph with $n$ nodes, where each edge occurs independently with probability $p$. Let $X_1, \ldots, X_n$ be the nodes of the graph and $d_{ij}$ the minimum number of edges on a path joining $X_i$ to $X_j$ for $1 \leq i, j \leq n$, where the minimum is taken over all paths joining $X_i$ to $X_j$. Thus $d_{ij}$ is the shortest distance between $X_i$ and $X_j$, assuming that all edges have length one.

Let $q = 1 - p$ and $b = 1/q$. The main result of this section is the following theorem.

**THEOREM 4.** (a) Consider $(1 + \epsilon) \log_3 n \leq k \leq n$, where $\epsilon > 0$ is fixed.
(i) If $n^{1+\delta} p \to \infty$ for all $\delta > 0$ fixed, then $z_{IP} = z_{LP}$ almost surely.
(ii) In general, we only have $\lim_{n \to \infty} \Pr(z_{IP} = z_{LP}) = 1$.

(b) Consider $2 \leq k \leq \log_3 n$ and $p \min(1, kp) = (\omega \log n)/n$, where $\omega \to \infty$. Then $(z_{IP} - z_{LP})/z_{IP} \leq 1/(1 + \epsilon)$ almost surely. In addition, if we let $kp \to \alpha$, $0 \leq \alpha \leq \infty$, and $p \to \beta$, $0 \leq \beta < 1$, where $\alpha$ and $\beta$ are fixed, then

$$\frac{z_{IP} - z_{LP}}{z_{IP}} \leq \frac{1 - (1 - \alpha)^+ a^a}{1 + a^a}$$ almost surely,

where $a = e$ if $\beta = 0$ and $(1 - \beta)^{-1/\beta}$ if $\beta > 0$.

See Figure 1.
PROOF OF THEOREM 4(a). This part of the theorem is a careful phrasing of a known result and is easy to prove. As \( d_{ij} \geq 1 \) for \( i \neq j \), we must have

\[
Z_{IP} \geq Z_{LP} \geq n - k. \tag{15}
\]

(i) follows from (15) if we can show that \( Z_{IP} = n - k \) almost surely.

But \( Z_{IP} = n - k \) if and only if there is a set \( K \subseteq X, |K| = k \), such that, for any \( X_j \in X - K \), there exists \( X_i \in K \) such that \( X_i \) and \( X_j \) are joined by an edge of \( G_n(p) \), i.e., \( K \) is a dominating set.

Let \( m = [2/\varepsilon] \) and \( K_i = \{ X_{i(-1)k+1}, \ldots, X_{ik} \} \) for \( i = 1, 2, \ldots, m \). If none of \( K_1, K_2, \ldots, K_m \) are dominating then one of the following events occurs:

- \( E_0 = \exists 1 \leq r \neq s \leq m \) and \( X_i \in K_r \) such that \( X_i \) is not adjacent in \( G_n(p) \) to any vertex of \( K_j \),

- \( E_1 = \bigcap_{i=1}^{m} F_i \) where \( F_i = \{ \exists X_i \in X - \bigcup_{j=1}^{m} K_j \) such that \( X_i \) is not adjacent in \( G_n(p) \) to any vertex of \( K_j \).

Now

\[
\Pr(E_0) \leq m^2 k (1 - p)^k
\]

\[
\leq m^2 kn^{-(1+\varepsilon)}
\]

\[
\leq \frac{m^2 \log n}{np} \quad \text{as} \quad \log n \leq \frac{\log n}{np}
\]

\[
= O(n^{-(1+\varepsilon/2)}) \quad \text{by assumption}.
\]

Furthermore,

\[
\Pr(E_1) = \prod_{i=1}^{m} \Pr(F_i) \quad \text{since the} \ F_i \ \text{are independent},
\]

\[
\leq \left( (n - km)(1 - p)^k \right)^m
\]

\[
\leq n^{-cm}
\]

\[
\leq n^{-2}
\]

and (i) follows.

(ii) \( \Pr(K_1 \text{ is not a dominating set}) \leq (n - k)(1 - p)^k \leq n^{-\varepsilon} \to 0. \)
Our proof of Theorem 4(b) will use the next two lemmas.

**Lemma 4.** Consider $1 \leq k \leq \log_4 n$. Assume $p \min(1, kp) \geq (\omega \log n)/n$, where $\omega \to \infty$. Then

$$z_{IP} = (1 + o(1))(n - k)(1 + q^k) \text{ almost surely.}$$

**Proof.** For $K \subseteq X$, let $N(K)$ be the neighbor set of $K$, i.e.

$$N(K) = \{ X_j \in X - K: \text{there exists an edge joining } X_j \text{ to a node of } K \}.$$  

We have

$$z_{IP} \geq \min_{|K| = k} \left(|N(K)| + 2(n - k - |N(K)|)\right)$$

$$= 2(n - k) - \max_{|K| = k} |N(K)|.$$  

We prove the lemma by showing that

$$\max_{|K| = k} |N(K)| = (1 + o(1))(n - k)(1 - q^k) \text{ almost surely, and (16)}$$

$$z_{IP} = (1 + o(1)) \min_{|K| = k} \left(|N(K)| + 2(n - k - |N(K)|)\right) \text{ almost surely. (17)}$$

Consider a fixed $K \subset X$, $|K| = k$. The quantity $|N(K)|$ is distributed as $B(n - k, 1 - q^k)$. Thus, by Lemma 1, for any small $\epsilon > 0$

$$\Pr[|N(K)| \leq (1 - \epsilon)(n - k)(1 - q^k)] \leq e^{-\frac{1}{2}\epsilon^2(n-k)(1-q^k)} \text{ and}$$

$$\Pr[|N(K)| \geq (1 + \epsilon)(n - k)(1 - q^k)] \leq e^{-\frac{1}{2}\epsilon^2(n-k)(1-q^k)}.$$

Thus we have

$$\Pr\left[\max_{|K| = k} |N(K)| \leq (1 - \epsilon)(n - k)(1 - q^k)\right] \leq e^{-\frac{1}{2}\epsilon^2(n-k)(1-q^k)}, \quad (18)$$

$$\Pr\left[\max_{|K| = k} |N(K)| \geq (1 + \epsilon)(n - k)(1 - q^k)\right] \leq \left(\frac{n}{k}\right) e^{-\frac{1}{2}\epsilon^2(n-k)(1-q^k)}. \quad (19)$$

To obtain (16) we put $\epsilon = 2(k \log(n/k)/(n - k)(1 - q^k))^{\frac{1}{2}}$. We can use $\left(\frac{n}{k}\right) \leq (ne/k)^k$ in (19). Then the right-hand sides in (18) and (19) both go to 0 sufficiently fast. Thus (16) is proved, provided that $\epsilon < 1$.

We consider two cases. Let $0 < \alpha < 1$ be a constant.

When $kp \leq \alpha$, $q^k = (1 - p)^k = [(1 - p)^{1/p}]^{kp} \leq (1/e)^{kp} \leq 1 - kp + (kp)^2/2$. So

$$\frac{\epsilon^2}{4} \leq \frac{k \log n}{(n - k)kp(1 - \alpha/2)} \to 0$$

since $(\log n)/np \to 0$. 


When \( kp > \alpha \), \( q^k = (1 - p)^k \leq e^{-kp} \leq e^{-\alpha} < 1 \). So

\[
\frac{\varepsilon^2}{4} \leq \frac{k \log n}{(n - k)(1 - e^{-\alpha})} \to 0
\]
since \((\log x)/x \to 0\) when \( x \to \infty \).

This completes the proof of (16).

To prove (17) it suffices to show that,

almost surely, every node in \( X - K_1 \) is joined by a path of length
\( \leq 2 \) to at least one node of \( K_1 \) where \( K_1 = \{ X_1, X_2, \ldots, X_k \} \).

(20)

The events

\[ A(j) = \{ X_j \text{ is joined to } K_1 \text{ by an edge} \}, \]

\[ B(j) = \{ X_j \text{ is joined to } K_1 \text{ via a node } X_i \neq X_j, X_i \notin K_1 \} \]

are independent for fixed \( j \) because they have no edges in common.

\[
\Pr(A(j)) = 1 - (1 - p)^k = p_0, \quad \text{say}
\]

\[
\Pr(B(j)) = 1 - (1 - p_0p)^{n-k-1}.
\]

Hence, if \( N \) is the number of nodes not within distance 2 of \( K_1 \), then

\[
\Pr(N > 0) \leq (n - k)(1 - p)^k(1 - p_0p)^{n-k-1}
\]

\[
\leq (n - k)(1 - p_0p)^{n-1}
\]

\[
\leq ne^{-(n-1)p_0p}.
\]

If \( kp \geq 1 \) then \( p_0 \geq 1 - e^{-1} \) and so \( \Pr(N > 0) \leq n^{-\omega/2} \) using \( p \geq (\omega \log n)/n \).

If \( kp < 1 \) then \( (1 - p)^k \leq 1 - kp + k^2p^2/2 \) and hence \( p_0 \geq kp/2 \) and then

\( \Pr(N > 0) \leq n^{-\omega/2} \) using \( kp^2 \geq (\omega \log n)/n \). This proves (20) and therefore (17) and the lemma.

**Lemma 5.** Consider \( 2 \leq k \leq \log_b n \). Assume \( p \geq (\omega \log n)/n \) and \( kp^2 \geq \omega/n \) where \( \omega \to \infty \). Then

\[
z_{LP} = \max(n - k, 2n - nkp(1 + o(1))) \quad \text{almost surely.}
\]

**Proof.** Given a node \( X_i \), let \( N_1(i) = \{ X_j : d_{ij} = 1 \} \) and \( N_2(i) = \{ X_j : d_{ij} = 2 \} \).

First we give probabilistic estimates of \( |N_1(i)| \) and \( |N_2(i)| \). We will show

\[
\min_i |N_1(i)| = (1 - o(1))np \quad \text{almost surely,}
\]

(21)

\[
\max_i |N_1(i)| = (1 + o(1))np \quad \text{almost surely, and}
\]

(22)

\[
\min_i |N_2(i)| \geq \min\left(\frac{n}{k}, (1 - o(1))nq\right) \quad \text{almost surely.}
\]

(23)
Note that $|N_1(i)|$ is distributed as $B(n - 1, p)$. So, by Lemma 1,

$$\Pr\left( \min_i |N_1(i)| \leq (1 - \epsilon)(n - 1)p \right) \leq ne^{-\frac{1}{2}(n-1)p}$$

$$\Pr\left( \max_i |N_1(i)| \geq (1 + \epsilon)(n - 1)p \right) \leq ne^{-\frac{1}{2}(n-1)p}.$$ Putting $\epsilon = 3(\log n/(n - 1)p)^{1/2}$ yields (21) and (22).

Now consider $|N_2(i)|$. We will assume $p \to 0$ (otherwise $N_1(i)$ is a dominating set by Theorem 4(a), and (23) follows). Conditional on $|N_1(i)|$, the quantity $|N_2(i)|$ is distributed as $B(n_2, p_2)$, where $n_2 = n - |N_1(i)| - 1$ and $p_2 = 1 - (1 - p)^{|N_1(i)|}$. By Lemma 1,

$$\Pr\left( \min_i |N_2(i)| \leq (1 - \epsilon)n_2 p_2 \right) \leq ne^{-\frac{1}{2}n_2 p_2}.$$ Set $\epsilon = 3(\log n/n_2 p_2)^{1/2}$. We have to show $\epsilon < 1$. Note that $n_2 = (1 - o(1))n$ and $p_2 = 1 - (1 - p)^{(1+o(1))np} \geq 1 - e^{-(1+o(1))np^2}$ almost surely.

If $np^2 \geq \delta > 0$ where $\delta$ is fixed, then

$$\frac{\epsilon^2}{4} \leq \frac{\log n}{(1 + o(1))n(1 - e^{-\delta})} \to 0.$$ If $np^2 = o(1)$, then

$$\frac{\epsilon^2}{4} \sim \frac{\log n}{n^2 p^2} = \frac{1}{\log n} \left( \frac{\log n}{np} \right)^2 \to 0.$$ So we have just shown that, almost surely,

$$\min_i |N_2(i)| \geq (1 - o(1))n_2 p_2.$$ Next we will use the fact that $kp^2 \geq \omega/n$ to show $n_2 p_2 \geq n/k$ almost surely.

If $np^2 \geq \delta, 0 < \delta < 1$ fixed, then almost surely $n_2 p_2 \geq (1 + o(1))n(1 - e^{-\delta}) \geq n/k$ for $k \geq 2$ and $\delta$ close enough to $1$.

If $np^2 \leq \delta < 1$, then $1 - e^{-(1+o(1))np^2} \geq np^2(1 - np^2/2)$. So

$$n_2 p_2 \geq (1 + o(1))n_2 p^2 \left( 1 - \frac{\delta}{2} \right) \geq (1 + o(1)) \frac{n \omega k}{k} \left( 1 - \frac{\delta}{2} \right) \geq \frac{n}{k}$$ almost surely.

This completes the proof of (23).

Now we are ready to get a probabilistic estimate of $z_{LP}$. First we obtain an upper bound by considering the solution

$$x_j = \frac{k}{n} \quad \text{for } j = 1, \ldots, n \quad \text{and} \quad y_{ij} \text{ defined in Lemma 2.} \quad (24)$$ Let $\delta = \min_i |N_1(i)|$ be the minimum degree of $G_n(p)$. Note that, if $\delta \geq n/k - 1$, then $z_{LP} = n - k$. For, using the solution (24), we have $d_i = \sum_{j=1}^{n} x_{ij} = 1 - k/n$ for $i = 1, \ldots, n$. On the other hand, if $\delta < n/k - 1$, then

$$d_i \leq \frac{k}{n} \delta + 2 \frac{k}{n} \left( \frac{n}{k} - 1 - \delta \right).$$
(y_{ij} only takes positive values for points X_i at distance one or two of X_i since, by (23), the number of points at distance 2 is at least min(n/k, (1 - o(1))nq) which is more than the n/k - 1 - δ points needed.) Therefore z_{LP} \leq n \sum_{i=1}^{n} d_i \leq 2n - k\delta, almost surely.

To obtain a probabilistic lower bound for z_{LP} we consider the dual bound given by Lemma 3. We put u_i = 2 - 1/n for i = 1, ..., n and let Δ denote the maximum degree of G_n(p). Then

\[ z_{LP} \geq n \left( 2 - \frac{1}{n} \right) - k(1 - \frac{1}{n}) = 2n - (1 + o(1))nk \] 

almost surely.

This completes the proof of Lemma 5. ■

PROOF OF THEOREM 4(b). It follows from Lemmas 4 and 5 that

\[ \frac{Z_{IP} - z_{LP}}{Z_{IP}} \sim \frac{(1 + q^k) - \max(1, 2 - kp)}{1 + q^k} \] 

almost surely

\[ = \frac{q^k - (1 - kp)^+}{q^k + 1}. \]

Setting a = (1 - p)^{-1/p} and kp = α, we get

\[ \frac{Z_{IP} - z_{LP}}{Z_{IP}} \sim \frac{1 - (1 - α)^+ a^α}{1 + a^α} \] 

almost surely.

It is easy to check that the maximum of this function is achieved when p → 0 and α = 1. Then its value is 1/(1 + ε). ■

An interesting range of parameters which is not considered in Theorem 4 is the case 2 \leq k \leq \log_b n and p \geq (ω \log n)/n \geq kp^2 where ω → ∞. In this range, the expressions for Z_{IP} and z_{LP} are more complicated than those found in Lemmas 4 and 5. However we conjecture that (Z_{IP} - z_{LP})/Z_{IP} → 0 almost surely.

In the range covered by Theorem 4(b), it is easy to identify conditions under which the ratio (Z_{IP} - z_{LP})/Z_{IP} is almost surely bounded away from 0. For example, consider

\[ ε \leq kp \leq 1/ε, \quad k \geq 2 \] 

and

\[ (\omega \log n)/n^{1/2} \leq p \leq 1 - ε \]

where ω → ∞ and 0 < ε < 1 is fixed.

Then

\[ k \log b = k \log \left( 1 + \frac{p}{1 - p} \right) \leq \frac{kp}{1 - p} \leq \frac{1}{ε^2}. \]

So, for n large enough, k \leq \log n/\log b = \log_b n. Therefore, by Theorem 4(b), there is a fixed value f(ε) > 0 such that

\[ \frac{Z_{IP} - z_{LP}}{Z_{IP}} \geq f(ε) \] 

almost surely.

In addition, we can show that, under these conditions, a branch and bound algorithm based on the LP bound z_{LP} almost surely requires close to complete enumeration.
THEOREM 5. Assume (25) and (26). A branch and bound procedure that branches by fixing a variable \( x_j \) to 0 or 1 at each node of the search tree which is not pruned, and uses the LP bound to prune the search tree, will almost surely expand at least \( n^{(1-o(1))(k-2)} \) nodes. (The number of feasible solutions of the k-median problem is \( \binom{n}{k} \).)

PROOF. We first note that, under the above assumptions, \( \epsilon \leq k \log b \leq 1/\epsilon^2 \) and therefore

\[
e^{-1/\epsilon^2} \leq q^k \leq e^{-\epsilon}.
\] (28)

In addition, the assumptions of Lemma 4 hold and \( k = o(n^{1/2}) \) so that

\[
z_{1p} \geq (1 - o(1))n(1 + q^k) \quad \text{almost surely.} \tag{29}
\]

Let \( z_{LP}(J_0, J_1) \) be the LP value of the subproblem where \( J_0 = \{ j : x_j \text{ is fixed to 0} \} \) and \( J_1 = \{ j : x_j \text{ is fixed to 1} \} \).

Let \( \alpha < 1 \) and \( \beta > 0 \) be fixed. We prove the theorem by showing that, for \( \beta \) chosen small enough, the following property holds almost surely.

(30) For any \( J_0, J_1 \subseteq \{1, 2, \ldots, n\} \) such that \( J_0 \cap J_1 = \emptyset \), \( |J_0| = [\beta n] \) and \( |J_1| = [\alpha k] \),

\[
z_{LP}(J_0, J_1) < z_{1p}.
\] (31)

This implies that the algorithm must explore at least

\[
\left( \begin{align*}
[\beta n] + [\alpha k] \\
[\alpha k]
\end{align*} \right) \geq \left( \frac{\beta n}{\alpha k} \right)^{ak} = n^{(1-o(1))ak} \text{ nodes.} \tag{32}
\]

To verify (32), imagine that setting \( x_j = 0 \) means branching to the left and setting \( x_j = 1 \) means branching to the right. (30)–(31) imply that any tree contains all possible paths which make \( [ak] \) right branches and \( [\beta n] \) left branches. The number of such paths is precisely the left hand side of (32).

We now turn to the proof of (31). As increasing \( J_0 \) or \( J_1 \) only serves to increase \( z_{LP} \) we can restrict our attention to \( |J_0| = [\beta n] \) and \( |J_1| = [\alpha k] \).

Using Lemma 1 we can easily prove that the following holds almost surely for \( G_n(p) \):

\[
J \subseteq \{1, 2, \ldots, n\} \text{ and } |J| = [ak] \text{ implies } |N(J)| \geq (1 - o(1))n(1 - q^{ak}) \quad \text{(see (18))}.
\] (33)

Furthermore, it is easy to see that

\[
\text{diam}(G_n(p)) = 2 \quad \text{almost surely,} \tag{34}
\]

where diam refers to the diameter of \( G_n(p) \).

Indeed

\[
\Pr(\text{there exists } i, j \in \{1, 2, \ldots, n\} \text{ such that } i, j \text{ are not joined by a path of length 2}) \leq \left( \begin{align*}
\frac{n}{2} \\
(1 - p^2)^{n-2}
\end{align*} \right)
\leq n^2e^{-(n-2)p^2}
\leq n^2e^{-(\omega \log n)(n-2)/n} \rightarrow 0.
\]
Thus (34) is proved \( \Pr[\text{diam}(G_n(p)) = 1] = p \binom{n}{2} \rightarrow 0 \). To obtain an upper bound on \( z_{\text{LP}}(J_0, J_1) \) let

\[
x_j = \begin{cases} 
0 & \text{if } j \in J_0, \\
1 & \text{if } j \in J_1, \\
\gamma & \text{if } j \notin J_0 \cup J_1,
\end{cases}
\]

where \( \gamma = (k - \lfloor ak \rfloor)/(n - \lfloor \beta n \rfloor - \lfloor ak \rfloor) \).

The values for \( y_{ij} \) are then chosen as follows:

- \( i \in J_1 \): \( y_{ii} = 1 \) and \( y_{ij} = 0, j \neq i, \)
- \( i \in N(J_1) \): \( y_{it} = 1 \) and \( y_{ij} = 0, j \neq t, \) where \( t \) is a node of \( J_1 \cap N(i) \).
- \( i \notin J_1 \cup N(J_1) \): the values are as defined in Lemma 2.

With this solution we find, using (34) that

\[
d_i = \begin{cases} 
0 & \text{if } i \in J_1, \\
1 & \text{if } i \in N(J_1),
\end{cases}
\]

\[
\leq \gamma(\delta - s_i) + 2(1 - \gamma(\delta - s_i)) \text{ if } i \notin J_1 \cup N(J_1),
\]

where \( s_i = |N(i) \cap J_0| \), \( \delta \) is the minimum node degree and \( \Delta \) is the maximum node degree in \( G_n(p) \).

To compute an upper bound on \( z_{\text{LP}} \), we will distinguish between the cases \( \gamma \delta < 1 \) and \( \gamma \delta > 1 \).

First assume that \( \gamma \delta > 1 \). We use the bound

\[
d_i \leq \gamma(\delta - s_i) + 2(1 - \gamma(\delta - s_i)) \leq 1 + \gamma s_i.
\]

\[
z_{\text{LP}} \leq |N(J_1)| + \sum_{i \in J_1 \cup N(J_1)} (1 + \gamma s_i)
\]

\[
\leq |N(J_1)| + n - |N(J_1)| + \gamma \Delta |J_0|
\]

\[
= n + \frac{\beta(k - \lfloor ak \rfloor) p}{1 - \beta} n + o(n).
\]

Since \( kp \) is bounded above by a constant as a consequence of (25), we simply choose \( \beta \) small enough to get our bound on \( z_{\text{LP}} \). Then (31) follows from (28) and (29).

Now assume that \( \gamma \delta \leq 1 \). We use the bound

\[
d_i \leq \gamma(\delta - s_i) + 2(1 - \gamma(\delta - s_i)) = 2 - \gamma \delta + \gamma s_i.
\]

\[
z_{\text{LP}} \leq |N(J_1)| + \sum_{i \in J_1 \cup N(J_1)} (2 - \gamma \delta + \gamma s_i)
\]

\[
= |N(J_1)| + (2 - \gamma \delta)(n - |N(J_1)|) + \gamma \Delta |J_0|
\]

\[
= (2 - \gamma \delta)n - (1 - \gamma \delta)|N(J_1)| + \gamma \Delta |J_0|
\]

\[
\leq \left[ 1 + q^{\lfloor ak \rfloor} \left( 1 - \frac{k - \lfloor ak \rfloor}{1 - \beta} p \right) + \frac{\beta(k - \lfloor ak \rfloor)}{1 - \beta} p \right] n + o(n)
\]
where the last inequality follows from the relations
\[
|N(J_1)| \geq (1 - o(1)) n (1 - q^{[a k]}),
\]
\[
\gamma \delta = (1 + o(1)) \frac{k - \lfloor ak \rfloor}{1 - \beta} p,
\]
\[
\Delta = (1 + o(1)) np.
\]
Therefore
\[
z_{IP} - z_{LP} \geq \left[ q^{[ak]} ((1 - p)^m - 1 + mp) - \frac{\beta mp}{1 - \beta} (1 - q^{[ak]}) \right] n + o(n)
\]
where \( m = k - \lfloor ak \rfloor \). Note that \( \gamma \delta - mp/(1 - \beta) \leq 1 \) implies \( mp < 1 \).

Next we show that \( S_m = (1 - p)^m - 1 + mp \) is bounded below by a positive constant. This will imply that \( z_{IP} - z_{LP} > 0 \) by choosing \( \beta \) small enough.

We assume that \( \alpha \) is chosen so that \( \alpha \leq 1 - 2/k \). This implies \( m \geq 2 \). Now
\[
S_m = S_{m-1} + p \left( 1 - (1 - p)^{m-1} \right)
\]
\[
\geq S_{m-1} + p \left( 1 - e^{-(m-1)p} \right)
\]
\[
\geq S_2 + p \sum_{i=3}^{m-1} (1 - e^{-(i-1)p})
\]
\[
\geq p^2 + p (m - m/2) (1 - e^{(m/2)-1)p}).
\]

If \( k \) is fixed, then \( p \) is bounded below by a constant as a consequence of (25). Therefore \( S_m \) is bounded below by a constant.

If \( k \to \infty \), then \( m \sim (1 - \alpha) k \). Thus \( mp \) and hence \( S_m \) is bounded below by a constant using (25).

This completes the proof of (31). Note that (32) and the condition \( \alpha \leq 1 - 2/k \) imply the bound \( n^{(1-o(1))(k-2)} \) announced in the statement of the theorem.

In [4], a different graphical model is associated with the variation of the \( k \)-median problem known as the \( k \)-plant location problem. The \( k \)-plant location problem is defined using two sets \( X = \{ X_1, \ldots, X_n \} \) and \( Y = \{ Y_1, \ldots, Y_m \} \). The quantity \( d_{ij} \) is defined for each \( 1 \leq i \leq m \) and \( 1 \leq j \leq n \). The problem consists of finding a set \( S \subseteq X \), \( |S| = k \), that minimizes \( \sum_{i=1}^{m} \min_{j \in S} \) \( d_{ij} \).

A \( k \)-plant location problem arises from a graph \( G \) by defining \( X \) as its node set, \( Y \) as its edge set and \( d_{ij} = 0 \) if \( X_j \) is incident with \( Y_i \), 1 otherwise. (The problem is to find \( k \) nodes that cover the maximum number of edges of \( G \).) It is shown in [4] that \( z_{IP} = z_{LP} \) almost surely when \( G = G_n(p) \) is a random graph with \( 0 < \epsilon \leq p \leq 1 - \epsilon \), \( \epsilon \) fixed, and \( k \ll n^\alpha \), \( \alpha < 1/6 \) fixed.

4. A tree model. This section is concerned with the following tree based model: we are given a random tree \( T_n \) with node set \( X = \{ X_1, \ldots, X_n \} \) where each of the \( n^{n-2} \) different trees is equally likely to occur. The distance \( d_{ij} \) is the number of edges in the unique path from \( X_i \) to \( X_j \) in \( T_n \). This section contains a probabilistic result (Theorem 7) and a deterministic one (Theorem 6).

Kolen [13] proved that \( z_{IP} = z_{LP} \) for every SPLP defined on a tree. For the \( k \)-median problem, this equality does not always hold as shown in Theorem 6.
we show in Theorem 7 that, for random trees on $n$ nodes, the number of values of $k$ such that $z_{LP} \neq z_{IP}$ is almost surely at least $cn$, for some constant $c > 0$.

**Theorem 6.** (a) For $k = 1$ or $k \geq \lfloor(n - 1)/2 \rfloor$, $z_{IP} = z_{LP}$ for every tree on $n$ nodes.  
(b) For $2 \leq k < \lfloor(n - 1)/2 \rfloor$, and $n \neq 8$, there is a tree on $n$ nodes such that $z_{IP} \neq z_{LP}$.  
(c) There is an infinite family of trees such that $(z_{IP} - z_{LP})/z_{IP} \rightarrow (k - 1)/2k$.

It would be interesting to perform a worst-case analysis of the $k$-median problem and its LP relaxation on trees. We conjecture that the ratio $(k - 1)/2k$ found in (c) is the worst-case bound.

**Proof of Theorem 6.** For the 1-median problem, it is well known that $z_{IP} = z_{LP}$ for every choice of $d_{ij}$, $1 \leq i, j \leq n$. For example, this result appears in Mukendi [18].

When $k \geq \lfloor n/2 \rfloor$, $z_{IP} = z_{LP} = n - k$ follows from the fact that every tree on $n$ nodes has a dominating set of cardinality at most $\lfloor n/2 \rfloor$. (A tree is bipartite and a color class dominates it.)

To complete the proof of Theorem 6(a), it suffices to consider the case where $n$ is even and $k = n/2 - 1$. By induction, one can show that the only trees which do not have a dominating set of size $k$ are constructed inductively from a path with 4 nodes by adding paths $P_i = (v'_i, v_2, v_1)$ where $v'_i$ is one of the nonleaf nodes of the current tree and $v'_2, v'_3$ are two new nodes. (See Figure 2(a).) From the construction $z_{IP} = n - k + 1 = n/2 + 2$. Using the dual values $u_j = 2$ if $X_j$ is a leaf, 1 if not, Lemma 3 yields $z_{LP} \geq n/2 + 2$. Therefore $z_{IP} = z_{LP}$.

To prove Theorem 6(b) when $n$ is odd, consider the tree of Figure 2(b). Let $p = (n - 1)/2$. An optimal solution of the $k$-median problem is to take $S = \{X_1, X_2, X_4, X_6, \ldots, X_{2(k-1)}\}$. Then $z_{IP} = 3p - 2(k - 1)$. We get a feasible solution of the LP relaxation by setting $x_i = (p - k)/(p - 1)$ and $x_{2i} = (k - 1)/(p - 1)$ for $i = 1, \ldots, p$. This yields $z_{LP} \leq (3p^2 - 2pk - p + k - 1)/(p - 1)$. Therefore $z_{IP} - z_{LP} \geq (k - 1)/(p - 1) > 0$.

To prove Theorem 6(b) when $n$ is even, $n \neq 8$, we first consider the case $k \geq 3$. Add a node $X_{2p+2}$ adjacent to $X_{2p}$ to the tree of Figure 2(b). Then it is optimum to choose $X_{2p}$ in $S$ and we can also choose $x_{2p} = 1$ in the LP solution. Removing $X_{2p}$, $X_{2p+1}$ and $X_{2p+2}$, we are back to the case where $n$ is odd and $k \geq 2$. Now consider the case $n \geq 10$ even and $k = 2$. Add three nodes to the graph of Figure 2(b), namely $X_{2p+1+i}$ adjacent to $X_{2i}$ for $i = 1, 2, 3$. Then $z_{IP} = 3p + 3$ but there is a better LP solution, namely $x_1 = 1$ and $x_2 = x_4 = x_6 = 1/3$. This yields $z_{LP} = 3p + 1$.

![Figure 2](image-url)
Finally, to prove Theorem 6(c), consider the tree of Figure 2(c). The node $X_1$ has degree $k + 1$ in the tree. Each branch incident with $X_1$ contains $b$ nonleaf nodes and $l$ leaf nodes where $b \to \infty$, $l \to \infty$ and $l$ grows much faster than $b$. We denote by $X_2, \ldots, X_{k+2}$ the $(k + 1)$ nodes of the tree which are incident with leaves. Then, an optimal solution of the $k$-median problem is $\{X_1, X_2, X_3, \ldots, X_k\}$.

$$z_{IP} = (k - 1)l + 2(b + 1)l + O(kb^2)$$

where the last term accounts for all the nonleaf nodes. Ignoring the lower order terms,

$$z_{IP} \sim 2bl.$$ 

To get an optimal LP solution, set $x_1 = 1/k$ and $x_j = (k - 1)/k$ for $j = 2, \ldots, k + 2$.

$$z_{LP} = (k + 1)lx \left(\frac{k - 1}{k}\right) + (k + 1)lx(b + 1)\frac{1}{k} + O(kb^2), \quad \text{i.e.}$$

$$z_{LP} \sim \frac{k + 1}{k}bl.$$ 

Therefore

$$\frac{z_{IP} - z_{LP}}{z_{IP}} \sim \frac{2 - (k + 1)/k}{2} = \frac{k - 1}{2k}. \quad \blacksquare$$

In the next theorem we consider all the $k$-median problems defined on a tree, namely all $1 \leq k \leq n$ where $n$ is the number of nodes in the tree.

**Theorem 7.** Let $T_n$ be a random tree. There exists a positive constant $c$ such that, almost surely, $z_{IP} \neq z_{LP}$ for at least $cn$ different values of $k$.

To prove this theorem we need a lemma (Lemma 6 below). Consider a random tree $T_n = (V_n, E_n)$ and a fixed tree $T = (V, E)$. Let $v \in V$. We say that $T_n$ contains a copy of $T$ suspended at $v$ if there exists $V' \subseteq V_n$ such that

(35) The subgraph of $T_n$ induced by $V'$ is isomorphic to $T$ under a mapping $\phi : V \to V'$.

(36) There is a unique edge of $T_n$ with exactly one end, say $v'$, in $V'$ and, in addition, $v' = \phi(v)$.

Let $m = |V|$ and $a = \text{the number of automorphisms of } T$. Then $m! / a$ is the number of distinct labeled graphs on $m$ nodes which are isomorphic copies of $T$.

**Lemma 6.** Almost surely $T_n$ contains at least $(1 - o(1))(n/e^m a)$ copies of $T$ suspended at $v$.

**Proof.** Let $N$ be the number of copies of $T$ suspended at $v$ contained in $T_n$. Now for a fixed copy of $T$ on a set of $m$ nodes, there are $(n - m)^{n - m - 1}$ ways of choosing a tree on the remaining $n - m$ nodes and then joining it to $v$. Thus

$$E(N) = \binom{n}{m} (m! / a) (n - m)^{n - m - 1} / n^{n - 2}$$

$$\sim n/e^{m a}.$$ 

Using the Markov inequality $\Pr(Y \geq \alpha) \leq E(Y) / \alpha$ with $Y = (N - E(N))^4$, $E(Y) = \mu_4$ and $\alpha = \lambda^4 \mu_4$ we get the Pearson extension of the Chebychev inequality:

$$\Pr\{|N - E(N)| \geq \lambda \mu_4^{1/4}\} \leq \frac{1}{\lambda^4}.$$
In terms of factorial moments \( \mu_i \) is given by

\[
\mu_4 = \mu_{[4]} - 4\mu_{[1]}\mu_{[3]} + 6\mu_{[1]}^2\mu_{[2]} - 3\mu_{[1]}^4
+ 6\mu_{[3]} - 12\mu_{[1]}\mu_{[2]} + 6\mu_{[1]}^3
+ 7\mu_{[2]} - 4\mu_{[1]}^2
+ \mu_{[1]}
\]

where \( \mu_{[i]} \) is the \( i \)th factorial moment:

\[
\mu_{[i]} = \frac{n!}{(m!)^i(n-im)!} \left( \frac{m!}{a} \right)^i \frac{(n-im)^{n-im-2}}{n^{n-2}} (n-im)^i.
\]

We find

\[
\mu_{[2]} = \mu_{[1]}^2 e^{-2m/n+O(1/n^2)}, \quad \mu_{[3]} = \mu_{[1]}^3 e^{-6m/n+O(1/n^2)}, \quad \mu_{[4]} = \mu_{[1]}^4 e^{-12m/n+O(1/n^2)}.
\]

In the expression for \( \mu_4 \) above, the first row is the powers of \( n^4 \). When we evaluate this row we find that terms in \( 1 \) and \( 1/n \) of the exponentials disappear simultaneously, leaving a term in \( \mu_{[1]}^4 O(1/n^2) \), i.e. \( O(n^2) \). Similarly in the next row (powers of \( n^3 \)) the terms in \( 1 \) of the exponentials disappear simultaneously leaving \( \mu_{[1]}^3 O(1/n) \), i.e. \( O(n^2) \). The last two rows are \( O(n^2) \).

Thus \( \mu_4 = O(n^2) \) and setting \( \lambda = n^{1+\delta} \) gives

\[
\Pr\left( \left| N - \left(1 + \epsilon \right) \frac{n}{e^{m/a}} \right| \geq n^{3/4} \right) = O\left( \frac{1}{n^{1+\delta}} \right).
\]

This completes the proof of Lemma 6. ■

**Proof of Theorem 7.** Consider the fixed tree \( T \) given in Figure 3. Let \( S_k \) be an optimal \( k \)-median solution in \( T_n \). We will let \( k \) increase from 1 to \( n \). Consider any copy of \( T \) suspended at \( v \) contained in \( T_n \), say \( (V', E') \). Note that, if \( |V' \cap S_k| \geq 1 \), then \( v' = \phi(v) \in S_k \). This implies that there exists a \( K \) such that for \( k \geq K, |V' \cap S_k| \) is a nondecreasing function of \( k \) which goes from 1 to 15 (= \( m \)).

Let \( z_{1p}(V') = \sum_{i \in V'} \min_{j \in S_k} d_{ij} \). When \( |V' \cap S_k| = 3 \), an optimal set \( V' \cap S_k \) is \( \{ v, X_1, X_2 \} \) with \( z_{1p}(V') = 14 \). However, consider the fractional solution \( x_1 = x_2 = x_3 \)
= x_d = \frac{1}{2}, x_j = 1 \text{ for the variable associated with node } v, \text{ and } x_j = 0 \text{ for the other nodes of } V'. \text{ Let } y_{ij} \text{ be defined as in Lemma 2 and } z_{\text{LP}}(V') = \sum_{i \in V} d_{ij} y_{ij}. \text{ The above fractional solution yields } z_{\text{LP}}(V') = 13.5 \text{ and therefore, when } |V' \cap S_4| = 3, z_{\text{IP}} > z_{\text{LP}}.

Since } T_n \text{ contains almost surely at least } (1 - o(1)) n/e^m a \text{ copies of } T \text{ suspended at } v, \text{ there are at least } k \text{ many values of } k \text{ for which } z_{\text{IP}} > z_{\text{LP}}.

5. The uniform cost model. In this section, we look briefly at the model where the } d_{ij} \text{'s are drawn independently from the } [0,1] \text{ uniform distribution, } 1 \leq i, j \leq n.

Here we do not assume } d_{ii} = 0, d_{ij} = d_{ji} \text{ or } d_{ij} = d_{ik} + d_{kj}, \text{ as we did in the other models. The quantity } d_{ij} \text{ is interpreted as the cost of assigning } X_i \text{ to } X_j.

The main result of this section states that, when } k \geq n(e - 1)/e, \text{ then } z_{\text{IP}} = z_{\text{LP}} \text{ almost surely, and when } k = o(n/\log n), \text{ then } (z_{\text{IP}} - z_{\text{LP}})/z_{\text{IP}} \sim (k - 1)/2k \text{ almost surely. The analysis is made possible by the fact that, in those ranges, the } k\text{-median problem is almost surely trivial to solve exactly or approximately.} \text{ (When } k \geq n(e - 1)/e \text{ there is an obvious optimal solution, and when } k = o(n \log n) \text{ every solution is close to optimum.)}

**Theorem 8.** (a) Suppose } k = o(n/\log n). \text{ Then}

\[ z_{\text{IP}} \sim n/(k + 1) \text{ almost surely,} \]
\[ z_{\text{LP}} \sim n/2k \text{ almost surely.} \]

(b) Suppose } k \geq (1 + o(1)) n(e - 1)/e. \text{ Then}

\[ z_{\text{IP}} = z_{\text{LP}} \text{ almost surely.} \]

**Proof.** Let } S \text{ be a fixed set of size } k. \text{ If we take } x_j = 1 \text{ for } j \in S \text{ as our solution to the integer program, then the } d_i = \min_{j \in S} d_{ij} \text{ are independently distributed as the minimum of } k \text{ uniform } [0,1] \text{ random variables, i.e. } \Pr(d_i \geq a) = a^k \text{ for } 0 \leq a \leq 1, \text{ and hence } E(d_i) = 1/(k + 1) \text{ for } i = 1, \ldots, n.

We first consider } k = O(n^{1/5}). \text{ Applying Lemma 1 to } D = d_1 + \cdots + d_n, \text{ we have}

\[ \Pr(|D - n/(k + 1)| \geq \epsilon n/(k + 1)) \leq 2e^{-\epsilon^2 n/3(k + 1)}. \]

Now put } \epsilon = n^{-1/5}. \text{ In addition, } \left(\frac{n}{k}\right)e^{-\epsilon^2 n/3(k + 1)} = o(e^{-n/3}), \text{ so } z_{\text{IP}} \sim n/k + 1 \text{ almost surely.}

We now consider } k/n^{1/6} \to \infty. \text{ Then } \omega = (n/k \log n)^{1/4} \to \infty. \text{ Set}

\[ \delta_i = \begin{cases} d_i & \text{if } d_i \leq \omega/k, \\ 0 & \text{otherwise,} \end{cases} \]

and note that

\[ E(\delta_i) = \frac{1}{k + 1} \left(1 - \left(1 - \frac{\omega}{k}\right)^k \right) \left(\omega - 1\right) = \frac{1 - o(1)}{k + 1}. \]

We rescale the } \delta_i \text{ to } [0,1] \text{ and apply Lemma 1.

\[ \Pr\left(\sum_{i=1}^n \delta_i/(\omega/k) \leq \left(1 - \frac{1}{\omega}\right)nE(\delta_i/(\omega/k))\right) \leq e^{-(n/2\omega^3)(1-o(1))}. \]
As \( \delta_i \leq d_i \) we deduce

\[
\Pr\left( z_{IP} \leq n \frac{1 - o(1)}{k + 1} \right) \leq \binom{n}{k} e^{-\left(n/2\omega\right)(1-o(1))}
\]

and hence if \( k = o(n/\log n) \)

\[
z_{IP} \geq \frac{(1 - o(1))n}{k + 1} \quad \text{almost surely.}
\]

On the other hand, taking \( S = \{1, 2, \ldots k\} \) we can show easily that

\[
z_{IP} \leq \frac{(1 + o(1))n}{k + 1} \quad \text{almost surely.}
\]

Now we prove the second part of Theorem 8(a). We put \( x_j = k/n \) as usual. Then \( d_i \) is dominated probabilistically by \( k/n \) times the sum of the \([n/k]\) smallest out of \( n \) independent \([0, 1]\) uniform random variables. Thus

\[
E(d_i) \leq (k/n) \sum_{t=1}^{[n/k]} t/(n + 1) = (1 + o(1))/2k.
\]

Applying Lemma 1 in the usual way shows that

\[
z_{LP} \leq (1 + o(1))n/2k \quad \text{almost surely.}
\]

On the other hand, consider the dual solution \( u_i = 1/k \) for \( i = 1, \ldots, n \). Then, by Lemma 3,

\[
z_{LP} \geq \sum_{i=1}^{n} u_i - k \max_{j=1, \ldots, n} \left( \sum_{i=1}^{n} (u_i - d_{ij})^+ \right).
\]

As in Theorem 2, for fixed \( j \), we consider random variables \( U_i = (u_i - d_{ij})^+ \). Setting \( u_i = 1/k \) we find \( E(U_i) = 1/2k^2 \). Rescaling the \( U_i \) to \([0, 1]\) and applying Lemma 1

\[
\Pr\left( \sum_{i=1}^{n} kU_i \geq (1 + \epsilon) \frac{n}{2k} \right) \leq e^{-\left((\epsilon^2/3)(n/2k)\right)}.
\]

Hence, if \( n/k = \theta^3 \log n \) where \( \theta \to \infty \), then taking \( \epsilon = 1/\theta \) yields

\[
z_{LP} \geq \frac{n}{k} - k(1 + \epsilon) \frac{n}{2k^2} \quad \text{almost surely}
\]

\[
= (1 - o(1)) \frac{n}{2k}.
\]

This completes the proof of Theorem 8(a).

Now consider the case where \( k \) is sufficiently large so that each point \( X_i \) can be assigned to the cheapest point \( X_{j(i)} \), defined by \( d_{j(i)} = \min_{j = i, \ldots, n} d_{ij} \). Then clearly \( z_{IP} = z_{LP} \).

For \( j = 1, \ldots, n \), let \( N_j \) be the number of points \( X_i \) assigned to \( X_j \) according to the above scheme. \( N_j \) is asymptotically distributed according to a Poisson process with mean 1; in particular \( \Pr(N_j = 0) \sim 1/e \). Therefore \( E(|\{j : N_j = 0\}|) \sim n/e \). To show
$Z = |\{ j : N_j = 0 \}| \leq (1 + o(1))n/e$ almost surely we use the generalized Markov inequality $\Pr(Z \geq a) \leq E(\phi(Z))/\phi(a)$ for any nonnegative monotone increasing $\phi$. We let $k = \lceil n^{1/3} \rceil$, $a = (1 + \epsilon)e/n$ where $\epsilon = n^{-1/4}$ and $\phi(a) = \max(0, a(a - 1) \ldots (a - k + 1))/k!$ and note that $\phi(Z)$ the number of $k$-sets $S$ for which $N_j = 0$, $j \in S$. This gives

$$
\Pr\left(Z \geq (1 + \epsilon)e/n\right) \leq \frac{n}{k} \frac{(1 - k/n)^n}{a(a - 1) \ldots (a - k + 1)/k!} = O((1 + \epsilon)^{-k}).
$$

This completes the proof of Theorem 8. ■

As for the Euclidean and graphical models, we can show

**Theorem 9.** Suppose $k = o((n/\log n)^{1/2})$ and $k \to \infty$. Then an LP based branch and bound algorithm almost surely explores at least $n^{(1 - o(1))k}$ nodes of the search tree.

**Proof.** Let $z_{LP}(J_0, J_1)$ be the LP value of the subproblem where $J_0 = \{ j : x_j$ is fixed to 0\} and $J_1 = \{ j : x_j$ is fixed to 1\}. Assume that $\alpha < 1$ is close to 1, that $\beta$ is large and that $\alpha$ and $\beta$ have been chosen so that $\alpha k$ and $\beta k$ are integer. To prove the theorem, it suffices to show that, almost surely,

(37) for any $J_0, J_1 \subseteq \{1, 2, \ldots, n\}$ such that $J_0 \cap J_1 \neq \emptyset$, $|J_1| \leq \alpha k$ and $|J_0| \leq n - (\beta - \alpha)k$, we have $z_{LP}(J_0, J_1) \leq z_{IP}$.

As increasing $J_0$ or $J_1$ only serves to increase $z_{LP}$, we can restrict our attention to $|J_1| = \alpha k$ and $|J_0| = n - (\beta - \alpha)k$. Let $L = \alpha k$, $K = \beta k$ and

$$
\gamma = \frac{(1 - \alpha)k}{n - |J_0| - |J_1|} = \frac{1 - \alpha}{\beta} = \frac{k - L}{K}.
$$

To obtain an upper bound on $z_{LP}(J_0, J_1)$, let

$$
\gamma = \begin{cases} 
0 & \text{if } j \in J_0, \\
1 & \text{if } j \in J_1, \\
\gamma & \text{if } j \notin J_0 \cup J_1.
\end{cases}
$$

Consider a fixed $i$ and suppose $c_{ij_1} \leq c_{ij_2} \leq \cdots \leq c_{ij_n}$. Let $t = \min\{ s : j_s \in J_1 \}$. Let $y_{ij}$ be given by Lemma 2 and $d_i = \sum_{j=1}^{n} d_{ij} y_{ij}$. The expected value of $d_i$, conditional on knowing the value of $t$, is

$$
\text{Exp}(d_i|t) = \gamma \sum_{i=1}^{t} \frac{i}{K + 1} + (1 - \gamma t) \frac{t}{K + 1}.
$$

and

$$
\text{Exp}(d_i|t) = \gamma \sum_{i=1}^{t-1} \frac{i}{K + 1} + \gamma (\gamma^{-1} - 1) \frac{|\gamma^{-1} - 1|}{2(K + 1)} \quad \text{if } t \leq |\gamma^{-1}|,
$$

(38)

and

$$
\text{Exp}(d_i|t) = \gamma \sum_{i=1}^{t-1} \frac{i}{K + 1} + \gamma (\gamma^{-1} - 1) \frac{|\gamma^{-1} - 1| + 1}{K + 1} \quad \text{if } t > |\gamma^{-1}|.
$$

(39)
Now

$$\Pr(t) = \binom{K-t}{L-1} \binom{K}{L}.$$  \hspace{1cm} (40)

Using (38)–(40), we get the expected value of $d_i$.

$$\text{Exp}(d_i) = \frac{1}{(K+1)\binom{K}{L}} \left[ \sum_{t=1}^{\lfloor \gamma^{-1} \rfloor} t \binom{K-t}{L-1} - \frac{\gamma}{2(K+1)\binom{K}{L}} \sum_{t=1}^{\lfloor \gamma^{-1} \rfloor} (t-1)t \binom{K-t}{L-1} \right]$$

$$+ \frac{\gamma(2\gamma^{-1} - \lfloor \gamma^{-1} \rfloor)(\lfloor \gamma^{-1} \rfloor + 1)}{2(K+1)\binom{K}{L}} \sum_{t>\lfloor \gamma^{-1} \rfloor} \binom{K-t}{L-1}.$$  

Now, as may be inductively verified,

$$\binom{L-1}{L-1} + \binom{L}{L-1} + \cdots + \binom{A}{L-1} = \binom{A+1}{L},$$

$$\binom{K-1}{L-1} + 2\binom{K-2}{L-1} + \cdots + A\binom{K-A}{L-1}$$

$$= \binom{K+1}{L+1} - A\binom{K-A}{L} - \binom{K+1-A}{L+1},$$

$$1.2\binom{K-2}{L-1} + 2.3\binom{K-3}{L-1} + \cdots + (A-1)A\binom{K-A}{L-1}$$

$$= 2\binom{K+1}{L+2} - A(A+1)\binom{K-A}{L}$$

$$- 2(A+1)\binom{K-A}{L+1} - 2\binom{K-A}{L+2}.$$  

Therefore

$$\text{Exp}(d_i) = \frac{1}{(K+1)\binom{K}{L}} \left[ \left( \binom{K+1}{L+1} - \lfloor \gamma^{-1} \rfloor \binom{K-\lfloor \gamma^{-1} \rfloor}{L} - \left( \binom{K+1-\lfloor \gamma^{-1} \rfloor}{L+1} \right) \right) \right]$$

$$- \frac{1}{2(K+1)\binom{K}{L}} \left[ 2\left( \binom{K+1}{L+2} - \lfloor \gamma^{-1} \rfloor(\lfloor \gamma^{-1} \rfloor + 1)\binom{K-\lfloor \gamma^{-1} \rfloor}{L} \right) \right]$$

$$- 2(\lfloor \gamma^{-1} \rfloor + 1)\binom{K-\lfloor \gamma^{-1} \rfloor}{L+1} - 2\binom{K-\lfloor \gamma^{-1} \rfloor}{L+2}$$

$$+ \frac{(2\gamma^{-1} - \lfloor \gamma^{-1} \rfloor)(\lfloor \gamma^{-1} \rfloor + 1)}{2(K+1)\binom{K}{L}} \binom{K-\lfloor \gamma^{-1} \rfloor}{L}.$$
Now
\[ \frac{(K + 1)(K)}{(L + 1)(K)} = \frac{1}{L + 1}, \quad \frac{\gamma(K + 1)(K + 2)}{(K + 1)(K)(L + 1)(L + 2)} = \frac{(k - L)(K - L)}{K(K + 1)(L + 1)(L + 2)} \]
and
\[ \left( K - \frac{\lfloor \gamma^{-1} \rfloor}{L} \right) \leq e^{-[(\gamma^{-1})/L]K}, \]

Hence, for \( \alpha \) and \( \beta \) fixed, where \( \alpha \) is close to 1 and \( \beta \) is large, and for \( k \to \infty \),
\begin{align*}
\text{Exp}(d_i) &= \frac{1}{ak} - \left( \frac{1 + O(\gamma)}{(1 - \alpha)k} + \frac{1}{ak} \right)\left( K - \frac{\lfloor \gamma^{-1} \rfloor}{L} \right) \left( 1 + O\left( \frac{1}{k} \right) \right) \\
&- \left[ \frac{(1 - \alpha)(\beta - \alpha)}{\beta \alpha^2 k} - \frac{1 + O(\gamma)}{2(1 - \alpha)k} + \frac{1 + O(\gamma)}{ak} + \frac{1 - \alpha}{\alpha^2 k} \right] \\
&\times \left( K - \frac{\lfloor \gamma^{-1} \rfloor}{L} \right) \left( 1 + O\left( \frac{1}{k} \right) \right) \\
&+ \frac{1 + O(\gamma)}{2(1 - \alpha)k} \left( K - \frac{\lfloor \gamma^{-1} \rfloor}{L} \right) \left( 1 + O\left( \frac{1}{k} \right) \right) \\
&= \frac{1}{k} \left[ \frac{1}{\alpha} - \frac{(1 - \alpha)(\beta - \alpha)}{\beta \alpha^2} + o\left( \frac{1}{1 - \alpha} e^{-\alpha/(1 - \alpha)} \right) \right] \left( 1 + O\left( \frac{1}{k} \right) \right) \\
&= \frac{1}{k} \left[ \frac{1}{\alpha} - \frac{1}{\alpha} - 1 \right] + \frac{1 - \alpha}{\alpha \beta} + o\left( \frac{1}{1 - \alpha} e^{-\alpha/(1 - \alpha)} \right) \left( 1 + O\left( \frac{1}{k} \right) \right).
\end{align*}

Let \( \beta = 2\alpha/(1 - \alpha) \). Then
\[ \text{Exp}(d_i) = \frac{1}{k} \left[ 1 - \frac{1}{2} \left( \frac{1}{\alpha} - 1 \right)^2 + o\left( \frac{1}{1 - \alpha} e^{-\alpha/(1 - \alpha)} \right) \right] \left( 1 + O\left( \frac{1}{k} \right) \right). \]

Now \( (1/\alpha - 1)^2/(1/(1 - \alpha)) e^{-\alpha/(1 - \alpha)} \to \infty \) as \( \alpha \to 1 \). Thus, by choosing \( \alpha \) close to 1, we get
\[ \text{Exp}(d_i) \leq \frac{1}{k} \left( 1 - \frac{1}{3} \left( \frac{1}{\alpha} - 1 \right)^2 \right). \]

Applying Lemma 1 with \( \epsilon = (1/\log n)^{1/2} \), we get
\[ z_{LP}(J_0, J_1) \leq \frac{n}{k} \left( 1 - \frac{1}{3} \left( \frac{1}{\alpha} - 1 \right)^2 \right) \left( 1 + o(1) \right) \]
with probability at least \( 1 - e^{-nB/3k \log n} \) where \( B = 1 - \frac{1}{3}(1/\alpha - 1)^2 \). We have to branch for all \( |J_1| \leq ak \) and \( |J_0| \leq n - (\beta - \alpha)k \) with probability at least
\[ 1 - \left( \frac{n}{\beta k} \right) \left( \frac{1}{ak} \right) e^{-nB/3k \log n} \to 1 \] since \( \left( \frac{n}{\beta k} \right) \left( \frac{1}{ak} \right) \leq n^{(\alpha + \beta)k} \) and \( n/k^2 \log n \to \infty \). This proves that, almost surely, (37) holds. As a consequence, the number of branches in the
search tree is at least \((n - \beta k) / \alpha k = n^{(1 - o(1))k}\).

6. Computational experience. The previous sections provide asymptotic results as \(n \to \infty\). In this section, we report our computational experience with medium-size \(k\)-median problems for the four probabilistic models introduced earlier. This computational experience is based on the solution of about 3300 random problems with \(n = 50\) points and an additional 950 random problems with \(n = 100\) points. The description of these problems is given later.

For each problem we computed \(z_{IP}\) and \(z_{LP}\). The value of \(z_{LP}\) was obtained by solving a Lagrangian dual by subgradient optimization as explained in [3]. In the process of computing \(z_{LP}\), this algorithm generates a feasible solution at each subgradient iteration. Of course, if it happens that the value of the best feasible solution generated equals \(z_{LP}\), the algorithm terminates since, then, \(z_{IP} = z_{LP}\). For most of the test problems with no gap \(z_{IP} - z_{LP}\), the algorithm terminated in less than 100 subgradient iterations, due to the above stopping criterion. If, after 100 subgradient iterations, there was still a gap between the best feasible solution (an upper bound on \(z_{IP}\)) and the best Lagrangian relaxation (a lower bound on \(z_{LP}\)), we resorted to branch and bound to find \(z_{IP}\). When the subgradient algorithm clearly converged to a value different from \(z_{IP}\), we accepted it as showing that \(z_{IP} \neq z_{LP}\). In the cases where the subgradient algorithm converged to a value close to \(z_{IP}\) we used the simplex algorithm to compute \(z_{LP}\). This allowed us to settle cases where there was a very small but positive gap \(z_{IP} - z_{LP}\).

Among the 4250 test problems that we generated we found about 3700 such that \(z_{IP} = z_{LP}\) and about 550 with a gap \(z_{IP} - z_{LP}\). Now we give a detailed description of these results.

The first set of experiments involves Euclidean problems. We decided to test whether approximating the Euclidean distances had an influence on the gap \(z_{IP} - z_{LP}\), since we suspected that data accuracy might be partly responsible for the discrepancy between the computational experience previously reported in the literature, namely few test problems were found to have gaps ([2], [3], [6], [10], [11], [19], [20], [23], [24]), and the results of §2 stating that asymptotically most instances should have small but positive gaps. To our surprise, data accuracy had little influence except maybe for the possibility that a very coarse approximation produces harder \(k\)-median problems. (These problems are more combinatorial, often have alternate optimal solutions and, in our experience, optimality was harder to prove.) We generated 10 problems, each with 50 points occurring at random in the unit square. Then, for \(i = 1, 2, 3, 4\) and 5, we multiplied each point coordinate by \(10^i\) and rounded it to the closest integer value. The Euclidean distances were then computed and rounded to the closest integer. The \(k\)-median problem and its LP relaxation were solved for each \(2 \leq k \leq 10\) and \(1 \leq i \leq 5\). For each such pair \(i, k\), Table 1 reports the number of problems (out of 10) with a gap \(z_{IP} - z_{LP}\).

The same two problems were responsible for all the gaps. The average value of \((z_{IP} - z_{LP}) / z_{IP}\) over the instances that had a gap was approximately 1.5% for \(i = 1\), 0.4% for \(i = 2\) and 0.1% for \(i = 3, 4\), and 5. Overall, the fraction of instances with a gap was about 5%. This is consistent with the computational experience reported in the literature. Clearly, the asymptotic behavior described in §2 is not felt for problems with \(n = 50\) points. It would be interesting to repeat the computational experiment for Euclidean \(k\)-median problems with about \(n = 1000\) points. Unfortunately our computer budget did not allow to do this.

The second set of experiments involves random trees. We generated 100 random trees, 50 of them with \(n = 50\) nodes and the other 50 with \(n = 100\) nodes, using the
method described in Even [7, p. 28]. First we assumed that all edge lengths were equal to 1 in the trees, and we solved the $k$-median problem and the LP relaxation for $2 \leq k \leq 11$ in each tree. For each pair $n, k$, Table 2 reports the number of problems (out of 50) with a gap $z_{IP} - z_{LP}$.

We also computed $z_{IP}$ and $z_{LP}$ for the same 100 trees assuming nonunit edge lengths. In this experiment, the nodes of the tree were random points in the unit square and the length of an edge in the tree was the Euclidean distance between its endpoints rounded using the scheme explained earlier with $i = 1$. The distance between two nodes of the tree was the length of the unique path joining them. Table 3 reports these results.

We did not find a significant difference in difficulty between the two tree models. Overall, the fraction of instances with a gap was about 4%.

Our third set of experiments involves random graphs. First we report the results when the edge lengths are equal to 1. Starting from a random tree on 50 nodes, we generated a sequence of graphs, adding 50 random edges at a time to the previous graph. Table 4 contains the value of $z_{LP}$ and $z_{IP}$ for each graph and $2 \leq k \leq 10$. Only one figure means that $z_{IP} = z_{LP}$. Note that when $z_{IP} = z_{LP} = n - k$ for some graph, it contains a dominating set and therefore every subsequent graph in the sequence also does.

### Table 1

**Euclidean model with $n = 50$. Number of instances with a gap.**

<table>
<thead>
<tr>
<th>$i$</th>
<th>$k$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total (out of 90)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Total (out of 50)</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>23 (out of 450)</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2

**Tree model with unit edge lengths. Number of instances with a gap.**

<table>
<thead>
<tr>
<th>$n$</th>
<th>$k$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total (out of 500)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>Total (out of 100)</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>35 (out of 1000)</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3

**Tree model with nonunit edge lengths. Number of instances with a gap.**

<table>
<thead>
<tr>
<th>$n$</th>
<th>$k$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total (out of 450)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>Total (out of 100)</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td>12</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>8</td>
<td>2</td>
<td>44 (out of 900)</td>
<td></td>
</tr>
</tbody>
</table>
### Table 4

<table>
<thead>
<tr>
<th>Number of edges</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>49</td>
<td>139</td>
<td>114</td>
<td>98</td>
<td>88</td>
<td>80</td>
<td>72</td>
<td>65</td>
<td>59</td>
<td>54</td>
</tr>
<tr>
<td>99</td>
<td>89</td>
<td>77</td>
<td>68.5/69</td>
<td>62</td>
<td>57</td>
<td>52</td>
<td>48</td>
<td>44.5/46</td>
<td>42</td>
</tr>
<tr>
<td>149</td>
<td>77</td>
<td>69</td>
<td>62</td>
<td>55.5/57</td>
<td>50</td>
<td>46</td>
<td>43</td>
<td>41</td>
<td>40</td>
</tr>
<tr>
<td>199</td>
<td>72</td>
<td>63</td>
<td>55</td>
<td>48</td>
<td>45</td>
<td>43</td>
<td>42</td>
<td>40</td>
<td>39</td>
</tr>
<tr>
<td>249</td>
<td>72</td>
<td>61</td>
<td>52</td>
<td>46</td>
<td>44/45</td>
<td>44</td>
<td>44</td>
<td>43</td>
<td>42</td>
</tr>
<tr>
<td>299</td>
<td>69</td>
<td>56</td>
<td>48</td>
<td>46</td>
<td>44</td>
<td>44</td>
<td>44</td>
<td>43</td>
<td>42</td>
</tr>
<tr>
<td>349</td>
<td>65</td>
<td>52.5/54</td>
<td>48</td>
<td>45/46</td>
<td>45</td>
<td>44/46</td>
<td>44</td>
<td>44</td>
<td>43</td>
</tr>
<tr>
<td>399</td>
<td>62</td>
<td>50</td>
<td>47/48</td>
<td>45</td>
<td>44/46</td>
<td>44</td>
<td>44</td>
<td>43</td>
<td>42</td>
</tr>
<tr>
<td>449</td>
<td>61</td>
<td>49</td>
<td>46/47</td>
<td>46</td>
<td>46/47</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>45</td>
</tr>
<tr>
<td>499</td>
<td>58</td>
<td>47.5/48</td>
<td>46/47</td>
<td>46</td>
<td>46/47</td>
<td>46</td>
<td>46</td>
<td>45</td>
<td>44</td>
</tr>
<tr>
<td>549</td>
<td>56</td>
<td>48</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>45</td>
</tr>
<tr>
<td>599</td>
<td>54</td>
<td>47/48</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>45</td>
<td>45</td>
<td>44</td>
</tr>
<tr>
<td>649</td>
<td>52</td>
<td>47</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>699</td>
<td>51</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>749</td>
<td>50</td>
<td>45</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>799</td>
<td>48.5/49</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
</tr>
<tr>
<td>849</td>
<td>48/49</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
</tr>
<tr>
<td>1199</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
</tr>
</tbody>
</table>

Among the instances where a dominating set did not exist, about 28% had a gap.

Next we turn to the graphical model with nonunit edge lengths. We started from 10 random trees on \( n = 50 \) nodes. We then added random edges, 50 at a time, until the graphs contained 849 edges. The edge lengths were computed using the same scheme as earlier. Namely, the nodes were assigned random integer coordinates in a square of size \( 10 \times 10 \) and the length of an edge was the Euclidean distance between its two endpoints, rounded to the closest integer. The distance between two nodes of the graph was taken to be the length of the shortest path joining them in the graph. Table 5 reports the number of instances with a gap (out of 10), as a function of the number of edges in the graph and \( k \).

For this model, the fraction of instances with a gap was about 12%. The average of \((z_{IP} - z_{LP})/z_{IP}\) taken over the instances with a gap was less than 1%. Note that the first line of Table 1 corresponds to the case of the graphical model where the number of edges is \((49 \times 50)/2 = 1250\) and, as such, could be added as a line of Table 5.

Finally, the fourth set of experiments deals with the uniform cost model. We generated 30 problems with random integer costs. In the first 10 problems the costs were in the range \([1, 10]\), in the next 10 in the range \([1, 100]\) and in the last 10 in the range \([1, 1000]\). For each problem the values of \(z_{IP}\) and \(z_{LP}\) were computed for \(2 \leq k \leq 10\). For each range and value of \(k\), Table 6 contains the number of instances with a gap (out of 10).

For this model, the fraction of instances with a gap was about 92% overall, 100% for \(k \leq 8\). This fits well with the results of §5. The value of the ratio \((z_{IP} - z_{LP})/z_{IP}\) was much larger than in the other models. It reached 18% for one of the problems with costs taken in the range \([1, 1000]\) and \(k = 3\). Note, however, that this is still far below the asymptotic value of 33% predicted by Theorem 8 when \(k = 3\).
### Table 5

**Graphical model with nonunit edge lengths. Number of instances with a gap.**

<table>
<thead>
<tr>
<th>number of edges</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total (out of 90)</th>
</tr>
</thead>
<tbody>
<tr>
<td>49</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>99</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>149</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>199</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>249</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>299</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>349</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>399</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>449</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>499</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>549</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>599</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>649</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>699</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>749</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>799</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>849</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Total (out of 170)</td>
<td>18</td>
<td>22</td>
<td>28</td>
<td>17</td>
<td>22</td>
<td>17</td>
<td>19</td>
<td>21</td>
<td>181</td>
<td></td>
</tr>
</tbody>
</table>

### Table 6

**Uniform cost model with n = 50. Number of instances with a gap.**

<table>
<thead>
<tr>
<th>range</th>
<th>k</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total (out of 90)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>6</td>
<td>6</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>Total (out of 30)</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>24</td>
<td>15</td>
<td>249</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 7. The simple plant location problem.

Although we proved our probabilistic results for the k-median problem, they can also be useful for the SPLP. To define an instance of SPLP, we need fixed costs \( f_j \), \( j = 1, \ldots, n \), in addition to the distances \( d_{ij} \), \( 1 \leq i, j \leq n \). For simplicity, we assume in this section that the fixed costs \( f_j \) are all identical, say \( f_j = f \).

**THEOREM 10.** Consider the Euclidean model in the plane and assume that \( n^{1-1/2} \leq f \leq n^{1-\epsilon} \) for some fixed \( \epsilon > 0 \). Then, for the SPLP,

\[
\frac{z_{IP} - z_{LP}}{z_{IP}} \sim 0.00189255 \ldots \quad \text{almost surely.}
\]

**Proof.** In this proof, \( z_{IP} \) and \( z_{LP} \) denote the optimum values of SPLP and its linear programming relaxation respectively. The solutions of the corresponding k-median problem (with same \( d_{ij} \)'s) and its relaxation are denoted by \( z_{IP}(k) \) and \( z_{LP}(k) \) respectively.
By definition $z_{LP} = \min_k (z_{LP}(k) + kf) = \min(z_1, z_2, z_3)$, where

$$z_1 = \min_{k < \omega} \left( z_{LP}(k) + kf \right),$$

$$z_2 = \min_{\omega \leq k \leq n / \log n} \left( z_{LP}(k) + kf \right), \text{ and}$$

$$z_3 = \min_{k > n / \log n} \left( z_{LP}(k) + kf \right).$$

First we compute $z_2$. From the proof of Theorem 2,

$$\Pr \left( z_{LP} \notin \left[ \frac{2n}{3\sqrt{k/\pi}} (1 - o(1)), \frac{2n}{3\sqrt{k/\pi}} (1 + o(1)) \right] \right) = O \left( ne^{-(2\sqrt{n \log n}/9)} \right)$$

and so

$$z_2 \leq \min_{\omega \leq k \leq n / \log n} \left\{ \frac{2n}{3\sqrt{k/\pi}} (1 + o(1)) + kf \right\} \text{ almost surely.}$$

Let $\alpha = 2/3 \sqrt{\pi}$. The minimum of the function $an / \sqrt{k} + kf$ is attained when $k = (an/2f)^{2/3}$. Note that, given our assumptions on $f$, this value is in the range $(\omega, n / n \log n)$ for a suitable $\omega$, say $\omega = \log n$. The minimum value of the function is $((27/4)\alpha^2 n^2 f)^{1/3}$. Therefore

$$z_2 = \left( \frac{27}{4} \alpha^2 n^2 f \right)^{1/3} (1 + o(1)) \text{ almost surely.}$$

Now consider $z_3$. With our choice of $\omega = n \log n$, we have $k > n / (\log n)^2$. Therefore, almost surely,

$$z_3 \geq \frac{n}{(\log n)^2} f$$

$$= \frac{n^{1/3} f^{2/3}}{(\log n)^2} \frac{z_2}{((27/4)\alpha^2 f^{1/3} (1 + o(1)) \geq z_2.}$$

Finally consider $z_1$. For all $k < \log n$, we have $z_{LP}(k) \geq z_{LP}(\log n)$. Therefore $z_1 \geq z_{LP}(\log n)$. This implies that, almost surely,

$$z_1 \geq \frac{2n}{3\sqrt{\pi \log n}} (1 + o(1)) = c \frac{n^{1/3} f^{-1/3}}{(\log n)^{1/2}} z_2 (1 + o(1)) \geq z_2,$$

where $c$ is a constant.

We have just proved that $z_{LP} \sim ((27/4)\alpha^2 n^2 f)^{1/3}$ almost surely.

Similarly, $z_{IP} = \min_k (z_{IP}(k) + kf)$. Following the proof of Papadimitriou [22], we can show that

$$z_{IP} = \min_k \left( \frac{\beta n}{\sqrt{k}} (1 + o(1)) + fk \right) \text{ almost surely,}$$

(41)

where $\beta = 0.3771967 \ldots$. The minimum in (41) is achieved when $k = (\beta n/2f)^{2/3}$ and
its value is \((27/4)\beta^2 n^2 f)^{1/3}(1 + o(1))\). So
\[
\frac{z_{\text{IP}} - z_{\text{LP}}}{z_{\text{IP}}} \sim \frac{\beta^{2/3} - \alpha^{2/3}}{\beta^{2/3}} \quad \text{almost surely.}
\]

Similarly, the next result can be shown using the proof of Theorem 8.

**Theorem 11.** Consider the uniform cost model and assume that \(n^{\epsilon - 1} \leq f \leq n^{1 - \epsilon}\) for some fixed \(\epsilon > 0\). Then
\[
\frac{z_{\text{IP}} - z_{\text{LP}}}{z_{\text{IP}}} \sim 1 - \frac{\sqrt{2}}{2} \quad \text{almost surely.}
\]

**8. Conclusion.** The LP relaxation (1)–(4) has been widely used in branch and bound algorithms for the \(k\)-median problem and has been reported to provide a tight bound in practice. Our analysis shows that such good results can indeed be expected in a probabilistic sense for some problem instances, but we also identify other instances where the LP relaxation is almost surely not tight. The probabilistic analysis is performed under four classical models in location theory, namely the Euclidean, network, tree, and uniform cost models. For example, let \(\omega = \omega(n) \to \infty\). When \(\omega \leq k \leq n/\omega \log n\) in the Euclidean model, \(z_{\text{LP}}/z_{\text{IP}} = 0.99716 \ldots + o(1)\) almost surely, and when \(\omega \leq k \leq n/\omega \log n\) in the uniform cost model, \(z_{\text{LP}}/z_{\text{IP}} = 0.5 + o(1)\) almost surely.

Our computational experience confirms that large gaps occur frequently in the uniform cost model whereas only small gaps were observed with the other models.

Another aspect of the probabilistic analysis performed in §§2, 3 and 5 is that, under various assumptions, branch and bound algorithms must almost surely expand a nonpolynomial number of nodes to solve \(k\)-median problems to optimality.

Finally, we mention as open problems the questions of describing the asymptotic behavior of \(z_{\text{LP}}/z_{\text{IP}}\) as \(n \to \infty\) when (i) \(k \geq n/\log n\) in the Euclidean model, (ii) each edge of the graph has a random length \(d_{ij}\) (drawn uniformly in the interval \([0, 1]\), say) in the network and tree models, (iii) \(n/\log n \leq k \leq n(e - 1)/e\) in the uniform cost model.

**References**


AHH: SEOUL NATIONAL UNIVERSITY, SEOUL, REPUBLIC OF KOREA
COOPER: POLYTECHNIC OF NORTH LONDON, LONDON, UNITED KINGDOM
CORNUEJOLS: GRADUATE SCHOOL OF INDUSTRIAL ADMINISTRATION, CARNEGIE-MELLON UNIVERSITY, PITTSBURGH, PENNSYLVANIA 15213
FRIEZE: DEPARTMENT OF MATHEMATICS, CARNEGIE-MELLON UNIVERSITY, PITTSBURGH, PENNSYLVANIA 15213