The $k$-th Total Path Length and the Total Steiner $k$-Distance for Digital Search Trees

Chung-Kuei Lee
Department of Applied Mathematics
National Chiao Tung University
Hsinchu, 300
Taiwan

Abstract

The total Steiner $k$-distance and the $k$-th total path length are the sum of the size of Steiner trees and ancestor-trees over sets of $k$ nodes of a given tree, respectively. They are useful statistics with many applications. Consequently, they have been analyzed for many different random trees, including increasing tree, binary search tree, generalized $m$-ary search tree and simply generated trees.

In this paper, we investigate the two parameters for digital search trees, which are fundamental data structures in computer science with wide applications. We derive the means, covariances and variances for the total Steiner $k$-distances and the $k$-th total path lengths by the "Poisson-Laplace-Mellin Method". Moreover, results about the limiting distributions are obtained as well.

1 Introduction

Digital search trees (DSTs) are fundamental data structure in computer science first introduced by Coffman and Eve [3]. They have attracted considerable attention due to their wide applications, especially their close connection to the famous Lempel-Ziv compression scheme [10]. In a DST, strings are stored in the nodes of a binary tree under the following rules. Given a series of binary strings, we place the first string in the root and the other strings are distributed to the right or the left subtree of the root depending on whether their first symbol is "0" (left) or "1" (right). This process will be recursively applied to the subtrees but with the removal of the first bits when comparison is completed. The resulting tree will be called a DST. See the below as an illustration.

Many parameters of DSTs, including depth, height, total path length, peripheral path length, Wiener index, fillup level and expected profile, have been analyzed, see [5], [7] and [8] for details and many references. In this paper, we are interested in two parameters, the $k$-th total path length and the total Steiner $k$-distance, which have not been analyzed yet. We equip DSTs with the so-called Bernoulli model which assumes that every bit of the string is independent and admits a Bernoulli distribution with the probability $p$ of being 0. In order to make the analysis more brief and simpler, we consider in this paper only the symmetric case. In other words, the probability $p$ for the Bernoulli model will be $1/2$ for the entire paper.

For a given tree $T$ with vertex set $V$ and a subset $M \subset V$, the smallest spanning tree containing $M$ is called the Steiner tree for $M$ in $T$ while the smallest subtree containing $M$ and the root is the so-called ancestor-tree for $M$ in $T$. The size of the Steiner tree for $M$ in $T$ (denoted by $S_M(T)$) and the size of the ancestor-tree for $M$ in $T$ (denoted by $D_M(T)$) are called the Steiner distance and the
Figure 1: A DST built from 7 keys.

$k$-th path length, respectively. Furthermore, for the given tree $T$ and integer $k \in \mathbb{N}$, the $k$-th total path length $P_k(T)$ and the Steiner $k$-distance $W_k(T)$ are defined as

$$P_k(T) = \sum_{|M|=k} D_M(T) \quad \text{and} \quad W_k(T) = \sum_{|M|=k} S_M(T).$$

Steiner trees and ancestor trees have numerous applications, e.g. in transportation and multiprocessor networks [16], circuit layouts, internet communication [17] and many others. Consequently, the Steiner distance and the $k$-th total path length are useful statistics. For example, when comparing the efficiency of communication potential of different networks, the Steiner distance can be used [4]. Moreover, the Steiner distance and $k$-th total path length have also applications to Multiple Quickselect algorithm [16] and the efficiency of certain traceroute algorithms [9].

In the last decade, several papers dedicated to the analysis of the two parameters in various random trees, including random increasing tree [16], random binary search tree [11], generalized random $m$-ary search tree [15], recursive trees [12, 14] and random simply generated trees [12, 14] have been published. In this paper, we use the "Laplace-Mellin Method" which was first proposed in [7] to obtain the means, variances and covariances of the $k$-th total path length and the total Steiner $k$-distance for DSTs under the Poisson model. With the help of the JS-admissibility language from [7] and then analytical depoissonization, which was first proposed by P. Jacquet and W. Szpankowski, we can get the means, variances and covariances under the above Bernoulli model directly from the results under Poisson model. Limit laws for the two parameter are derived as well. In the rest of the paper, we use $P_n^{[k]}$ and $S_n^{[k]}$ to denote the $k$-th total path length and total $k$-th Steiner distance of random digital search trees built on $n$ strings, respectively. Also, we use the common notation for the constant $Q_m = \prod_{j=1}^{m} \left(1 - 2^{-j}\right)$ and $Q_\infty = \lim_{m \to \infty} Q_m$. The main results of this paper are:

**Theorem 1.** We have that for $k \geq 2$,

$$\mathbb{E} \left( P_n^{[k]} \right) \sim \mathbb{E} \left( S_n^{[k]} \right) \sim n^k \log_2 n \frac{1}{(k-1)!}.$$

Moreover, the variance and covariance of $P_n^{[k]}$ and $S_n^{[k]}$ are given by

$$\text{Var} \left( P_n^{[k]} \right) \sim \text{Var} \left( S_n^{[k]} \right) \sim n^{2k-1} - \frac{2^{2-2k}}{Q_{k-1}^2} \left( C_{kps} + \varpi_{kps}(\log_2 n) \right),$$

$$\text{Cov} \left( P_n^{[k_1]}, P_n^{[k_2]} \right) \sim \text{Cov} \left( S_n^{[k_1]}, S_n^{[k_2]} \right) \sim n^{k_1+k_2-1} - \frac{2^{2-k_1-k_2}}{Q_{k_1-1}Q_{k_2-1}} \left( C_{kps} + \varpi_{kps}(\log_2 n) \right),$$

where $C_{kps}$ and $\varpi_{kps}$ are defined in Theorem 2.1 of [7].
Theorem 2. Let
\[ X_n^{[k]} = \frac{P_n^{[k]} - \mathbb{E}(P_n^{[k]})}{\sqrt{\text{Var}(P_n^{[k]})}} \quad \text{and} \quad Y_n^{[k]} = \frac{S_n^{[k]} - \mathbb{E}(S_n^{[k]})}{\sqrt{\text{Var}(S_n^{[k]})}}. \]
We have that for any \( k \geq 2 \),
\[ (X_n^{[1]}, \ldots, X_n^{[k-1]}, Y_n^{[k]}) \xrightarrow{d} (X, \ldots, X), \]
where \( X \) is the standard normal distributed random variable and \( \xrightarrow{d} \) denotes weak convergence.

Remark 1. The asymptotics of \( S_n^{[k]} \) can be explained intuitively. It is well-known that the expected value of the depth of a node is of the order \( \log n \). For a Steiner tree, the size will be more or less the sum of the depth of the \( k \) chosen nodes. Thus, for \( k \) chosen nodes, the expected size of the Steine tree will be of order \( k \log_2 n \). Since there are \( \binom{n}{k} \) ways to choose the \( k \) nodes, the mean of the total Steiner \( k \)-distance will be roughly \( \binom{n}{k} k \log_2 n \sim \frac{n^k \log_2 n}{(k-1)!} \).

Remark 2. In fact, we can find more terms in the asymptotic of the means, variances and covariances for \( P_n^{[k]} \) and \( S_n^{[k]} \) by the same method applied in the following sections. For example, let \( \chi_m = 2m\pi i/\log 2 \), we have that
\[ \mathbb{E} \left( P_n^{[k]} \right) \sim \mathbb{E} \left( S_n^{[k]} \right) \sim \frac{n^k \log n}{(k-1)!} + \frac{n^k}{(k-1)!} \left( c_k + e_k + \frac{1}{\log 2} \sum_{m \in \mathbb{Z} \setminus \{0\}} G_k(\chi_m)n^{\chi_m} \right). \]
where \( G_k(\chi_m) \) is defined as
\[ G_k(\chi_m) = \Gamma(k + 1 - \chi_m)\Gamma(-1 - \chi_m) \]
and the constant \( c_k \) is given by
\[ c_k = \frac{H_k - 1}{\log 2} + \frac{1}{2} - \sum_{j \geq 1} \frac{1}{2^j - 1} + \frac{(k-1)!d_k}{\log 2}. \]
In the expression, \( H_k \) is the \( k \)-th harmonic number and \( d_k \) is defined recursively as \( d_1 = 0 \) and
\[ d_k = \frac{1}{2^k-1} - \sum_{r=1}^{k-1} \frac{d_{k-r}}{r!} + \frac{2^{k-1}}{2^{k-1} - 1(k-1)!}. \]
Also, the sequence \( \{e_k\}_{k \geq 1} \) is defined recursively as \( e_1 = 0 \) and
\[ e_k = \frac{1}{2^k-1} - \sum_{r=1}^{k-2} \frac{k!}{r!} e_{k-r} + \frac{2^k-1}{2^{k-1} - 1} \quad \text{for} \quad k \geq 2. \]

We state the main result in the form of Theorem 1 because the leading term is the most interesting part and it would be enough for proving the central limit theorem. Also, computing more terms can be extremely complicated. As we have seen from the above statements, the first several terms of the results for the \( k \)-th total path length and the total Steiner \( k \)-distance are the same. This is not surprising, intuitively speaking, because in most cases, the \( k \)-subset of vertices will contain vertices from both subtrees and hence the ancestor tree and the Steiner tree will be the same. This is similar to the distance between two random nodes (see [1, 2]) which is also twice the depth, because in most cases the root is included in the path from one node to the other.
Remark 3. Note that the Steiner $k$-distance is a generalization of the Wiener index, namely, for $k = 2$ we obtain the Wiener index. Thus, Theorem 2 of [8] is actually a special case of Theorem 2 of this paper with $k = 2$.

The paper is organized as follows. In Section 2, we derive the means, covariances and variance of the $k$-th total path length for DSTs. In Section 3, we give the asymptotics of the means, covariances and variances of the total Steiner $k$-distance. We also explained how to prove Theorem 2 in Section 3.

**Notation.** For a given function $f$, we use $\mathcal{L}[f; s]$ and $\mathcal{M}[f; \omega]$ to denote the Laplace transform and Mellin transform of $f$, respectively. Moreover, $\epsilon$ is an arbitrarily small positive number.

## 2 $k$-th Total Path Length

In this section, we start with the recurrence under the Bernoulli model and then use it to get the differential-functional equation of the Poisson model. The rest of the analysis will focus on the Poisson model, since the depoissonization is standard, see [7]. The method we use in the analysis of the Poisson model is the "Laplace-Mellin Method" which uses a combination of Laplace and Mellin transform; see [8] for a summary of the method and [6] for a comprehensive introduction to the Mellin transform and its properties.

### 2.1 Mean of the $k$-th Total Path Length of DSTs

First, we start with deriving a distributional recurrence relation for the $k$-th total path length. Recall the notation $P^{[k]}_B$ for the $k$-th total path length from the introduction. Moreover, we will use the notation $B_n \overset{d}{=} \text{Binom}(n, \frac{1}{2})$. Let a DST with $n + 1$ nodes given. Depending on how the $k$ nodes are chosen, there are 4 cases:

1. **All $k$ nodes are from one subtree.**
   The contribution to the $k$-th total path length will be
   
   $$ P^{[k]}_{B_n} + P^{[k]}_{n-B_n} + \binom{B_n}{k} + \binom{n-B_n}{k}, $$

   where $P^{[k]}_{B_n}$ is independent of $P^{[k]}_{n-B_n}$ and $P^{[k]}_{B_n} \overset{d}{=} P^{[k]}_{n-B_n}$.

2. **The $k$ nodes are chosen from both subtrees and the root is not chosen.**
   We will have the contribution
   
   $$ \sum_{r=1}^{k-1} \left( \binom{n-B_n}{k-r} P^{[r]}_{B_n} + \binom{B_n}{r} P^{[k-r]}_{n-B_n} + 2 \binom{B_n}{r} \binom{n-B_n}{k-r} \right). $$

3. **The root is chosen, the other $k-1$ nodes are all from one subtree.**
   It will contribute
   
   $$ P^{[k-1]}_{B_n} + P^{[k-1]}_{n-B_n} + \binom{n-B_n}{k-1} + \binom{B_n}{k-1}. $$

4. **The root is chosen, the other $k-1$ nodes are from both subtrees.**
   The contribution will be
   
   $$ \sum_{r=1}^{k-2} \left( \binom{n-B_n}{k-r-1} P^{[r]}_{B_n} + \binom{B_n}{r} P^{[k-r-1]}_{n-B_n} + 2 \binom{B_n}{r} \binom{n-B_n}{k-r-1} \right). $$


Combining all four cases, we get that for \( n + 1 \geq k \geq 1 \):

\[
P^{[k]}_{n+1} = P^{[k]}_{B_n} + P^{[k]}_{B_{n-B_n}} + P^{[k-1]}_{B_n} + \sum_{r=1}^{k-1} \left( \binom{n-B_n}{k-r} P^{[r]}_{B_n} + \binom{B_n}{r} P^{[k-r-1]}_{n-B_n} \right)
\]

\[+ 2\binom{n}{k} + 2 \binom{n}{k-1} + \sum_{r=1}^{k-2} \left( \binom{n-B_n}{k-r} P^{[r]}_{B_n} + \binom{B_n}{r} P^{[k-r-1]}_{n-B_n} \right)\]

\[- \binom{n-B_n}{k} - \binom{B_n}{k} - \binom{n-B_n}{k-1} - \binom{B_n}{k-1}.\]

Note that from the above equation, we see that the \( k \)-th total path length depends on the 1-st, 2-nd, \ldots, \((k-1)\)-th total path length. Thus, we actually have a system of recurrences. The initial conditions are \( P^{[0]}_{n} = 0 \) for all \( n \) and \( P^{[k]}_{n} = 0 \) for all \( k > n \).

Let \( \tilde{f}^{[k]}(z) = e^{-z} \sum_{n \geq 0} \mathbb{E}(P^{[k]}_{n}) \frac{z^n}{n!} \) which is the mean in the Poisson model. Then, from the recurrence relation above, we get

\[
\tilde{f}^{[k]}(z) + \tilde{f}^{[k]}(\frac{z}{2}) = 2\tilde{f}^{[k]}(\frac{z}{2}) + 2 \sum_{r=1}^{k-1} \frac{(-1)^r}{r!} \tilde{f}^{[k-r]}(\frac{z}{2}) + 2 \sum_{r=1}^{k-2} \frac{(-1)^r}{r!} \tilde{f}^{[k-r-1]}(\frac{z}{2})
\]

\[+ 2 \left( \frac{z^k - (\frac{z}{2})^k}{k!} + \frac{z^{k-1} - (\frac{z}{2})^{k-1}}{(k-1)!} \right).\]

Note that when \( k = 1 \), the above equation will be exactly the same as the one derived in [7] and hence the order of \( \tilde{f}^{[1]}(z) \) is known. Thus, by induction and the closure properties of JS-admissibility from [7], we get that

\[
\tilde{f}^{[k]}(z) = \begin{cases} \mathcal{O}(z^{k+\epsilon}), & \text{as } z \to \infty; \\ \mathcal{O}(z^k), & \text{as } z \to 0^+ 
\end{cases}
\]

uniformly for \( z \) with \( |\arg z| \leq \frac{\pi}{2} - \epsilon \), where \( \epsilon > 0 \) is an arbitrary small constant. Applying Laplace transform, we get the differential-functional equation

\[(1+s)\mathcal{L}[\tilde{f}^{[k]}; s] = 4\mathcal{L}[\tilde{f}^{[k]}; 2s] + 4\mathcal{L}[\tilde{f}^{[k-1]}; 2s] + 4 \sum_{l=1}^{k-1} \frac{(-1)^l}{l!} \mathcal{L}^{(l)}[\tilde{f}^{[k-l]}; 2s] + 4 \sum_{l=1}^{k-2} \frac{(-1)^l}{l!} \mathcal{L}^{(l)}[\tilde{f}^{[k-l-1]}; 2s] + 2 \left( \frac{1+s}{s^{k+1}} - \frac{1 + 2s}{2^{k+1} s^{k+1}} \right),\]

where \( \mathcal{L}^{(l)}[\tilde{f}^{[k-l]}; s] \) is the \( l \)-th differentiation of \( \mathcal{L}[\tilde{f}^{[k-l]}; s] \). Let

\[
Q(-s) = \prod_{j \geq 1} \left( 1 - \frac{s}{2j} \right) \quad \text{and} \quad \mathcal{H}^{(l)}[\tilde{f}^{[k]}; s] = \frac{\mathcal{L}[\tilde{f}^{[k]}; s]}{Q(-s)}
\]

and divide both sides of above equation by \( Q(-2s) \). This yields

\[
\mathcal{H}^{(l)}[\tilde{f}^{[k]}; s] = 4\mathcal{H}^{(l)}[\tilde{f}^{[k]}; 2s] + 4\mathcal{H}^{(l)}[\tilde{f}^{[k-1]}; 2s] + 4 \sum_{l=1}^{k-1} \frac{(-1)^l}{l!} \mathcal{H}^{(l)}[\tilde{f}^{[k-l]}; 2s] + 4 \sum_{l=1}^{k-2} \frac{(-1)^l}{l!} \mathcal{H}^{(l)}[\tilde{f}^{[k-l-1]}; 2s] + 2 \left( \frac{1}{s^{k+1} Q(-s)} - \frac{1 + 2s}{2^{k+1} s^{k+1} Q(-2s)} \right)
\]

\[- \frac{1}{2^{k+1} Q(-2s)} - 4 \sum_{l=1}^{k-1} \sum_{r=0}^{l-1} \frac{(-1)^l}{l! (l-r)!} 2^{r-l} \mathcal{L}^{(r)}[\tilde{f}^{[k-l-1]}; 2s] h^{(l-r)}(s),
\]

5
where \( h(s) = \frac{1}{Q(2s)} \) and \( h^{(n)}(s) \) is the \( n \)-th derivative of \( h(s) \). From the bound for \( Q(-2s) \) obtained in [7] 
\[
Q(-2s) = \begin{cases} 
\mathcal{O}(|s|^b), & \text{as } s \to \infty; \\
1 + s + \mathcal{O}(|s|^2), & \text{as } s \to 0,
\end{cases}
\]
we obtain the bounds 
\[
\mathcal{L}[\tilde{f}^{[k]}; s] = \begin{cases} 
\mathcal{O}(|s|^{-b}), & \text{as } s \to \infty; \\
\mathcal{O}(|s|^{-(k+\epsilon)}), & \text{as } s \to 0^+ 
\end{cases}
\]
and 
\[
h^{(n)}(s) = \begin{cases} 
\mathcal{O}(|s|^b), & \text{as } s \to \infty; \\
\mathcal{O}(1), & \text{as } s \to 0^+
\end{cases}
\]
uniformly for \( s \) with \( |\arg(s)| \leq \pi - \epsilon \), where \( b \) is an arbitrary large constant. We let 
\[
R^{[k]}(s) = -4 \sum_{l=1}^{k-1} \frac{(-1)^l}{l!} 2^{r-l} \mathcal{L}^{(r)}[\tilde{f}^{[k-l]}; 2s] h^{(l-r)}(s)
\]
Then, by Ritt’s Theorem (Theorem 4.2 of [13]), we derive the bounds 
\[
R^{[k]}(s) = \begin{cases} 
\mathcal{O}(|s|^{-b}), & \text{as } s \to \infty; \\
\mathcal{O}(|s|^{-(k+\epsilon)}), & \text{as } s \to 0^+
\end{cases}
\]
uniformly for \( s \) with \( |\arg z| \leq \pi - \epsilon \). Thus, we may apply Mellin transform:
\[
\mathcal{M}[\tilde{z}^{[k]}; \omega] = \frac{2^{2-\omega}}{1 - 2^{2-\omega}} \mathcal{M}[\tilde{z}^{[k-1]}; \omega] + \frac{2^{2-\omega}}{1 - 2^{2-\omega}} \sum_{l=1}^{k-1} \frac{\prod_{r=1}^{l} (\omega - i)}{l!} \mathcal{M}[\tilde{z}^{[k-l]}; \omega - l] 
\]
\[
+ \frac{2^{2-\omega}}{1 - 2^{2-\omega}} \sum_{l=1}^{k-1} \frac{\prod_{r=1}^{l} (\omega - i)}{l!} \mathcal{M}[	ilde{z}^{[k-l-1]}; \omega - l] 
\]
\[
+ \frac{2}{1 - 2^{2-\omega}} \frac{Q(2^{\omega-k-1})}{Q(1)} \Gamma(k - \omega) \Gamma(\omega - k + 1)(1 - 2^{-k}) 
\]
\[
+ \frac{2}{1 - 2^{2-\omega}} \frac{Q(2^{\omega-k-1})}{Q(1)} \Gamma(k + 1 - \omega) \Gamma(\omega - k)(1 - 2^{1-k}) + \mathcal{M}[R^{[k]}; \omega] 
\]
where for convenience, we use the notation \( \mathcal{M}[\tilde{z}^{[k]}; s; \omega] \) for \( \mathcal{M}[\tilde{z}^{[k]}; \omega] \). The fundamental strip of the above expression will be the half plane \( \Re(\omega) > k + 1 \). To apply inverse Mellin transform, we need to figure out all the singularities of the above expression. Since the case \( k = 1 \) is already solved in [7] and the general case \( k \) will be determined by \( 1, \ldots, k - 1 \), we get that for \( k \geq 2 \) the expression can be simplified as
\[
\mathcal{M}[\tilde{z}^{[k]}; \omega] = \frac{2^{2-\omega}}{1 - 2^{2-\omega}} \left( \sum_{r=1}^{k-1} \frac{\prod_{r=1}^{l} (\omega - i)}{r!} \mathcal{M}[\tilde{z}^{[k-r]}; \omega - r] 
\]
\[
+ \frac{Q(2^{\omega-k-1})}{Q(1)} \Gamma(k - \omega) \Gamma(\omega - k + 1)(1 - 2^{-k}) \right) + \tilde{g}_k(\omega)
\]
where \( \tilde{g}_k(\omega) \) is the sum of all the remaining terms in the expression. From the bound we derived for \( R^{[k]}(s) \) and \( \mathcal{L}[\tilde{f}^{[k]}; s] \) and the properties of the Mellin transform [6], we get that if \( \alpha \) is a singularity of \( \tilde{g}_k(\omega) \), then \( \Re(\alpha) \leq k \). From [7], we have that
\[
\mathcal{M}[\tilde{z}^{[1]}; \omega] = \frac{G_1(\omega)}{1 - 2^{2-\omega}}.
\]
where
\[ G_1(\omega) = \frac{Q(2^{\omega-2})}{Q(1)} \Gamma(\omega) \Gamma(1 - \omega). \]

Plugging this into the recurrence and iterating, we get that for \( k \geq 2 \)
\[ \mathcal{M}[\mathcal{L}^{[k]}; \omega] = \prod_{i=1}^{k-1} (\omega - i) G_1(\omega - k + 1) A_k(\omega) + T_k(\omega) G_1(\omega - k + 1) + g_k(\omega) \]
where \( g_k(\omega) \) is defined recursively by \( g_1(\omega) = 0, g_2(\omega) = \bar{g}_2(\omega) \) and
\[ g_k(\omega) = \frac{2^{2-\omega}}{1 - 2^{2-\omega}} \sum_{r=1}^{k-1} \prod_{i=1}^{r} (\omega - i) A_{k-r}(\omega - r - r) \bar{g}_k(\omega). \]

Again, by similar argument as above, we have that if \( \alpha \) is a singularity of \( g_k(\omega) \), then \( \Re(\alpha) \leq k \). The function \( A_k(\omega) \) is defined recursively as \( A_1(\omega) = 1, A_2(\omega) = \frac{1}{2^{\omega-2} - 1} \) and
\[ A_k(\omega) = \frac{2^{2-\omega}}{1 - 2^{2-\omega}} \sum_{r=1}^{k-1} A_{k-r}(\omega - r) \frac{r!}{r}. \]

Also, \( T_k(\omega) \) is defined recursively as \( T_1(\omega) = 0, T_2(\omega) = \frac{6}{4(1 - 2^{2-\omega})} \) and
\[ T_k(\omega) = \frac{2^{2-\omega}}{1 - 2^{2-\omega}} \sum_{r=1}^{k-1} \prod_{i=1}^{r} (\omega - i) T_{k-r}(\omega - r) + \frac{2(1 - 2^{-k})}{1 - 2^{2-\omega}}. \]

Note that one can easily prove that
\[ A_k(k + 1 + \chi_m) = A_k(k + 1) = \frac{1}{(k - 1)!} \]
for \( \chi_m = \frac{2i\pi m}{\log 2}, m \in \mathbb{Z} \) by induction. Moreover, the Laurent series of \( A_k(\omega) \) at \( \omega = k + 1 + \chi_r \) is given as
\[ A_k(\omega) = \frac{1}{(k - 1)!} + d_k(\omega - k - 1) + \mathcal{O}(\omega - k - 1)^2, \]
where \( \{d_k\}_{k \geq 1} \) is a sequence which is defined recursively as \( d_1 = 0 \) and
\[ d_k = \frac{1}{2^{k-1} - 1} \sum_{r=1}^{k-1} \frac{d_{k-r}}{r!} - \frac{2^{k-1}}{2^{k-1} - 1} \frac{\log 2}{(k - 1)!}. \]

Because we have the explicit form of \( G_1(\omega) \), we rewrite the expression as
\[ \mathcal{M}[\mathcal{L}^{[k]}; \omega] = \frac{Q(2^{\omega-k-1})}{Q(1)} \Gamma(\omega) \Gamma(k - \omega) A_k(\omega) + g_k(\omega). \]

Finally, applying inverse Mellin transform and collecting residues, we get that
\[ \mathcal{F}[\mathcal{L}^{[k]}; s] = k s^{-(k+1)} \log_2 \frac{1}{s} + s^{-(k+1)} \left( c' + e_k + \frac{1}{\log 2} \sum_{m \in \mathbb{Z} \setminus \{0\}} G_k(\chi_m) s^{-\chi_m} \right) + \mathcal{O}(|s|^{-k-\epsilon}) \]
where \( e_k = T_k(k + 1) \),
\[ G_k(\chi_m) = \frac{\Gamma(k + 1 - \chi_r) \Gamma(-1 - \chi_r)}{(k - 1)!} \]
and
\[ c'_k = k \left( \frac{H_k - 1}{\log 2} + \frac{1}{2} - \sum_{j \geq 1} \frac{(k - 1)!}{2^j - 1} + \frac{(k - 1)!d_k}{\log 2} \right). \]

Note that the asymptotic hold uniformly as \(|s| \to 0\) with \(|\arg(s)| \leq \pi - \epsilon\). Finally, we apply Proposition 1 of [7] and obtain that, as \( z \to \infty \),
\[ \tilde{f}[k](z) = z^k \log z + \frac{z^k}{(k - 1)!} \left( c_k + T_k(k + 1) + \frac{1}{\log 2} \sum_{m \in \mathbb{Z} \setminus \{0\}} G_k(\chi_m) z^\chi_m \right) + \mathcal{O}(z^{k-1+\epsilon}) \]
where
\[ c_k = \gamma - 1 \frac{1}{\log 2} + \frac{1}{2} - (k - 1)! \sum_{j \geq 1} \frac{1}{2^j - 1} + \frac{(k - 1)!d_k}{\log 2}. \]

### 2.2 Variance and Covariance of the \( k \)-th Total Path Length

Next, let us consider the variance. Here we introduce the poissonized variance and covariance as
\[
\tilde{V}^{[k]}(z) = \tilde{f}[k]^2(z) - \tilde{f}[k](z)^2 - z \tilde{f}[k]'(z)^2,
\]
\[
\tilde{C}^{[k_1,k_2]}(z) = \tilde{f}[k_1,k_2]^2(z) - \tilde{f}[k_1,z] \tilde{f}[k_2,z](z) - z \tilde{f}[k_1]'(z) \tilde{f}[k_2]'(z),
\]
where
\[
\tilde{f}[k]^2(z) = e^{-z} \sum_{n \geq 0} \mathbb{E} \left( P_n[k]^2 \right) \frac{z^n}{n!} \quad \text{and} \quad \tilde{f}[k_1,k_2]^2(z) = e^{-z} \sum_{n \geq 0} \mathbb{E} \left( P_n[k_1] P_n[k_2] \right) \frac{z^n}{n!}.
\]

For detailed explanation of why we choose them this way, see [7]. Note that when \( k_1 = k_2 = k \), \( \tilde{V}^{[k]}(z) = \tilde{C}^{[k_1,k_2]}(z) \). Thus, we will consider only \( \tilde{C}^{[k_1,k_2]}(z) \) in this section.

From the given definition, we derive that
\[
\tilde{C}^{[k_1,k_2]}(z) + \tilde{C}^{[k_1,k_2]}'(z) = \tilde{f}[k_1,k_2]^2(z) + \tilde{f}[k_1,k_2]'(z) - \tilde{f}[k_1](z) \tilde{f}[k_2]'(z) - z \tilde{f}[k_1]'(z) \tilde{f}[k_2]'(z)
\]
\[
- \tilde{f}[k_1]'(z) \tilde{f}[k_2]^2(z) - \tilde{f}[k_1]^2(z) \tilde{f}[k_2]'(z) - z \tilde{f}[k_1]'(z) \tilde{f}[k_2]^2(z)
\]
\[
- z \tilde{f}[k_1]^2(z) \tilde{f}[k_2]'(z) - z \tilde{f}[k_1]'(z) \tilde{f}[k_2]'(z).
\]

From the recurrence of \( P_n[k_1+1] \), we derive the differential-functional equations of \( \tilde{f}[k] \) and \( \tilde{f}[k_1,k_2] \) and plug them into the above equation. Thus, by the same argument we used in the mean case, we find the bounds
\[
\tilde{C}^{[k_1,k_2]}(z) = \begin{cases} 
\mathcal{O}(z^{k_1+k_2-1+\epsilon}), & \text{as } z \to \infty; \\
\mathcal{O}(z^{\max(k_1,k_2)}), & \text{as } z \to 0^+
\end{cases}
\]
uniformly for \( z \) with \(|\arg z| \leq \frac{\pi}{2} - \epsilon\). With the help of computer algebra systems, we get that
\[
\tilde{C}^{[k_1,k_2]}(z) + \tilde{C}^{[k_1,k_2]}'(z) = 2 \sum_{r_1=1}^{k_1} \sum_{r_2=1}^{k_2} \left( \frac{z}{2} \right)^{k_1+k_2-r_1-r_2} \tilde{C}^{[k_1,k_2]} \left( \frac{z}{2} \right) + \tilde{g}[k_1,k_2](z).
\]

Because the exact expression of \( \tilde{g}[k_1,k_2](z) \) is way too complicated, we do not list the whole expression here. For the later computation, we only need the property that \( \tilde{g}[k_1,k_2](z) = \mathcal{O}(z^{k_1+k_2-2}) \) as \( z \to \infty \).
Similar to our analysis of the mean, we apply Laplace transform to the differential-functional equations and divide both sides by $Q(-2s)$. Let $k' = k_1 + k_2$, then

$$\mathcal{L}[\tilde{C}^{[k_1,k_2]}; s] = 4 \sum_{r_1=1}^{k_1} \sum_{r_2=1}^{k_2} (-1)^{k'-r_1-r_2} \mathcal{L}([k'-r_1-r_2];\tilde{C}[r_1,r_2]; 2s) + R_2^{[k_1,k_2]}(s),$$

where

$$R_2^{[k_1,k_2]}(s) = (-1)^{k'} \frac{\mathcal{L}[g_2^{[m_1,m_2]}; s]}{Q(-2s)} - 4 \sum_{r_1=1}^{k_1} \sum_{r_2=1}^{k_2} (-1)^{k'-r_1-r_2} \left( k' - r_1 - r_2 \right) \frac{h_j(s)}{2j} \mathcal{L}([k'-r_1-r_2-k];\tilde{C}[r_1,r_2]; 2s).$$

Before we proceed to apply Mellin transform, we derived similar bounds as in the analysis of the mean:

$$\mathcal{L}[\tilde{C}^{[k_1,k_2]}; s] = \begin{cases} \mathcal{O}(|s|^{-b}), & \text{as } s \to \infty; \\ \mathcal{O}(|s|^{-(k'+\epsilon)}), & \text{as } s \to 0^+, \end{cases}$$

where $b$ is a constant which can be arbitrarily large. Note that the bounds hold uniformly for $|\arg s| \leq \pi - \epsilon$. Now, we apply Mellin transform on both sides of the above equalities.

Again, we use the simplified notation $\mathcal{M}[\tilde{L}^{[k_1,k_2]}; \omega] = \mathcal{M}[\mathcal{L}[\tilde{C}^{[k_1,k_2]}]; \omega]$. Then, the equation becomes

$$\mathcal{M}[\tilde{L}^{[k_1,k_2]}; \omega] = 2^{2-\omega} \sum_{r_1=1}^{k_1} \sum_{r_2=1}^{k_2} \mathcal{M}[\tilde{L}^{[r_1,r_2]}; \omega + r_1 + r_2 - k'] \prod_{i=1}^{k'-r_1-r_2} (\omega - i) + \mathcal{M}[R_2^{[k_1,k_2]}; \omega]$$

for $\Re(\omega) > k'$. From [7], we already have that

$$\mathcal{M}[\tilde{L}^{[1]}; s; \omega] = \frac{H_1(\omega)}{1 - 2^{2-\omega}},$$

where

$$H_1(\omega) = Q_\infty \sum_{j,h,l \geq 0} (-1)^{j} 2^{-\left(\frac{j+1}{2}\right)+j(\omega-2)} Q_j Q_h Q_l 2^{h+l} \varphi(\omega; 2^{-j-h} + 2^{-j-l})$$

with

$$\varphi(\omega; x) = \int_0^\infty \frac{s^{\omega-1}}{(s+1)(s+x)^2} ds = \begin{cases} \frac{\pi(1 + x^{\omega-2}(\omega - 2)\zeta + 1 - \omega)}{2\sin(\pi\omega)}, & \text{if } x \neq 1; \\ \frac{x(\omega - 1)\sin(\pi\omega)}{2\sin(\pi\omega)}, & \text{if } x = 1. \end{cases}$$

Consequently, we can express $\mathcal{M}[\tilde{L}^{[k_1,k_2]}; \omega]$ in terms of $H_1(\omega)$

$$\mathcal{M}[\tilde{L}^{[k_1,k_2]}; \omega] = A_{k_1,k_2}(\omega) \frac{H_1(\omega + 2 - k')}{1 - 2^{2-\omega}} \prod_{i=1}^{k'-2} (\omega - i) + g_2^{[k_1,k_2]}(\omega),$$

where $A_{r_1,r_2}(\omega)$ satisfies the recurrence

$$A_{k_1,k_2}(\omega) = \frac{2^{2-\omega}}{1 - 2^{2-\omega}} \left( \sum_{r_1=1}^{k_1} \sum_{r_2=1}^{k_2} A_{r_1,r_2}(\omega + r_1 + r_2 - k') \right)$$
with the initial condition $A_{1,1}(\omega) = 1$. Note that $\tilde{g}^{[k_1,k_2]}_2(\omega)$ has no singularities with real part larger than $k' - 1$. From above recurrence, we can easily prove that for all $k \in \mathbb{Z}$

$$A_{k_1,k_2}(k_1 + k_2 + \chi_k) = \frac{2(k_1-1)(k_1-2)/2 \cdot 2(k_2-1)(k_2-2)/2}{\prod_{j=1}^{k_1-1}(2j - 1) \prod_{i=1}^{k_2-1}(2i - 1)} = \frac{2^{2-k'}}{Q_{k_1-1}Q_{k_2-1}}$$

by induction. For convenience, we set $C_{k_1,k_2} = A_{k_1,k_2}(k_1 + k_2)$. Applying inverse Mellin transform and collecting residues, we get

$$\mathcal{L}[\tilde{C}^{[k_1,k_2]}; \omega] = \frac{s-k'}{\log 2} \sum_{r \in \mathbb{Z}} C_{k_1,k_2} H_1(2 + \chi_r) s^{-\chi_r} \prod_{i=2}^{k_1+k_2-2} (i + \chi_r) + O(|s|^{1-k'})$$

uniformly as $|s| \to 0$ with $|\arg s| \leq \pi - \epsilon$. Finally, we apply inverse Laplace transform and Proposition 1 of [10] and obtain that, as $z \to \infty$,

$$\tilde{C}^{[k_1,k_2]}(z) = z^{k_1+k_2-1} C_{k_1,k_2} (C_{kps} + \overline{\omega}_{kps}(\log 2 \, n)) + O(|z|^{2k-2+\epsilon}).$$

In particular,

$$\tilde{V}^{[k]}(z) = \frac{z^{2k-1}}{\log 2} C_{k,k} (C_{kps} + \overline{\omega}_{kps}(\log 2 \, n)) + O(|z|^{2k-2+\epsilon})$$

as $z \to \infty$.

**Remark 4.** Note that from the expression of $C_{k_1,k_2}$, we have $C_{k_1,k_2} = C_{k_1,k_1} C_{k_2,k_2}$. Thus,

$$\rho(P_n^{[k_1]}, P_n^{[k_2]}) = \frac{\text{Cov}(P_n^{[k_1]}, P_n^{[k_2]})}{\sqrt{\text{Var}(P_n^{[k_1]}) \text{Var}(P_n^{[k_2]})}} \sim \frac{n^{2k_1+2k_2-2} C_{m,m}^2 (C_{kps} + \overline{\omega}_{kps}(\log 2 \, n))^2}{n^{2k_1+2k_2-2} C_{m,m} C_{m-1,m-1} (C_{kps} + \overline{\omega}_{kps}(\log 2 \, n))^2} = 1.$$ 

**Remark 5.** Since we already know that $P_n^{[1]}$ satisfies a central limit theorem [10], together with the result in the above remark and applying similar argument as of [8], we obtain that

$$\left( \frac{P_n^{[1]} - \mathbb{E}(P_n^{[1]})}{\sqrt{\text{Var}(P_n^{[1]})}}, \ldots, \frac{P_n^{[k]} - \mathbb{E}(P_n^{[k]})}{\sqrt{\text{Var}(P_n^{[k]})}} \right) \overset{d}{\to} (X, \ldots, X),$$

where $X$ is a standard normal distributed random variable and $\overset{d}{\to}$ denotes weak convergence.

### 3 Total Steiner $k$-distance

Let $S_{n}^{[k]}$ be the Steiner $k$-distance. Then, using the same idea as for the $k$-th total path length, we consider four cases:

1. **All $k$ nodes are from one subtree.**
   $$S_{B_n}^{[k]} + S_{n-B_n}^{[k]}.$$

2. **The $k$ nodes are chosen from both subtrees and the root is not chosen.**
   $$\sum_{l=1}^{k-1} \binom{n-B_n}{k-l} P_0^{[l]} + \binom{B_n}{l} P_{n-B_n}^{[k-l]} + 2 \binom{B_n}{l} \binom{n-B_n}{k-l}.$$
3. The root is chosen, the other $k-1$ nodes are all from one subtree.

$$P_{B_n}^{[k-1]} + P_{B_n}^{[k-1]s} + \left( \frac{n - B_n}{k - 1} \right) + \left( \frac{B_n}{k - 1} \right).$$

4. The root is chosen, the other $k-1$ nodes are from both subtrees.

$$\sum_{l=1}^{k-2} \left( \frac{n - B_n}{k - l - 1} \right) P_{B_l}^{[l]} + \left( B_n \right) P_{B_n}^{[l-1]s} + 2 \left( B_n \right) \left( \frac{n - B_n}{k - l - 1} \right).$$

Note that as for the $k$-th total path length, here we have a system of recurrences for the Steiner $k$-distance. Similar to the analysis of the $k$-th total path length, we let $\tilde{g}^{[k]}(z)$ be the Poisson generating function of the mean of the total Steiner $k$-distance, $\tilde{W}^{[k_1,k_2]}(z)$ be the Poissonized covariance of the total $k_1$-th Steiner distance and the total $k_2$-th total path length and $\tilde{V}_S^{[k]}(z)$ be the variance of the $k$-th Steiner distance. With the help from computer algebra systems, we get the differential-functional equations

$$\tilde{g}^{[k]}(z) + \tilde{g}^{[k]}(z) = 2\tilde{g}^{[k]}(z) + 2 \sum_{r=1}^{k-2} \left( \frac{\tilde{g}^{[r-1]}(z)}{r} \right) + \frac{2}{(k-1)!},$$

$$\tilde{W}^{[k_1,k_2]}(z) + \tilde{W}^{[k_1,k_2]}(z) = 2 \sum_{r=1}^{k_2} \left( \frac{\tilde{W}^{[k_1,k_2]}(z)}{r} \right) + 2 \sum_{r_1,r_2} \left( \tilde{W}^{[k_1]r_1} \tilde{W}^{[k_2]r_2} \right) + \tilde{h}^{[k_1,k_2]}(z).$$

and

$$\tilde{V}_S^{[k]}(z) + \tilde{V}_S^{[k]}(z) = 2\tilde{V}_S^{[k]}(z) + 4 \sum_{r=1}^{k-2} \left( \frac{\tilde{W}^{[k_1,k_2]}(z)}{r} \right) + 4 \sum_{r=1}^{k-2} \left( \tilde{W}^{[k_1,k_2]}(z) \right) + \tilde{h}_S^{[k]}(z).$$

We use $\tilde{h}_S^{[k_1,k_2]}(z)$ and $\tilde{h}_S^{[k]}(z)$ to denote the lower order terms. Because the rest of the analysis will be very similar to the one with the $k$-th total path length, we skip the details and list only the results

$$\mathbb{E} \left( S_n^{[k]} \right) = \frac{n^k}{(k-1)!} + \frac{n^k}{(k-1)!} + C_k(\chi_n) \left( n^{k-1} + \mathcal{O}(n^{k-1+\epsilon}) \right),$$

$$\text{Cov} \left( S_n^{[k_1]}, P_n^{[k_2]} \right) = \frac{n^{k_1+k_2-1}}{\log 2} C_{k_1,k_2} (C_{k_1} + \mathcal{O}(n^2 \log n)) + \mathcal{O}(n^{k_1+k_2-2}),$$

$$\text{Var} \left( S_n^{[k]} \right) = \frac{n^{2k-1}}{\log 2} C_{k,k} (C_{k_1} + \mathcal{O}(n^2 \log n)) + \mathcal{O}(n^{2k-2}).$$

Since the leading terms are exactly the same as for the $k$-th total path length, the same arguments as for $P_n^{[k]}$ gives us the results stated in Theorem 2.

**Acknowledgements**

The author thanks Michael Fuchs for many helpful suggestions. Part of this work was carried out when the author visited the Institute of Discrete Mathematics and Geometry, Vienna University of
Technology. The author wants to thank the department for hospitality and support. Moreover, the author acknowledges partial support by the Office of International Affairs, National Chiao Tung University and the National Science Council, Taiwan.

References


