Title	Computing the Distribution Function of the Stochastic Longest Path Length in a DAG with Continuously Distributed Edge Lengths (Theoretical Computer Science and Its Applications)
Author(s)	Ando, Ei; Ono, Hirotaka; Sadakane, Kunihiko; Yamashita, Masafumi
Citation	数理解析研究所講究録 (2009), 1649: 252-259
Issue Date	2009-05
URL	http://hdl.handle.net/2433/140729
Right	
Type	Departmental Bulletin Paper
Textversion	publisher

数理解析研究所講究録
第1649巻 2009年 252-259

連続確率分布枝重み付きDAGに対する最長路長さ分布の計算
Computing the Distribution Function of the Stochastic Longest Path Length in a DAG with Continuously Distributed Edge Lengths

Ei Ando† Hirotaka Ono†‡
Kunihiko Sadakane† Masafumi Yamashita†‡

†Department of Computer Science and Communication Engineering,
Graduate School of Information Science and Electrical Engineering,
Kyushu University.
‡Institute of Systems, Information Technologies and Nanotechnologies.

Abstract. Consider the longest path problem for directed acyclic graphs (DAGs), where a mutually independent random variable is associated with each of the edges as its edge length. Given a DAG G and any distributions that the random variables obey, let $F_{\mathrm{MAX}}(x)$ be the distribution function of the longest path length. We first represent $F_{\mathrm{MAX}}(x)$ by a repeated integral that involves $n-1$ integrals, where n is the order of G. We next present an algorithm to symbolically execute the repeated integral, provided that the random variables obey the standard exponential distribution. Although there can be $\Omega(2^n)$ paths in G, its running time is bounded by a polynomial in n, provided that k, the cardinality of the maximum anti-chain of the incidence graph of G, is bounded by a constant. We finally propose an algorithm that takes x and $\epsilon > 0$ as inputs and approximates the value of repeated integral of x, assuming that the edge length distributions satisfy some natural conditions. : (1) The length of each edge $(v_i, v_j) \in E$ is non-negative, (2) the Taylor series of its distribution function $F_{ij}(x)$ converges to $F_{ij}(x)$, and (3) there is a constant σ that satisfies $\sigma^p \leq \left|\left(\frac{d}{dx}\right)^p F_{ij}(x)\right|$ for any non-negative integer p. It runs in polynomial time in n, and its error is bounded by ϵ, when x, ϵ, σ and k can be regarded as constants.

1 Introduction

Let $G = (V, E)$ be a directed acyclic graph (DAG), where V and E are the sets of of n vertices and m edges, respectively. Each edge (v_i, v_j) is associated with a random variable X_{ij} representing its length. Although the longest path problem for DAGs is solvable in linear time when edge lengths are constant values, the same problem with stochastic edge lengths is formidable. Actually, there are at least two different problem formulations; to find a path that has the highest probability of being the longest [12], or to compute the distribution function $F_{\mathrm{MAX}}(x)$ of the longest path length [1–5, 7, 8, 10, 11]. In this paper, we adopt the second formulation.

The longest path problem in G with uncertain edge lengths is known as the classic problems such as Program Evaluation and Review Technique (PERT) [6] or Critical Path Planning (CPP) [9]. In these problems, the lower and the upper bounds of the edge lengths (the activity duration) are given as static values, and their goal is to obtain the lower and upper bounds on the longest path length in G, the duration of the whole project. However, we assume, in this paper, that edge lengths are random variables; we are not to determine the edge lengths but to cope with the resulting edge lengths that realize with some probability.

Delay analysis of logical circuits is a killer application of this problem, and besides Monte Carlo simulations, many heuristic approximation algorithms have been proposed so far (see e.g., [3, 5, 7]). They run fast but their general drawback is that they do not have a theoretical approximation guarantee. To theoretically guarantee an approximation ratio, some authors of this paper proposed an algorithm to construct a primitive function that approximates $F_{\mathrm{MAX}}(x)$ [1, 2].

Computing the exact distribution function has also a long research history. Martin [11] proposed a series-parallel reduction based method, assuming that each edge length obeys a polynomial distribution.

Kulkarni and Adlakha [10] proposed an algorithm that is based on the analysis of continuous time Markov chain. Both algorithms unfortunately take an exponential time with respect to the graph size. Indeed, when edge lengths obey discrete distributions, the problem is #P-complete [8], and is NP-hard even for the series-parallel graphs [4].

We first show that $F_{\mathrm{MAX}}(x)$ is represented by a repeated integral that involves $n-1$ integrals, for any instance of the problem. The problem of computing $F_{\mathrm{MAX}}(x)$ for any x is thus reducible to the problem of evaluating the repeated integral for x. The evaluation of the repeated integral is possible by making use of standard numerical methods at the expense of accuracy and time.

In this paper, we pursuit the possibility of exact computation using the repeated integral. That only $n-1$ integrals are involved might give us a chance to symbolically compute it in polynomial in n, although there can be $\Omega(2^n)$ paths in G (and the above NP-hardness results essential suggest that any algorithm would need to evaluate each of the $\Omega(2^n)$ paths).

Assuming that the random variables obey the standard exponential distribution, we show that there is an algorithm to transform the repeated integral into a product of primitive functions. It runs in polynomial time in n, provided that k, the cardinality of the maximum anti-chain of the incidence graph of G, is bounded by a constant.

We (of course) cannot present a polynomial time algorithm that works for any distribution of edge length. Naive numerical methods to approximate the repeated integral, on the other hand, need sufficiently long computation time and do not guarantee approximation performance. We thus assume that the distribution function F_{ij} associated with any edge (v_i, v_j) satisfies the following three natural conditions: (1) The length of each edge is positive (i.e., $F_{ij}(x) = 0$ for $x \leq 0$), (2) there is a constant σ that satisfies $\left|\left(\frac{d}{dx}\right)^p F_{ij}(x)\right| \leq \sigma^p$ for any non-negative integer p, and (3) the Taylor series of $F_{ij}(x)$ converges to $F_{ij}(x)$, and then present, for any $\epsilon > 0$, an approximation algorithm that evaluates $F_{\mathrm{MAX}}(x)$ (i.e., the repeated integral) with an error less than ϵ. It runs in polynomial time in n, when x, ϵ, σ and k can be regarded as constants.

This paper is organized as follows: After giving basic definitions and formulas in Section 2, we derive the repeated integral form of $F_{\mathrm{MAX}}(x)$ in Section 3. Section 4 is devoted to the first case in which an exact formula is derived assuming the standard exponential distribution, and Section 5 proposes an approximation algorithm for the second case. Section 6 concludes this paper.

2 Preliminaries

Let $G = (V, E)$ be a directed acyclic graph with vertex set $V = \{v_1, v_2, \ldots, v_n\}$ and directed edge set $E \subseteq V \times V$ of m edges. We assume that each edge $(v_i, v_j) \in E$ is associated with its length X_{ij} that is a random variable. A *source* (resp. *terminal*) of G is a vertex in V such that its in-degree (resp. out-degree) is 0. We define the *(directed) incidence graph* of $G = (V, E)$ as a directed graph G' with vertex set $V' = V \cup E$ and edge set $E' = \{(v_i, e), (e, v_j) | e = (v_i, v_j) \in E\} \subseteq (V \times E) \cup (E \times V)$. We denote the incidence graph of G by $L(G)$. A subset A of V is called an *antichain* of G if each $v_a \in A$ is not reachable from any other vertex $v_b \in A$. If $(v_i, v_j) \in E$, two vertices v_i and v_j are *neighbors* to each other, v_i is a *parent* of v_j, and v_j is a *child* of v_j. By $N(W)$ we denote the set of all neighbors of vertices in W. Let $\mathcal{P}$ be the set of all source-terminal paths. The longest path length X_{MAX} of G is given as $X_{\mathrm{MAX}} = \max_{\pi \in \mathcal{P}} \left\{ \sum_{(v_i, v_j) \in \pi} X_{ij} \right\}$.

Let X be a random variable. The probability $P(X \leq x)$ is called the *(cumulative) distribution function* of X. The *density function* of X is the derivative of $P(X \leq x)$ with respect to x. We say X *obeys the standard exponential distribution* if the distribution function $P(X \leq x)$ is given by $P(X \leq x) = 1 - \exp(-x)$ if $x \geq 0$ and $P(X \leq x) = 0$ if $x < 0$.

Let X_1 and X_2 be two mutually independent random variables. Let $f_1(x)$ and $f_2(x)$ be the density functions of X_1 and X_2, respectively. The sum $X_1 + X_2$ is also a random variable whose distribution function is given as

$$P(X_1 + X_2 \leq x) = \int_{\mathbf{R}} P(X_1 + t \leq x \mid X_2 = t) f_2(t) dt = \int_{\mathbf{R}} F_1(x - t) f_2(t) dt, \tag{1}$$

where $F_1(x)$ and $F_2(x)$ are the distribution functions of X_1 and X_2, respectively. The distribution function of $\max\{X_1, X_2\}$ is given as $P(\max\{X_1, X_2\} \leq x) = P(X_1 \leq x \wedge X_2 \leq x) = F_1(x)F_2(x)$.

3 Repeated Integral Representation of $F_{\mathrm{MAX}}(x)$

In this section, we show that the distribution function $F_{\mathrm{MAX}}(x)$ of the longest path length is represented by a repeated integral that involves $n-1$ integrals. By definition,

$$F_{\mathrm{MAX}}(x) = P(X_{\mathrm{MAX}} \leq x) = P\left(\bigwedge_{\pi \in \mathcal{P}} \left(\sum_{e \in \pi} X_e \leq x\right)\right). \tag{2}$$

Although this formula is compact, this fact does not directly implies an efficient computability, since it would take into account all source-terminal paths in G, which can be as many as $\Omega(2^n)$. Next theorem shows that $F_{\mathrm{MAX}}(x)$ is represented by a repeated integral that involves $n-1$ integrals. Thus $F_{\mathrm{MAX}}(x)$ can be computed by executing only $n-1$ integrals, which may be dramatically more efficient than the calculation of Eq. (2). Let $H(x)$ be a function that satisfies $H(x) = 1$ if $x \geq 0$ and $H(x) = 0$ if $x < 0$. Let $\mathbf{1}(x)$ be a constant function that maps every x to 1. Note that if $P(X \leq x) = H(x)$ (resp. $P(X \leq x) = \mathbf{1}(x)$) for any x, X is always equal to 0 (resp. $-\infty$).

Theorem 1. *Let $G = (V, E)$ is a DAG. Without loss of generality, we assume that $V = \{v_1, v_2, \ldots, v_n\}$ is topologically ordered. For any edge $(v_i, v_j) \in E$, let $F_{ij}(x)$ be the distribution function that X_{ij} obeys. We associate a function $F_{ij}(x)$ with each edge $(v_i, v_j) \notin E$ as follows: If (v_i, v_j) connects two sources or two terminals, then $F_{ij}(x) = H(x)$; otherwise, $F_{ij}(x) = \mathbf{1}(x)$. Then the distribution function $F_{\mathrm{MAX}}(x)$ is given as*

$$P(X_{\mathrm{MAX}} \leq x) = \int_{\mathbf{R}^{n-1}} H(x - z_1) \prod_{1 \leq i \leq n-1} \left(\frac{d}{dz_i} \prod_{i+1 \leq j \leq n} F_{ij}(z_i - z_j)\right) dz_i. \tag{3}$$

Proof. Given a DAG $G = (V, E)$, we first add edges as many as possible in such a way that the added edges do not change the topological order. This yields a complete graph with acyclic orientations, which is denoted by $\overrightarrow{K}_n = (V, E_K)$, where $E_K = \{(v_i, v_j) \mid 1 \leq i < j \leq n\}$. Notice that $\overrightarrow{K}_n$ has a unique source v_0 and a unique terminal v_n. For each of the edge $(v_i, v_j) \in E_K$, we associate a random variable X_{ij} that represents the length of (v_i, v_j), and assume that X_{ij} obeys $F_{ij}(x)$. We observe that $F_{\mathrm{MAX}}(x)$ is exactly the same for G and $\overrightarrow{K}_n$. Note that any path in G is also a path in $\overrightarrow{K}_n$. Consider any path π in $\overrightarrow{K}_n$ connecting a source v_S (of G) and a terminal v_T (of G) that does not exist in G. If π contains an edge $(v_i, v_j) \notin E$ such that $F_{ij}(x) = \mathbf{1}(x)$, then $\sum_{e \in \pi} X_e \leq x$ holds for all x (see the note above for intuition), which implies that such π is ignorable in Eq. (2). Suppose that π does not contain an edge $(v_i, v_j) \notin E$ such that $F_{ij}(x) = \mathbf{1}(x)$. Let π' is the path constructed from π by removing all edges connecting two sources or two terminals. Then π' is a path connecting a source and a terminal in G, and $\sum_{e \in \pi} X_e \leq x$ iff $\sum_{e \in \pi'} X_e \leq x$ (see the note above for intuition).[1] Thus $F_{\mathrm{MAX}}(x)$ is exactly the same for G and $\overrightarrow{K}_n$. In what follows, we assume $G = \overrightarrow{K}_n$.

Define several notations: $\mathcal{P}(i, j)$ is the set of all paths from v_i to v_j, $\mathcal{P}_k(i, j)$ is the set of all v_i-v_j paths that *do not* pass a vertex in $U_k = \{v_k, v_{k+1}, \ldots, v_{n-1}\}$, and $Z_{n-1} = X_{n-1,n}$ is the longest path length from v_{n-1} to the unique terminal v_n of $\overrightarrow{K}_n$. We would like to use Z_{n-1} rather than $X_{n-1,n}$ in order to illustrate the transformations that should follow (5), but is not explicitly explained here for the limitation of the space. Since $Z_{n-1} = X_{n-1,n}$ and X_{ij}'s are mutually independent, we have

$$P(X_{\mathrm{MAX}} \leq x) = P\left(\bigwedge_{\pi \in \mathcal{P}_{n-1}(1,n)} \left(\sum_{(v_i, v_j) \in \pi} X_{ij} \leq x\right) \wedge \bigwedge_{\pi \in \mathcal{P}(1,n-1)} \left(\sum_{(v_i, v_j) \in \pi} X_{ij} + Z_{n-1} \leq x\right)\right). \tag{4}$$

[1] Several different paths $\pi_1, \pi_2, \cdots, \pi_\ell$ may correspond to a single π'. Even in such a case, the AND of conditions $\sum_{e \in \pi_i} X_e \leq x$ for $i = 1, 2, \cdots, \ell$ is reduced to condition $\sum_{e \in \pi'} X_e \leq x$.

Let $G_{n-1}(x) = F_{n-1,n}(x)$ and $g_{n-1}(x)$ be the distribution function of Z_{n-1} and its density function, respectively. Since $dG_{n-1}(z_{n-1}) = g_{n-1}(z_{n-1})dz_{n-1}$, like the derivation of Eq.(1), by introducing an integral, the right-hand side of Eq. (4) is represented as

$$\int_{\mathbf{R}} P\left(\underbrace{\bigwedge_{\pi\in\mathcal{P}_{n-1}(1,n)} \left(\sum_{(v_i,v_j)\in\pi} X_{ij} \le x \right)}_{(A)} \wedge \underbrace{\bigwedge_{\pi\in\mathcal{P}(1,n-1)} \left(\sum_{(v_i,v_j)\in\pi} X_{ij} + z_{n-1} \le x \right)}_{(B)} \right) dG_{n-1}(z_{n-1}). \quad (5)$$

We then calculate the contribution of each edge by repeating the transformation of representing (and replacing) the contribution by an integral. For $Z_k = \max_{k+1\le l\le n}\{X_{kl} + z_l\}$, we divide the paths from v_1 to v_n into two groups according to whether or not they pass v_k. We then introduce one more integral to aggregate the probability that Z_k takes a constant value z_k. We can consider z_k as the dummy variable of a convolution. Note that $z_n = 0$ by definition.[2] Now for each of $Z_{n-1}, Z_{n-2}, \ldots, Z_2$, an integral has been introduced with respect to z_i, and (4) is transformed into

$$\int_{\mathbf{R}^{n-2}} P\left(\bigwedge_{2\le l\le n} \bigwedge_{\pi\in\mathcal{P}_2(1,l)} \left(\sum_{(v_i,v_j)\in\pi} X_{ij} + z_l \le x \right) \wedge \bigwedge_{\pi\in\mathcal{P}(1,2)} \left(\sum_{(v_i,v_j)\in\pi} X_{ij} + z_2 \le x \right) \right) \prod_{1\le i\le n-1} G_i(z_i, \ldots, z_{n-1}) dz_i, \quad (6)$$

where $G_i(z_i, \ldots, z_{n-1}) = \dfrac{d}{dz_i} \displaystyle\prod_{i+1\le j\le n} P(X_{ij} + z_j \le z_i) = \dfrac{d}{dz_i} \prod_{i+1\le j\le n} F_{ij}(z_i - z_j)$.

By definition, $\mathcal{P}(1,2) = \{(v_1, v_2)\}$ and $\mathcal{P}_2(1,l) = \{(v_1, v_l)\}$. Hence (6) is equal to

$$\int_{\mathbf{R}^{n-2}} \prod_{2\le l\le n} F_{1l}(x - z_l) \prod_{2\le i\le n-1} \left(\frac{d}{dz_i} \prod_{i+1\le j\le n} F_{ij}(z_i - z_j) \right) dz_i, \quad (7)$$

which implies the theorem by $\int_{\mathbf{R}} H(x - z_1)\frac{d}{dz_1} G_1(z_1, z_2, \ldots, z_{n-1})dz_1 = G_1(x, z_2, \ldots, z_{n-1})$ where $G_1(z_1, \ldots, z_{n-1}) = \prod_{2\le l\le n} F_{1l}(z_1 - z_l)$.

□

It is worth noting that Theorem 1 is applicable, even if the length c_{ij} of each edge (v_i, v_j) is a constant value. In this case, the step function $H(x - c_{ij})$ is given as the distribution function that X_{ij} obeys. Let d_i be the (definite) longest path length from v_i to v_n. Then the step function $F_{\mathrm{MAX}}(x) = H(x - d_1)$ is obtained by Theorem 1.

In the following, we call dummy variable z_i *the corresponding variable* of v_i. Let $Q_1(z_1, z_2, \ldots, z_{n-1}; x) = H(x - z_1)$ and

$$Q_{l+1}(z_{l+1}, \ldots, z_{n-1}; x) = \int_{\mathbf{R}} Q_l(z_l, z_{l+1}, \ldots, z_{n-1}; x) G_l(z_l, z_{l+1}, \ldots, z_{n-1}) dz_l.$$

Theorem 1 states that we can calculate $Q_n(x) = F_{\mathrm{MAX}}(x)$ by repeating integrals.

4 Exact Computation of the Repeated Integral

This section considers the case in which the edge lengths are given by mutually independent random variables that obey the standard exponential distribution function. We present an algorithm to compute each of $Q_1, Q_2, \ldots, Q_n$ symbolically in this order by expanding the integrand into a sum of products before calculating each integral. Let k be the cardinality of the maximum anti-chain of $L(G)$. By bounding the number of different terms that can appear during the symbolic calculation, we show that its running time is a polynomial in the size of G, if k is bounded by a constant.

[2] Notice that we define Z_k after $z_{k+1}, z_{k+2}, \ldots, z_{n-1}$; if we define Y_i as the length of the longest path from v_k to v_n at a time, then Y_i's are dependent on each other, which implies that the above proof cannot be applied to Y_i's.

Proposition 1. *Let $W_i = \{v_j \mid 1 \leq j \leq i\}$. If $v_l \in W_i \setminus \{v_i\}$ or $(u, v_l) \notin E$ for any $u \in W_i$, then $Q_i(z_i, \ldots, z_{n-1}; x)$ does not depend on z_l.*

Proof. Since z_l is a dummy variable of an integral if $l < i$, it is obvious that z_l never show up in $Q_i(z_i, \ldots, z_{n-1}; x)$ after the integrals are computed.

Suppose otherwise that $l > i$. Then $v_l \notin N(W_i) \setminus W_i$. By Theorem 1, $G_i(z_i, \ldots, z_{n-1})$ does not depend on z_l. □

Let $m = |E|$, $n = |V|$ and V_i be the set of children of v_i.

Theorem 2. *Let $G = (V, E)$ be a DAG such that the cardinality of the maximum anti-chain of $L(G)$ is at most k. Assume that each random variable X_{ij}, which represents the length of edge (v_i, v_j), obeys the standard exponential distribution. Then the distributed function $F_{\mathrm{MAX}}(x)$ of the longest path length in G is computable in $O((k+1)!n^{k+2}(2m+1)^{k+1})$ time.*

Proof. We first show how we calculate $Q_{i+1}(z_{i+1}, \ldots, z_{n-1}; x)$ from $Q_i(z_i, \ldots, z_{n-1}; x)$ by symbolically executing the integral with respect to z_i. For example, $Q_3(z_3, \ldots, z_{n-1}; x)$ is given by

$$Q_3(z_3, \ldots, z_{n-1}; x) = \int_{\mathbf{R}} Q_2(z_2, \ldots, z_{n-1}; x) G_2(z_2, \ldots, z_{n-1}) dz_2. \tag{8}$$

Since $H(x) = 0$ for all $x < 0$, $Q_3(z_3, \ldots, z_{n-1}; x) = 0$ if $x < 0$. When $x \geq 0$, since $H(x) = 1$,

$$Q_3(z_3, \ldots, z_{n-1}; x) = \int_a^b \prod_{v_j \in V_1} (1 - \exp(-(x - z_j))) \frac{d}{dz_2} \prod_{v_l \in V_2} (1 - \exp(-(z_2 - z_l)))) dz_2, \tag{9}$$

where $a = \max_{v_\ell \in V_2} z_\ell$ and $b = x$, since otherwise the contribution to the integral becomes 0 because of the effect of H. Since each of z_l's can take the maximum, at most $|V_2|$ different formulas appear, corresponding to different $a = z_l$, as possible results of $Q_3(z_3, \ldots, z_{n-1}; x)$. Once a is fixed to a z_l, executing symbolic integration of the right-hand side of Eq. (9) is easy, since possible terms appearing in the integrand have a form of $c_1 \exp(-c_2 z_2)$ for some constant c_1 and c_2. In general, we can derive $Q_{i+1}(z_{i+1}, \ldots, z_{n-1}; x)$ from $Q_i(z_i, \ldots, z_{n-1}; x)$ in the same way.

To estimate the time complexity of the algorithm, let us estimate the number of terms that are possible to appear in the execution. By Proposition 1, the number of variables appeared in $Q_i(z_i, \ldots, z_{n-1}; x)$ is at most $k+1$ for any i. As explained, to obtain $Q_3(z_3, \ldots, z_{n-1}; x)$, we need to consider at most $k+1$ different cases corresponding to different $a = z_l$. It is easy to see that to obtain $Q_4(z_4, \ldots, z_{n-1}; x)$, for each of the cases for Q_3, we need to consider at most k different cases. Although this leads to that there may be $O(k^i)$ cases for Q_i in general, the number of variables on which Q_i depends is at most $k+1$ by Proposition 1, which implies that, in general in Q_i, there can be no more than $(k+1)!$ distinct cases. To complete the proof, we show that at most $n^{k+1}(2m+1)^{k+1}$ terms are possible to appear, for each of at most $(k+1)!$ cases.

Let us consider the number of the terms in the integrand in each case. Since it is easy to see that each term is a product of x^{α_0}, $z_j^{\alpha_j}$, $\exp(\beta_0 x)$ and $\exp(\beta_j z_j)$, where α_j's and β_j's are integers, we bound the number of terms by the number of possible terms. By the form of Theorem 1, we can see that the maximum degrees of z_j's and x that appear in the terms in $Q_i(z_i, \ldots, z_{n-1}; x) G_i(z_i, \ldots, z_{n-1})$ of each case can only increase by one in one integral and hence α_j's are non-negative integer and less than n. Similarly, we can also see that the degrees β_j's of $\exp(z_j)$'s and $\exp(x)$ can only increase or decrease by one in a multiplication of two distribution functions and hence β_j's are integers between $-m$ and m. Therefore, the integrand in each cases consists of at most $n^{k+1}(2m+1)^{k+1}$ terms, which amounts to that the calculation of each $Q_{i+1}(z_{i+1}, \ldots, z_{n-1}; x)$ from $Q_i(z_i, \ldots, z_{n-1}; x)$ takes $O((k+1)!n^{k+1}(2m+1)^{k+1})$ time. □

Corollary 1. *A closed form of $F_{\mathrm{MAX}}(x)$ consisting of primitive functions is obtained in polynomial time if k is bounded by a constant, provided that the edge lengths obey the standard exponential distribution.*

5 Approximation of the Repeated Integral

In this section, we assume that the cardinality of the incidence graph $L(G)$ of a given DAG G is bounded by a constant k. We show that the distribution function $F_{\mathrm{MAX}}(x)$ of the longest path lengths can be approximately calculated in polynomial time in n, if the length of each edge $(v_i, v_j) \in E$ is non-negative and the Taylor series of its distribution function $F_{ij}(x)$ converges to $F_{ij}(x)$. Here by Taylor polynomial of $f(z_1, \ldots, z_n; x)$, we mean the Taylor polynomial that is generated by $f(z_1, \ldots, z_n; x)$ at $x = z_1 = z_2 = \cdots = z_n = 0$.

We must be careful for the order of computing the Taylor polynomial of the repeated integral that is shown in Theorem 1. Let p be the order of the Taylor polynomial. The most intuitive idea is that we compute the Taylor polynomial of the whole integrand

$$H(x - z_1) \prod_{1 \le i \le n-1} \frac{d}{dz_i} \prod_{i+1 \le j \le n} F_{ij}(z_i - z_j), \tag{10}$$

treating it as a function of n variables (i.e., $x, z_1, z_2, \ldots, z_{n-1}$). However, this intuitive way of computing the Taylor polynomial is not efficient for obtaining the value of $F_{\mathrm{MAX}}(x)$ with an error less than ϵ; the running time may be more than exponential with respect to the size of G even if k is a constant. Let us describe the p-th derivatives of (10) as a sum of products of $F_{ij}(z_i - z_j)$ or its derivatives of some order. Since one differentiating operation of (10) or its derivatives creates $2k$ times as many terms as in the original if we do not replace the distribution functions $F_{ij}(x)$ by particular definition of the edge lengths' distribution functions, there may be $O((2k)^p)$ terms in the p-th derivative of (10). Then, it can be shown that the order p of the Taylor polynomial needs to be almost linear to the size of the given DAG G to keep the error less than a constant ϵ even if k is a constant, which implies that the running time may be more than exponential of m and n.

In order to lower the running time, we approximate $Q_i(z_i, \ldots, z_{n-1}; x)$ by $A_i^p(z_i, \ldots, z_{n-1}; x)$ that is computed by the following procedures: (1) $A_2^p(z_2, \ldots, z_{n-1}; x)$ is the Taylor polynomial of order p generated by $Q_2(z_2, \ldots, z_{n-1}; x) = \prod_{v_j \in V_1} F_{1j}(x - z_j)$, and (2) $A_i^p(z_i, \ldots, z_{n-1}; x)$ is the Taylor polynomial of order p generated by

$$\int_{\mathbf{R}} A_{i-1}^p(z_{i-1}, \ldots, z_{n-1}; x) G_{i-1}(z_{i-1}, \ldots, z_{n-1}) dz_{i-1}. \tag{11}$$

This integral can be calculated using integration by parts, which yields a sum of products of polynomials and some anti-derivatives of $G_{i-1}(z_{i-1}, \ldots, z_{n-1}) = \prod_{i \le j \le n} F_{i-1,j}(z_{i-1} - z_j)$. The procedure (2) can be repeated for $i = 3, 4, \ldots, n$.

Since all edge lengths are non-negative by assumption, the anti-derivative of $G_{i-1}(z_{i-1}, \ldots, z_{n-1})$ of positive order is equal to 0 at the origin $x = z_i = z_{i+1} = \cdots = z_{n-1} = 0$, which allows us to compute $A_i(z_i, \ldots, z_{n-1}; x)$ as the Taylor polynomial of order p without knowing the analytic form of the anti-derivatives $G_{i-1}(z_{i-1}, \ldots, z_n)$.

In the next theorem, we show that the time to compute $A_n^p(x)$ where p is large enough to keep the error less than ϵ is polynomial of the size of G, assuming that x, ϵ and the maximum size k of an antichain in $L(G)$ is a constant. We also assume the existence of a constant σ, that satisfies $\sigma^p \ge \left|\left(\frac{d}{dx}\right)^p F_{ij}(x)\right|$ for any non-negative integer p and any edge $(v_i, v_j) \in E$.

Notice that σ must be bounded by a constant for the assumption that x is bounded by a constant. If there is an algorithm A that gives the value of $F_{\mathrm{MAX}}(x)$ in the same time regardless of σ, we can consider "compressed edge length" $X'_e = X_e/s$, where X_e is the length of e and $s \ge 1$. Then we can define the "compressed" distribution functions $F'_{\mathrm{MAX}}(x) = P(\bigwedge_{\pi \in \mathcal{P}}(\sum_{e \in \pi} X'_e \le x))$ of the longest path length. Since $F'_{\mathrm{MAX}}(x) = F_{\mathrm{MAX}}(sx)$, the value of $F_{\mathrm{MAX}}(sx)$ can be obtained for any s in the same running time, which can be used for obtaining the value of $F_{\mathrm{MAX}}(x)$ for arbitrary x. Therefore, it is essential to bound σ by a constant as well as x.

Theorem 3. *Let $G = (V, E)$ be a DAG and assume that the cardinality of the anti-chain of its incidence graph $L(G)$ is at most k. Let $F_{ij}(x)$ be the distribution function of the length of an edge (v_i, v_j) that*

is defined in Theorem 2. Let σ be a value such that $\sigma^p \geq |(\frac{d}{dx})^p(F_{ij}(x))|$ for any non-negative integer p and any edge $(i,j) \in E$. We further assume that the Taylor series of $F_{ij}(x)$ converges to $F_{ij}(x)$ itself and that the time complexity of computing the p-th derivative of $F_{ij}(x)$ is $O(\exp(p))$. Then $A_n^p(x)$ such that $|A_n^p(x) - F_{\mathrm{MAX}}(x)| \leq \epsilon$ holds is calculated in time $O((k+1)!(p+1)^k k^{p+1}\exp(p))$, where $p = O(k^2 x\sigma + \ln n + \ln 1/\epsilon)$.

Proof. By the similar argument in the previous section, it can be shown that the time to compute $A_n^p(x)$ is $O((k+1)!n(p+1)^k k^{p+1}\exp(p))$. We have $O((k+1)!)$ cases for computing integral of $A_i^p(z_i,\ldots,z_{n-1};x)$. In each cases, the integrand is the sum of products $C(\alpha)\prod_{v_j\in V_i} y_j^{\alpha_j}$ over all possible $\alpha = \{\alpha_1,\ldots,\alpha_k\}$, where y_j is the corresponding variable of the j-th vertex in $N(W_i)\setminus W_i$ (i.e., y_j is equal to one of z_h where $v_h \in N(W_i)\setminus W_i$), where $W_i = \{v_1,\ldots,v_i\}$. Since α_j is a non-negative integer at most p, the number of the terms in each case of integral is at most $(p+1)^k$. Since differentiating a term in the resulting form of (11) p times with respect to one of $z_{i+1},\ldots,z_{n-1},x$ creates at most k^p terms that consist of at most k dummy variables as the factors each, the total running time of computing $A_i^p(z_i,\ldots,z_{n-1};x)$ as the Taylor polynomial is at most $O((k+1)!(p+1)^k k^{p+1})$. Since, by assumption, the time complexity of computing p-th derivative of $F_{ij}(x)$ is $O(\exp(p))$, the running time of computing $A_n^p(x)$ is $O((k+1)!n(p+1)^k k^{p+1}\exp(p))$.

Now we concentrate on proving that $p = O(k^2 x\sigma + \ln n + \ln 1/\epsilon)$ is sufficient for satisfying $|A_n^p(x) - F_{\mathrm{MAX}}(x)| \leq \epsilon$. For each edge (v_i,v_j), we consider a random variable $X'_{ij} = X_{ij}/(k\sigma)$. Let $F'_{\mathrm{MAX}}(x)$ be the distribution function of the longest path length which is defined for the case where edge lengths are given as X'_{ij} instead of X_{ij}. Since $F_{\mathrm{MAX}}(x) = F'_{\mathrm{MAX}}(k\sigma x)$, we consider the normalized edge length X'_{ij} and the normalized distribution function $F'_{\mathrm{MAX}}(x)$ instead of X_{ij} and $F_{\mathrm{MAX}}(x)$ in the following. For the simplicity, we give the proof for the case $\sigma = 1/k$. The proof for the general σ can be given by replacing x in the following by $kx\sigma$.

Let ϵ_i be the difference between $A_i^p(z_i,\ldots,z_n;x)$ and $Q_i(z_i,\ldots,z_{n-1};x)$. We first bound the error that is created when the Taylor polynomial of $Q_2(z_2,\ldots,z_{n-1};x)$ is computed. By generalizing the evaluation of the Taylor polynomials in [13], it is easy to show that $|\epsilon_2| \leq \dfrac{M(\sum_{v_j\in V_2} z_j)^{p+1}}{(p+1)!}$ where M is an upper bound on the $(p+1)$-th derivatives of $Q_2(z_2,\ldots,z_{n-1};x) = \prod_{2\leq j\leq n} F_{1j}(x-z_j)$ at the origin where $z_2 = z_3 = \cdots = z_{n-1} = x = 0$. By the above normalization, it is easy to show that M is less than 1. Since the dummy variables $z_1,\ldots,z_{n-1}$ of integrals are non-negative and less than x, by Proposition 1, we have

$$|\epsilon_2| \leq \frac{(xk)^{p+1}}{(p+1)!}. \tag{12}$$

Let us bound the error that is created when $A_{i+1}^p(z_{i+1},\ldots,z_{n-1};x)$ is computed as the Taylor polynomial of the convolution of $A_i^p(z_i,\ldots,z_{n-1};x)$ and $\prod_{i+1\leq j\leq n-1} F_{ij}(z_i - z_j)$ with respect to z_i. By definition, we have

$$\epsilon_{i+1} = A_{i+1}^p(z_{i+1},\ldots,z_{n-1};x) - \int_{\mathrm{R}} (A_i^p(z_i,\ldots,z_{n-1};x) + \epsilon_i) G_i(z_i,\ldots,z_{n-1})dz_i. \tag{13}$$

Since $A_{i+1}^p(z_{i+1},\ldots,z_{n-1};x)$ is the Taylor polynomial of $\int_{\mathrm{R}} A_i^p(z_i,\ldots,z_{n-1};x)G_i(z_i,\ldots,z_{n-1})dz_i$, we have

$$|\epsilon_{i+1}| \leq \frac{(kx)^{p+1}}{(p+1)!} + \int_{\mathbf{R}} |\epsilon_i| \frac{d}{dz_i} \prod_{i+1\leq j\leq n-1} F_{ij}(z_i - z_j)dz_i = \frac{(kx)^{p+1}}{(p+1)!} + |\epsilon_i|. \tag{14}$$

This leads to $|\epsilon_n| \leq (n-1)\dfrac{(kx)^{p+1}}{(p+1)!}$ by (12) and (14) for $i = 2,3,\ldots,n$.

If $kx \leq 1$, the error ϵ_n converges to 0 very quickly. If $kx > 1$, $p = O(\ln(n-1) + kx + \ln 1/\epsilon)$ is sufficient to have $|\epsilon_n| = |A_n^p(x) - Q_n(x)|$ less than ϵ. □

We immediately obtain the following corollary.

Corollary 2. *If x, ϵ, k and σ are constants, the proposed algorithm computes the value of $F_{\mathrm{MAX}}(x)$ within error ϵ in a polynomial time of the size of G.*

6 Conclusion

In this paper, we have investigated the longest path problem for DAGs G, where the edge lengths are given as mutually independent random variables. We have shown that the distribution function $F_{\mathrm{MAX}}(x)$ of the longest path length is given as a form of repeated integral that involves $n-1$ integrals, where n is the order of G. We can thus approximately evaluate $F_{\mathrm{MAX}}(x)$ for any fixed x by applying numerical methods to the form, at the expense of accuracy and time.

We however suggest that an important application of the repeated integral is in symbolic computation of $F_{\mathrm{MAX}}(x)$. Because only $n-1$ integrals are involved with, it may give us a chance to symbolically compute it in polynomial in n, although there are $\Omega(2^n)$ paths in G. In fact, we have shown that a representation of $F_{\mathrm{MAX}}(x)$ by a combination of primitive functions is obtained in $O((k+1)!n^{k+2}(2m+1)^{k+1})$ time, provided that the edge lengths obey the standard exponential distribution, where k is the maximum anti-chain cardinality of the incidence graph $L(G)$. Recall that the problem is NP-hard even for series-parallel graphs when the edge lengths obey discrete distributions. A natural open question is thus to find another class of distribution functions for which there is a polynomial algorithm to symbolically execute the repeated form.

Since naive numerical methods to approximate the repeated integral need sufficiently long time and do not have performance guarantees, by making use of the Taylor polynomials, we have proposed an approximation algorithm to compute $F_{\mathrm{MAX}}(x)$ with error smaller than ϵ for any given x and ϵ, assuming that the distributions that the edge lengths obey satisfy the following three natural conditions; 1) $F_e(x) = 0$ for $x \leq 0$, 2) the Taylor series of $F_e(x)$ converges to $F_e(x)$ itself, and 3) for any non-negative integer p, there is a constant σ satisfying $\sigma^p \geq \left|\left(\frac{d}{dx}\right)^p F_e(x)\right|$. It takes a polynomial time in n, when each of k, x, ϵ and σ can be regarded as constants.

References

1. E. Ando, T. Nakata, M. Yamashita, Approximating the longest path length of a stochastic DAG by a normal distribution in linear time, *Journal of Discrete Algorithms* (2009), doi:10.1016/j.jda.2009.01.001
2. E. Ando, H. Ono, K. Sadakane, M. Yamashita, A Generic Algorithm for Approximately Solving Stochastic Graph Optimization Problems, submitted for publication.
3. E. Ando, M. Yamashita, T. Nakata, Y. Matsunaga, The Statistical Longest Path Problem and Its Application to Delay Analysis of Logical Circuits, proc.TAU, 2002, pp. 134–139.
4. M. O. Ball, C. J. Colbourn, and J. S. Proban, Network Reliability, Handbooks in Operations Research and Management Science, Vol 7: Network Models, M. O. Ball, T. L. Magnanti, C. L. Monma, and G. L. Nemhauser (eds.), Elsevier Science B. V. (1995) 673–762.
5. M. Berkelaar, Statistical delay calculation, a linear time method. *Proceedings of the International Workshop on Timing Analysis (TAU'97)*, 1997, pp. 15–24,
6. C. E. Clark, The PERT model for the distribution of an activity time, *Operations Research* 10 (1962) 405–406.
7. M. Hashimoto and H. Onodera, A performance optimization method by gate sizing using statistical static timing analysis. *IEICE Trans. Fundamentals*, E83-A, 12, 2000, pp. 2558–2568.
8. J. N. Hagstrom, Computational Complexity of PERT Problems, *NETWORKS*, Vol. 18, 1988, pp. 139–147.
9. J. E. Kelley, Jr., Critical-path planning and scheduling: Mathematical basis, *Operations Research* 10 (1962) 912–915.
10. V. G. Kulkarni and V. G. Adlakha, Markov and Markov-Regenerative PERT Networks, *Operations Research*, Vol. 34, 1986, pp. 769–781.
11. J. J. Martin, Distribution of the time through a directed, acyclic network, *Operations Research* Vol. 13, 1965, pp. 46–66.
12. E. Nikolova, Stochastic Shortest Paths Via Quasi-convex Maximization, *Proceedings of 14th Annual European Symposium on Algorithms (ESA2006)*, LNCS 4168, 2006, pp.552–563.
13. G. B. Thomas,Jr., Thomas' Calculus International Edition, Pearson Education, 2005, pp.965–1066.