

Computing Diverse Shortest Paths Efficiently: A Theoretical and Experimental Study

Tesshu Hanaka¹, Yasuaki Kobayashi², Kazuhiro Kurita³, See Woo Lee², Yota Otachi¹

¹ Nagoya University, Nagoya, Japan Email: {hanaka, otachi}@nagoya-u.jp

² Kyoto University, Kyoto, Japan Email: {kobayashi, slee06}@iip.ist.i.kyoto-u.ac.jp

³ National Institute of Informatics, Tokyo, Japan Email: kurita@nii.ac.jp

Abstract

Finding diverse solutions in combinatorial problems recently has received considerable attention (Baste et al. 2020; Fomin et al. 2020; Hanaka et al. 2021). In this paper we study the following type of problems: given an integer k , the problem asks for k solutions such that the sum of pairwise (weighted) Hamming distances between these solutions is maximized. Such solutions are called *diverse solutions*. We present a polynomial-time algorithm for finding diverse shortest st -paths in weighted directed graphs. Moreover, we study the diverse version of other classical combinatorial problems such as diverse weighted matroid bases, diverse weighted arborescences, and diverse bipartite matchings. We show that these problems can be solved in polynomial time as well. To evaluate the practical performance of our algorithm for finding diverse shortest st -paths, we conduct a computational experiment with synthetic and real-world instances. The experiment shows that our algorithm successfully computes diverse solutions within reasonable computational time.

Introduction

When solving a real-world problem, we usually interpret it as an instance of a mathematical optimization problem. We then seek a *single* best solution that optimizes a given objective function while satisfying the set of constraints. Unfortunately, the *single* solution obtained this way may be inadequate for practical usage, since real-life intricacies (side constraints) are oversimplified or even ignored in the optimization problem for the sake of computational feasibility.

To illustrate, consider the task of finding a fastest route between location s and t , which may seem like an obvious instance of the shortest st -path problem in graphs. However, numerous important real-life considerations – such as traffic congestion, constructions, or availabilities of petrol stations in the route, to name a few – cannot be factored into the graph data structure on which Dijkstra’s or Bellman-Ford’s algorithms are run. Furthermore, real-life constraints are often “...aesthetic, political, or environmental” in nature (Baste et al. 2020); such qualitative characteristics are difficult or downright impossible to express mathematically.

In light of this issue, the *multiplicity* of the solutions has gained attention. One of the best known approaches

to achieve the solution multiplicity is the *k-best enumeration* (Hamacher and Queyranne 1985; Eppstein 2016) (also known as *top-k enumeration*). The *k*-best enumeration version of an algorithm finds k feasible solutions $\mathcal{S} = \{S_1, S_2, \dots, S_k\}$ such that no feasible solution outside \mathcal{S} is strictly better than that inside \mathcal{S} . This approach allows us to choose the best-suited solution amongst k *sufficiently good* feasible solutions, in a given real-life situation. There are many *k*-best enumeration algorithms for various optimization problems. See (Hamacher and Queyranne 1985; Eppstein 2016) for surveys.

One potential drawback of the *k*-best enumeration approach is the lack of *diversity* of solutions. Most of *k*-best enumeration algorithms, such as Lawler’s framework (Lawler 1972), recursively generate solutions from a single optimal solution $X = \{x_1, x_2, \dots, x_t\}$ by finding a solution including $\{x_1, \dots, x_{i-1}\}$ and excluding x_i for each $1 \leq i \leq t$. This implies that solutions tend to be similar to each other in nature (Akgün, Erkut, and Batta 2000), which may potentially defeat the purpose of finding multiple solutions.

Motivated by this, (explicitly) optimizing diversity of solutions has received considerable attention in the literature. There are many results for finding diverse solutions for Constrained Satisfaction Problems or Mixed Integer Programming (Danna and Woodruff 2009; Hebrard et al. 2005; Nadel 2011; Petit and Trapp 2015, 2019) and diversifying query results (Drosou and Pitoura 2010; Liu et al. 2018; Qin, Yu, and Chang 2012; Vieira et al. 2011). According to Baste et al. (2019), Michael Fellows proposed the *Diverse X Paradigm*, where X is a placeholder for an optimization problem, and to investigate these problems from the theoretical perspective, particularly fixed-parameter tractability. Based on this proposal, they initiated a theoretical study of the Diverse X Paradigm and gave several results through the lens of fixed-parameter tractability. In particular, they showed that several diverse versions of combinatorial problems, such as VERTEX COVER, FEEDBACK VERTEX SET, and d -HITTING SET, are fixed-parameter tractable parameterized by the solution size plus the number of solutions (Baste et al. 2019). Their follow-up research also discussed the fixed-parameter tractability of diverse versions of several combinatorial problems on bounded-treewidth graphs (Baste et al. 2020).

Diversity measures: Before describing our results, we need to define known diversity measures and discuss known results relevant to our results. There are mainly two diversity measures in these theoretical studies. Let U be a finite set. Let S_1, \dots, S_k be (not necessarily disjoint) subsets of U . We define

$$d_{\text{sum}}(S_1, \dots, S_k) = \sum_{1 \leq i < j \leq k} |S_i \triangle S_j|,$$

where \triangle is the symmetric difference of two sets (we may call it the Hamming distance between two sets). We also define

$$d_{\text{min}}(S_1, \dots, S_k) = \min_{1 \leq i < j \leq k} |S_i \triangle S_j|.$$

Related work: There are several recent work closely related to our results. Fomin et al. (2020) showed that the problem of finding two maximum matchings M_1, M_2 maximizing its symmetric difference $|M_1 \triangle M_2|$ in bipartite graphs can be solved in polynomial time, whereas it is NP-hard on general graphs (Fomin et al. 2020; Holyer 1981; Hanaka et al. 2021). To overcome this intractability, they devised an FPT-algorithm with respect to parameter $|M_1 \triangle M_2|$ for general graphs, that is, it runs in time $f(d)n^{O(1)}$, where n is the number of vertices of an input graph and $d = |M_1 \triangle M_2|$. Moreover, Fomin et al. (2021) studied the problems of finding k solutions S_1, S_2, \dots, S_k for several combinatorial problems related to (linear) matroids and matchings such that the weighted version of $d_{\text{min}}(S_1, S_2, \dots, S_k)$ is at least d . They showed that these problems are fixed-parameter tractable parameterized by $k + d$ (i.e., the running time of these algorithms is $f(k, d)n^{O(1)}$, where f is some computable function and n is the input size). On the negative side, they showed that finding k bases of a matroid maximizing the weighted version of d_{min} is NP-hard even on uniform matroids. Contrary to this hardness result, Hanaka et al. (2021) showed that finding k bases of a matroid maximizing d_{sum} is solvable in polynomial time.¹ They also proposed a general framework to obtain diverse solutions with running time exponentially depending on k and the cardinality of solutions.

Finding diverse short(est) paths is a fruitful research area and a considerable amount of effort has been dedicated to it so far (Akgün, Erkut, and Batta 2000; Chondrogiannis et al. 2020; Liu et al. 2018; Voss, Moll, and Kavraci 2015). Akgün, Erkut, and Batta (2000) experimentally compared some existing methods and their method based on the dispersion problem (Kuby 1987). Chondrogiannis et al. (2020) studied the problem of finding a set of k paths P_1, P_2, \dots, P_k between two vertices s and t such that these paths are “dissimilar” to each other and for every path P between s and t with $P \neq P_i$ for any $1 \leq i \leq k$, either there is a path P_i that is similar to P or the length of P is not shorter than that of P_i for any $1 \leq i \leq k$. They showed that this problem is NP-hard and gave an exponential-time algorithm for this problem. By (Liu et al. 2018), a similar problem was discussed

¹In their problem setting, the independence oracle of matroids is given as input and is assumed to be evaluated in polynomial time.

and a heuristic algorithm was given. Voss, Moll, and Kavraci (2015) gave a heuristic algorithm that optimizes some diversity measure of paths based on the Fréchet distance. To the best of our knowledge, no shortest st -path algorithms that run in polynomial time with a theoretical guarantee of diversity are known in the literature.

Our results: In this paper, we expand the tractability border of the diverse version of classical combinatorial optimization problems. Let $\mathcal{S} \subseteq 2^U$ be a set of solutions. The common goal of our problems is to find a set of k solutions $S_1, S_2, \dots, S_k \in \mathcal{S}$ maximizing $d_w(S_1, \dots, S_k)$, where d_w is the weighted version of the diversity measure d_{sum} (see the next section for the definition). We show that if \mathcal{S} consists of either (1) the shortest st -paths in a directed graph, (2) the bases of a matroid, (3) the arborescences of a directed graphs, or (4) the bipartite matchings of size p , then the problem can be solved in polynomial time. The algorithm for (2) is an extension of the algorithm of (Hanaka et al. 2021) to the weighted counterpart and that for (4) is an extension of (Fomin et al. 2020) that allows to find more than two diverse bipartite matchings in polynomial time. Contrary to the tractability of (1), we show that the problem of finding two disjoint “short” paths is NP-hard.

As will be elaborated on Lemma 1, this positive outcome is achieved by understanding our problem as a generalization of packing problems. In the packing problem (with respect to \mathcal{S}), we are asked to find *mutually disjoint* sets S_1, S_2, \dots, S_k in \mathcal{S} . Suppose that each solution in \mathcal{S} has the same cardinality. Hanaka et al. (2021) showed that by appropriately duplicating and weighting each (duplicated) element in U , the problem of maximizing $d_{\text{sum}}(S_1, S_2, \dots, S_k)$ can be reduced to a certain weighted packing problem. Our lemma formalizes their idea and states that finding diverse solutions can be reduced to a certain packing problem as well even if solutions are weighted and have different cardinalities. As an application of this lemma, we show that finding diverse shortest st -paths can be reduced to the minimum cost flow problem, which can be solved in polynomial time.

We conducted a computational experiment to measure the practical performance of our algorithm for the diverse version of the shortest path problem. Our experiment shows that the proposed algorithm runs in reasonable time and computes more diverse shortest paths than those computed by the k -best enumeration algorithm.

Preliminaries

Let $D = (V, E)$ be a (directed) graph. We denote by $V(D)$ and $E(D)$ the sets of vertices and edges of D , respectively.

Let U be a finite set and let $w : U \rightarrow \mathbb{Z}$. Let S_1, \dots, S_k be (not necessarily disjoint) subsets of U . We define

$$d_w(S_1, \dots, S_k) = \sum_{1 \leq i < j \leq k} w(S_i \triangle S_j),$$

where $w(X) = \sum_{x \in X} w(x)$. This notation extends the diversity measure d_{sum} defined in the introduction, where $d_{\text{sum}}(S_1, \dots, S_k) = d_w(S_1, \dots, S_k)$ with $w(x) = 1$ for all $x \in U$. Let k be a positive integer and let $\mathcal{S} \subseteq 2^U$ be

a family of subsets of U . We expand each element e in U into k copies: Let $U^* = \{e_1, \dots, e_k : e \in U\}$. We define a function $f : U^* \rightarrow U$ such that $f(e_i) = e$ for all $e_i \in U^*$. We say that $S^* \subseteq U^*$ is a k -packing of U^* with respect to \mathcal{S} if S^* can be partitioned into S_1, \dots, S_k such that $\{f(e^*) : e^* \in S_i\} \in \mathcal{S}$ for all $1 \leq i \leq k$.

We consider a weight function $w^* : U^* \rightarrow \mathbb{Z}$ such that $w^*(e_i) = w(e) \cdot (k - 2i + 1)$ for $e \in U$ and $1 \leq i \leq k$. Then, we have the following lemma.

Lemma 1. *There are S_1, S_2, \dots, S_k in \mathcal{S} with $d_w(S_1, \dots, S_k) \geq t$ if and only if there is a k -packing S^* of U^* with respect to \mathcal{S} such that $w^*(S^*) \geq t$.*

Proof. Suppose that there are S_1, S_2, \dots, S_k in \mathcal{S} with $d_w(S_1, \dots, S_k) \geq t$. For each $e \in U$, we denote by $m(e)$ the number of occurrences of e in the collection $\{S_1, \dots, S_k\}$. Then, we have

$$\begin{aligned} d_w(S_1, \dots, S_k) &= \sum_{e \in U} (w(e) \cdot m(e) \cdot (k - m(e))) \\ &= \sum_{e \in U} (w(e) \cdot \sum_{1 \leq i \leq m(e)} (k - 2i + 1)) \\ &= \sum_{e \in U} \sum_{1 \leq i \leq m(e)} w^*(e_i) \\ &= w^*(S^*). \end{aligned}$$

Conversely, assume that there is a k -packing S^* of U^* with respect to \mathcal{S} with $w^*(S^*) \geq t$. We assume moreover that, for each $e \in U$, S^* contains consecutive elements e_1, \dots, e_m for some m , that is, $\{e_1, \dots, e_m\} \subseteq S^*$ and $\{e_{m+1}, \dots, e_k\} \cap S^* = \emptyset$. This assumption is legitimate as $w(e_i) > w(e_j)$ for $1 \leq i < j \leq k$. We denote the multiplicity m of e by $m(e)$. Let $\{S_1^*, \dots, S_k^*\}$ be a partition of S^* such that $S_i^* = \{f(e^*) : e^* \in S_i^*\} \in \mathcal{S}$ for $1 \leq i \leq k$. For each $e \in U$, the contribution of e to $w^*(S^*)$ is indeed $w(e) \cdot \sum_{1 \leq i \leq m(e)} (k - 2i + 1)$. Hence, we have

$$\begin{aligned} w^*(S^*) &= \sum_{e \in U} \sum_{1 \leq i \leq m(e)} w^*(e_i) \\ &= d_w(S_1, \dots, S_k) \end{aligned}$$

as in the “only-if” direction. \square

Let us note that a similar idea is used in (Hanaka et al. 2021), however, their proof relies on the fact that solutions have the same cardinality and the objective function is to maximize the sum of *unweighted* Hamming distances.

For several cases, we reduce the problem of finding a largest weight k -packing to the minimum cost flow problem. In the minimum cost flow problem, given a directed graph $D = (V, E)$, $s, t \in V$, a cost function $w : E \rightarrow \mathbb{Q}$, a capacity function $c : E \rightarrow \mathbb{N}_{\geq 0}$, and an integer k , the goal is to compute a flow $f : E \rightarrow \mathbb{Q}$ with flow value k such that (1) $f(e) \leq c(e)$ for all $e \in E$, (2) $\sum_{(v,w) \in E} f((v,w)) = \sum_{(w,v) \in E} f((w,v))$ for all $v \in V \setminus \{s, t\}$, and (3) $\sum_{e \in E} w(e)f(e)$ is minimized, where the flow value is defined as $\sum_{(s,v) \in E} f((s,v))$. This problem can be solved in polynomial time.

Theorem 1 (Orlin 1993). *The minimum cost flow problem can be solved in time $O(|E|^2 \log^2 |V|)$.*

Diverse Shortest st -Paths

As an application of Lemma 1, we consider the following problem.

Definition 1. *Let $D = (V, E)$ be a directed graph with specified vertices $s, t \in V$. Let $\ell : E \rightarrow \mathbb{N}_{>0}$ be a length function on edges. Let \mathcal{P} be the set of all shortest paths from s to t in (D, ℓ) . Given an integer k and a weight function $w : E \rightarrow \mathbb{Z}$, DIVERSE SHORTEST st -PATHS asks for k paths $P_1, \dots, P_k \in \mathcal{P}$ such that $d_w(E(P_1), \dots, E(P_k))$ is maximized.*

Theorem 2. *DIVERSE SHORTEST st -PATHS can be solved in $O(k^2 |E|^2 \log^2 |V|)$ time.*

We first compute the shortest distance label $\text{dist} : V \rightarrow \mathbb{N}_{\geq 0}$ from s in time $O(|E| + |V| \log |V|)$ by Dijkstra’s single source shortest path algorithm. For each edge $e = (u, v) \in E$ with $\text{dist}(u) \neq \text{dist}(v) - \ell(e)$, we remove it from D . We also remove vertices that are not reachable from s and not reachable to t in the removed graph. This does not change any optimal solution since every path in \mathcal{P} does not include such vertices and edges. Then, the obtained graph, denoted by $D' = (V', E')$, has no directed cycles.

Observation 1. *Every path in \mathcal{P} is a path from s to t in D' . Moreover, every path from s to t in D' belongs to \mathcal{P} .*

From this directed acyclic graph D' , we construct a weighted directed multigraph D^* by replacing each edge $e = (u, v)$ with k copies e_1, \dots, e_k and setting $w^*(e_i)$ to $w(e) \cdot (k - 2i + 1)$ for $e \in E'$ and $1 \leq i \leq k$. By Lemma 1, it is sufficient to find a maximum weight k -packing of $E(D^*)$ with respect to \mathcal{P} .

Lemma 2. *Let D^* and $w : E(D^*) \rightarrow \mathbb{N}_{\geq 0}$ be as above. Then, there is a polynomial-time algorithm that finds a maximum weight k -packing of $E(D^*)$ with respect to \mathcal{P} .*

Proof. We reduce the k -packing problem to the minimum-cost flow problem, which can be solved in polynomial time (by Theorem 1). The source and the sink vertices of D^* are defined as s and t , respectively. For each $e \in E(D^*)$, we assign the capacity value of 1 (to prevent edge sharing) and the cost value of $-w^*(e)$. The flow requirement is set to k . Then we can find a flow $f : E^* \rightarrow \mathbb{R}_{\geq 0}$ maximizing $\sum_{e \in E^*} f(e) \cdot w^*(e)$ in polynomial time. Moreover, it is well known that f is integral, that is, $f(e) \in \mathbb{N}_{\geq 0}$ for every $e \in E^*$, as the capacity is integral. Since all of the edges in E^* has a capacity of 1, f can be decomposed into k edge-disjoint st -paths P_1, P_2, \dots, P_k , which implies that the maximum weight k -packing with respect to \mathcal{P} can be found in polynomial time as well. \square

Since D^* has at most $|V|$ vertices and at most $k|E|$ edges, Theorem 2 follows.

We note that our algorithm can be used for finding diverse shortest st -paths in *undirected* graphs. For undirected graphs, we first compute the shortest distance label dist . Then, for each (undirected) edge $e = \{u, v\}$, we orient e as

(u, v) if $\text{dist}(u) = \text{dist}(v) - \ell(e)$ and remove e otherwise. Then, the obtained graph D' is indeed a directed acyclic graph and satisfies Observation 1. Hence we can apply the same algorithm to D' .

In practical situations, it suffices to find diverse “nearly shortest” paths. More specifically, given the same instance of DIVERSE SHORTEST st -PATHS and a positive threshold θ , we are asked to find a set of k paths P_1, P_2, \dots, P_k from s to t in (D, ℓ) such that the length of P_i is at most θ and $d_w(E(P_1), E(P_2), \dots, E(P_k))$ is maximized. Here, the length of a path P_i is defined as $\sum_{e \in E(P_i)} \ell(e)$. Let us note that paths are simple, that is, every vertex appears at most once in the path. The problem, which we call DIVERSE SHORT st -PATHS, indeed generalizes DIVERSE SHORTEST st -PATHS: If θ is the shortest path distance from s to t in D , the problem corresponds to DIVERSE SHORTEST st -PATHS. Unfortunately, this generalization is intractable.

Theorem 3. DIVERSE SHORT st -PATHS is NP-hard even if $k = 2$ and $w(e) = \ell(e) = 1$ for all $e \in E$.

Proof. We perform a polynomial-time reduction from the st -Hamiltonian path problem, which is well known to be NP-complete (Karp 1972). In this problem, we are given a directed graph $D = (V, E)$ and $s, t \in V$ with $s \neq t$ and asked to determine whether D has a Hamiltonian path from s to t . From D , we construct a directed graph D' by adding a path P from s to t of length $|V|$. Now, we claim that D has a Hamiltonian path from s to t if and only if D' has two st -paths P_1 and P_2 of length at most $\theta = |V|$ with $|E(P_1) \Delta E(P_2)| \geq 2|V| - 1$.

Suppose that D has a Hamiltonian path P' from s to t . Then, P' is disjoint from P and hence $|E(P') \Delta E(P)| = |E(P')| + |E(P)| = 2|V| - 1$.

Conversely, assume that D' has two st -paths P_1 and P_2 of length at most $|V|$ with $|E(P_1) \Delta E(P_2)| \geq 2|V| - 1$. Since every st -path P' distinct from P has length at most $|V| - 1$, one of the two paths, say P_1 , must be P . Moreover, such a path P' is disjoint from P as otherwise P' must contain s or t as an internal vertex. This implies that P_2 is also an st -path in D . As $|E(P_1) \Delta E(P_2)| \geq 2|V| - 1$, the length of P_2 is at least $|V| - 1$, implying that P_2 is a Hamiltonian path from s to t in D . \square

Diverse Matroid Bases and Arborescences

In this section, we discuss some tractable problems related to matroids. Let E be a finite set. A *matroid* \mathcal{M} on E is a pair (E, \mathcal{I}) with $\mathcal{I} \subseteq 2^E$ satisfying the following axioms: (1) $\emptyset \in \mathcal{I}$; (2) For $X \in \mathcal{I}$, every $Y \subseteq X$ is contained in \mathcal{I} ; (3) For $X, Y \in \mathcal{I}$ with $|X| < |Y|$, there is $e \in Y$ such that $X \cup \{e\} \in \mathcal{I}$. Every set in \mathcal{I} is called an *independent set* of a matroid $\mathcal{M} = (E, \mathcal{I})$. A *base* of \mathcal{M} is an inclusion-wise maximal independent set of \mathcal{M} . The first part of this section is devoted to develop a polynomial-time algorithm for the following problem.

Definition 2. Given a matroid $\mathcal{M} = (E, \mathcal{I})$ with a weight function $w : E \rightarrow \mathbb{Z}$ and an integer k , WEIGHTED DIVERSE MATROID BASES asks for k bases B_1, \dots, B_k of \mathcal{M} such that $d_w(B_1, \dots, B_k)$ is maximized.

In (Hanaka et al. 2021), they consider a special case of WEIGHTED DIVERSE MATROID BASES where each element in the ground set E has a unit weight and give a polynomial-time algorithm for it, assuming that the independence oracle \mathcal{I} can be evaluated in polynomial time. This result is obtained by reducing the problem to that of finding disjoint bases of a matroid, which can be solved in polynomial time.

Theorem 4 ((Edmonds 1968; Nash-Williams 1967)). Let $\mathcal{M} = (E, \mathcal{I})$ be a matroid and let $w : E \rightarrow \mathbb{Z}$. Suppose that the membership of \mathcal{I} can be checked in polynomial time. Then, the problem of deciding whether there is a set of mutually disjoint k bases B_1, \dots, B_k of \mathcal{M} can be solved in polynomial time. Moreover, if the answer is affirmative, we can find such bases that maximize the total weight (i.e., $\sum_{1 \leq i \leq k} \sum_{e \in B_k} w(e)$) in polynomial time.

By applying Lemma 1, we have a polynomial-time algorithm for WEIGHTED DIVERSE MATROID BASES as well.

Theorem 5. WEIGHTED DIVERSE MATROID BASES can be solved in polynomial time.

Proof. The proof is almost analogous to that in (Hanaka et al. 2021). Let $\mathcal{M} = (E, \mathcal{I})$ be a matroid and let $e \in E$. Define $\mathcal{J} = \mathcal{I} \cup \{(F \setminus \{e\}) \cup \{e'\} : F \in \mathcal{I} \wedge e \in F\}$. Then, $(E \cup \{e'\}, \mathcal{J})$ is also a matroid (Hanaka et al. 2021; Nagamochi et al. 1997). We define k copies e_1, e_2, \dots, e_k for each $e \in E$ and $E^* = \{e_1, \dots, e_k : e \in E\}$. Then, the pair $\mathcal{M}^* = (E^*, \mathcal{I}^*)$ is a matroid if \mathcal{I}^* consists of all sets $F \subseteq 2^{E^*}$ such that F contains at most one copy of e_1, \dots, e_k for each $e \in E$ and $\bigcup_{e_i \in F} f(e_i) \in \mathcal{I}$, where $f(e_i) = e$ for $e \in E$ and $1 \leq i \leq k$.

To find a set of k bases B_1, \dots, B_k of \mathcal{M} maximizing $d_w(B_1, \dots, B_k)$, by Lemma 1, it suffices to find a maximum weight k -packing with respect to the base family of \mathcal{M}^* under a weight function w^* with $w^*(e_i) = w(e) \cdot (k - 2i + 1)$ for $e \in E$ and $1 \leq i \leq k$, which can be solved in polynomial time by Theorem 4. \square

Theorem 5 allows us to find diverse spanning trees in undirected graphs as the set of edges in a spanning tree forms a base of a graphic matroid (Oxley 2006). In this section, we develop a polynomial-time algorithm for a directed version of this problem. Let $D = (V, E)$ be a directed graph and let $r \in V$. We say that a subgraph T of D is an *arborescence* (with root r) if for every $v \in V$, there is exactly one directed path from r to v in T . In other words, an arborescence is a spanning subgraph of D in which each vertex except r has in-degree one and its underlying undirected graph is a tree. The problem we consider here is defined as follows.

Definition 3. Given an edge-weighted directed graph $D = (V, E)$ with weight function $w : E \rightarrow \mathbb{Z}$, $r \in V$, and an integer k , WEIGHTED DIVERSE ARBORESCENCES asks for k arborescences T_1, \dots, T_k of D with root r such that $d_w(E(T_1), \dots, E(T_k))$ is maximized.

Note that we cannot apply Theorem 5 to this problem as we do not know whether the set of arborescences can be defined as the family of bases of a matroid. However, Edmonds

(1975) gave a polynomial-time algorithm for finding k arc-disjoint arborescences with the maximum total weight.

Theorem 6. WEIGHTED DIVERSE ARBORESCENCES can be solved in polynomial time.

The proof of Theorem 6 is almost analogous to that in Theorem 5. We define a directed graph D^* with vertex set V from $D = (V, E)$ such that for $e = (u, v) \in E$, we add k parallel edges e_1, \dots, e_k directed from u to v to D^* . Then, we set $w^*(e_i) = w(e) \cdot (k - 2i + 1)$ for $e \in E$ and $1 \leq i \leq k$. By Lemma 1, it is sufficient to find a maximum weight k -packing of $E(D^*)$ with respect to the family of arborescences of D^* , which can be found in polynomial time by the following result.

Theorem 7 (Edmonds 1975). Given an edge-weighted directed multigraph $D = (V, E)$ with weight function $w : E \rightarrow \mathbb{Z}$, $r \in V$, and an integer k , the problem of finding k edge-disjoint arborescences T_1, \dots, T_k with root r maximizing $w(\bigcup_{1 \leq i \leq k} E(T_i))$ is solved in polynomial time.

Diverse Bipartite Matchings

A matching of a graph $G = (V, E)$ is a set $M \subseteq E$ of edges such that no two edges share their end vertices. In this section, we consider the following problem.

Definition 4. Let $G = (A \cup B, E, w)$ be an edge-weighted bipartite graph with $w : E \rightarrow \mathbb{Z}$, where A and B are the color classes of G . Let k, p be positive integers. We denote by \mathcal{M} the collection of all matchings of G with cardinality exactly p . DIVERSE BIPARTITE MATCHINGS asks for k matchings $M_1, \dots, M_k \in \mathcal{M}$ maximizing $d_w(M_1, \dots, M_k)$.

As mentioned in the introduction, the problem of finding two edge-disjoint perfect matchings in a general graph is known to be NP-complete (Holyer 1981; Fomin et al. 2020; Hanaka et al. 2021), while the problem of finding two maximum matchings M_1 and M_2 maximizing $|M_1 \triangle M_2|$ can be solved in polynomial time on bipartite graphs (Fomin et al. 2020). In this section, we design a polynomial-time algorithm for DIVERSE BIPARTITE MATCHINGS by applying Lemma 1, which generalizes the aforementioned polynomial-time algorithm of (Fomin et al. 2020).

We construct a bipartite multigraph G^* from G by replacing each edge $e = \{a, b\} \in E$ with k parallel edges e_1, \dots, e_k . We set $w^*(e_i) = w(e) \cdot (k - 2i + 1)$ for each $e \in E$ and $1 \leq i \leq k$. By Lemma 1, it suffices to show that there is a polynomial-time algorithm that computes a maximum weight k -packing of E^* with respect to \mathcal{M} , where \mathcal{M} is the collection of matchings M of G with cardinality exactly p . This problem can be solved in polynomial time by reducing to the minimum cost flow problem as follows.

Let $G^* = (A \cup B, E^*)$ be bipartite and let $w^* : E^* \rightarrow \mathbb{Z}$. We construct a directed acyclic graph from G^* by orienting each edge $\{a, b\}$ of G^* from a to b , where $a \in A$ and $b \in B$. Each arc (a, b) in G^* has capacity one and cost $-w^*(\{a, b\})$. We also add a source vertex s , a sink vertex t , and then arcs (s, a) for $a \in A$ and (b, t) for $b \in B$. The arcs incident to the source or the sink have capacity k and have cost zero. Now, we set the flow requirement from s to t to kp . Thanks to the

integral theorem of the minimum cost flow problem, by Theorem 1, we can find in polynomial time a maximum weight subgraph H^* of G^* such that H^* has exactly kp edges and each vertex has degree at most k . From this subgraph H^* , we need to construct a maximum weight k -packing of E^* with respect to \mathcal{M} . The following lemma ensures that it is always possible.

Lemma 3. Let H^* be a bipartite graph with kp edges. Suppose the maximum degree of a vertex in H^* is at most k . Then, the edges of H^* can be partitioned into k matchings of cardinality exactly p . Moreover, such a partition can be computed in polynomial time from H^* .

Proof. Every bipartite graph of maximum degree at most k has a proper edge-coloring with k colors, and such an edge coloring can be computed in polynomial time (Cole and Hopcroft 1982; Gabow and Kariv 1982). We can assume that each color is used at least once by recoloring an edge whose color is used at least twice. Then, the edges of H^* can be decomposed into k non-empty matchings M_1, \dots, M_k . If $|M_1| = \dots = |M_k| = p$, we are done. Suppose that there is a pair of matchings M_i and M_j with $|M_i| > p$ and $|M_j| < p$. Then $M_i \triangle M_j$ is a subgraph of H^* of maximum degree at most two. As $|M_i| > |M_j|$, the subgraph contains an augmenting path $P = (v_1, \dots, v_\ell)$ with $\{v_\ell, v_{\ell+1}\} \in M_i$ for odd ℓ and $\{v_\ell, v_{\ell+1}\} \in M_j$ for even ℓ . Let $M'_i = (M_i \setminus E(P)) \cup (E(P) \cap M_j)$ and $M'_j = (M_j \setminus E(P)) \cup (E(P) \cap M_i)$. Then, we have matchings M'_i and M'_j such that $|M'_i| = |M_i| - 1$, $|M'_j| = |M_j| + 1$, and $M_i \cup M_j = M'_i \cup M'_j$. By repeating this argument, we have a desired set of matchings. \square

Therefore, $E(H^*) = M_1 \cup \dots \cup M_k$ is a maximum weight k -packing of $E(H^*)$ with respect to \mathcal{M} , which implies the following theorem.

Theorem 8. DIVERSE BIPARTITE MAXIMUM MATCHING can be solved in polynomial time.

Computational Experiments

In order to assess the practical performance of our algorithm for DIVERSE SHORTEST st -PATHS, a computational experiment was conducted on a computer equipped with Intel(R) Xeon(R) Gold 5122 processor (3.60GHz) and 92GB RAM. Our diverse shortest st -paths algorithm² was implemented using Java with JGraphT library (Michail et al. 2020)³. We compared our algorithm with Eppstein’s k -shortest paths algorithm (1998), which is included in JGraphT by default.

Datasets: We used one type of synthetic graphs (Grid) and two types of real-world graphs (SNAP and DIMACS).

- **Grid:** We generated $p \times p$ -grid graphs with vertex set $V = \{1, 2, \dots, p\} \times \{1, 2, \dots, p\}$ and edge set $E = \{(i, j), (i', j')\} : |i - i'| + |j - j'| = 1\}$ for $p \in \{40, 50, \dots, 140\}$. We set s and t to the “top-left” corner vertex $(1, 1)$ and the “bottom-right” corner vertex (p, p) ,

²The code is available at <https://github.com/Dotolation/diverse-graph-algo>.

³<https://jgrapht.org/>

respectively. Each edge has a unit weight and hence every edge belongs to a shortest st -path.

- **SNAP:** We selected three directed unweighted graphs from Stanford Large Network Dataset Collection⁴: wiki-Vote ($|V| = 7,115$, $|E| = 103,689$), soc-Slashdot0922 ($|V| = 82,168$, $|E| = 870,161$), and ego-Twitter ($|V| = 81306$, $|E| = 1,768,149$). These graphs are directed and unweighted. As the diameters (i.e., the maximum distance between any pair of vertices) of these graphs are relatively small, the shortest path distance between s and t is indeed small (see **Results**).
- **DIMACS:** We selected two directed edge-weighted graphs from the 9th DIMACS Implementation Challenge – Shortest Paths⁵: NY ($|V| = 264,346$, $|E| = 733,846$) and FLA ($|V| = 1,070,376$, $|E| = 2,712,798$). Each edge weight of DIMACS graphs were rounded to the nearest 100 (any value below 100 was set to 100) for rudimentary data smoothing, which may increase the number of shortest st -paths.

Method: In principle, the source-sink vertex pair used in each test instance was selected randomly (seed = 2021). However, given an st -pair, if the number of the shortest st paths were less than $3k$, the pair was excluded from our experiment, since the point of our discussion is to evaluate the ability to pick k solutions in a diverse manner from the much larger solution set. Likewise, we also excluded st pairs whose average shortest path length (unweighted) falls short of 3 as it is difficult to assess the diversity of such short paths. Using the randomly selected n vertex pairs with $n = 400$ for each instance, our algorithm and the k -best algorithm were executed. We evaluate *processing time* (in milliseconds) and the *diversity measure* (the pair-wise Hamming distance of k shortest paths) for these algorithms.

A different method is used when testing Grid graphs. Each Grid graph created with $p \in \{40, 50, \dots, 140\}$ was tested once using the source vertex $(1, 1)$ and the sink vertex (p, p) . The processing time and diversity were gauged at three distinct values of k ($k \in \{10, 50, 100\}$).

Results: For each real-world graph, $n = 400$ random st -pairs were sampled to test both algorithms; the number k of solutions was fixed to 10.

Table 1: Average st -path length and count ($n = 400$, $k = 10$).

Name	Length	Count
SNAP (Wiki-Vote)	4.28	58.85
SNAP (soc-Slashdot0922)	4.59	100.14
SNAP (ego-Twitter)	5.40	136.56
DIMACS (NY)	523.41	1307.32
DIMACS (FLA)	1496.00	2205.34

The SNAP dataset consists of social networks characterized by high vertex-to-edge ratios and low diameters. Unsur-

⁴<http://snap.stanford.edu/index.html>

⁵<https://www.diag.uniroma1.it/challenge9/>

Table 2: Average processing time (ms) ($n = 400$, $k = 10$).

Name	Ours	k -Best
SNAP (Wiki-Vote)	20.65	17.24
SNAP (soc-Slashdot0922)	603.95	583.30
SNAP (ego-Twitter)	1027.89	1021.89
DIMACS (NY)	1244.96	1223.76
DIMACS (FLA)	6319.66	6287.20

prisingly, their average (unweighted) shortest st -path length was significantly low as Table 1 demonstrates. Note that our experiment excluded the st -pairs whose shortest path distance is below 3. This is contrasted with DIMACS graphs having very high average path length, which is at least 100 times greater than those of SNAP. DIMACS graphs also have significantly higher shortest path counts (i.e., the number of shortest st -paths) as well.

As stated on Table 2, the k -best enumeration approach was marginally superior to our diverse approach in terms of the processing time. However, the difference might be small enough to be ignored in practical situations, and this observation held true for both SNAP and DIMACS datasets. Let us note that the theoretical running time bounds are quite different: our algorithm runs in time $O(k^2|E|^2 \log^2 |V|)$ (Theorem 2), while the k -shortest paths algorithm runs in time $O(|E| + |V|(k + \log |V|))$ (Eppstein 1998). Regardless the difference between the computational time, both algorithms seems to run reasonably fast on moderately large real-life graphs, such as DIMACS (FLA) that has over a million vertices and two million edges.

Table 3: Average diversity of solutions ($n = 400$, $k = 10$).

Name	Ours	k -Best
Wiki-Vote	3.198×10^2	2.891×10^2
soc-Slashdot0922	3.133×10^2	2.791×10^2
ego-Twitter	4.150×10^2	3.415×10^2
NY	2.280×10^6	1.059×10^6
FLA	3.610×10^6	1.406×10^6

On the other hand, our algorithm consistently demonstrated greater diversity than the k -best algorithm (Table 3). The difference in the diversity measure was especially pronounced in the DIMACS dataset (in particular DIMACS (FLA)), which has noticeably longer average shortest path length. In case of the SNAP graphs, the difference in the diversity measure was observable, but significantly less dramatic. In essence, the effectiveness of our algorithm seems to correlate with the average length and the total number of shortest st -paths, which is in line with our prediction.

Figure 1 shows the running time of our algorithm and the k -best algorithm on Grid graphs with varying values of k across grid dimensions p .⁶ Our diverse algorithm was con-

⁶It might be hard to distinguish the lines for the k -best algorithm in Figure 1 as their running times are almost the same for $k = 10, 50, 100$.

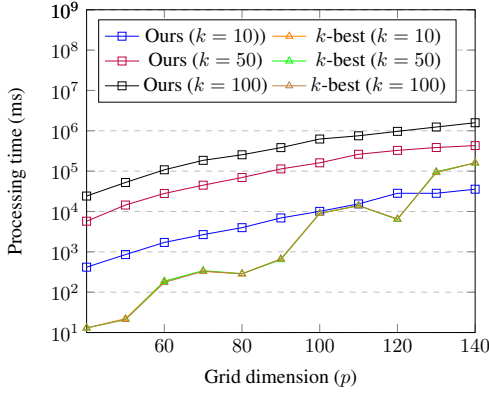


Figure 1: Grid dimension vs. processing time

siderably slower than the k -best algorithm, regardless the value of k . The growth of k resulted in a significant slowdown of the diverse algorithm, whereas it had negligible impact on the processing time of the k -best algorithm. As every single edge in a grid graph is included in a shortest path from s to t , no edge or vertex is discarded in the preprocessing step. (This phenomenon is unlikely in most real-world graphs.) This increases the computational burden of our algorithm on the minimum cost flow step, which explains the experimental outcome.

However, our algorithm vastly outperformed the k -best approach in regard to diversity maximization for all values of k (Figure 2). As the diversity is measured by using edge-wise Hamming distance, the growth of k has a strong impact on the diversity measure by the virtue of adding more edges to compare. Despite this, the diverse algorithm with $k = 50$ has outperformed the k -best counterpart with $k = 100$ for every grid dimension with $p \geq 40$.

In summary, our experiment suggests that our diverse algorithm gives significantly more diverse solutions compared to the conventional k -best enumeration algorithm. Although its slow speed on synthetic grid graphs may raise a question of practicality, it should be emphasized that our algorithm is efficient on large real-life data with over 2 million edges, with the processing time on par with the JGraphT implementation of k -best algorithm.

Concluding Remarks

In this paper, we study the problem of finding k shortest st -paths P_1, P_2, \dots, P_k in edge-weighted directed graphs maximizing $d_w(E(P_1), E(P_2), \dots, E(P_k))$, namely DIVERSE SHORTEST st -PATHS. We show that this problem is polynomial-time solvable, while a slightly “relaxed” version DIVERSE SHORT st -PATHS is NP-hard. The positive result is shown by reducing to a maximum weight packing problem, which also proves the tractability of several problems, such as WEIGHTED DIVERSE MATROID BASES, WEIGHTED DIVERSE ARBORESCENCES, and DIVERSE BIPARTITE MATCHINGS. Our result of DIVERSE BIPARTITE MATCHINGS extends the result of (Fomin et al. 2020), in which they gave a polynomial-time algorithm for find-

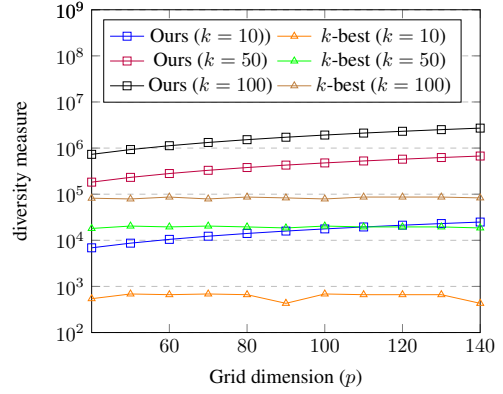


Figure 2: Grid dimension vs. diversity measure

ing two maximum bipartite matchings M_1, M_2 maximizing $d(M_1, M_2)$, to general $k \geq 2$. We also conducted an experiment to assess the practical performance of our algorithm for DIVERSE SHORTEST st -PATHS. Our experiment shows that the practical running time of our algorithm is comparable with that of the known k -best enumeration algorithm (Eppstein 1998) for real-world graph. For synthetic graphs, we used grid graphs, in which every edge belongs to a shortest path between the top-left corner and the bottom-right corner. Our algorithm works slower than the k -best enumeration algorithm on grid graphs. However, the running time of our algorithm is still reasonable for moderate-sized instances, while the diversity of paths is significantly larger than that computed by the k -best enumeration algorithm.

There are several open questions related to our results. We gave a polynomial-time algorithm for DIVERSE SHORTEST st -PATHS. However, it might be more acceptable in some applications to use d_{\min} as its diversity measure. It would be interesting to know the computational complexity of the problem of finding k shortest st -paths P_1, P_2, \dots, P_k maximizing $d_{\min}(E(P_1), E(P_2), \dots, E(P_k))$. Contrary to this tractability, we show that DIVERSE SHORT st -PATHS is NP-hard. In our proof, we use a relatively large threshold $\theta = |V|$ of the length of st -paths. Another interesting question is whether DIVERSE SHORT st -PATHS is still hard even if θ is close to the shortest st -path distance, which might be a reasonable restriction in practice. WEIGHTED DIVERSE ARBORESCENCES and DIVERSE BIPARTITE MATCHINGS are special cases of the following problem: Given two matroids $\mathcal{M}_1 = (E, \mathcal{I}_1)$ and $\mathcal{M}_2 = (E, \mathcal{I}_2)$ of the same ground set E and an integer k , we are asked to find k common bases B_1, B_2, \dots, B_k of \mathcal{M}_1 and \mathcal{M}_2 maximizing $d_w(B_1, B_2, \dots, B_k)$. Here, a subset $B \subseteq E$ is a common base of \mathcal{M}_1 and \mathcal{M}_2 if B is a base of both matroids. This problem is at least as hard as the problem of packing k common bases of two matroids, implying that this problem cannot be solved with a polynomial number of independence oracle queries (Bérczi and Schwarcz 2021). This raises the following question: Is there another natural problem that is a special case of finding diverse common bases and solvable in polynomial time?

Acknowledgments

The authors thank anonymous referees for valuable comments. This work is partially supported by JSPS Kakenhi Grant Numbers JP18H04091, JP18K11168, JP18K11169, JP19H01133, JP20H05793, JP20K19742, JP21H03499, JP21K11752, JP21K17707 JP21H05852 JP21K17812, and JST CREST Grant Number JPMJCR18K3.

References

- Akgün, V.; Erkut, E.; and Batta, R. 2000. On finding dissimilar paths. *Eur. J. Oper. Res.*, 121(2): 232–246.
- Baste, J.; Fellows, M. R.; Jaffke, L.; Masarík, T.; de Oliveira Oliveira, M.; Philip, G.; and Rosamond, F. A. 2020. Diversity of Solutions: An Exploration Through the Lens of Fixed-Parameter Tractability Theory. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, 1119–1125. ijcai.org.
- Baste, J.; Jaffke, L.; Masarík, T.; Philip, G.; and Rote, G. 2019. FPT Algorithms for Diverse Collections of Hitting Sets. *Algorithms*, 12(12): 254.
- Bérczi, K.; and Schwarcz, T. 2021. Complexity of packing common bases in matroids. *Math. Program.*, 188(1): 1–18.
- Chondrogiannis, T.; Bouros, P.; Gamper, J.; Leser, U.; and Blumenthal, D. B. 2020. Finding k-shortest paths with limited overlap. *VLDB J.*, 29(5): 1023–1047.
- Cole, R.; and Hopcroft, J. E. 1982. On Edge Coloring Bipartite Graphs. *SIAM J. Comput.*, 11(3): 540–546.
- Danna, E.; and Woodruff, D. L. 2009. How to select a small set of diverse solutions to mixed integer programming problems. *Oper. Res. Lett.*, 37(4): 255–260.
- Drosou, M.; and Pitoura, E. 2010. Search Result Diversification. *SIGMOD Rec.*, 39(1): 41–47.
- Edmonds, J. 1968. Matroid partition. *Mathematics of the Decision Sciences*, 11: 335–345.
- Edmonds, J. 1975. Some Well-Solved Problems in Combinatorial Optimization. In Roy, B., ed., *Combinatorial Programming: Methods and Applications*, 285–301. Dordrecht: Springer Netherlands. ISBN 978-94-011-7557-9.
- Eppstein, D. 1998. Finding the k Shortest Paths. *SIAM J. Comput.*, 28(2): 652–673.
- Eppstein, D. 2016. *k*-Best Enumeration. In *Encyclopedia of Algorithms*, 1003–1006.
- Fomin, F. V.; Golovach, P. A.; Jaffke, L.; Philip, G.; and Saurabh, D. 2020. Diverse Pairs of Matchings. In *31st International Symposium on Algorithms and Computation, ISAAC 2020, December 14-18, 2020, Hong Kong, China (Virtual Conference)*, 26:1–26:12.
- Fomin, F. V.; Golovach, P. A.; Panolan, F.; Philip, G.; and Saurabh, S. 2021. Diverse Collections in Matroids and Graphs. In Bläser, M.; and Monmege, B., eds., *38th International Symposium on Theoretical Aspects of Computer Science (STACS 2021)*, volume 187 of *Leibniz International Proceedings in Informatics (LIPIcs)*, 31:1–31:14. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-180-1.
- Gabow, H. N.; and Kariv, O. 1982. Algorithms for Edge Coloring Bipartite Graphs and Multigraphs. *SIAM J. Comput.*, 11(1): 117–129.
- Hamacher, H.; and Queyranne, M. 1985. K best solutions to combinatorial optimization problems. *Annals of Operations Research*, 4: 123–143.
- Hanaka, T.; Kobayashi, Y.; Kurita, K.; and Otachi, Y. 2021. Finding Diverse Trees, Paths, and More. In *AAAI*, 3778–3786. AAAI Press.
- Hebrard, E.; Hnich, B.; O’Sullivan, B.; and Walsh, T. 2005. Finding Diverse and Similar Solutions in Constraint Programming. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 1, AAAI’05*, 372–377. AAAI Press. ISBN 157735236x.
- Holyer, I. 1981. The NP-Completeness of Edge-Coloring. *SIAM J. Comput.*, 10(4): 718–720.
- Karp, R. M. 1972. Reducibility Among Combinatorial Problems. In *Complexity of Computer Computations*, The IBM Research Symposia Series, 85–103. Plenum Press, New York.
- Kuby, M. J. 1987. Programming Models for Facility Dispersion: The *p*-Dispersion and Maximum Dispersion Problems. *Geographical Analysis*, 19(4): 315–329.
- Lawler, E. L. 1972. A Procedure for Computing the K Best Solutions to Discrete Optimization Problems and Its Application to the Shortest Path Problem. *Management Science*, 18(7): 401–405.
- Liu, H.; Jin, C.; Yang, B.; and Zhou, A. 2018. Finding Top-k Shortest Paths with Diversity. *IEEE Trans. Knowl. Data Eng.*, 30(3): 488–502.
- Michail, D.; Kinable, J.; Naveh, B.; and Sichi, J. V. 2020. JGraphT—A Java Library for Graph Data Structures and Algorithms. *ACM Trans. Math. Softw.*, 46(2).
- Nadel, A. 2011. Generating Diverse Solutions in SAT. In Sakallah, K. A.; and Simon, L., eds., *Theory and Applications of Satisfiability Testing - SAT 2011 - 14th International Conference, SAT 2011, Ann Arbor, MI, USA, June 19-22, 2011. Proceedings*, volume 6695 of *Lecture Notes in Computer Science*, 287–301. Springer.
- Nagamochi, H.; Zeng, D.; Kabutoya, N.; and Ibaraki, T. 1997. Complexity of the Minimum Base Game on Matroids. *Math. Oper. Res.*, 22(1): 146–164.
- Nash-Williams, C. S. J. A. 1967. An application of matroids to graph theory. In *Theory of Graphs; Proceedings of an International Symposium in Rome 1966*, 263–265. Gordon and Breach, New York.
- Orlin, J. B. 1993. A Faster Strongly Polynomial Minimum Cost Flow Algorithm. *Oper. Res.*, 41(2): 338–350.
- Oxley, J. G. 2006. *Matroid Theory (Oxford Graduate Texts in Mathematics)*. USA: Oxford University Press, Inc. ISBN 0199202508.
- Petit, T.; and Trapp, A. C. 2015. Finding Diverse Solutions of High Quality to Constraint Optimization Problems. In Yang, Q.; and Wooldridge, M. J., eds., *Proceedings of the Twenty-Fourth International Joint Conference on Artificial*

Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015, 260–267. AAAI Press.

Petit, T.; and Trapp, A. C. 2019. Enriching Solutions to Combinatorial Problems via Solution Engineering. *INFORMS J. Comput.*, 31(3): 429–444.

Qin, L.; Yu, J. X.; and Chang, L. 2012. Diversifying Top-k Results. *Proc. VLDB Endow.*, 5(11): 1124–1135.

Vieira, M. R.; Razente, H. L.; Barioni, M. C. N.; Hadjieleftheriou, M.; Srivastava, D.; Traina, C.; and Tsotras, V. J. 2011. On query result diversification. In *2011 IEEE 27th International Conference on Data Engineering*, 1163–1174.

Voss, C.; Moll, M.; and Kavraki, L. E. 2015. A heuristic approach to finding diverse short paths. In *ICRA*, 4173–4179. IEEE.