

# Optimizing Safe Flow Decompositions in DAGs

Shahbaz Khan   

Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, India

Alexandru I. Tomescu   

Department of Computer Science, University of Helsinki, Finland

---

## Abstract

---

Network flow is one of the most studied combinatorial optimization problems having innumerable applications. Any flow on a directed acyclic graph  $G$  having  $n$  vertices and  $m$  edges can be decomposed into a set of  $O(m)$  paths. The applications of such a flow decomposition range from network routing to the assembly of biological sequences. However, in some applications, each solution (decomposition) corresponds to some particular data that generated the original flow. Given the possibility of multiple optimal solutions, no optimization criterion ensures the identification of the correct decomposition. Hence, recently flow decomposition was studied [RECOMB22] in the Safe and Complete framework, particularly for RNA Assembly. The proposed solution reported *all* the *safe* paths, i.e., the paths which are subpath of every possible solution of flow decomposition.

They presented a characterization of the safe paths, resulting in an  $O(mn + out_R)$  time algorithm to compute all safe paths, where  $out_R$  is the size of the raw output reporting each safe path explicitly. They also showed that  $out_R$  can be  $\Omega(mn^2)$  in the worst case but  $O(m)$  in the best case. Hence, they further presented an algorithm to report a concise representation of the output  $out_C$  in  $O(mn + out_C)$  time, where  $out_C$  can be  $\Omega(mn)$  in the worst case but  $O(m)$  in the best case.

In this work, we study how different safe paths interact, resulting in optimal output-sensitive algorithms requiring  $O(m + out_R)$  and  $O(m + out_C)$  time for computing the existing representations of the safe paths. Our algorithm uses a novel data structure called *Path Tries*, which may be of independent interest. Further, we propose a new characterization of the safe paths resulting in the *optimal* representation of safe paths  $out_O$ , which can be  $\Omega(mn)$  in the worst case but requires optimal  $O(1)$  space for every safe path reported. We also present a near-optimal algorithm to compute all the safe paths in  $O(m + out_O \log n)$  time. The new representation also establishes tighter worst case bounds  $\Theta(mn^2)$  and  $\Theta(mn)$  bounds for  $out_R$  and  $out_C$  (along with  $out_O$ ), respectively.

Overall we further develop the theory of safe and complete solutions for the flow decomposition problem, giving an optimal algorithm for the explicit representation, and a near-optimal algorithm for the optimal representation of the safe paths.

**2012 ACM Subject Classification** Mathematics of computing → Graph algorithms; Mathematics of computing → Network flows; Theory of computation → Network flows; Networks → Network algorithms

**Keywords and phrases** safety, flows, networks, directed acyclic graphs

**Digital Object Identifier** 10.4230/LIPIcs.ESA.2022.72

**Related Version** *Full Version:* <https://arxiv.org/abs/2102.06480>

**Funding** This work was partially funded by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 851093, SAFE BIO) and partially by the Academy of Finland (grants No. 322595, 328877).

**Acknowledgements** We would like to thank Manuel Cáceres from University of Helsinki, for helpful discussions and for highlighting several errors in our previous approaches.



© Shahbaz Khan and Alexandru I. Tomescu;  
licensed under Creative Commons License CC-BY 4.0  
30th Annual European Symposium on Algorithms (ESA 2022).

Editors: Shiri Chechik, Gonzalo Navarro, Eva Rotenberg, and Grzegorz Herman; Article No. 72; pp. 72:1–72:17



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

## 1 Introduction

Network flow is one of the most studied problems in theoretical computer science with innumerable applications. For a flow network with a unique source  $s$  and a unique sink  $t$ , every valid flow can be decomposed into a set of weighted  $s$ - $t$  paths and cycles [7]. For a directed acyclic graph (DAG) such a decomposition contains only paths. Such path (and cycle) view of a flow indicates how information optimally passes from  $s$  to  $t$ , being a key step in network routing problems (e.g. [10, 6, 9, 15]), transportation problems (e.g. [16, 17]), or in the more recent and prominent application of reconstructing biological sequences (*RNA transcripts*, see e.g. [18, 23, 8, 5, 22, 27], or *viral quasi-species genomes*, see e.g. [2, 1]).

Finding the minimum flow decomposition (i.e., having the minimum number of paths and cycles) is NP-hard, even if the flow network is a DAG [25]. This hardness result led to research on approximation algorithms [9, 21, 19, 15, 3], and FPT algorithms [12]. Practical approaches usually employ the standard *greedy width* heuristic [25], repeatedly removing an  $s$ - $t$  path carrying the most amount of flow. Recently, another pseudo-polynomial-time heuristic was proposed [20] for biological data, which tries to iteratively simplify the graph such that the flow decomposition problem can be solved locally at some vertices.

In the routing and transportation applications, an optimal flow decomposition indicates how to send some information from  $s$  to  $t$ , and thus any optimal decomposition is satisfactory. However, this is not the case in the prominent application of reconstructing biological sequences, since each flow path represents a reconstructed sequence: a different optimal set of flow paths encodes different biological sequences, which may differ from the real ones. For a concrete example, consider the following application. In complex organisms, a gene may produce more RNA molecules (*RNA transcripts*, i.e., strings over an alphabet of four characters), each having a different abundance. Currently, given a sample, one can read the RNA transcripts and find their abundances using *high-throughput sequencing* [26]. This technology produces short overlapping substrings of the RNA transcripts. The main approach for recovering the RNA transcripts from such data is to build an edge-weighted DAG from these fragments and to transform the weights into flow values by various optimization criteria, and then to decompose the resulting flow into an “optimal” set of weighted paths (i.e., the RNA transcripts and their abundances in the sample) [14]. Clearly, if there are multiple optimal flow decomposition solutions, then the reconstructed RNA transcripts may not match the original ones, and thus be incorrect. Thus, the best possible solution is to find *whatever* can be *safely* reported as being *correct*.

### 1.1 Problem Definition and Related Work

Recently, Ma et al. [13] were the first to address the issue of multiple solutions to the flow decomposition problem, under a probabilistic framework. Later, they [28] solve a problem (*AND-Quant*), which, in particular, leads to a quadratic-time algorithm for the following problem: given a flow in a DAG, and edges  $e_1, e_2, \dots, e_k$ , decide if in *every* flow decomposition there is always a decomposed flow path passing through all of  $e_1, e_2, \dots, e_k$ . Thus, by taking the edges  $e_1, e_2, \dots, e_k$  to be the edges of a path  $p$ , the AND-Quant problem can decide if a path  $p$  (i.e., a given biological sequence) appears in all flow decompositions. This indicates that  $p$  is likely part of some original RNA transcript.

Another popular approach to address the issue of multiple solutions is the *safety* framework, which was introduced by Tomescu and Medvedev [24] for the genome assembly problem from bioinformatics. For a problem admitting multiple solutions, a partial solution is said to be *safe* if it appears in all solutions to a problem. For the flow decomposition problem, a

path  $p$  is safe if for *any* flow decomposition into paths  $\mathcal{P} = \{p_1, \dots, p_k\}$ , it holds that  $p$  is a subpath of some  $p_i$ . Considering the weight, a path  $p$  is further called *w-safe* if, in *any* flow decomposition,  $p$  is a subpath of some path(s) in  $\mathcal{P}_f$  whose total weight is at least  $w$ .

Khan et al. [11] built upon the AND-Quant problem by addressing flow decomposition under the *safety* framework. They presented a local characterization of safe flow paths as compared to the global characterization of AND-Quant. It was directly adaptable to give an optimal verification algorithm, and a simple enumeration algorithm enumerating all safe paths in  $O(mn + out)$  time by applying the characterization on a candidate flow decomposition using the standard *two pointer algorithm*<sup>1</sup>. They presented the maximal safe paths  $out$  in two formats, the raw output  $out_R$  reported each safe path explicitly, and a concise representation  $out_C$  which combined the safe paths occurring contiguously in the candidate flow decomposition. Using a worst case example they also proved that the size of  $out_R$  can be  $\Omega(mn^2)$  in the worst case and  $O(m)$  in the best case, whereas that of  $out_C$  can be  $\Omega(mn)$  in the worst case and  $O(m)$  in the best case. However, in their solution the concise representation of the solution depends on the underlying candidate solution used, which hence does not optimize the concise representation. Moreover, they did not address whether the concise representation is the most succinct approach to represent the safe paths.

## 1.2 Our results

Our main contributions can be described as follows:

1. **Merge-Diverge Property of safe paths.** We develop the theory of safe paths for flow decomposition further by studying the conditions for interaction of safe paths. We prove that two safe paths cannot *merge* at a vertex (or a set of vertices) and later *diverge*.
2. **Optimal output-sensitive enumeration algorithms for the current representations.** We use the *merge-diverge* property to present optimal output-sensitive algorithms for enumerating all safe paths explicitly in  $O(m + out_R)$  time and their optimal concise representation in  $O(m + out_C)$  time. Our algorithms uses a novel application of the Trie on paths, referred as *Path Tries* which may be of independent interest.
3. **Optimal representation of safe paths.** We present a novel characterization of safe paths  $out_O$  allowing us to represent a safe path optimally, requiring  $O(1)$  space for every reported path.
  - **Remark 1.** In the worst case both concise representation  $out_C$  [11] and our optimal representation  $out_O$  may require  $\Omega(mn)$  space, however space required per reported path can be much larger for  $out_C$  than the optimal  $O(1)$  of  $out_O$ .
4. **Near optimal algorithm for the optimal output format.** We present an algorithm to report all safe paths using the optimal representation in  $O(m + out_O \log n)$  time.
5. **Tighter worst case bounds on  $out_R$  and  $out_C$ .** Our characterization allows us to prove matching upper bounds for the worst case lower bounds [11] on  $out_R$  and  $out_C$ .

## 2 Preliminary

Consider a directed acyclic flow graph  $G = (V, E)$  with  $|V| = n$  vertices and  $|E| = m$  edges, where each edge  $e$  has a flow (or weight)  $f(e)$  passing through it. For simplicity we assume the graph is connected giving  $m \geq n - 1$ . For each vertex  $u$ ,  $f_{in}(u)$  and  $f_{out}(u)$  denotes the

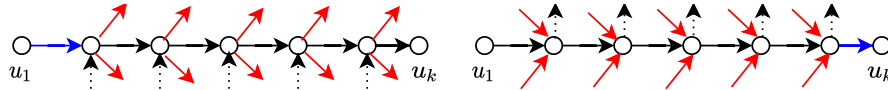
<sup>1</sup> Along a sample solution, the keeping the left end at the start, the right end is moved along the solution as long as the path is safe (evaluated using verification algorithm). This is reported as a maximal safe. Then right end is extended by an edge making the path unsafe, followed by moving the left pointer right until it is safe again. The process is repeated to report the next maximal safe path, and so on.

total flow on its incoming edges and total flow on its outgoing edges, respectively. A vertex  $v$  in the graph is called a *source* if  $f_{in}(v) = 0$  and a *sink* if  $f_{out}(v) = 0$ . The set of sources and sinks of the graph  $G$  is denoted by  $Source(G)$  and  $Sink(G)$  respectively. Every other vertex  $v$  satisfies the *conservation of flow*  $f_{in}(v) = f_{out}(v)$ , making the graph a *flow graph*.

For the vertex  $u$ ,  $f_{max}(\cdot, u)$  (or  $f_{max}(u, \cdot)$ ) denotes the maximum value of flow on the incoming edges (or outgoing edges) of  $u$ . The corresponding edge is represented by  $e_{max}(\cdot, u)$  (or  $e_{max}(u, \cdot)$ ) and its other endpoint (except  $u$ ) is represented by  $v_{max}(\cdot, u)$  (or  $v_{max}(u, \cdot)$ ). Note that in case multiple incoming edges (or outgoing edges) have the maximum flow value, we prefer the edge whose other endpoint (except  $u$ ) appears first in the topological order, making  $e_{max}(\cdot, u)$  (or  $e_{max}(u, \cdot)$ ) and  $v_{max}(\cdot, u)$  (or  $v_{max}(u, \cdot)$ ) distinct. Hence, it is referred as *preferred* maximum incoming (or outgoing) edge/vertex. Further, we represent  $e_{max}^*(\cdot, u)$  (or  $e_{max}^*(u, \cdot)$ ) as the *unique* maximum incoming (or outgoing) edge if  $f_{max}(\cdot, u)$  (or  $f_{max}(u, \cdot)$ ) corresponds to exactly one edge making it equal to  $e_{max}(\cdot, u)$  (or  $e_{max}(u, \cdot)$ ) in such a case, and null otherwise. We similarly define its other endpoint (except  $u$ )  $v_{max}^*(\cdot, u)$  (or  $v_{max}^*(u, \cdot)$ ) which is called *unique* maximum incoming (or outgoing) vertex.

For a path  $p$  in the graph,  $|p|$  represents the number of its edges. A vertex  $u$  is called as being on the *left* of a vertex  $v$  on the path, if  $v$  is reachable from  $u$  on the path. Similarly, in such a case the vertex  $v$  is called as being on the *right* of a vertex  $u$  on the path. For any path  $p$  (or edge) we define its *left extension* to be a path created from  $p$  by repeatedly prepending the path with the unique maximum incoming edge of the first vertex of the (updated) path. Similarly, we define the *right extension* of a path to be a path created by repeatedly adding the unique maximum outgoing edge of the last vertex of the (updated) path.

The *flow decomposition* of  $G$  is a set of weighted *paths*  $\mathcal{P}_f$  such that the flow on each edge in the  $G$  equals the sum of the weights of the paths containing it. A path  $p$  is called *w-safe* if, in every possible flow decomposition,  $p$  is a subpath of some paths in  $\mathcal{P}_f$  whose total weight is at least  $w$ . A  $w$ -safe path with  $w > 0$ , is called a *safe flow path*, or simply *safe path*. A safe path is *left maximal* (or *right maximal*) if extending it to the left (or right) with any edge makes it unsafe. A safe path is *maximal* if it is both left and right maximal. The safety of a path can be characterized by its *excess flow* (see Figure 1) and properties of safe paths, described as follows.



■ **Figure 1** The excess flow of a path is the incoming or outgoing flow (*blue*) that passes through the path despite the flow (*red*) leaking at its internal vertices (reproduced from [11]).

► **Definition 2** (Excess flow [11]). *The excess flow  $f_p$  of a path  $p = \{u_1, u_2, \dots, u_k\}$  is*

$$f_p = f(u_1, u_2) - \sum_{\substack{u_i \in \{u_2, \dots, u_{k-1}\} \\ v \neq u_{i+1}}} f(u_i, v) = f(u_{k-1}, u_k) - \sum_{\substack{u_i \in \{u_2, \dots, u_{k-1}\} \\ v \neq u_{i-1}}} f(v, u_i)$$

where the former and later equations are called *diverging* and *converging criterion*, respectively.

► **Theorem 3** (Safe flow paths [11]). *Safety of flow decomposition satisfy the following.*

- (a) *A path  $p$  is  $w$ -safe iff its excess flow  $f_p \geq w > 0$ .*
- (b) *The converging and diverging criteria for a path  $p = \{u_1, \dots, u_k\}$  are equivalent to*

$$f_p = \sum_{i=1}^{k-1} f(u_i, u_{i+1}) - \sum_{i=2}^{k-1} f_{out}(u_i) = \sum_{i=1}^{k-1} f(u_i, u_{i+1}) - \sum_{i=2}^{k-1} f_{in}(u_i).$$

- (c) Adding an edge  $(u, v)$  to the start or the end of a path in the flow graph, reduces its excess flow by  $f_{in}(v) - f(u, v)$ , or  $f_{out}(u) - f(u, v)$ , respectively.

Additionally, we use the following data structure for answering the level ancestor queries.

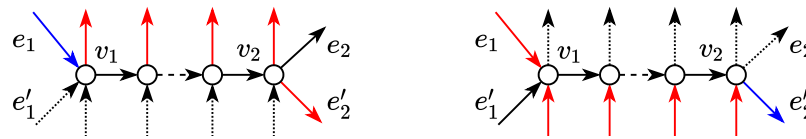
- **Theorem 4** (Level Ancestors [4]). A given tree with  $n$  vertices can be preprocessed in  $O(n)$  time to report the level ancestor  $LA(v, d)$  for a vertex  $v$  at a depth  $d$  in  $O(1)$  time.

### 3 Interaction of Safe Paths

The previous work [11] focused on properties of safe paths useful for applying the characterization directly in verification and enumeration algorithms. We now explore further properties of safe walks particularly related to the interaction of safe paths and its consequences.

- **Lemma 5** (Merge Diverge). Two safe paths cannot merge (through distinct edges) at an intermediate vertex (or vertices) and then diverge (through distinct edges).

**Proof.** Let two safe paths  $p$  and  $p'$  merge at a vertex  $v_1$ , entering  $v_1$  respectively by distinct edges  $e_1$  and  $e'_1$ , and then diverge at a vertex  $v_2$ , leaving  $v_2$  respectively by distinct edges  $e_2$  and  $e'_2$  (see Figure 2).



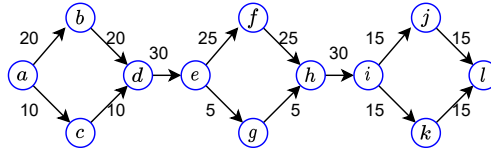
■ **Figure 2** The diverging and converging criterion applied to path  $p$  and  $p'$  respectively.

Using Theorem 3(c) we know that removing an edge from the end of a path increases the excess flow and hence remains safe. Thus, the subpaths  $\{e_1 \cdots, e_2\}$  and  $\{e'_1, \cdots, e'_2\}$  of safe paths  $p$  and  $p'$  respectively, are also safe. By diverging criterion of the safe path  $p$  we have  $f(e_1) > f(e'_2)$ . On the other hand, by converging criterion of the safe path  $p'$  we have  $f(e'_2) > f(e_1)$ , which is a contradiction. ◀

This *merge-diverge* property has an interesting consequence on the structure of a safe path having an edge that is not a unique maximum outgoing edge of a vertex.

- **Lemma 6.** Any safe path having an edge  $e(u, v) \neq e_{max}^*(u, \cdot)$ , can be extended to the left using only unique maximum incoming edges.

**Proof.** Consider a path  $p$  containing  $e_1(u, v) \neq e_{max}^*(u, \cdot)$ , which extends to the left of  $u$  using edges containing an edge which is not a unique maximum incoming edge  $e_2(x, y) \neq e_{max}^*(\cdot, y)$ . Now, using Theorem 3(c) we know subpath created by removing edges from the end is safe, as removing such edges only increases the excess flow. Hence, the subpath  $p : \{e_2 \cdots e_1\}$  is safe. Further, Theorem 3(c) also implies that a path  $p'$  replacing  $(x, y)$  with an alternate  $e'_2 = e_{max}(\cdot, y)$ , and  $(u, v)$  with an alternate  $e'_1 = e_{max}(u, \cdot)$  is also safe, as we replace an edge with another having at least the same weight. Note that  $e'_1$  and  $e'_2$  always exists since  $e_1$  and  $e_2$  are not unique maximum edges. Thus, both  $p : \{e_2 \cdots e_1\}$  and  $p' : \{e'_2 \cdots e'_1\}$  are safe which merge at  $y$  and then diverge at  $u$  using distinct edges, which is a contradiction. ◀



■ **Figure 3** Problem with the simplistic approach. While processing the vertex  $e$ , we get two safe paths  $p_1 : \langle a, b, d, e \rangle$  and  $p_2 : \langle a, c, d, e \rangle$ . As we continue processing to reach  $h$ , both  $p_1$  and  $p_2$  are extended through  $f$  to get  $p_{11} : \langle a, b, d, e, f, h, i \rangle$  and  $p_{21} : \langle a, c, d, e, f, h, i \rangle$ . However, when extending through  $g$  both get trimmed to give the same  $p_{12}, p_{22} : \langle d, e, g, h, i \rangle$ . Further, the paths  $p_{11}$  and  $p_{22}$  are again extended through  $j$  and  $k$  (recall the two pointer algorithm) to get four paths,  $p_{111}, p_{211} : \langle d, e, f, h, i, j, l \rangle$  and  $p_{112}, p_{212} : \langle d, e, f, h, i, k, l \rangle$ . Moreover, paths  $p_{12}, p_{22}$  can also be extended through  $j$  and  $k$  to get  $p_{121}, p_{221} : \langle h, i, j, l \rangle$  and  $p_{122}, p_{222} : \langle h, i, k, l \rangle$ . We thus obtain duplicate paths representing the same safe paths from different sources, and some non-maximal paths ( $p_{122}, p_{222}$ ) which are subpath of the other reported paths.

## 4 Optimal computation of Raw Safe paths

The essential bottle-neck of the previous approach [11] was the use of a candidate flow decomposition, on whose subpaths the safety criteria was evaluated. The computation of a candidate flow decomposition itself requires  $O(mn)$  time making it suboptimal. In order to avoid it we are required to process the graph in a structured manner. Given the graph is a DAG, the topological ordering of the graph serves this purpose.

A simple approach is to follow the topological order and maintain all maximal safe paths ending at the currently processed vertex explicitly. And use the two pointer algorithm to extend it as we continue processing the vertices in the topological order. However, to avoid duplicate and non-maximal results we need to identify the common suffixes of the safe paths, which can be processed accordingly (see Figure 3). Fortunately, for strings the data structure Trie (considered on reversed strings) serves exactly for the same purpose which motivates us to use Tries for storing all the left maximal safe paths ending at a vertex as follows.

### 4.1 Data structures $\mathcal{T}_u$ and $\mathcal{L}_u$

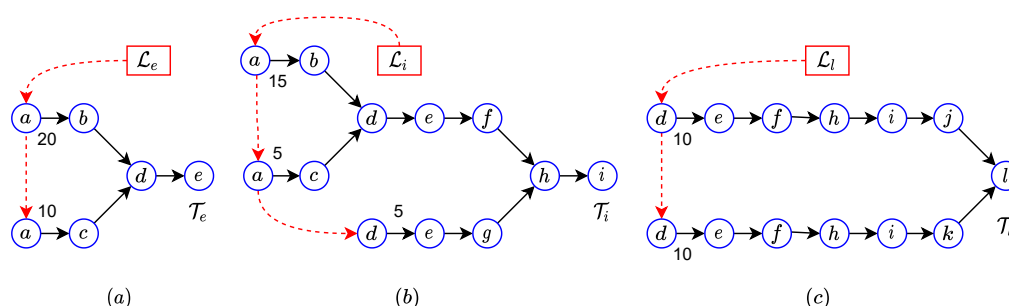
We build a Trie structure  $\mathcal{T}_u$  treating the reverse paths ending at a vertex  $u$  as strings, such that the common suffixes of the paths are combined. Note that a vertex  $v$  can appear multiple times in the Path Trie, if multiple paths containing  $v$  do not share  $v$  in their common suffix.

All left maximal safe paths ending at a vertex  $u$  are hence maintained in  $\mathcal{T}_u$ . Additionally, we maintain a linked list  $\mathcal{L}_u$  of the leaves of  $\mathcal{T}_u$  along with the path's corresponding excess flow, i.e.  $\mathcal{L}_u = \{(v_1, f_1), (v_2, f_2), \dots\}$ . Consider Figure 4, we show the path tries at the vertices  $e, i$  and  $l$  for the graph shown in Figure 3. Note that the leaves represent the left maximal safe paths without repetition or storing subpaths as in the simplistic approach.

### 4.2 Algorithm

The main idea behind our approach is to uniquely extend each safe path ending at a vertex to its preferred maximum out-neighbour in constant time associated with each edge. For the rest of the out-neighbours, we can build their safe paths from scratch at the expense of the path length. This requires us to process the vertices in the topological order of the graph, such that all the safe paths ending at a vertex are computed before it is processed. We maintain the left maximal safe paths ending at a vertex  $u$  in the Path Trie  $\mathcal{T}_u$  and the list of safe paths  $\mathcal{L}_u$ . Our algorithm uses optimal  $O(m + out_R)$  time, where  $out_R$  is the size of the raw output, i.e., each safe path stored explicitly, which is optimal.





■ **Figure 4** Path Trie structure storing the left maximal paths ending at vertices (a)  $\mathcal{T}_e$  with  $\mathcal{L}_e$ , (b)  $\mathcal{T}_i$  and  $\mathcal{L}_i$ , and (c)  $\mathcal{T}_l$  and  $\mathcal{L}_l$ .

Algorithm 1 describes our approach, where vertices are processed in topological order such that while processing a vertex  $u$ , all the left maximal safe paths ending at  $u$  (along with their excess flow) are stored in  $\mathcal{T}_u$  and  $\mathcal{L}_u$ . While processing  $u$  all its safe paths are evaluated for a possible extension to its preferred maximum outgoing neighbour  $v^* = v_{max}(u, \cdot)$ , which always exists except when  $u$  is a sink. Note that  $v^*$  is not the unique maximum outgoing neighbour ( $v_{max}^*(u, \cdot)$ ) rather preferred ( $v_{max}(u, \cdot)$ ), so that  $\mathcal{T}_u$  can be used to extend to some vertex in case there is no unique maximum. Additionally, when  $u$  is a source, we have an empty  $\mathcal{T}_u$ , so we have no safe paths to extend. Hence, for the other cases we check all paths in  $\mathcal{L}_u$  for possible extension to  $v^*$  using Theorem 3(c). For this we add the complete  $\mathcal{T}_u$  as a child of  $\mathcal{T}_{v^*}$ . Thereafter, we need to trim the prefixes of paths in  $\mathcal{T}_{v^*}$  which are not safe and compute the list of safe paths  $\mathcal{L}_{v^*}$ . Further, we need to add the safe paths using every other outgoing edge  $(u, v)$  ( $v \neq v^*$ ) in the corresponding  $\mathcal{T}_v$ .

We now process each path in  $\mathcal{L}_u$ . The paths which are not safe on extending to  $(u, v^*)$  are clearly right-maximal (and hence maximal), and hence are reported in the solution SOL. We extend all the paths in  $\mathcal{L}_u$  with  $(u, v^*)$ , and add their maximal suffixes which are safe to  $\mathcal{T}_{v^*}$ . Note that the entire path may be safe or at least the edge  $(u, v^*)$  is safe. So we start trimming  $\mathcal{T}_{v^*}$  until the path is safe. This is done by maintaining in  $f_x$  the excess flow of the path from  $x$  to  $v^*$ , where  $f_x$  is updated using Theorem 3(c). When  $(u, v^*)$  is added  $f_x$  may become negative in which case the path is trimmed from the left until  $f_x$  is positive. Now, since in a Trie an edge can be shared by multiple paths having a common suffix, we trim the edge only if it is a leaf, and similarly add to  $\mathcal{L}_{v^*}$  only if it starts from a leaf. Hence, multiple paths from  $\mathcal{L}_v$  will not add the same safe path to  $\mathcal{L}_{v^*}$  as it will start from a leaf only when the last such path is processed. This also avoids adding a non-maximal path which is a subpath of another safe path.

Finally, we need to add the safe paths to non-preferred maximum outgoing neighbours, which by Lemma 6 is always on a single path containing the unique maximum incoming edges. We thus compute the single safe left extension for all such neighbours explicitly using Theorem 3(c). We do this again by maintaining the excess flow of the path from  $x$  to  $v$  in  $f_x$ . As we start the path is a single edge with  $f_x$  necessarily positive, where we continue adding the preferred maximum incoming edge to the left until  $f_x$  is negative. Note that we do not insist on a unique maximum incoming edge as required by Lemma 6 as the flow  $f_x$  will itself become negative if the preferred maximum incoming edge is not unique.

■ **Algorithm 1** Optimally Computing Raw representation of Safe Paths.

---

```

Compute Topological Order of  $G$ 
forall  $u \in V$  in topological order do
  | COMPUTE-SAFE( $u$ )
COMPUTE-SAFE( $u$ ):
if  $u \notin \text{Sink}(G) \cup \text{Source}(G)$  then
  |  $v^* \leftarrow v_{\max}(u, \cdot)$ 
else  $v^* \leftarrow \text{null}$ 
if  $u \in \text{Source}(G)$  then Initialize  $\mathcal{T}_u$  with  $u$ 
forall  $(u, v) \in G, v \neq v^*$  do // new paths
  | Add  $(u, v)$  to  $\mathcal{T}_v$ 
  |  $x \leftarrow u, f_x \leftarrow f(u, v)$ 
  | while  $x \notin \text{Source}(G)$  and  $f_x - f_{\text{in}}(x) +$ 
  |    $f_{\max}(\cdot, x) > 0$  do
  |   | Add  $e_{\max}(\cdot, x)$  to  $\mathcal{T}_v$ 
  |   |  $f_x \leftarrow f_x - f_{\text{in}}(x) + f_{\max}(\cdot, x)$ 
  |   |  $x \leftarrow v_{\max}(\cdot, x)$ 
  | Add  $(x, f_x)$  to  $\mathcal{L}_v$ 
if  $v^* \neq \text{null}$  then
  | Make  $\mathcal{T}_u$  as child of  $v^*$  in  $\mathcal{T}_{v^*}$ 
forall  $(x, f_x) \in \mathcal{L}_u$  do // Process  $\mathcal{T}_u$ 
  | if  $v^* = \text{null}$  or  $f_x - f_{\text{out}}(u) + f(u, v^*) \leq 0$ 
  |   then
  |   |  $p \leftarrow$  Extract path from  $x$  to  $u$  in  $\mathcal{T}_u$ 
  |   | Add  $(p, f_x)$  to SOL
  | if  $v^* \neq \text{null}$  then
  |    $f_x \leftarrow f_x - f_{\text{out}}(u) + f(u, v^*)$ 
  |   while  $f_x \leq 0$  and  $x$  is leaf of  $\mathcal{T}_{v^*}$  do
  |   |  $y \leftarrow$  Parent of  $x$  in  $\mathcal{T}_{v^*}$ 
  |   |  $f_x \leftarrow f_x + f_{\text{in}}(y) - f(x, y)$ 
  |   | Remove  $(x, y)$  from  $\mathcal{T}_{v^*}$ 
  |   |  $x \leftarrow y$ 
  |   if  $x$  is a leaf in  $\mathcal{T}_{v^*}$  then
  |   | Add  $(x, f_x)$  to  $\mathcal{L}_{v^*}$ 

```

---

### 4.3 Correctness

We prove the correctness of the algorithm by induction over the topological order of the graph. The underlying invariant is as follows:

*After COMPUTE-SAFE( $u$ ) is executed, all left maximal safe paths having starting vertex and the internal vertices with topological order up to  $u$ , are stored in corresponding  $\mathcal{T}_v$  and  $\mathcal{L}_v$ . Also, all maximal safe paths ending at vertices with topological order up to  $u$ , are reported in SOL.*

The base case is trivially true when no vertices are processed, as no safe paths exist. Now, when we start processing  $u$ , using the invariant we know all the *left maximal* safe paths not having  $u$  as internal vertex, and all the *maximal* safe paths ending at the vertex with topological order less than  $u$  are already in SOL. So we need to process only the *left maximal* paths having  $u$ , which necessarily have the last internal vertex as  $u$ , and all the *maximal* safe paths ending at  $u$  must be added to SOL. The prefix (not necessarily proper) of both these kinds of paths up to  $u$ , are clearly safe (using Theorem 3(c)) and hence are present in  $\mathcal{T}_u$  and  $\mathcal{L}_u$  by the invariant.

Now, all the *left maximal* safe paths are checked for a possible extension to  $v^*$  by construction, and for the remaining out-neighbours we explicitly add the single safe path possible (Lemma 6). Further, all the paths in  $\mathcal{L}_u$  are checked for being maximal and added to SOL in such a case. Note that processing the vertices in the topological order ensures that all safe paths ending at  $u$  have internal vertices already processed so that the complete path is present in  $\mathcal{T}_u$  and  $\mathcal{L}_u$  before it is processed.

### 4.4 Analysis

The total time required by the algorithm can be associated with the edges of the graph  $m$  or the total length of safe paths reported, i.e. size of the raw representation of the output  $out_R$ . Computing the topological order of the graph requires  $O(m)$  time. Now, for each vertex  $u$ ,



processing the paths in  $\mathcal{L}_u$  either extends it to  $v^*$  in  $O(1)$  time or reporting a safe path  $p$  and extending its subpath to  $v^*$  in  $O(|p|)$  time. In the former case, the length of the safe path is increased by one (adding  $u$ ), and the latter case is associated with the length of the reported path (as each safe path is reported exactly once). For residual out-neighbours, the time required is proportional to the size of added safe path. Hence we have the following:

► **Theorem 7.** *Given a flow graph (DAG) having  $n$  vertices and  $m$  edges, the set of all safe paths can be optimally reported in its raw representation in  $O(m + out_R)$  time.*

## 5 Optimal computation of the optimal Concise Representation

Previous work [11] presented a simple algorithm for computing the concise representation of the solution. Hence instead of reporting each safe path individually which may have overlaps among each other, they combined several overlapping safe paths to report a single path  $p$  along with indices representing the subpaths of  $p$  which are maximally safe.

However, their concise representation was dependent on the underlying candidate path decomposition, which may be suboptimal. We shall now present an optimal algorithm for computing the optimal concise representation of the solution. Our algorithm again uses Path Tries  $\mathcal{T}_u$  with a modified version  $\mathcal{L}_u^+$  of the list of safe paths to store the concise representation.

### 5.1 Data structures $\mathcal{T}_u$ and $\mathcal{L}_u^+$

We again build a Trie structure on the reverse paths, whose common suffixes are combined. Similar to the previous algorithm, it stores all the left maximal safe paths ending at  $u$ .

Now, the concise representation of safe paths are reported in the form  $\{(p_1, I_1), (p_2, I_2), \dots\}$ , where each path  $p_i$  has maximal safe subpaths denoted by intervals  $I_i = \{(l_1, r_1, f_1), (l_2, r_2, f_2), \dots\}$ . Each  $(l_j, r_j)$  and  $f_j$  denote the corresponding end vertices of the maximal safe path on  $p_i$  and its excess flow, respectively. While processing  $u$ , the partial results are maintained in the list  $\mathcal{L}_u^+ = \{(p_1, I_1), (p_2, I_2), \dots\}$  where  $p_i$  are the partially built concise representation of the safe paths, where the reported paths contain  $u$ . Further, for each  $I_i$  the last interval  $(l_j, r_j, f_j)$  has  $r_j = u$  representing the left maximal safe path ending at  $u$ , whereas the remaining intervals are maximal. Note that while processing  $u$ ,  $p_i$  does not include the last path from  $l_j$  to  $u$ .

### 5.2 Algorithm

The main idea behind our approach is to always attempt to extend a path  $p_i$  of the concise representation with the interval corresponding to a safe path that overlaps the most with  $p_i$ . Clearly, while extending the left maximal safe paths on  $u$  to its out-neighbours, the maximum such overlap with  $p_i$  would correspond to the safe path ending at the preferred maximum outgoing neighbour  $v^*$ , which is hence added to  $p_i$ . However, in case multiple paths  $p_i, p_j$  in the concise representation add exactly the same safe path  $p_{v^*}$  corresponding to  $v^*$ , it is not optimal to add  $p_{v^*}$  to both  $p_i$  and  $p_j$ . In such a case  $p_{v^*}$  can be added to anyone such path (say  $p_i$ ), and  $p_j$  will add the maximum overlapping path  $p_v$  corresponding to some other out-neighbour  $v$  of  $u$ , if it exists. If no such neighbour exists, we will report  $p_j$  in the solution having the last interval ending at  $u$ . And in case the safe path  $p_{v'}$  of some out-neighbour  $v'$  of  $u$  is not accommodated in the existing paths of the concise representation, we add a new path  $p_{v'}$  to the concise representation. The optimality of our concise representation is guaranteed by our choice of maximum overlap, ensuring a new path in the concise representation is always of the minimum length.

Consider Algorithm 2, similar to the previous algorithm we process each vertex in the topological order, and when a vertex  $u$  is processed its  $\mathcal{T}_u$  and  $\mathcal{L}_u^+$  have already been computed by its incoming neighbours. Similar to the previous approach for each out-neighbour  $v$  of  $u$ , that is not the preferred maximum out-neighbour of  $u$  exactly a single safe path exists containing the unique maximum incoming edges (Lemma 6). We compute it similar to the previous algorithm for each  $v$  from scratch, and update its corresponding  $\mathcal{T}_v$ , inserting the path  $p_v$  temporarily to  $\mathcal{L}_v^+$  as a new path. This path can potentially be added to some existing path in  $\mathcal{L}_u^+$ , for which we mark the starting vertex of  $p_v$  in  $\mathcal{T}_u$ .

Now, for the unique maximum outgoing neighbour  $v^*$  of  $u$ , we add  $\mathcal{T}_u$  as a child of  $v^*$  in  $\mathcal{T}_{v^*}$  and attempt to extend each path  $p_k$  in  $\mathcal{L}_u^+$  to  $v^*$  using Theorem 3(c). Similar to the previous algorithm, this is accompanied by trimming the leaves of last interval of  $p_k$  from  $\mathcal{T}_u$  if they are not safe in  $\mathcal{T}_{v^*}$ . Again, if the start of the safe path  $x$  for  $v^*$  is no longer a leaf, then the same safe subpath is shared by some other safe path in  $\mathcal{L}_u^+$ . In such a case we do not extend  $p_k$  to  $v^*$ , rather either (a) extend it to the out-neighbour  $v$  of  $u$  having the maximum overlap (lowest vertex marked in  $\mathcal{T}_u$  along the last interval of  $p_k$ ) which is not a unique maximum out-neighbour of  $u$ , or (b) terminate  $p_k$  at  $u$  including it as its last interval. Thus, while processing  $\mathcal{T}_u$  for each path in  $\mathcal{L}_u^+$  we deal with five distinct cases (see Algorithm 2).

- (a)  **$v^*$  is null because  $u \in \text{Source}(G)$ :** The list  $\mathcal{L}_u^+$  is empty and each out-neighbour  $v$  of  $u$  is addressed as non-preferred maximum out-neighbour adding the corresponding edge to their  $\mathcal{T}_v$  and  $\mathcal{L}_v^+$  computing from scratch.
- (b)  **$v^*$  is null because  $u \in \text{Sink}(G)$ :** For all paths in  $\mathcal{L}_u^+$ , the left limit  $x$  reaches  $u$  (as  $f_x$  is always negative and no out-neighbour exists to mark a vertex in  $\mathcal{T}_u$ ), which is hence updated to include the last interval of  $I_k$  and added to SOL.
- (c) **Path  $p_k$  is extended to include  $v^*$ :** The safe path is unique to  $p_k$  and hence the left limit of safe path  $x$  is always a leaf, terminating as soon as  $f_x > 0$  and added to  $\mathcal{L}_{v^*}^+$  accordingly. Note that no vertex before reaching  $f_x > 0$  could have been marked, as left limit of  $v^*$  would be the lowest among all out-neighbours of  $u$ .
- (d) **Path  $p_k$  is extended to include some  $v \neq v^*$ :** This is possible only if the safe path for  $v^*$  is not unique to  $p_k$ , i.e.,  $x$  is no longer a leaf. Then maximum overlap is the lowest vertex along  $x$  to  $u$  path, to which  $p_k$  is added accordingly.
- (e) **Path  $p_k$  is not extended and reported in Sol:** This is possible again when safe path for  $v^*$  is not unique, so  $x$  is no longer a leaf and  $x$  reaches  $u$  similar to case (b).

### 5.3 Correctness and Analysis

The optimality of  $out_C$  is ensured by appending a path in  $out_C$  with the safe path having the maximum overlap with the existing path. This is ensured by processing the marked vertices bottom-up, which represent the start vertex of the safe paths corresponding to the out-neighbour  $v$  of  $u$ , which is not a unique maximum out-neighbour of  $u$ . This guarantees that in case the path cannot be uniquely extended to  $v^*$ , the vertex  $v$  with the maximum overlap is selected resulting in an optimal concise representation.

The total time taken while processing  $u$  is dominated by the processing of  $\mathcal{L}_u^+$  and building the safe paths for those out-neighbours of  $u$  which are not unique maximum out-neighbours of  $u$ , from scratch. Consider the cases (a), (b), (c) and (e), the time taken in processing  $\mathcal{L}_u^+$  can be easily associated with the length of the path  $p_k$  in  $\mathcal{L}_u^+$  since it will be reported exactly once (cases (b) and (e)), and removed from the last interval exactly once (case (c)). Case (a) is also easy to associate as it increases the corresponding paths in  $\mathcal{L}_v^+$ , and hence can be associated with its length.

The only hard case is (d) as it computes the safe path for  $v$  from scratch, but extends it on an existing  $p_k \in \mathcal{L}_u^+$  which takes more time than the increase in  $p_k$ . However, note that this is possible only when the left limit  $x$  is no longer a leaf, which implies that the extra processed path is a common suffix of multiple paths in  $\mathcal{L}_u^+$ . And hence the suffix was accounted for only once for multiple paths, and now when  $p_k$  is detached from the common suffix the cost of the processed path can be associated with that of the detached path (previously unaccounted being a part of the common suffix). Thus, all the steps can be accounted for with the length of  $out_C$  and we get the following.

► **Theorem 8.** *Given a flow graph (DAG) having  $n$  vertices and  $m$  edges, the optimal concise representation  $out_C$  of the safe paths can be optimally reported in  $O(m + out_C)$  time.*

■ **Algorithm 2** Optimally Computing concise Representation of Safe Paths.

---

```

Compute Topological Order of  $G$ 
forall  $u \in V$  in topological order do
  | COMPUTE-SAFE-COMPR( $u$ )

COMPUTE-SAFE-COMPR( $u$ ): if  $u \notin Sink(G) \cup Source(G)$  then
  |  $v^* \leftarrow v_{max}(u, \cdot)$ 
else  $v^* \leftarrow null$ 
if  $u \in Source(G)$  then Initialize  $\mathcal{T}_u$  with  $u$ 
forall  $(u, v) \in G, v \neq v^*$  do // new paths
  | Add  $(u, v)$  to  $\mathcal{T}_v$ 
  |  $x \leftarrow u$  in  $\mathcal{T}_u, f_x \leftarrow f(u, v)$ 
  | while  $x \neq leaf$  of  $\mathcal{T}_u$  and  $f_x - f_{in}(x) + f_{max}(\cdot, x) > 0$  do
  |   | Add  $e_{max}(\cdot, x)$  to  $\mathcal{T}_v$ 
  |   |  $f_x \leftarrow f_x - f_{in}(x) + f_{max}(\cdot, x)$ 
  |   |  $x \leftarrow v_{max}(\cdot, x)$  in  $\mathcal{T}_u$ 
  | Add  $(\emptyset, \{(x, v, f_x)\})$  to  $\mathcal{L}_v^+$ 
  | Push  $v$  to  $Mark[x]$ 
  | Add  $x$  to  $\mathcal{M}$ 
if  $v^* \neq null$  then
  | Add  $\mathcal{T}_u$  as child of  $v^*$  in  $\mathcal{T}_{v^*}$ 

forall  $(p_k, I_k) \in \mathcal{L}_u^+$  do // Process  $\mathcal{T}_u$ 
  |  $(l_i, u, f_i) \leftarrow$  Last of  $I_k, x \leftarrow l_i$  in  $\mathcal{T}_u$ 
  | if  $v^* = null$  then  $f_x \leftarrow -\infty$ 
  | else  $f_x \leftarrow f_i - f_{out}(u) + f(u, v^*)$ 
  | while  $f_x \leq 0$  and  $Mark[x] = \emptyset$  and  $x \neq u$  do
  |   |  $y \leftarrow$  Parent of  $x$  in  $\mathcal{T}_u$ 
  |   | if  $y$  is not a leaf in  $\mathcal{T}_u$  then
  |   |   |  $f_x \leftarrow f_x + f_{in}(y) - f(x, y)$ 
  |   |   | Remove  $(x, y)$  from  $\mathcal{T}_u$ 
  |   |  $x \leftarrow y$ 
  |  $p \leftarrow$  Path from  $l_i$  to  $x$  in  $\mathcal{T}_u$ 
  |  $p_k \leftarrow p_k \cup \{p \setminus \{x\}\}$ 
  | if  $f_x > 0$  then
  |   | if  $l_i \neq x$  then Add  $(x, v^*, f_x)$  to  $I_k$ 
  |   | else Last of  $I_k \leftarrow (l_i, v^*, f_x)$ 
  |   | Add  $(p_k, I_k)$  to  $\mathcal{L}_{v^*}^+$ 
  | else
  |   | if  $Mark[x] \neq \emptyset$  then
  |   |   |  $v \leftarrow$  Pop from  $Mark[x]$ 
  |   |   |  $(\emptyset, I_v) \leftarrow$  Pop from  $\mathcal{L}_v^+$ 
  |   |   | Add  $(p_k, I_v \cup I_k)$  to  $\mathcal{L}_v^+$ 
  |   | else Add  $(p_k \cup \{x\}, I_k)$  to SOL

forall  $x \in \mathcal{M}$  do Clear  $Mark[x]$ 
Clear  $\mathcal{M}$ 

```

---

## 6 Optimal Representation of Safe paths

The raw representation of the safe paths  $out_R$  can take  $\Theta(mn^2)$  space and hence time in the worst case. The previous work [11] presented a concise representation of the safe paths reporting a combination of the overlapping safe paths along with its indices, requiring total  $\Theta(mn)$  space. However, it may not be optimal as the total size of this concise representation may be much larger than the number of safe paths. We thus present an optimal representation of the safe path whose size requires  $\Theta(1)$  space for every safe path reported.

## 6.1 Representative edge with left and right extensions

Lemma 6 presents an interesting property about safe paths being extendible in a preferred way to the left for edges which are not unique maximum outgoing edges of some vertex. We extend the notion further by considering a representative edge for each safe path such that the maximal path can always be generated by extending it to the left along unique maximum incoming edges and to the right along unique maximum outgoing edges as follows.

► **Theorem 9** (Representative edge). *Given a flow graph (DAG), every safe path  $p$  can be described using a representative edge  $e_p$ , such that  $p$  can be constructed by extending  $e_p$  to the left along the unique maximum incoming edges and to the right along the unique maximum outgoing edges.*

**Proof.** Given a maximal safe path  $p$ , let  $e(x, y) \in p$  be the leftmost edge such that  $e(x, y) \neq e_{max}^*(\cdot, y)$ . If no such edge exists, we define the last edge as the representative edge of  $p$ , where rest of  $p$  is along its unique maximum incoming edges (left extension of  $e$ ) proving the existence of the representative edge.

Now, by definition the prefix of  $p$  on the left of  $e$  is along the unique maximum incoming edges (left extension of  $e$ ), so we only need to prove that the suffix of  $p$  after  $e$  is along the unique maximum outgoing edges (or the right extension of  $e$ ). We shall prove it by contradiction, hence assume there exist an edge  $e'(a, b) \in p$  to the right of  $e$  such that  $e'(a, b) \neq e_{max}^*(a, \cdot)$ . Clearly, the path  $e_{max}^*(\cdot, y) \cup p[y, a] \cup e_{max}^*(a, \cdot)$  is also safe using Theorem 3(c). However, this violates the merge-diverge property (Lemma 5) contradicting our assumption and proving the existence of  $e$  as the representative edge of  $p$ . ◀

► **Remark 10.** Every safe path contains a representative edge which is either the last edge which is also unique maximum incoming edge, or an edge which is not a unique maximum incoming edge. For the sake of uniformity, in case multiple edges satisfy this property (not being unique maximum incoming edge) for a safe path  $p$ , we consider  $e_p$  to be the rightmost such edge.

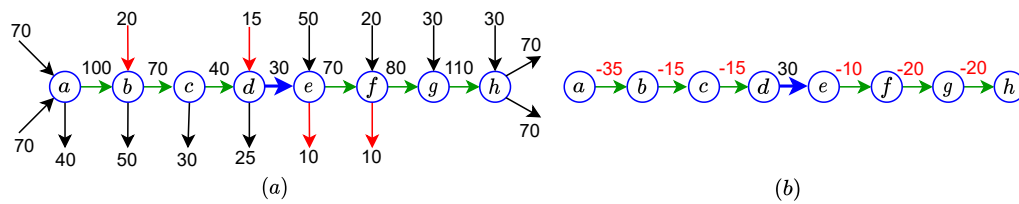
Note that the safe paths when represented using such a representation requires  $O(1)$  space per safe path to store the representative edge and its two endpoints, where a single representative edge may store multiple pairs of endpoints representing individual safe paths. We assume that the unique maximum incoming and outgoing edges of each vertex are known which can be pre-computed in  $O(m)$  time and stored using  $O(n)$  space.

## 6.2 Approach

As described previously in Remark 10, the representative edge  $e$  can be either (a) an edge which is not a unique maximum incoming edge, or (b) a unique maximum incoming edge. The former case is non-trivial as the edge can represent multiple safe paths, whereas the latter is trivial as the edge can represent at most one (can be zero) maximal safe path ending at  $e$ . So here we describe only the non-trivial case as trivial can be computed similarly.

Consider Figure 5 (a), where the edge  $(d, e)$  is a edge which is not a unique maximum incoming edge of  $e$ . The path  $\langle a, b, c, d \rangle$  is its left extension (along unique maximum incoming edges), and the path  $\langle e, f, g, h \rangle$  is its right extension (along unique maximum outgoing edges). Now, once we compute the left and right extensions we can easily compute all the safe paths represented by  $(d, e)$  using two pointer approach in time proportional to length of the path. We get the maximal safe paths  $\langle b, c, d, e, f \rangle$  and  $\langle d, e, f, g, h \rangle$ .

However, assuming we have pre-computed the left and right extensions we can compute all the safe paths using binary search along the path in time  $O(\log n)$  times the number of safe paths. This can be done by pre-computing the loss for extension along each edge



■ **Figure 5** Graphs describing (a) the left and right extensions of a representative edge, and (b) the cumulative losses on extension along each edge.

(Theorem 3(c)) and storing the cumulative value on the edge. Now, we find the safe paths as follows. Given the flow on  $(d, e)$  is 30 we search for the leftmost minimum value greater than  $-30$  which we find as  $(b, c)$  with value  $-15$  giving a left maximal safe path. We make it right maximal (and hence maximal) by searching for the rightmost minimum value greater than  $-30 + 15 = -15$  which we find as  $(e, f)$ , giving our maximal safe path from  $b$  to  $f$  using two binary searches. Now, to find the next maximal we extend it to right including  $(f, g)$  and again find the left maximal on  $-30 + 20$  as  $(d, e)$ , and thereafter right maximal on  $-30 + 0$  as  $(g, h)$ , giving the second maximal safe path from  $d$  to  $h$  using another two binary searches.

### 6.3 Data structures

As described in our approach above we require pre-computed left and right extensions for each edge along with the cumulative losses on each edge for efficient search of the maximal safe paths. This can be efficiently computed by building all possible left and right extensions separately which will form two forests corresponding to all unique maximum incoming and outgoing edges. Further for  $O(1)$  time access to elements on these forests for binary search we require the classical Level ancestor data structure.

#### 1. Unique maximum incoming and outgoing forests $\mathcal{F}_i$ and $\mathcal{F}_o$ .

For each vertex in the graph, we add the unique maximum incoming edge (if exists) to  $\mathcal{F}_i$ . Clearly, each vertex has at most one incoming neighbour (parent) making  $\mathcal{F}_i$  a forest. Similarly, for each vertex, we add the *reverse* of the unique maximum outgoing edge (if exists) and  $\mathcal{F}_o$ . Again, each vertex has at most one incoming neighbour as a parent (as we added reverse edges), making  $\mathcal{F}_o$ .

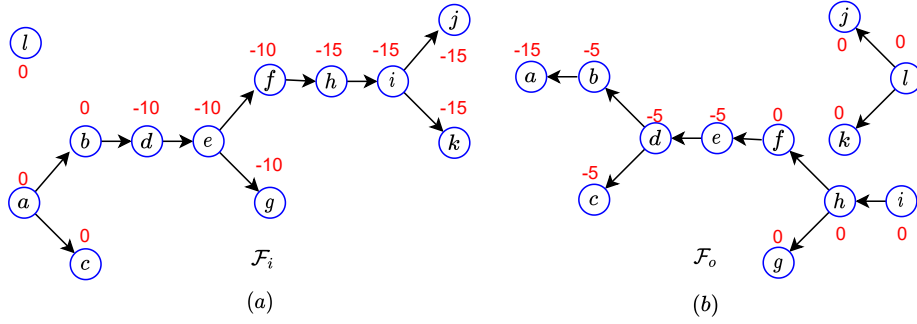
Now, for each vertex  $v$  in the forest  $\mathcal{F}_i$  we store the cumulative loss  $c_i[v]$  (using Theorem 3(c)) along the path  $v$  to the root of its tree  $root[\mathcal{F}_i(v)]$ . The cumulative loss for any subpath from  $v$  to  $u$  (where  $u$  is ancestor of  $v$  in  $\mathcal{F}_i$ ), can be simply computed as  $c_i[u] - c_i[v]$ . Similarly, we store the cumulative loss for each vertex  $v$  on  $\mathcal{F}_o$  in  $c_o[v]$ . The corresponding forests  $\mathcal{F}_i$  and  $\mathcal{F}_o$  for Figure 3 are shown in Figure 6. Clearly, these structures can be computed in  $O(m)$  time.

#### 2. Level ancestors $LA_i$ and $LA_o$

We use Theorem 4 to compute the data structure on  $\mathcal{F}_i$  and  $\mathcal{F}_o$  using  $O(n)$  time, for reporting the level ancestors  $LA_i(v, d)$  and  $LA_o(v, d)$  of a vertex  $v$  at depth  $d$  in  $O(1)$  time.

### 6.4 Algorithm

We now describe how our approach (described on paths) can be used to compute the optimal representation of all the maximal safe paths in  $O(m + out_O \log n)$  time.



■ **Figure 6** For the graph in Figure 3 we show (a) the unique maximum incoming forest  $\mathcal{F}_i$  and (b) the unique maximum outgoing forest  $\mathcal{F}_o$ . Note that  $l$  (in  $\mathcal{F}_i$ ) and  $i$  (in  $\mathcal{F}_o$ ) do not have a parent in the absence of corresponding unique maximum edges.

We first address computing all non-trivial safe paths. Consider each edge which is not unique maximum incoming edge, say  $e(u, v) \neq e_{max}^*(\cdot, u)$ . We use the two pointer algorithm as described in [11] however in each step we perform a binary search to compute each safe path in  $O(\log n)$  time. The binary search computes the excess flow on a path using the values of  $c_o[v], c_i[v]$ , and probes an element by considering ancestors of  $u$  in  $\mathcal{F}_i$  and ancestors of  $v$  in  $\mathcal{F}_o$  which can be directly accessed in  $O(1)$  time using  $LA_i$  and  $LA_o$  structures. The optimal output reports a list of pairs of end vertices for maximal safe paths represented by  $e(u, v)$ . However, we avoid this algorithm if the maximal safe path containing  $e$  is the single edge  $e$ , which is typically not reported. This can be evaluated in  $O(1)$  time using Theorem 3(c) in both left and right directions.

For computing the trivial paths, we need to compute all maximal safe paths containing only unique maximum incoming edges. We simply look at the leaves of  $\mathcal{F}_i$  and search for the left maximal path ending on it, and thereafter continue the search using a modified two pointer algorithm (using binary search) on its ancestors. However, we need to ensure two aspects. Firstly, we do not perform a binary search in case there exist no maximal safe path, which is only possible if (a)  $f(u, v) > |c_i[u]|$ , implying the entire path to root is safe, and (b)  $f(e_{max}(v, \cdot)) > |c_i[v]|$  implying the path till  $v$  is not maximal. In case only (a) is true we report a single safe path from the root to  $v$ , and if both are true we skip the leaf. Secondly, the two pointer algorithm on internal vertices may repeat the safe paths reported by other leaves as the internal vertices may be shared. Hence we mark the internal vertices which have been processed so as to avoid repetition of processing and results.

## 6.5 Implementation details

For the sake of completeness we now describe the two pointer algorithm using binary search on  $\mathcal{F}_i$  and  $\mathcal{F}_o$  in detail.

The algorithm computes all pairs of end points for safe paths represented by  $e(u, v)$  as follows. It first computes the start of left maximal safe path ending at  $v$  by performing a binary search on the ancestors of  $u$  in  $\mathcal{F}_i$ . It computes the highest ancestor  $l$  of  $u$  such that excess flow of  $l$  to  $v$  is positive, i.e.  $c_i[l] < c_i[u] + f(e)$ . This search involves accessing the ancestor at mid-depth directly in  $O(1)$  time using  $LA_i(u, d[u]/2)$  and so on for the whole binary search. Thereafter, the algorithm computes the end  $r$  of right maximal path starting from  $l$  in ancestors of  $v$  in  $\mathcal{F}_o$  such that excess flow of the path  $\langle l, \dots, u, v, \dots, r \rangle$  is positive, i.e. highest ancestor of  $v$  such that  $c_o[r] < f(e) + c_i[u] + c_o[v] - c_i[l]$ . We record



$[r, l]$  with the corresponding flow  $f(e) + c_i[u] - c_i[l] + c_o[v] - c_o[r]$  as a safe path for  $e$ . Then we find the next start of the left maximal path ending at ancestor of  $r$  in  $\mathcal{F}_o$ , i.e., using  $f(e) + c_o[v] - c_o[r]$  instead of  $f(e)$  in the process described above. This continues until the left maximal reaches  $v$  or the right maximal reaches the root of  $\mathcal{F}_o$  containing  $v$ .

## 6.6 Correctness and Analysis

The correctness of our algorithm follows from that of the two pointer algorithm described in [11]. By construction, we show that the algorithm requires  $O(1)$  time to check whether a safe path exists corresponding to a representative edge, and thereafter  $O(\log n)$  time to report each safe path represented by the edge. Since any explicit representation of the safe paths would require  $O(1)$  space for every safe path,  $out_O$  is the number of safe paths. We thus get the following result.

► **Theorem 11.** *Given a flow graph (DAG) having  $n$  vertices and  $m$  edges, the optimal representation  $out_O$  of the safe paths can be reported in  $O(m + out_O \log n)$  time.*

## 7 Space bounds for different representations of safe paths

Previous work [11] presented a worst case example demonstrating that  $out_R$  and  $out_C$  can respectively be  $\Omega(mn^2)$  and  $\Omega(mn)$  in the worst case and  $O(m)$  in the best case. The worst case example graph they presented also gives a bound of  $\Omega(mn)$  in the worst case and  $O(m)$  in the best case for  $out_O$ . In the light of the new characterization, we shall now understand these bounds in more detail.

Using Theorem 9, we know that every safe path can be represented as an edge with a subpath of its left and right extensions. Now, in the worst case each of the  $m$  edges can have left and right extensions of length  $O(n)$  each, making a complete path of  $O(n)$  size. Using the two pointer algorithm we know, that the number of maximal safe paths on a path  $p$  are  $|p|$ . Hence, for each edge we can have  $O(n)$  safe paths, each of possibly  $O(n)$  edges in the worst case.

This establishes an upper bound of  $O(mn^2)$  on the size of  $out_R$  matching the  $\Omega(mn^2)$  bound of [11], proving the tight bound of  $\Theta(mn^2)$  on  $out_R$  in the worst case. Further, since each edge with its extensions creates a valid path for  $out_C$ , having  $O(n)$  length and  $O(n)$  indices on every path, we also get  $O(mn)$  bound for  $out_C$  and  $out_O$ , resulting in tight worst case bound of  $\Theta(mn)$  for both  $out_C$  and  $out_O$ .

## 8 Conclusion

We study the optimization of the solutions for the safety of flow paths in a given flow graph (DAG), which has applications in various domains, including the more prominent assembly of biological sequences. The previous work characterized such paths giving an optimal verification algorithm but suboptimal enumeration algorithms, which required computing a candidate flow decomposition taking  $\Omega(mn)$  time even when the reported solution is small.

We present output-sensitive optimal algorithms for reporting the safe paths when represented in the raw format reporting each path explicitly, and optimal concise representation previously described. This is achieved by exploiting a crucial property related to the interaction of safe paths and a novel data structure Path Tries, which may be of independent interest. Further, we characterized an optimal representation of the safe paths, requiring  $O(1)$  space for every safe path reported. We also presented a near optimal algorithm to

compute the optimal representation of the safe paths. The new characterization additionally allows us to understand the space bounds of various representations of all safe paths, where we match the existing lower bounds with worst case upper bounds.

In the future, it would be interesting to see an optimal output-sensitive algorithm for computing even the optimal representation of the safe paths (dropping the  $O(\log n)$  factor). It would also be interesting to see if similar properties or algorithms can be used to solve related problems as path covers, or path decomposition for general graphs.

---

## References

- 1 Jasmijn A. Baaijens, Bastiaan Van der Roest, Johannes Köster, Leen Stougie, and Alexander Schönhuth. Full-length de novo viral quasispecies assembly through variation graph construction. *Bioinform.*, 35(24):5086–5094, 2019. doi:10.1093/bioinformatics/btz443.
- 2 Jasmijn A. Baaijens, Leen Stougie, and Alexander Schönhuth. Strain-aware assembly of genomes from mixed samples using flow variation graphs. In *Research in Computational Molecular Biology - 24th Annual International Conference, RECOMB 2020, Padua, Italy, May 10-13, 2020, Proceedings*, pages 221–222, 2020.
- 3 Georg Baier, Ekkehard Köhler, and Martin Skutella. The k-splittable flow problem. *Algorithmica*, 42(3-4):231–248, 2005. doi:10.1007/s00453-005-1167-9.
- 4 Michael A. Bender and Martin Farach-Colton. The level ancestor problem simplified. *Theor. Comput. Sci.*, 321(1):5–12, 2004. doi:10.1016/j.tcs.2003.05.002.
- 5 Elsa Bernard, Laurent Jacob, Julien Mairal, and Jean-Philippe Vert. Efficient RNA isoform identification and quantification from rna-seq data with network flows. *Bioinform.*, 30(17):2447–2455, 2014. doi:10.1093/bioinformatics/btu317.
- 6 Rami Cohen, Liane Lewin-Eytan, Joseph Seffi Naor, and Danny Raz. On the effect of forwarding table size on sdn network utilization. In *IEEE INFOCOM 2014-IEEE conference on computer communications*, pages 1734–1742. IEEE, 2014.
- 7 D. R. Ford and D. R. Fulkerson. *Flows in Networks*. Princeton University Press, USA, 2010.
- 8 Thomas Gatter and Peter F Stadler. Ryūtō: network-flow based transcriptome reconstruction. *BMC bioinformatics*, 20(1):190, 2019. doi:10.1186/s12859-019-2786-5.
- 9 Tzvika Hartman, Avinatan Hassidim, Haim Kaplan, Danny Raz, and Michal Segalov. How to split a flow? In *2012 Proceedings IEEE INFOCOM*, pages 828–836. IEEE, 2012.
- 10 Chi-Yao Hong, Srikanth Kandula, Ratul Mahajan, Ming Zhang, Vijay Gill, Mohan Nanduri, and Roger Wattenhofer. Achieving high utilization with software-driven wan. In *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*, pages 15–26, 2013.
- 11 Shahbaz Khan, Milla Kortelainen, Manuel Cáceres, Lucia Williams, and Alexandru I. Tomescu. Safety and completeness in flow decompositions for RNA assembly. In *26th Annual International Conference, RECOMB 2022, San Diego, CA, USA, May 22-25, 2022*, pages 177–192, 2022.
- 12 Kyle Kloster, Philipp Kuinke, Michael P O’Brien, Felix Reidl, Fernando Sánchez Villaamil, Blair D Sullivan, and Andrew van der Poel. A practical fpt algorithm for flow decomposition and transcript assembly. In *2018 Proceedings of the Twentieth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 75–86. SIAM, 2018.
- 13 Cong Ma, Hongyu Zheng, and Carl Kingsford. Exact transcript quantification over splice graphs. In *20th International Workshop on Algorithms in Bioinformatics, WABI 2020, September 7-9, 2020, Pisa, Italy (Virtual Conference)*, pages 12:1–12:18, 2020.
- 14 Veli Mäkinen, Djamel Belazzougui, Fabio Cunial, and Alexandru I. Tomescu. *Genome-Scale Algorithm Design: Biological Sequence Analysis in the Era of High-Throughput Sequencing*. Cambridge University Press, 2015. doi:10.1017/CB09781139940023.
- 15 Brendan Mumey, Samareh Shahmohammadi, Kathryn McManus, and Sean Yaw. Parity balancing path flow decomposition and routing. In *2015 IEEE Globecom Workshops (GC Wkshps)*, pages 1–6. IEEE, 2015.

- 16 Jan Peter Ohst. *On the Construction of Optimal Paths from Flows and the Analysis of Evacuation Scenarios*. PhD thesis, University of Koblenz and Landau, Germany, 2015.
- 17 Nils Olsen, Natalia Kliewer, and Lena Wolbeck. A study on flow decomposition methods for scheduling of electric buses in public transport based on aggregated time–space network models. *Central European Journal of Operations Research*, 2020. doi:10.1007/s10100-020-00705-6.
- 18 Mihaela Perteau, Geo M Perteau, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell, and Steven L Salzberg. Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nature biotechnology*, 33(3):290–295, 2015. doi:10.1038/nbt.3122.
- 19 Krzysztof Pieńkosz and Kamil Kołtyś. Integral flow decomposition with minimum longest path length. *European Journal of Operational Research*, 247(2):414–420, 2015. doi:10.1016/j.ejor.2015.06.012.
- 20 Mingfu Shao and Carl Kingsford. Theory and a heuristic for the minimum path flow decomposition problem. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(2):658–670, 2017.
- 21 Vorapong Suppakitpaisarn. An approximation algorithm for multiroute flow decomposition. *Electronic Notes in Discrete Mathematics*, 52:367–374, 2016. INOC 2015 – 7th International Network Optimization Conference.
- 22 Alexandru I. Tomescu, Travis Gagie, Alexandru Popa, Romeo Rizzi, Anna Kuosmanen, and Veli Mäkinen. Explaining a weighted DAG with few paths for solving genome-guided multi-assembly. *IEEE ACM Trans. Comput. Biol. Bioinform.*, 12(6):1345–1354, 2015. doi:10.1109/TCBB.2015.2418753.
- 23 Alexandru I Tomescu, Anna Kuosmanen, Romeo Rizzi, and Veli Mäkinen. A novel min-cost flow method for estimating transcript expression with rna-seq. *BMC bioinformatics*, 14(S5):S15, 2013. doi:10.1186/1471-2105-14-S5-S15.
- 24 Alexandru I. Tomescu and Paul Medvedev. Safe and complete contig assembly through omnitigs. *Journal of Computational Biology*, 24(6):590–602, 2017. Preliminary version appeared in RECOMB 2016.
- 25 Benedicte Vatinlen, Fabrice Chauvet, Philippe Chrétienne, and Philippe Mahey. Simple bounds and greedy algorithms for decomposing a flow into a minimal set of paths. *European Journal of Operational Research*, 185(3):1390–1401, 2008. doi:10.1016/j.ejor.2006.05.043.
- 26 Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009. doi:10.1038/nrg2484.
- 27 Lucia Williams, Gillian Reynolds, and Brendan Mumeay. Rna transcript assembly using inexact flows. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1907–1914. IEEE, 2019.
- 28 Hongyu Zheng, Cong Ma, and Carl Kingsford. Deriving ranges of optimal estimated transcript expression due to nonidentifiability. *J. Comput. Biol.*, 29(2):121–139, 2022. doi:10.1089/cmb.2021.0444.