

RESEARCH

Open Access



# Revisiting the complexity of and algorithms for the graph traversal edit distance and its variants

Yutong Qiu<sup>1†</sup>, Yihang Shen<sup>1†</sup> and Carl Kingsford<sup>1\*</sup>

## Abstract

The graph traversal edit distance (GTED), introduced by Ebrahimpour Boroojeny et al. (2018), is an elegant distance measure defined as the minimum edit distance between strings reconstructed from Eulerian trails in two edge-labeled graphs. GTED can be used to infer evolutionary relationships between species by comparing de Bruijn graphs directly without the computationally costly and error-prone process of genome assembly. Ebrahimpour Boroojeny et al. (2018) propose two ILP formulations for GTED and claim that GTED is polynomially solvable because the linear programming relaxation of one of the ILPs always yields optimal integer solutions. The claim that GTED is polynomially solvable is contradictory to the complexity results of existing string-to-graph matching problems. We resolve this conflict in complexity results by proving that GTED is NP-complete and showing that the ILPs proposed by Ebrahimpour Boroojeny et al. do not solve GTED but instead solve for a lower bound of GTED and are not solvable in polynomial time. In addition, we provide the first two, correct ILP formulations of GTED and evaluate their empirical efficiency. These results provide solid algorithmic foundations for comparing genome graphs and point to the direction of heuristics. The source code to reproduce experimental results is available at <https://github.com/Kingsford-Group/gtednewilp/>.

**Keywords** Integer linear programming, Genome graphs, Flow network, Graph comparison

## Introduction

Graph traversal edit distance (GTED) [1] is an elegant measure of the similarity between the strings represented by edge-labeled Eulerian graphs. For example, given two de Bruijn assembly graphs [2], computing GTED between them measures the similarity between two genomes without the computationally intensive and possibly error-prone process of assembling the genomes.

Using an estimation of GTED between assembly graphs of Hepatitis B viruses, Ebrahimpour Boroojeny et al. [1] group the viruses into clusters consistent with their taxonomy. This can be extended to inferring phylogeny relationships in metagenomic communities or comparing heterogeneous disease samples such as cancer. There are several other methods to compute a similarity measure between strings encoded by two assembly graphs [3–6]. GTED has the advantage that it does not require prior knowledge on the type of the genome graph or the complete sequence of the input genomes. The input to the GTED problem is two unidirectional, edge-labeled Eulerian graphs, which are defined as:

**Definition 1** (*Unidirectional, edge-labeled Eulerian Graph*). A unidirectional, edge-labeled Eulerian graph is

<sup>†</sup>Yutong Qiu and Yihang Shen contributed equally to this work.

\*Correspondence:  
Carl Kingsford  
carlk@cs.cmu.edu

<sup>1</sup>Ray and Stephanie Lane Computational Biology Department, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh 15213, PA, USA



a connected directed graph  $G = (V, E, \ell, \Sigma)$ , with node set  $V$ , edge multi-set  $E$ , constant-size alphabet  $\Sigma$ , and single-character edge labels  $\ell : E \rightarrow \Sigma$ , such that  $G$  contains an Eulerian trail that traverses every edge  $e \in E$  exactly once. The unidirectional condition means that all edges between the same pair of nodes are in the same direction.

Such graphs arise in genome assembly problems (e.g. the de Bruijn subgraphs). Computing GTED is the problem of computing the minimum edit distance between the two most similar strings represented by Eulerian trails in each input graph. A trail in a graph is a walk that contains distinct edges and may contain repeated nodes.

**Problem 1** (*Graph Traversal Edit Distance* (GTED) [1]). Given two unidirectional, edge-labeled Eulerian graphs  $G_1$  and  $G_2$ , compute

$$\text{GTED}(G_1, G_2) \triangleq \min_{\substack{t_1 \in \text{trails}(G_1) \\ t_2 \in \text{trails}(G_2)}} \text{edit}(\text{str}(t_1), \text{str}(t_2)). \quad (1)$$

Here,  $\text{trails}(G)$  is the collection of all Eulerian trails in graph  $G$ ,  $\text{str}(t)$  is a string constructed by concatenating labels on the Eulerian trail  $t = (e_0, e_1, \dots, e_n)$ , and  $\text{edit}(s_1, s_2)$  is the edit distance between strings  $s_1$  and  $s_2$ .

Ebrahimpour Boroojeny et al. [1] claim that GTED is polynomially solvable by proposing an integer linear programming (ILP) formulation of GTED and arguing that the constraints of the ILP make it polynomially solvable. This result, however, conflicts with several complexity results on string-to-graph matching problems. Kupferman and Vardi [7] show that it is NP-complete to determine if a string exactly matches an Eulerian tour in an edge-labeled Eulerian graph. Additionally, Jain et al. [8] show that it is NP-complete to compute an edit distance between a string and strings represented by a labeled graph if edit operations are allowed on the graph. On the other hand, polynomial-time algorithms exist to solve string-to-string alignment [9] and string-to-graph alignment [8] when edit operations on graphs are not allowed.

We resolve the conflict among the results on complexity of graph comparisons by revisiting the complexity of and the proposed solutions to GTED. We prove that computing GTED is NP-complete by reducing from the HAMILTONIAN PATH problem, reaching an agreement with other related results on complexity. Further, we point out with a counter-example that the optimal solution of the ILP formulation proposed by Ebrahimpour Boroojeny et al. [1] does not solve GTED.

We give two ILP formulations for GTED. The first ILP has an exponential number of constraints and can be solved by subtour elimination iteratively [10, 11]. The

second ILP has a polynomial number of constraints and shares a similar high-level idea of the global ordering approach [11] in solving the TRAVELING SALESMAN problem [12].

In Qiu and Kingsford [13], Flow-GTED (FGTED), a variant of GTED is proposed to compare two sets of strings (instead of two strings) encoded by graphs. FGTED is equal to the edit distance between the most similar sets of strings spelled by the decomposition of flows between a pair of predetermined source and sink nodes. The similarity between the sets of strings reconstructed from the flow decomposition is measured by the Earth Mover's Edit Distance [13, 14]. FGTED is used to compare pan-genomes, where both the frequency and content of strings are essential to represent the population of organisms. Qiu and Kingsford [13] reduce FGTED to GTED, and via the claimed polynomial-time algorithm of GTED, argued that FGTED is also polynomially solvable. We show that this claim is false by proving that FGTED is also NP-complete.

While the optimal solution to the ILP proposed in Ebrahimpour Boroojeny et al. [1] does not solve GTED, it does compute a lower bound to GTED that we call closed-trail cover traversal edit distance (CCTED). We characterize the cases when GTED is equal to CCTED. In addition, we point out that solving this ILP formulation finds a minimum-cost matching between closed-trail decompositions in the input graphs, which may be used to compute the similarity between repeats in the genomes. Ebrahimpour Boroojeny et al. [1] claim their proposed ILP formulation is solvable in polynomial time by arguing that the constraint matrix of the linear relaxation of the ILP is always totally unimodular. We show that this claim is false by proving that the constraint matrix is not always totally unimodular and showing that there exists optimal fractional solutions to its linear relaxation. We prove that, in fact, CCTED is also NP-complete.

We evaluate the efficiency of solving ILP formulations for GTED and CCTED on simulated genomic strings and show that it is impractical to compute GTED or CCTED on larger genomes.

In summary, we revisit two important problems in genome graph comparisons: Graph Traversal Edit Distance (GTED) and its variant FGTED. We show that both GTED and FGTED are NP-complete, and provide the first correct ILP formulations for GTED. We also show that the ILP formulation proposed by Ebrahimpour Boroojeny et al. [1], i.e. CCTED, is a lower bound to GTED and is also NP-complete. We evaluate the efficiency of the ILPs for GTED and CCTED on genomic sequences. These results provide solid algorithmic foundations for continued algorithmic innovation on the task

of comparing genome graphs and point to the direction of heuristics that estimate GTED and CCTED efficiently.

### GTED and FGTED are NP-complete

#### Conflicting results on computational complexity of GTED and string-to-graph matching

The natural decision versions of all of the computational problems described above and below are clearly in NP. Under the assumption that  $P \neq NP$ , the results on the computational complexity of GTED and string-to-graph matching claimed in Boroojeny et al. [1] and Kupferman and Vardi [7], respectively, cannot be both true.

Kupferman and Vardi [7] show that the problem of determining if an input string can be spelled by concatenating edge labels in an Eulerian trail in an input graph is NP-complete. We call this problem EULERIAN TRAIL EQUALING WORD. We show in Theorem 1 that we can reduce ETEW to GTED, and therefore if GTED is polynomially solvable, then ETEW is polynomially solvable. The complete proof is in Appendix "Reduction from Hamiltonian Path to GTED".

**Problem 2** (Eulerian Trail Equaling Word [7]). Given a string  $s \in \Sigma^*$ , an edge-labeled Eulerian graph  $G$ , find an Eulerian trail  $t$  of  $G$  such that  $\text{str}(t) = s$ .

**Theorem 1** If  $\text{GTED} \in P$  then  $\text{ETEW} \in P$ .

**Proof sketch** We first convert an input instance  $\langle s, G \rangle$  of ETEW into an input instance  $\langle G_1, G_2 \rangle$  to GTED by (a) creating graph  $G_1$  that only contains edges that reconstruct string  $s$  and (b) modifying  $G$  into  $G_2$  by extending the anti-parallel edges so that  $G_2$  is unidirectional. We show that if  $\text{GTED}(G_1, G_2) = 0$ , there must be an Eulerian trail in  $G$  that spells  $s$ , and if  $\text{GTED}(G_1, G_2) > 0$ ,  $G$  must not contain an Eulerian trail that spells  $s$ .  $\square$

Hence, an (assumed) polynomial-time algorithm for GTED solves ETEW in polynomial time. This contradicts Theorem 6 of Kupferman and Vardi [7] of the NP-completeness of ETEW (under  $P \neq NP$ ).

#### Reduction from Hamiltonian Path to GTED and FGTED

We resolve the contradiction by showing that GTED is NP-complete. The details of the proof are in Appendix Reduction from Hamiltonian Path to GTED.

**Theorem 2** GTED is NP-complete.

**Proof sketch** We reduce from the HAMILTONIAN PATH problem, which asks whether a directed, simple graph  $G$  contains a path that visits every vertex exactly once. Here simple means no self-loops or parallel edges. The

reduction is almost identical to that presented in Kupferman and Vardi [7], and from here until noted later in the proof the argument is identical except for the technicalities introduced to force unidirectionality (and another minor change described later).

Let  $\langle G = (V, E) \rangle$  be an instance of HAMILTONIAN PATH, with  $n = |V|$  vertices. We first create the Eulerian closure of  $G$ , which is defined as  $G' = (V', E')$  where

$$V' = \{v^{in}, v^{out} : v \in V\} \cup \{w\}. \tag{2}$$

Here, each vertex in  $V$  is split into  $v^{in}$  and  $v^{out}$ , and  $w$  is a newly added vertex.  $E'$  is the union of the following sets of edges and their labels:

- $E_1 = \{(v^{in}, v^{out}) : v \in V\}$ , labeled  $a$ ,
- $E_2 = \{(u^{out}, v^{in}) : (u, v) \in E\}$ , labeled  $b$ ,
- $E_3 = \{(v^{out}, v^{in}) : v \in V\}$ , labeled  $c$ ,
- $E_4 = \{(v^{in}, u^{out}) : (u, v) \in E\}$ , labeled  $c$ ,
- $E_5 = \{(u^{in}, w) : u \in V\}$ , labeled  $c$ ,
- $E_6 = \{(w, u^{in}) : u \in V\}$ , labeled  $b$ .

$G'$  is an Eulerian graph by construction but contains anti-parallel edges. We further create  $G''$  from  $G'$  by adding dummy nodes so that each pair of anti-parallel edges is split into two parallel, length-2 paths with labels  $x\#$ , where  $x$  is the original label.

We also create a graph  $C$  that has the same number of edges as  $G''$  and spells out a string

$$q = a\#(b\#a\#)^{n-1}(c\#)^{2n-1}(c\#b\#)^{|E|+1}. \tag{3}$$

We then argue that  $G$  has a Hamiltonian path if and only if  $G''$  spells out the string  $q$ , which uses the same line of arguments and graph traversals as in Kupferman and Vardi [7]. We then show that  $\text{GTED}(G'', C) = 0$  if and only if  $G''$  spells  $q$ .  $\square$

Following a similar argument, we show that FGTED is also NP-complete, and its proof is in Appendix FGTED is NP-complete.

**Theorem 3** FGTED is NP-complete.

### Revisiting the correctness of the proposed ILP solutions to GTED

In this section, we revisit two proposed ILP solutions to GTED by Boroojeny et al. [1] and show that the optimal solution to these ILP is not always equal to GTED.

#### Alignment graph

The previously proposed ILP formulations for GTED are based on the alignment graph constructed from input

graphs. The high-level concept of an alignment graph is similar to the dynamic programming matrix for the string-to-string alignment problem [9].

**Definition 2** (Alignment graph). Let  $G_1, G_2$  be two unidirectional, edge-labeled Eulerian graphs. The alignment graph  $\mathcal{A}(G_1, G_2) = (V, E, \delta)$  is a directed graph that has vertex set  $V = V_1 \times V_2$  and edge multi-set  $E$  that equals the union of the following:

- Vertical edges  $[(u_1, u_2), (v_1, v_2)]$  for  $(u_1, v_1) \in E_1$  and  $u_2 \in V_2$ ,
- Horizontal edges  $[(u_1, u_2), (u_1, v_2)]$  for  $u_1 \in V_1$  and  $(u_2, v_2) \in E_2$ ,
- Diagonal edges  $[(u_1, u_2), (v_1, v_2)]$  for  $(u_1, v_1) \in E_1$  and  $(u_2, v_2) \in E_2$ .

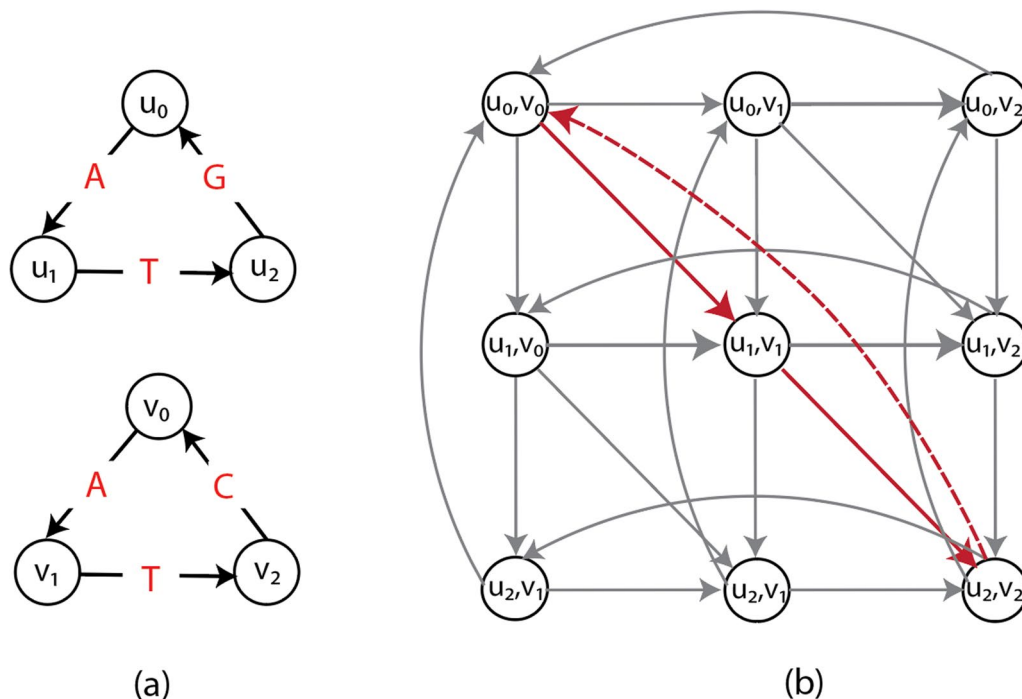
Each edge is associated with a cost by the cost function  $\delta : E \rightarrow \mathbb{R}$ .

Each diagonal edge  $e = [(u_1, u_2), (v_1, v_2)]$  in an alignment graph can be projected to  $(u_1, v_1)$  and  $(u_2, v_2)$  in  $G_1$  and  $G_2$ , respectively. Similarly, each vertical edge can be

projected to one edge in  $G_1$ , and each horizontal edge can be projected to one edge in  $G_2$ .

An example of an alignment graph is shown in Fig. 1b. The horizontal edges correspond to gaps in strings represented by  $G_1$ , vertical edges correspond to gaps in strings represented by  $G_2$ , and diagonal edges correspond to the matching between edge labels from the two graphs. In the rest of this paper, we assume that the costs for horizontal and vertical edges are 1, and the costs for the diagonal edges are 1 if the diagonal edge represents a mismatch and 0 if it is a match. The cost function  $\delta$  can be defined to capture the cost of matching between edge labels or inserting gaps. This definition of alignment graph is also a generalization of the alignment graph used in string-to-graph alignment [8].

We define that the edge projection function,  $\pi_i$ , projects an edge from the alignment graph to an edge in the input graph  $G_i$ . If the alignment edge is a vertical or horizontal edge, it is projected to one edge in only one input graph. We also define that the path projection function,  $\Pi_i$ , projects a trail in the alignment graph to a trail in the input graph  $G_i$ . For example, let a trail in the alignment graph be  $p = (e_1, e_2, \dots, e_m)$ , and  $\Pi_i(p) = (\pi_i(e_1), \pi_i(e_2), \dots, \pi_i(e_m))$  is a trail in  $G_i$ .



**Fig. 1** **a** An example of two edge labeled Eulerian graphs  $G_1$  (top) and  $G_2$  (bottom). **b** The alignment graph  $\mathcal{A}(G_1, G_2)$ . The cycle with red edges is the path corresponding to  $GTED(G_1, G_2)$ . Red solid edges are matches with cost 0 and red dashed-line edge is mismatch with cost 1

**The first previously proposed ILP for GTED**

Lemma 1 in Ebrahimpour Boroojeny et al. [1] provides a model for computing GTED by finding the minimum-cost trail in the alignment graph. We reiterate it here for completeness.

**Lemma 1** [1]. For any two edge-labeled Eulerian graphs  $G_1$  and  $G_2$ ,

$$\begin{aligned} \text{GTED}(G_1, G_2) = \text{minimize}_c \quad & \delta(c) \\ \text{subject to} \quad & c \text{ is a trail in } \mathcal{A}(G_1, G_2), \\ & \Pi_i(c) \text{ is an Eulerian trail in } G_i \text{ for } i = 1, 2, \end{aligned} \tag{4}$$

where  $\delta(c)$  is the total edge cost of  $c$ , and  $\Pi_i(c)$  is the projection from  $c$  to  $G_i$ .

An example of such a minimum-cost trail is shown in Fig. 1b for input graphs in Fig. 1a. Ebrahimpour Boroojeny et al. [1] provide the following ILP formulation and claim that it is a direct translation of Lemma 1:

$$\text{minimize}_{x \in \mathbb{N}^{|E|}} \sum_{e \in E} x_e \delta(e) \tag{5}$$

$$\text{subject to} \quad Ax = 0 \tag{6}$$

$$\sum_{e \in E} x_e I_i(e, f) = 1 \quad \text{for } i = 1, 2 \text{ and for all } f \in E_i, \tag{7}$$

where

$$A_{ue} = \begin{cases} -1 & \text{if } e = (u, v) \in E \text{ for some vertex } v \in V \\ 1 & \text{if } e = (v, u) \in E \text{ for some } u \in V \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Here,  $E$  is the edge set of  $\mathcal{A}(G_1, G_2)$ .  $A$  is the negative incidence matrix of  $\mathcal{A}(G_1, G_2)$  of size  $|V| \times |E|$ .  $I_i(e, f)$  is an indicator function that is 1 if edge  $e$  in  $E$  projects to edge  $f$  in the input graph  $G_i$  (and 0 otherwise). We define the domain of each  $x_e$  to include all non-negative integers. However, due to constraints (7), the values of  $x_e$  are limited to either 0 or 1. We describe this ILP formulation with the assumption that both input graphs have closed Eulerian trails, which means that each node has equal numbers of incoming and outgoing edges. We show that any input graph that contains open Eulerian trails can be converted to a graph with closed Eulerian trails in section "New ILP solutions to GTED".

The edges where  $x_e = 1$  in optimal solutions to the ILP in (5)–(8) form a subgraph of the alignment graph. The ILP in (5)–(8) allows the solutions that form in

disjoint cycles in the alignment graph. However, the projection of such disjoint cycles does not correspond to a single string represented by either of the input graphs. Therefore, when solutions to the ILP in (5)–(8) form disjoint cycles in the alignment graph, the optimal objective value of the ILP is not equal to GTED.

We give an example of disjoint cycles in the ILP formulated based on an input alignment graph. Construct two input graphs as shown in Fig. 2a. Specifically,  $G_1$

spells circular permutations of TTTGAA and  $G_2$  spells circular permutations of TTTAGA. It is clear that  $\text{GTED}(G_1, G_2) = 2$  under Levenshtein edit distance. On the other hand, as shown in Fig. 2a, an optimal solution in  $\mathcal{A}(G_1, G_2)$  contains two disjoint cycles with nonzero  $x_e$  values that have a total edge cost equal to 0. This solution is a feasible solution to the ILP in (5)–(8). It is also an optimal solution because the objective value is zero, which is the lower bound on the ILP in (5)–(8). This optimal objective value, however, is smaller than  $\text{GTED}(G_1, G_2)$ . Therefore, the ILP in (5)–(8) does not solve GTED since it allows the solution to be a set of disjoint components.

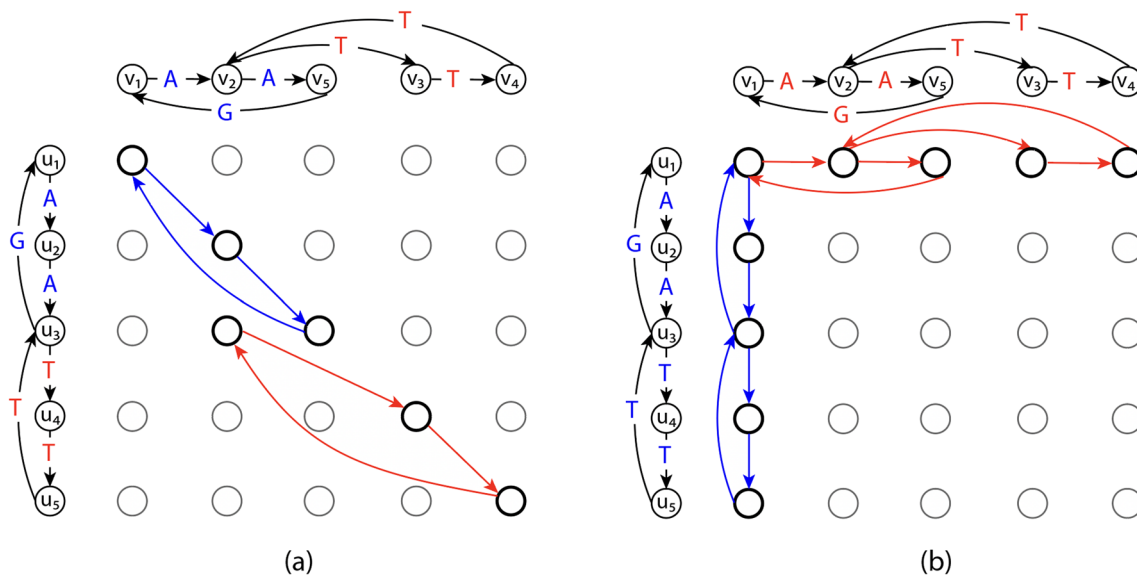
**The second previously proposed ILP formulation of GTED**

We describe the second proposed ILP formulation of GTED by Ebrahimpour Boroojeny et al. [1].

Following Ebrahimpour Boroojeny et al. [1], we use simplices, a notion from geometry, to generalize the notion of an edge to higher dimensions. A  $k$ -simplex is a  $k$ -dimensional polytope which is the convex hull of its  $k + 1$  vertices. For example, a 1-simplex is an undirected edge, and a 2-simplex is a triangle. We use the orientation of a simplex, which is given by the ordering of the vertex set of a simplex up to an even permutation, to generalize the notion of the edge direction [15, p. 26]. We use square brackets  $[\cdot]$  to denote an oriented simplex. For example,  $[v_0, v_1]$  denotes a 1-simplex with orientation  $v_0 \rightarrow v_1$ , which is a directed edge from  $v_0$  to  $v_1$ , and  $[v_0, v_1, v_2]$  denotes a 2-simplex with orientation corresponding to the vertex ordering  $v_0 \rightarrow v_1 \rightarrow v_2 \rightarrow v_0$ . Each  $k$ -simplex has two possible unique orientations, and we use the signed coefficient to connect their forms together, e.g.  $[v_0, v_1] = -[v_1, v_0]$ .

For each pair of graphs  $G_1$  and  $G_2$  and their alignment graph  $\mathcal{A}(G_1, G_2)$ , we define an oriented 2-simplex set  $T(G_1, G_2)$  which is the union of:





**Fig. 2** **a** The subgraph in the alignment graph induced by an optimal solution to the ILP in (5)–(8) and the ILP in (11)–(12) with input graphs on the left and top. The red and blue edges in the alignment graph are edges matching labels in red and blue font, respectively, and are part of the optimal solution to the ILP in (5)–(8). The cost of the red and blue edges are zero. **b** The subgraph induced by  $x^{init}$  with  $s_1 = u_1$  and  $s_2 = v_1$  according to the ILP in (11)–(12). The rest of the edges in the alignment graph are omitted for simplicity

- $[(u_1, u_2), (v_1, u_2), (v_1, v_2)]$  for all  $(u_1, v_1) \in E_1$  and  $(u_2, v_2) \in E_2$ , or
- $[(u_1, u_2), (u_1, v_2), (v_1, v_2)]$  for all  $(u_1, v_1) \in E_1$  and  $(u_2, v_2) \in E_2$ .

We use the boundary operator [15, p. 28], denoted by  $\partial$ , to map an oriented  $k$ -simplex to a sum of oriented  $(k - 1)$ -simplices with signed coefficients.

$$\partial[v_0, v_1, \dots, v_k] = \sum_{i=0}^k (-1)^i [v_0, \dots, \hat{v}_i, \dots, v_k], \quad (9)$$

where  $\hat{v}_i$  denotes the vertex  $v_i$  is to be deleted. Intuitively, the boundary operator maps the oriented  $k$ -simplex to a sum of oriented  $(k - 1)$ -simplices such that their vertices are in the  $k$ -simplex and their orientations are consistent with the orientation of the  $k$ -simplex. For example, when  $k = 2$ , we have:

$$\begin{aligned} \partial[v_0, v_1, v_2] &= [v_1, v_2] - [v_0, v_2] + [v_0, v_1] \\ &= [v_1, v_2] + [v_2, v_0] + [v_0, v_1]. \end{aligned} \quad (10)$$

We reiterate the second ILP formulation proposed in Ebrahimpour Boroojeny et al. [1]. Given an alignment graph  $\mathcal{A}(G_1, G_2) = (V, E, \delta)$  and the oriented 2-simplex set  $T(G_1, G_2)$ ,

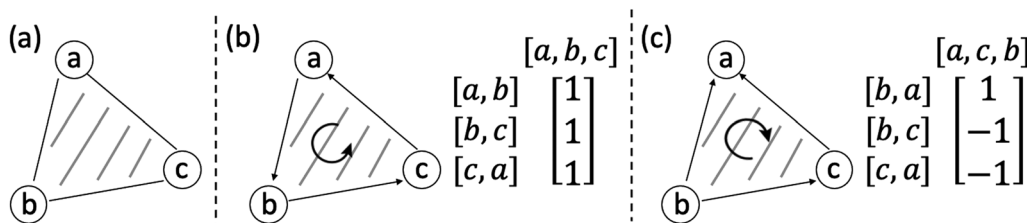
$$\begin{aligned} &\text{minimize} && \sum_{e \in E} x_e \delta(e) \\ &x \in \mathbb{N}^{|E|}, y \in \mathbb{Z}^{|T(G_1, G_2)|} \\ &\text{subject to} && x = x^{init} + [\partial]y \end{aligned} \quad (11)$$

Entries in  $x$  and  $y$  correspond to 1-simplices and 2-simplices in  $E$  and  $T(G_1, G_2)$ , respectively.  $[\partial]$  is a  $|E| \times |T(G_1, G_2)|$  boundary matrix where each entry  $[\partial]_{i,j}$  is the signed coefficient of the oriented 1-simplex (the directed edge) in  $E$  corresponding to  $x_i$  in the boundary of the oriented 2-simplex in  $T(G_1, G_2)$  corresponding to  $y_j$ . The index  $i, j$  for each 1-simplex or 2-simplex is assigned based on an arbitrary ordering of the 1-simplices in  $E$  or the 2-simplices in  $T(G_1, G_2)$ . An example of the boundary matrix is shown in Fig. 3.  $\delta(e)$  is the cost of each edge.  $x^{init} \in \mathbb{R}^{|E|}$  is a vector where each entry corresponds to a 1-simplex in  $E$  with  $|E_1| + |E_2|$  nonzero entries that represent one Eulerian trail in each input graph.  $x^{init}$  is a feasible solution to the ILP. Let  $s_1$  be the source of the Eulerian trail in  $G_1$ , and  $s_2$  be the sink of the Eulerian trail in  $G_2$ . Each entry in  $x^{init}$  is defined by

$$x_e^{init} = \begin{cases} 1 & \text{if } e = [(u_1, s_2), (v_1, s_2)] \text{ or } e = [(s_1, u_2), (s_1, v_2)], \\ 0 & \text{otherwise.} \end{cases} \quad (12)$$

If the Eulerian trail is closed in  $G_i$ ,  $s_i$  can be any vertex in  $V_i$ . An example of  $x^{init}$  is shown in Fig. 2b.

We provide a complete proof in Appendix "Equivalence between two ILPs proposed by [1]" that the ILP in (5)–(8) is equivalent to the ILP in (11)–(12). Therefore, the



**Fig. 3** **a** A graph that contains an unoriented 2-simplex with three unoriented 1-simplices. **b, c** The same graph with two different ways of orienting the simplices and the corresponding boundary matrices

example we provided in section "The first previously proposed ILP for GTED" is also an optimal solution to the ILP in (11)–(12) but not a solution to GTED. Thus, the ILP in (11)–(12) does not always solve GTED.

### New ILP solutions to GTED

To ensure that our new ILP formulations are applicable to input graphs regardless of whether they contain an open or closed Eulerian trail, we add a source node  $s$  and a sink node  $t$  to the alignment graph. Figure 4 illustrates three possible cases of input graphs.

1. If only one of the input graphs has closed Eulerian trails, wlog, let  $G_1$  be the input graph with open Eulerian trails. Let  $a_1$  and  $b_1$  be the start and end of the Eulerian trail that have odd degrees. Add edges  $[s, (a_1, v_2)]$  and  $[(b_1, v_2), t]$  to  $E$  for all nodes  $v_2 \in V_2$  (Fig. 4a). Let the labels on the newly added edges be the empty character  $\epsilon$ .
2. If both input graphs have closed Eulerian trails, let  $a_1$  and  $a_2$  be two arbitrary nodes in  $G_1$  and  $G_2$ , respectively. Add edges  $[s, (a_1, v_2)], [s, (v_1, a_2)], [(a_1, v_2), t]$  and  $[(v_1, a_2), t]$  for all nodes  $v_1 \in V_1$  and  $v_2 \in V_2$  to  $E$  (Fig. 4b).

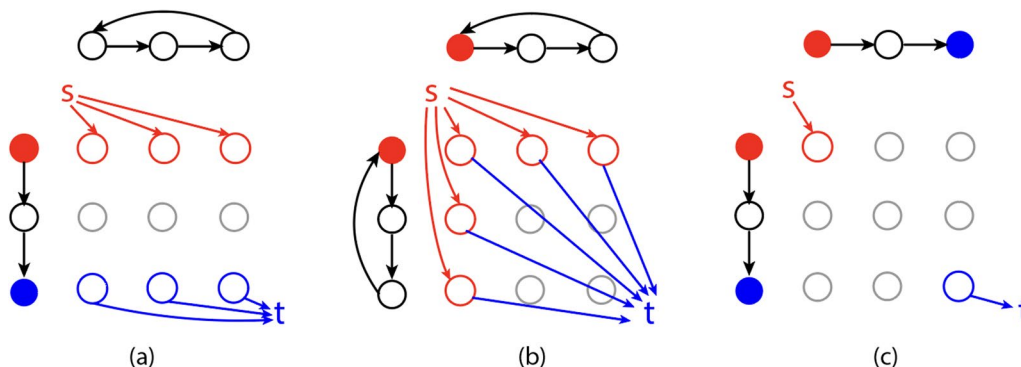
3. If both input graphs have open Eulerian trails, add edges  $[s, (a_1, a_2)]$  and  $[t, (b_1, b_2)]$ , where  $a_i$  and  $b_i$  are start and end nodes of the Eulerian trails in  $G_i$ , respectively (Fig. 4c).

According to Lemma 1, we can solve  $GTED(G_1, G_2)$  by finding a trail in  $\mathcal{A}(G_1, G_2)$  that satisfies the projection requirements. This is equivalent to finding a  $s$ - $t$  trail in  $\mathcal{A}(G_1, G_2)$  that satisfies constraints:

$$\sum_{(u,v) \in E} x_{uv} I_i((u,v), f) = 1 \quad \text{for all } (u,v) \in E, f \in G_i, u \neq s, v \neq t, \tag{13}$$

where  $I_i(e, f) = 1$  if the alignment edge  $e$  projects to  $f$  in  $G_i$ , and  $x_{uv}$  is the ILP variable for edge  $(u, v) \in E$ . An optimal solution to GTED in the alignment graph must start and end with the source and sink node because they are connected to all possible starts and ends of Eulerian trails in the input graphs.

Since a trail in  $\mathcal{A}(G_1, G_2)$  is a flow network, we use the following flow constraints to enforce the equality between the number of in- and out-edges for each node in the alignment graph except the source and sink nodes.



**Fig. 4** Modified alignment graphs based on input types. **a**  $G_1$  has open Eulerian trails while  $G_2$  has closed Eulerian trails. **b** Both  $G_1$  and  $G_2$  have closed Eulerian trails. **c** Both  $G_1$  and  $G_2$  have open Eulerian trails. Solid red and blue nodes are the source and sink nodes of the graphs with open Eulerian trails. "s" and "t" are the added source and sink nodes. Colored edges are added alignment edges directing from and to source and sink nodes, respectively

$$\sum_{(s,u) \in E} x_{su} = 1 \tag{14}$$

$$\sum_{(v,t) \in E} x_{vt} = 1 \tag{15}$$

$$\sum_{(u,v) \in E} x_{uv} = \sum_{(v,w) \in E} x_{vw} \quad \text{for all } v \in V \tag{16}$$

Constraints (13) and (16) are equivalent to constraints (7) and (6), respectively. Therefore, we rewrite the ILP in (5)–(8) in terms of the modified alignment graph.

$$\begin{aligned} & \underset{x \in \mathbb{N}^{|E|}}{\text{minimize}} && \sum_{e \in E} x_e \delta(e) \\ & \text{subject to} && \text{constraints (13)–(16)}. \end{aligned} \tag{lower bound ILP}$$

As we show in section "The first previously proposed ILP for GTED", constraints (13)–(16) do not guarantee that the ILP solution is one trail in  $\mathcal{A}(G_1, G_2)$ , thus allowing several disjoint covering trails to be selected in the solution and fails to model GTED correctly. We show in section "Closed-trail cover traversal edit distance" that the solution to this ILP is a lower bound to GTED.

According to Lemma 1 in Dias et al. [11], a subgraph of a directed graph  $G$  with source node  $s$  and sink node  $t$  is a  $s$ - $t$  trail if and only if it is a flow network and every strongly connected component (SCC) of the subgraph has at least one edge outgoing from it. Thus, in order to formulate an ILP for the GTED problem, it is necessary to devise constraints that prevent disjoint SCCs from being selected in the alignment graph. In the following, we describe two approaches for achieving this.

### Enforcing one trail in the alignment graph via constraint generation

Section 3.2 of Dias et al. [11] proposes a method to design linear constraints for eliminating disjoint SCCs, which can be directly adapted to our problem. Let  $\mathcal{C}$  be the collection of all strongly connected subgraphs of the alignment graph  $\mathcal{A}(G_1, G_2)$ . We use the following constraint to enforce that the selected edges form one  $s$ - $t$  trail in the alignment graph:

$$\text{If } \sum_{(u,v) \in E(C)} x_{uv} = |E(C)|, \text{ then } \sum_{(u,v) \in \varepsilon^+(C)} x_{uv} \geq 1 \quad \text{for all } C \in \mathcal{C}, \tag{17}$$

where  $E(C)$  is the set of edges in the strongly connected subgraph  $C$  and  $\varepsilon^+(C)$  is the set of edges  $(u, v)$  such that  $u$  belongs to  $C$  and  $v$  does not belong to  $C$ .  $\sum_{(u,v) \in E(C)} x_{uv} = |E(C)|$  indicates that  $C$  is in the subgraph of  $\mathcal{A}(G_1, G_2)$  constructed by all edges  $(u, v)$  with positive  $x_{uv}$ , and  $\sum_{(u,v) \in \varepsilon^+(C)} x_{uv} \geq 1$  guarantees that there exists an out-going edge of  $C$  that is in the subgraph.

We use the same technique as Dias et al. [11] to linearize the "if-then" condition in (17) by introducing a new variable  $\beta$  for each strongly connected component:

$$\sum_{(u,v) \in E(C)} x_{uv} \geq |E(C)|\beta_C \quad \text{for all } C \in \mathcal{C} \tag{18}$$

$$\sum_{(u,v) \in E(C)} x_{uv} - |E(C)| + 1 - |E(C)|\beta_C \leq 0 \quad \text{for all } C \in \mathcal{C} \tag{19}$$

$$\sum_{(u,v) \in \varepsilon^+(C)} x_{uv} \geq \beta_C \quad \text{for all } C \in \mathcal{C} \tag{20}$$

$$\beta_C \in \{0, 1\} \quad \text{for all } C \in \mathcal{C} \tag{21}$$

To summarize, given any pair of unidirectional, edge-labeled Eulerian graphs  $G_1$  and  $G_2$  and their alignment graph  $\mathcal{A}(G_1, G_2) = (V, E, \delta)$ , GTED( $G_1, G_2$ ) is equal to the optimal solution of the following ILP formulation:

$$\begin{aligned} & \underset{x \in \{0,1\}^{|E|}}{\text{minimize}} && \sum_{e \in E} x_e \delta(e) \\ & \text{subject to} && \text{constraints (13)–(16) and} \\ & && \text{constraints (18)–(21)}. \end{aligned} \tag{exponential ILP}$$

This ILP has an exponential number of constraints as there is a set of constraints for every strongly connected subgraph in the alignment graph. To solve this ILP more efficiently, we can use the procedure similar to the iterative constraint generation procedure in Dias et al. [11] (Algorithm 1). Initially, solve the ILP with only constraints (13)–(16). Create a subgraph,  $G'$ , induced by edges with positive  $x_{uv}$ . For each disjoint SCC in  $G'$  that does not contain the sink node, add constraints (18)–(21) for edges in the SCC and solve the new ILP. Iterate until no disjoint SCCs are found in the solution.



**Algorithm 1** Iterative constraint generation algorithm to solve (exponential ILP)

---

```

1: Input Two unidirectional, edge-labeled Eulerian graphs and their alignment graph
2:  $\mathcal{C} \leftarrow \emptyset$ 
3: while true do
4:   Solve the ILP (exponential ILP) with  $\mathcal{C}$ 
5:   if the ILP variables  $x_{uv}$  induce a strongly connected component  $C$  not
     satisfying (17) then
6:      $\mathcal{C} = \mathcal{C} \cup \{C\}$ 
7:   else
8:     return the optimal ILP value and the corresponding optimal solution  $x$ 
9:   end if
10: end while

```

---

**A compact ILP for GTED with polynomial number of constraints**

In the worst cases, the number of iterations to solve (exponential ILP) via constraint generation is exponential. As an alternative, we introduce a compact ILP with only a polynomial number of constraints. The intuition behind this ILP is that we can impose a partially increasing ordering on all the edges so that the selected edges forms a  $s$ - $t$  trail in the alignment graph. This idea is similar to the Miller-Tucker-Zemlin ILP formulation of the TRAVELLING SALESMAN problem (TSP) [12].

We add variables  $d_{uv}$  that are constrained to provide a partial ordering of the edges in the  $s$ - $t$  trail and set the variables  $d_{uv}$  to zero for edges that are not selected in the  $s$ - $t$  trail. Intuitively, there must exist an ordering of edges in a  $s$ - $t$  trail such that for each pair of consecutive edges  $(u, v)$  and  $(v, w)$ , the difference in their order variable  $d_{uv}$  and  $d_{vw}$  is 1. Therefore, for each node  $v$  that is not the source or the sink, if we sum up the order variables for the incoming edges and outgoing edges respectively, the difference between the two sums is equal to the number of selected incoming/outgoing edges. Lastly, the order variable for the edge starting at source is 1, and the order variable for the edge ending at sink is the number of selected edges. This gives the ordering constraints as follows:

$$\text{If } x_{uv} = 0, \text{ then } d_{uv} = 0 \text{ for all } (u, v) \in E \tag{22}$$

$$\sum_{(v,w) \in E} d_{vw} - \sum_{(u,v) \in E} d_{uv} = \sum_{(v,w) \in E} x_{vw} \text{ for all } v \in V \setminus \{s, t\} \tag{23}$$

$$\sum_{(s,u) \in E} d_{su} = 1 \tag{24}$$

$$\sum_{(v,t) \in E} d_{vt} = \sum_{(u,v) \in E} x_{uv} \tag{25}$$

We enforce that all variables  $x_e \in \{0, 1\}$  and  $d_e \in \mathbb{N}$  for all  $e \in E$ .

The “if-then” statement in Eq. (22) can be linearized by introducing an additional binary variable  $y_{uv}$  for each edge [11, 16]:

$$-x_{uv} - |E|y_{uv} \leq -1 \tag{26}$$

$$d_{uv} - |E|(1 - y_{uv}) \leq 0 \tag{27}$$

$$y_{uv} \in \{0, 1\}. \tag{28}$$

Here,  $y_{uv}$  is an indicator of whether  $x_{uv} \geq 0$ . The coefficient  $|E|$  is the number of edges in the alignment graph and also an upper bound on the ordering variables. When  $y_{uv} = 1$ ,  $d_{uv} \leq 0$ , and  $y_{uv}$  does not impose constraints on  $x_{uv}$ . When  $y_{uv} = 0$ ,  $x_{uv} \geq 1$ , and  $y_{uv}$  does not impose constraints on  $d_{uv}$ .

**Correctness of (compact ILP) for GTED**

To show that the optimal objective value of (compact ILP) is equal to GTED, we show that the optimal solutions to (compact ILP) always form one connected component.

**Lemma 2** *Let  $x_e$  and  $d_e$  be ILP variables. Let  $G'$  be a subgraph of  $\mathcal{A}(G_1, G_2)$  that is induced by edges with  $x_e = 1$ . If  $x_e$  and  $d_e$  satisfy constraints (13)–(25) for all  $e \in E$ ,  $G'$  is connected with one trail from  $s$  to  $t$  that traverses each edge in  $G'$  exactly once.*

**Proof** We prove the lemma in 2 parts: (1) all nodes except  $s$  and  $t$  in  $G'$  have an equal number of in- and out-edges, (2)  $G'$  contains only one connected component.

The first statement holds because the edges of  $G'$  form a flow from  $s$  to  $t$ , and is enforced by constraints (16).

We then show that  $G'$  does not contain isolated subgraphs that are not reachable from  $s$  or  $t$ . Due to constraint (16), the only possible scenario is that the isolated subgraph is strongly connected. Suppose for contradiction that there is a strongly connected component,  $C$ , in  $G'$  that is not reachable from  $s$  or  $t$ .

The sum of the left hand side of constraint (23) over all vertices in  $C$  is

$$\sum_{v \in C} \left( \sum_{(u,v) \in C} d_{uv} - \sum_{(v,w) \in C} d_{vw} \right) = \sum_{v \in C} \sum_{(u,v) \in C} d_{uv} - \sum_{v \in C} \sum_{(v,w) \in C} d_{vw} \tag{29}$$

$$= \sum_{(u,v) \in E(C)} d_{uv} - \sum_{(v,w) \in E(C)} d_{vw} = 0. \tag{30}$$

However, the right-hand side of the same constraints is always positive. Hence we have a contradiction. Therefore,  $G'$  has only one connected component.  $\square$

Due to Lemma 1 and Lemma 2, given input graphs  $G_1$  and  $G_2$  and the alignment graph  $\mathcal{A}(G_1, G_2)$ , GTED( $G_1, G_2$ ) is equal to the optimal objective of

$$\begin{aligned} & \text{minimize} && \sum_{e \in E} x_e \delta(e) \\ & \text{subject to} && \text{constraints (13)–(16),} \\ & && \text{constraints (23)–(25)} \\ & && \text{and constraints (26)–(28).} \\ & && \text{(compact ILP)} \end{aligned}$$

**Closed-trail cover traversal edit distance**

While the (lower bound ILP) and the ILP in (11)–(12) do not solve GTED, the optimal solution to these ILPs is a lower bound of GTED. These ILP formulations also solve an interesting variant of GTED, which is a local similarity measure between two genome graphs. We call this variant Closed-trail cover traversal edit distance (CCTED). In the following, we provide the formal definition of the CCTED problem and then show that the (lower bound ILP) is the correct ILP formulation for solving CCTED.

We first introduce the min-cost item matching problem between two multi-sets. Let two multi-sets of items be  $S_1$  and  $S_2$ , and, wlog, let  $|S_1| \leq |S_2|$ . Let  $c : (S_1 \cup \{\epsilon\}) \times S_2 \rightarrow \mathbb{N}$  be the cost of matching either an empty item  $\epsilon$  or an item in  $S_1$  with an item in  $S_2$ . Given  $S_1, S_2$  and the cost function  $c$ , min-cost matching problem finds a matching,  $\mathcal{M}_c(S_1, S_2)$ , such that each item in  $S_1 \cup \{\epsilon\}^{|S_2|-|S_1|}$  is matched with exactly one distinct item in  $S_2$  and the total cost of the matching,  $\sum_{(s_1, s_2) \in \mathcal{M}_c(S_1, S_2)} c(s_1, s_2)$ , is minimized.

The min-cost item matching problem is similar to the Earth Mover’s Distance defined in [17], except that only integral units of items can be matched and the cost of matching an empty item with another item is not constant. Similar to the Earth Mover’s Distance, the min-cost item matching problem can be computed using the ILP formulation of the min-cost max-flow problem [13, 14]. When the cost is the edit distance, the cost to match  $\epsilon$  with a string is equal to the length of the string.

Define traversal edit distance,  $edit_t(t_1, t_2)$  as the edit distance between the strings constructed from a pair of trails  $t_1$  and  $t_2$ . In other words,  $edit_t(t_1, t_2) = edit(\text{str}(t_1), \text{str}(t_2))$ . CCTED is defined as:

**Problem 3** (Closed-trail cover traversal edit distance (CCTED)). Given two unidirectional, edge-labeled Eulerian graphs  $G_1$  and  $G_2$  with closed Eulerian trails, compute

$$\begin{aligned} & \text{CCTED}(G_1, G_2) \\ & \triangleq \min_{\substack{C_1 \in \text{CC}(G_1), \\ C_2 \in \text{CC}(G_2)}} \sum_{(t_1, t_2) \in \mathcal{M}_{edit_t}(C_1, C_2)} edit(\text{str}(t_1), \text{str}(t_2)), \end{aligned} \tag{31}$$

Here,  $\text{CC}(G)$  denotes the collection of all possible sets of edge-disjoint, closed trails in  $G$ , such that every edge in  $G$  belongs to exactly one of these trails. Each element of  $\text{CC}(G)$  can be interpreted as a cover of  $G$  using such trails.  $\mathcal{M}_{edit_t}(C_1, C_2)$  is a min-cost matching between two covers using the traversal edit distance as the cost.

CCTED is likely a more suitable metric comparison between genomes that undergo large-scale rearrangements. This analogy is to the relationship between the synteny block comparison [3] and the string edit distance computation, where the former is more often used in interspecies comparisons and in detecting segmental duplications [18, 19] and the latter is more often seen in intraspecies comparisons.

Following similar ideas as Lemma 1, we can compute CCTED by finding a set of closed trails in the alignment graph such that the total cost of alignment edges is minimized, and the projection of all edges in the collection of selected trails is equal to the multi-set of input graph edges.

**Lemma 3** For any two edge-labeled Eulerian graphs  $G_1$  and  $G_2$ ,

$$\text{CCTED}(G_1, G_2) = \underset{C}{\text{minimize}} \sum_{c \in C} \delta(c) \tag{32}$$

subject to  $C$  is a set of closed trails in  $\mathcal{A}(G_1, G_2)$ ,

$$\bigcup_{e \in C} \Pi_i(e) = E_i \quad \text{for } i = 1, 2, \tag{33}$$

where  $C$  is a collection of trails and  $\delta(c)$  is the total cost of edges in trail  $c$ .

**Proof** Given any pair of covers  $C_1 \in CC(G_1)$  and  $C_2 \in CC(G_2)$  and their min-cost matching based on the edit distance  $\mathcal{M}_{\text{edit}_t}(C_1, C_2)$ , we can project each pair of matched closed trails to a closed trail in the alignment graph. For a matching between a trail and the empty item  $\epsilon$ , we can project it to a closed trail in the alignment graph with all vertical edges if the trail is from  $G_1$  or horizontal edges if the trail is from  $G_2$ . The total cost of the projected edges must be greater than or equal to the objective (32). On the other hand, every collection of trails  $C$  that satisfy constraint (33) can be projected to a cover in each of the input graphs, and  $\sum_{c \in C} \delta(c) \geq CCTED(G_1, G_2)$ . Hence equality holds.  $\square$

**The ILP formulation for CCTED**

We show that the ILP in (5)–(8) proposed by Ebrahimipour Boroojeny et al. [1] solves CCTED.

**Theorem 4** Given two input graphs  $G_1$  and  $G_2$ , the optimal objective value of the ILP in (5)–(8) based on  $\mathcal{A}(G_1, G_2)$  is equal to CCTED( $G_1, G_2$ ).

**Proof** As shown in the proof of Lemma 3, any pair of edge-disjoint, closed-trail covers in the input graph can be projected to a set of closed trails in  $\mathcal{A}(G_1, G_2)$ , which satisfied constraints (6)–(8). The objective of this feasible solution, which is the total cost of the projected closed trails, equals CCTED. Therefore, CCTED( $G_1, G_2$ ) is greater than or equal to the objective of the ILP in (5)–(8).

Conversely, we can transform any feasible solutions of the ILP in (5)–(8) to a pair of covers of  $G_1$  and  $G_2$ . We can do this by transforming one closed trail at a time from the subgraph of the alignment graph,  $\mathcal{A}'$  induced by edges with ILP variable  $x_{uv} = 1$ . Let  $c$  be a closed trail in  $\mathcal{A}'$ . Let  $c_1 = \Pi_1(c)$  and  $c_2 = \Pi_2(c)$  be two closed trails in  $G_1$  and  $G_2$  that are projected from  $c$ . We can construct an alignment between  $\text{str}(c_1)$  and  $\text{str}(c_2)$  from  $c$  by adding match or insertion/deletion columns for each match or insertion/deletion edges in  $c$  accordingly. The cost of the alignment is equal to the total cost of edges in  $c$  by the construction of the alignment graph. We can then

remove edges in  $c$  from the alignment graph and edges in  $c_1$  and  $c_2$  from the input graphs, respectively. The remaining edges in  $\mathcal{A}'$  and  $G_1$  and  $G_2$  still satisfy the constraints (6)–(8). Repeat this process and we get a total cost of  $\sum_{e \in E} x_e \delta(e)$  that aligns pairs of closed trails that form covers of  $G_1$  and  $G_2$ . This total cost is greater than or equal to CCTED( $G_1, G_2$ ).  $\square$

**CCTED is a lower bound of GTED**

Since the constraints for (lower bound ILP) are a subset of (exponential ILP), a feasible solution to (exponential ILP) is always a feasible solution to (lower bound ILP). Since two ILPs have the same objective function, CCTED( $G_1, G_2$ )  $\leq$  GTED( $G_1, G_2$ ) for any pair of graphs. Moreover, when the solution to (lower bound ILP) forms only one connected component, the optimal value of (lower bound ILP) is equal to GTED.

**Theorem 5** Let  $\mathcal{A}'(G_1, G_2)$  be the subgraph of  $\mathcal{A}(G_1, G_2)$  induced by edges  $(u, v) \in E$  with  $x_{uv}^{opt} = 1$  in the optimal solution to (lower bound ILP). There exists  $\mathcal{A}'(G_1, G_2)$  that has exactly one connected component if and only if  $c^{opt} = \text{GTED}(G_1, G_2)$ .

**Proof** We first show that if  $c^{opt} = \text{GTED}(G_1, G_2)$ , then there exists  $\mathcal{A}'(G_1, G_2)$  that has one connected component. A feasible solution to (exponential ILP) is always a feasible solution to (lower bound ILP), and since  $c^{opt} = \text{GTED}(G_1, G_2)$ , an optimal solution to (exponential ILP) is also an optimal solution to (lower bound ILP), which can induce a subgraph in the alignment graph that only contains one connected component.

Conversely, if  $x^{opt}$  induces a subgraph in the alignment graph with only one connected component, it satisfies constraints (18)–(21) and therefore is feasible to the ILP for GTED (exponential ILP). Since  $c^{opt} \leq \text{GTED}(G_1, G_2)$ , this solution must also be optimal for GTED( $G_1, G_2$ ).  $\square$

In practice, we may estimate GTED approximately by the solution to (lower bound ILP). As we show in section Empirical evaluation of the ILP formulations for GTED and its lower bound, the time needed to solve (lower bound ILP) is much less than the time needed to solve GTED. However, in adversarial cases,  $c^{opt}$  could be zero but GTED could be arbitrarily large. We can determine if the  $c^{opt}$  is a lower bound on GTED or exactly equal to GTED by checking if the subgraph induced by the solution to (lower bound ILP) has multiple connected components.

### NP-completeness of CCTED

We prove that the CCTED problem (Problem 3) is NP-complete by reducing from the Eulerian Trail Equaling Word problem [7].

**Theorem 6** Computing CCTED is NP-complete.

**Proof** Let Eulerian graph  $G = (V, E, \ell, \Sigma)$  and  $s$  be an instance of the EULERIAN TOUR EQUALING WORD problem. Construct two graphs,  $G_1$  and  $G_2$ . If  $G$  contains open Eulerian trails, add an edge directing from the sink of the graph to the source of the graph. Let the label of the added edge be # that does not appear in  $\Sigma$ . Let the modified graph be  $G_1$ . If  $G$  contains closed Eulerian trails, let  $G_1$  be the same as  $G$ . Let  $G_2$  be a graph that contains one cycle with  $|E_1|$  edges, where  $E_1$  is the edge set of  $G_1$ . Assign labels to the edges in  $G_2$  such that the cycle in  $G_2$  spells  $s$  if  $G$  contains closed Eulerian trails,  $s\#$  otherwise.

If  $\text{CCTED}(G_1, G_2) = 0$ ,  $G_2$  must contain at least one closed Eulerian trail that spells some circular permutation of  $s\#$ . If CCTED is not zero, it means that  $s$  must not match Eulerian trails in  $G$ .  $\square$

### Empirical evaluation of the ILP formulations for GTED and its lower bound

#### Implementation of the ILP formulations

We implement the algorithms and ILP formulations for (exponential ILP), (compact ILP) and (lower bound ILP). In practice, the multi-set of edges of each input graph may contain many duplicates of edges that have the same start and end vertices due to repeats in the strings. We reduce the number of variables and constraints in the implemented ILPs by merging the edges that share the same start and end nodes and record the multiplicity of

each edge. Each  $x$  variable is no longer binary but a non-negative integer that satisfies the modified projection constraints (13):

$$\sum_{(u,v) \in E} x_{uv} I_i((u,v), f) = M_i(f) \tag{34}$$

for all  $(u,v) \in E, f \in G_i, u \neq s, v \neq t$ ,

where  $M_i(f)$  is the multiplicity of edge  $f$  in  $G_i$ . Let  $C$  be the strongly connected component in the subgraph induced by positive  $x_{uv}$ , now  $\sum_{(u,v) \in E(C)} x_{uv}$  is no longer upper bounded by  $|E(C)|$ . Therefore, constraints (19) are changed to

$$\sum_{(u,v) \in E(C)} x_{uv} - |E(C)| + 1 - W(C)\beta_C \leq 0 \quad \text{for all } C \in \mathcal{C},$$

$$W(C) = \sum_{(u,v) \in E(C)} \max \left( \sum_{f \in G_1} M_1(f) I_1((u,v), f), \sum_{f \in G_2} M_2(f) I_2((u,v), f) \right), \tag{35}$$

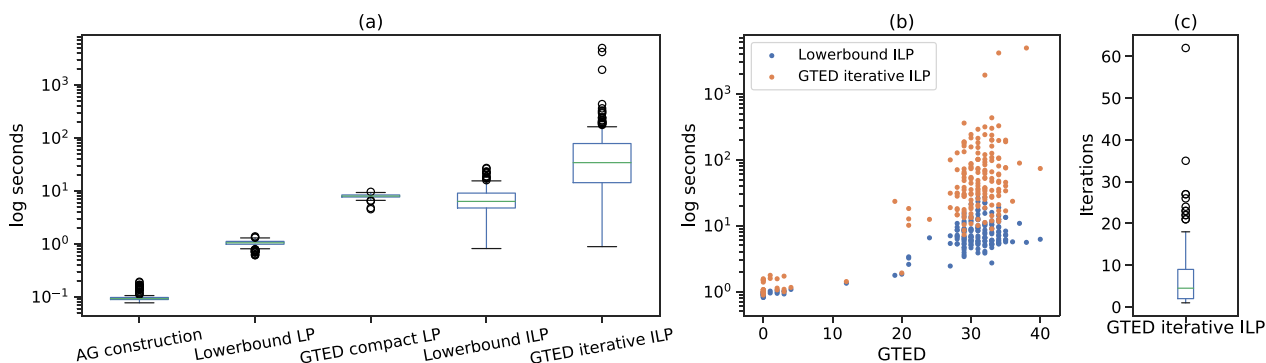
where  $W(C)$  is the maximum total multiplicities of edges in the strongly connected subgraph in each input graph that is projected from  $C$ .

Likewise, constraints (27) that set the upper bounds on the ordering variables also need to be modified as the upper bound of the ordering variable  $d_{uv}$  for each edge no longer represents the order of one edge but the sum of orders of copies of  $(u, v)$  that are selected, which is at most  $|E|^2$ . Therefore, constraint (27) is changed to

$$d_{uv} - |E|^2(1 - y_{uv}) \leq 0. \tag{36}$$

The rest of the constraints remain unchanged.

We ran all our experiments on a server with 48 cores (96 threads) of Intel(R) Xeon(R) CPU E5-2690 v3 @ 2.60GHz and 378 GB of memory. The system was running Ubuntu 18.04 with Linux kernel 4.15.0. We solve all



**Fig. 5** **a** The distribution of wall-clock running time for constructing alignment graphs, solving the ILP formulations for GTED and its lower bound, and their linear relaxations on the log scale. **b** The relationship between the time to solve (lower bound ILP), (exponential ILP) iteratively and GTED. **c** The distribution of the number of iterations to solve exponential ILP. The box plots in each plot show the median (middle line), the first and third quartiles (upper and lower boundaries of the box), the range of data within 1.5 inter-quartile range between Q1 and Q3 (whiskers), and the outlier data points

the ILP formulations and their linear relaxations using the Gurobi solver [20] using 32 threads.

### GTED on simulated TCR sequences

We construct 20 de Bruijn graphs with  $k = 4$  using 150-character sequences extracted from the V genes from the IMGT database [21]. We solve the linear relaxation of (compact ILP), (exponential ILP) and (lower bound ILP) and their linear relaxation on all 190 pairs of graphs. We do not show results for solving (compact ILP) for GTED on this set of graphs as the running time exceeds 30 min on most pairs of graphs.

To compare the time to solve the ILP formulations when GTED is equal to the optimal objective of (lower bound ILP), we only include 168 out of 190 pairs where GTED is equal to the lower bound (GTED is slightly higher than the lower bound in the remaining 22 pairs). On average, it takes 26 s wall-clock time to solve (lower bound ILP), and 71 s to solve (exponential ILP) using the iterative algorithm. On average, it takes 9 s to solve the LP relaxation of (compact ILP) and 1 s to solve the LP relaxation of (lower bound ILP). The time to construct the alignment graph for all pairs is less than 0.2 s. The distribution of wall-clock running time is shown in Fig. 5a. The time to solve (exponential ILP) and (lower bound ILP) is generally positively correlated with the GTED values (Fig. 5b). On average, it takes 7 iterations for the iterative algorithm to find the optimal solution that induces one strongly connected subgraph (Fig. 5c).

In summary, it is fastest to compute the lower bound of GTED. Computing GTED exactly by solving the proposed ILPs on genome graphs of size 150 is already time consuming. When the sizes of the genome graphs are fixed, the time to solve for GTED and its lower bound increases as GTED between the two genome graphs increases. In the case where GTED is equal to its lower

bound, the subgraph induced by some optimal solutions of (lower bound ILP) contains more than one strongly connected component. Therefore, in order to reconstruct the strings from each input graph that have the smallest edit distance, we generally need to obtain the optimal solution to the ILP for GTED. In all cases, the time to solve the (exponential ILP) is less than the time to solve the (compact ILP).

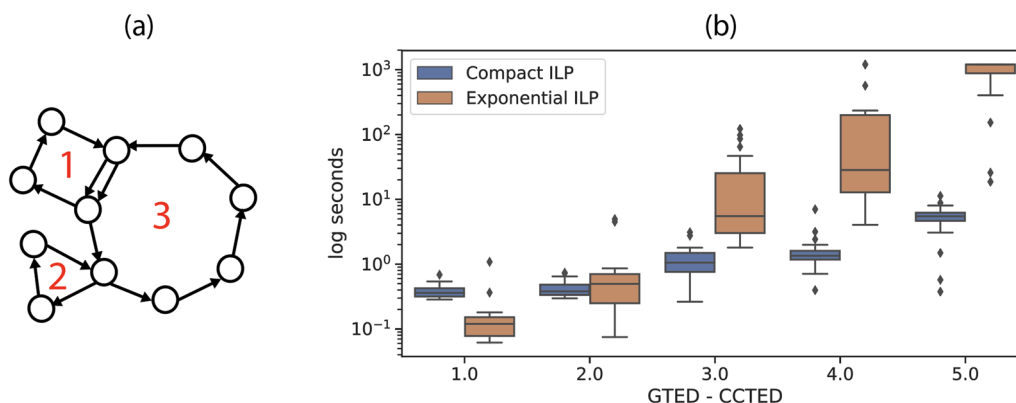
### GTED on difficult cases

Repeats, such as segmental duplications and translocations [22, 23] in the genomes increase the complexity of genome comparisons. We simulate such structures with a class of graphs that contain  $n$  simple cycles of which  $n - 1$  peripheral cycles are attached to the  $n$ -th central cycle at either a node or a set of edges (Fig. 6a). The input graphs in Fig. 2 belong to this class of graphs that contain 2 cycles. This class of graphs simulates the complex structural variants in disease genomes or the differences between genomes of different species.

We generate pairs of 3-cycle graphs with varying sizes and randomly assign letters from  $\{A, T, C, G\}$  to edges. We compute the CCTED and GTED using (lower

**Table 1** The average wall-clock time to solve lower bound ILP, exponential ILP, compact ILP and the number of iterations for pairs of 3-cycle graphs for each GTED – CCTED

GTED - CCTED	lower bound ILP runtime (s)	GTED iterative runtime (s)	Iterations	GTED compact runtime (s)
1.0	0.06	0.17	3.55	0.39
2.0	0.05	0.87	13.00	0.43
3.0	0.08	25.41	67.60	1.24
4.0	0.07	205.59	179.10	1.70
5.0	0.08	943.68	502.85	5.37



**Fig. 6** **a** An example of a 3-cycle graph. Cycle 1 and 2 are attached to cycle 3. **b** The distribution of wall-clock time to solve the compact ILP and the iterative exponential ILP on 100 pairs of 3-cycle graphs



bound ILP) and (compact ILP), respectively. We group the generated 3-cycle graph pairs based on the value of (GTED – CCTED) and select 20 pairs of graphs randomly for each (GTED – CCTED) value ranging from 1 to 5. The maximum number of edges in all selected graphs is 32.

We show the difficulty of computing GTED using the iterative algorithm on the 100 selected pairs of 3-cycle graphs. We terminate the ILP solver after 20 min. As shown in Fig. 6, as the difference between GTED and CCTED increases, the wall-clock time to solve (exponential ILP) for GTED increases faster than the time to solve (compact ILP) for GTED. For pairs of graphs with (GTED – CCTED) = 5, on average it takes more than 15 min to solve (exponential ILP) with more than 500 iterations. On the other hand, it takes an average of 5 s to solve (compact ILP) for GTED and no more than 1 s to solve for the lower bound. The average time to solve each ILP is shown in Table 1.

In summary, on the class of 3-cycle graphs introduced above, the difficulty of solving GTED via the iterative algorithm increases rapidly as the gap between GTED and CCTED increases. Although (exponential ILP) is solved more quickly than (compact ILP) for GTED when the sequences are long and the GTED is equal to CCTED (Section 6.2), (compact ILP) may be more efficient when the graphs contain overlapping cycles such that the gap between GTED and CCTED is larger.

## Conclusion

We point out the contradictions in the result on the complexity of labeled graph comparison problems and resolve the contradictions by showing that GTED, as opposed to the results in Ebrahimpour Boroojeny et al. [1], is NP-complete. On one hand, this makes GTED a less attractive measure for comparing graphs since it is unlikely that there is an efficient algorithm to compute the measure. On the other hand, this result better explains the difficulty of finding a truly efficient algorithm for computing GTED exactly. In addition, we show that the previously proposed ILP of GTED [1] does not solve GTED and give two new ILP formulations of GTED.

While the previously proposed ILP of GTED does not solve GTED, it solves for a lower bound of GTED, and we show that this lower bound can be interpreted as a more “local” measure, CCTED, of the distance between labeled graphs. Further, we characterize the LP relaxation of the ILP in (11)–(12) and show that, contrary to the results in Ebrahimpour Boroojeny et al. [1], the LP in (11)–(12) does not always yield optimal integer solutions.

As shown previously [1, 13], it takes more than 4 h to solve (lower bound ILP) for graphs that represent

viral genomes that contain  $\approx 3000$  bases with a multi-threaded LP solver. Likewise, we show that computing GTED using either (exponential ILP) or (compact ILP) is already slow on small genomes, especially on pairs of simulated genomes that are different due to segmental duplications and translations. The empirical results show that it is currently impossible to solve GTED or CCTED, its lower bound, directly using this approach for bacterial- or eukaryotic-sized genomes on modern hardware. The results here should increase the theoretical interest in GTED along the directions of heuristics or approximation algorithms as justified by the NP-hardness of finding GTED.

## Appendix 1 Proofs for the NP-completeness of GTED

### Reduction from ETEW to GTED

We provide below the complete proof for Theorem 1.

**Theorem 1** *If*  $GTED \in P$  *then*  $ETEW \in P$ .

*Proof* Let  $(s, G)$  be an instance of ETEW. Construct a directed, acyclic graph (DAG),  $C$ , that has only one path. Let the path in  $C$  be  $P = (e_1, \dots, e_{|s|})$  and the edge label of  $e_i$  be  $s[i]$ . Clearly,  $C$  is a unidirectional, edge-labeled Eulerian graph,  $P$  is the only Eulerian trail in  $C$ , and  $str(P) = s$ .

For the graph  $G = (V_G, E_G, \ell_G, \Sigma)$  from the ETEW instance, which may not be unidirectional, create another graph  $G'$  that contains all of the nodes and edges in  $G$  except the anti-parallel edges. Let  $\Sigma_{G'} = \Sigma \cup \{\epsilon\}$ , where  $\epsilon$  is a character that is not in  $\Sigma$ . For each pair of anti-parallel edges  $(u, v)$  and  $(v, u)$  in  $G$ , add four edges  $(u, w_1), (w_1, v), (v, w_2), (w_2, u)$  by introducing new vertices  $w_1, w_2$  to  $G'$ . Let  $\ell_{G'}(u, w_1) = \ell_G(u, v)$  and  $\ell_{G'}(w_2, u) = \ell_G(v, u)$ . Let  $\ell_{G'}(w_1, v) = \ell_{G'}(v, w_2) = \epsilon$  for every newly introduced vertex.  $G'$  has at most twice the number of edges as  $G$  and is Eulerian and unidirectional.

Define the cost of changing a character from  $a$  to  $b$   $\text{cost}(a, b)$  for  $a, b \in \Sigma \cup \{-\}$  to be 0 if  $a = b$  and 1 otherwise. “-” is the gap character indicating an insertion or a deletion. Define  $\text{cost}(a, \epsilon)$  with  $a \in \Sigma$  to be 1. Define  $\text{cost}(-, \epsilon)$  to be 0.

Use the (assumed) polynomial-time algorithm for GTED to ask whether  $GTED(C, G') \leq 0$  under edit distance  $\Sigma$ . If yes, then let  $(s_1, s_2)$  be the 0-cost alignment of the strings spelled out by the trails in  $C$  and  $G'$ , respectively. The non-gap characters of  $s_1$  must spell out  $s$  since there is only one Eulerian trail in  $C$ . Because the alignment cost is 0, any - (gap) characters in  $s_1$  must be

aligned with  $\epsilon$  characters in  $s_2$  and any non-gap characters in  $s_1$  must be aligned to the same character in  $s_2$ . The trail in  $G'$  that spells  $s_2$  can be transformed to a trail that spells  $s_3$  by collapsing the edges with  $\epsilon$  character labels, and  $s_3 = s_1$ .

If  $GTED(C, G') > 0$ ,  $G$  must not contain an Eulerian trail that spells  $s$ . Otherwise, such a trail could be extended to a trail introducing some  $\epsilon$  characters that could be aligned to  $s$  with zero cost by aligning gaps with  $\epsilon$  characters.

Hence, an (assumed) polynomial-time algorithm for GTED solves ETEW in polynomial time.  $\square$

**Reduction from Hamiltonian Path to GTED**

We provide below the complete proof for Theorem 2.

**Theorem 2** *GTED is NP-complete.*

*Proof* We reduce from the HAMILTONIAN PATH problem, which asks whether a directed, simple graph  $G$  contains a path that visits every vertex exactly once. Here simple means no self-loops or parallel edges. Let  $\langle G = (V, E) \rangle$  be an instance of HAMILTONIAN PATH, with  $n = |V|$  vertices. The reduction is almost identical to that presented in Kupferman and Vardi [7], and from here until noted later in the proof the argument is identical except for the technicalities introduced to force unidirectionality (and another minor change described later). The first step is to construct the Eulerian closure of  $G$ , which is defined as  $G' = (V', E')$  where

$$V' = \{v^{in}, v^{out} : v \in V\} \cup \{w\}, \tag{37}$$

and  $E'$  is the union of the following sets of edges and their labels:

- $E_1 = \{(v^{in}, v^{out}) : v \in V\}$ , labeled  $a$ ,
- $E_2 = \{(u^{out}, v^{in}) : (u, v) \in E\}$ , labeled  $b$ ,
- $E_3 = \{(v^{out}, v^{in}) : v \in V\}$ , labeled  $c$ ,
- $E_4 = \{(v^{in}, u^{out}) : (u, v) \in E\}$ , labeled  $c$ ,
- $E_5 = \{(u^{in}, w) : u \in V\}$ , labeled  $c$ ,
- $E_6 = \{(w, u^{in}) : u \in V\}$ , labeled  $b$ .

Since  $G'$  is connected and every outgoing edge in  $G'$  has a corresponding antiparallel incoming edge,  $G'$  is Eulerian. It is not unidirectional, so we further create  $G''$  from  $G'$  by adding dummy nodes to each pair of antiparallel edges and labelling the length-2 paths so created with  $x\#$ , where  $x$  is the original label of the split edge ( $a$ ,  $b$ , or  $c$ ) and  $\#$  is some new symbol (shared between all the new edges). We call these length-2 paths introduced to achieve unidirectionality “split edges”.

We now argue that  $G$  has a Hamiltonian path iff  $G''$  has an Eulerian trail that spells out

$$q = a\#(b\#a\#)^{n-1}(c\#)^{2n-1}(c\#b\#)^{|E|+1}. \tag{38}$$

If such an Eulerian trail exists, then the trail starts with spelling the string  $a\#(b\#a\#)^{n-1}$ , which corresponds to a Hamiltonian trail in  $G$  since it visits exactly  $n$  “vertex split edges” (type  $E_1$ , labeled  $a\#$ ) and each vertex split edge can be used only once (since it is an Eulerian trail). Further, successively visited vertices must be connected by an edge in  $G$  since those are the only  $b\#$  split edges in  $G''$  (except those leaving  $w$ , but  $w$  must not be involved in spelling out  $a\#(b\#a\#)^{n-1}$ , since entering  $w$  requires using a split edge labeled  $c\#$ ).

For the other direction, if a  $G$  has a Hamiltonian path  $v_1, \dots, v_n$ , then walking that sequence of vertices in  $G''$  will spell out  $a\#(b\#a\#)^{n-1}$ . This path will cover all  $E_1$  edges and the  $E_2$  edges that are on the Hamiltonian path. Retracing the path so far in reverse will use  $2n - 1$  split edges labeled  $c\#$ , consuming the  $(c\#)^{2n-1}$  term in  $q$  and covering all nodes’ reverse vertex edges  $E_3$  (since the path is Hamiltonian). The reverse path also covers the  $E_4$  edges corresponding to reverse Hamiltonian path edges. Our Eulerian trail is now “at” node  $v_1^{in}$ .

What remains is to complete the Eulerian walk covering (a) edges and their antiparallel counterparts corresponding to edges in  $G$  that were not used in the Hamiltonian path, and (b) the edges adjacent to node  $w$ . To do this, define  $\text{pred}(v)$  be the vertices  $u$  in  $G$  for which edge  $(u, v)$  exists and  $u$  is not the predecessor of  $v$  along the Hamiltonian path. For each  $u \in \text{pred}(v_1)$ , traverse the split edge labeled  $c\#$  to  $u^{out}$  then traverse the forward split edge labeled  $b\#$  back to  $v_1^{in}$ . This results in a string  $(c\#b\#)^{|\text{pred}(v_1)|}$ . Once the predecessors of  $v_1$  are exhausted, traverse the split edge labeled  $c\#$  from  $v_1^{in}$  into node  $w$  and then traverse the split edge labeled  $b\#$  to  $v_2^{in}$ . This again generates a  $c\#b\#$  string. Repeat the process, covering the edges of  $v_2$ ’s predecessors and returning to  $w$  to move to the next node along the Hamiltonian path for each node  $v_3, \dots, v_n$ . After covering the predecessors of  $v_n^{in}$ , go to  $v_1^{in}$  through the remaining edges in  $E_5$  and  $E_6$ ,  $(v_n^{in}, w)$  and  $(w, v_1^{in})$ , which completes the Eulerian tour. This covers all the edges of  $G''$ . The word spelled out in this last section of the Eulerian trail is a sequence of repetitions of  $c\#b\#$ , with one repetition for each edge that is not in the Hamiltonian path  $(|E| - n + 1)$  and all of the edges in  $E_5$  and  $E_6$  for entering and leaving each node  $(2n)$ , with a total of  $|E| + 1$  repetitions, which is the final  $(c\#b\#)^{|E|+1}$  term in  $q$ .

This ends the slight modification of the proof in Kupferman and Vardi [7], where the differences are (a) the

introduction of the # characters and (b) using the exponent  $|E| + 1$  of the final part of  $q$  instead of  $|E| + n + 1$  as in Kupferman and Vardi [7] since we create  $w$ -edges only to  $v_1^{in}$  vertices. (This second change has no material effect on the proof, but reduces the length of the string that must be matched).

Now, given an instance  $\langle G = (V, E) \rangle$  of HAMILTONIAN PATH, with  $n = |V|$  vertices, we construct  $G''$  as above (obtaining a unidirectional Eulerian graph) and create graph  $C$  that only represents string  $q$ . Note that  $|\Sigma| = 4$  and  $G''$  and  $C$  can be constructed in polynomial time.  $GTED(G'', C) = 0$  if and only if an Eulerian path in  $G''$  spells out  $q$ , since there can be no indels or mismatches. By the above argument, an Eulerian tour that spells out  $q$  exists if and only if  $G$  has a Hamiltonian path.  $\square$

**FGTED is NP-complete**

**Problem 4** (Flow Graph Traversal Edit Distance (FGTED) [13]). Given unidirectional, edge-labeled Eulerian graphs  $G_1$  and  $G_2$ , each of which has distinguished  $s_1, s_2$  source and  $t_1, t_2$  sink vertices, compute

$$FGTED(G_1, G_2) \triangleq \min_{\substack{D_1 \in \text{flow}(G_1, s_1, t_1) \\ D_2 \in \text{flow}(G_2, s_2, t_2)}} \text{emedit}(\text{strset}(D_1), \text{strset}(D_2)), \tag{39}$$

where  $\text{flow}(G_i, s_i, t_i)$  is the collection of all possible sets of  $s_i$ - $t_i$  trail decomposition of saturating flow from  $s_i$  to  $t_i$ ,  $\text{strset}(D)$  is the multi-set of strings constructed from trails in  $D$ .

**Theorem 3** *FGTED is NP-complete.*

*Proof* Let  $G = (V, E)$  be an instance of the HAMILTONIAN CYCLE problem. Let  $n = |V|$  be the number of vertices in  $G$ . Construct the Eulerian closure of  $G$  and split the anti-parallel edges. Let the new graph be  $G' = (V', E')$ . Attach a source  $s$  and a sink node  $t$  to an arbitrary node  $v_1^{in}$  by adding edge  $(s, v_1^{in})$  and  $(v_1^{in}, t)$  with labels  $s$  and  $t$ , respectively.

Construct a string  $q$ , such that

$$q = sa\#(b\#a\#)^{n-1}(c\#)^{2n-1}(c\#b\#)^{|E|+1}t. \tag{40}$$

Create a graph  $Q$  that only contains one path with labels on the edges of the path that spell the string  $q$ . The union of the set of trails in any flow decomposition of  $G'$  is equal to a set of Eulerian trails,  $\mathcal{E}$ , that starts at  $s$  and ends at  $t$ . All Eulerian trails in  $\mathcal{E}$  are also closed Eulerian trails of  $G' \setminus \{s, t\}$  that starts and ends at  $v_1^{in}$ .

Using the same line of argument in the proof of Theorem 2, an Eulerian trail in  $G'$  that spells  $q$  is equivalent to a Hamilton Cycle in  $G$ . In addition,  $FGTED(Q, G') = 0$  if and only if all Eulerian trails in  $\mathcal{E}$  spell out  $q$ . Therefore, if  $FGTED(Q, G') = 0$ , then there is a Hamiltonian Cycle in  $G$ . Otherwise, then there must not exist a Hamiltonian Cycle in  $G$ .  $\square$

**Appendix 2 Equivalence between two ILPs proposed by [1]**

The analysis provided by Ebrahimpour Boroojeny et al. [1] states that the LP relaxation of the ILP in (5)–(8) does not always yield integer solutions, but the LP relaxation of the ILP in (11)–(12) always yields integer solutions. This suggests that the two LP relaxations have difference feasibility regions for  $x$ . We show that these two LP relaxations are actually equivalent in Theorem 7. Further, we show that the ILP in (5)–(8) and the ILP in (11)–(12) are also equivalent. Since the ILP in (5)–(8) does not solve for  $GTED(G_1, G_2)$  as shown in Section 3.2,

we conclude that the ILP in (11)–(12) also does not solve  $GTED(G_1, G_2)$ .

**Theorem 7** *Given two unidirectional, edge-labeled Eulerian graphs  $G_1, G_2$ , the feasibility region of  $x$  in the LP relaxation of the ILP in (11)–(12) is the same as the feasibility region of  $x$  in the LP relaxation of the ILP in (5)–(8).*

Let  $\mathcal{A}(G_1, G_2) = (V, E, \delta)$  be the alignment graph of  $G_1 = (V_1, E_1, \ell_1, \Sigma_1)$  and  $G_2 = (V_2, E_2, \ell_2, \Sigma_2)$ , and let  $T(G_1, G_2)$  be its two-simplex set. First, we have the following result:

**Lemma 4** *Let  $[y_i] \in \mathbb{R}^{|T(G_1, G_2)|}$  be a vector such that the  $j$ -th entry of  $[y_i], [y_i]$  is equal to 0 for all  $j \neq i$ . The vector  $x' = x + [\delta][y_i]$  satisfies the constraints (6)–(7) if the vector  $x$  satisfies the constraints (6)–(7).*

*Proof* Let  $\sigma_i \in T(G_1, G_2)$  be the 2-simplex corresponding to the entry  $i$  of  $[y_i]$ . Based on the construction of  $T(G_1, G_2)$ ,  $\sigma_i$  has two forms:  $[(u_1, u_2), (v_1, u_2), (v_1, v_2)]$  or  $[(u_1, u_2), (u_1, v_2), (v_1, v_2)]$ . Without loss of generality, we assume  $\sigma_i = [(u_1, u_2), (v_1, u_2), (v_1, v_2)]$ . We can prove this lemma by using the same way when  $\sigma_i = [(u_1, u_2), (u_1, v_2), (v_1, v_2)]$ . Since

$$\partial \sigma_i = [(u_1, u_2), (v_1, u_2)] + [(v_1, u_2), (v_1, v_2)] - [(u_1, u_2), (v_1, v_2)],$$

We have

$$[\partial][y_i] = [y_i]_i[x_{e_1}] + [y_i]_i[x_{e_2}] - [y_i]_i[x_{e_3}],$$

where  $e_1 = [(u_1, u_2), (v_1, u_2)]$ ,  $e_2 = [(v_1, u_2), (v_1, v_2)]$ ,  $e_3 = [(u_1, u_2), (v_1, v_2)]$ , and  $[x_e] \in \mathbb{R}^{|E|}$  is a vector such that all the entries are 0 except that the one corresponding to edge  $e$  is 1. we also let  $[x_v] \in \mathbb{R}^{|V|}$  be a vector such that all the entries are 0 except that the one corresponding to vertex  $v$  is 1. Therefore, we have

$$Ax' = Ax + [y_i]_i[x_{v_2}] - [y_i]_i[x_{v_1}] + [y_i]_i[x_{v_3}] - [y_i]_i[x_{v_2}] - [y_i]_i[x_{v_3}] + [y_i]_i[x_{v_1}] = Ax,$$

where  $v_1 = (u_1, u_2)$ ,  $v_2 = (v_1, u_2)$ , and  $v_3 = (v_1, v_2)$ .

Hence,  $x'$  satisfies the constraint (6) if  $x$  satisfies the constraint (6).

In addition, since  $\sum_{e \in E} x'_e I_i(e, f) = \sum_{e \in E} x_e I_i(e, f) + [y_i]_i I_i(e_1, f) + [y_i]_i I_i(e_2, f) - [y_i]_i I_i(e_3, f)$ , and:

- $I_1(e_1, (u_1, v_1)) = 1$  and  $I_i(e_1, f) = 0$  for other  $f \in G_i$ ,
- $I_2(e_2, (u_2, v_2)) = 1$  and  $I_i(e_2, f) = 0$  for other  $f \in G_i$ ,
- $I_1(e_3, (u_1, v_1)) = 1$ ,  $I_2(e_3, (u_2, v_2)) = 1$ , and  $I_i(e_3, f) = 0$  for other  $f \in G_i$ ,

we have:

- $[y_i]_i I_1(e_1, (u_1, v_1)) + [y_i]_i I_1(e_2, (u_1, v_1)) - [y_i]_i I_1(e_3, (u_1, v_1)) = [y_i]_i + 0 - [y_i]_i = 0$ ,
- $[y_i]_i I_2(e_1, (u_2, v_2)) + [y_i]_i I_2(e_2, (u_2, v_2)) - [y_i]_i I_2(e_3, (u_2, v_2)) = 0 + [y_i]_i - [y_i]_i = 0$ ,
- $[y_i]_i I_i(e_1, f) + [y_i]_i I_i(e_2, f) - [y_i]_i I_i(e_3, f) = 0 + 0 - 0 = 0$  for any other  $i = 1, 2$  and  $f \in E_i$ .

Therefore,  $\sum_{e \in E} x'_e I_i(e, f) = \sum_{e \in E} x_e I_i(e, f)$ , meaning that  $x'$  satisfies the constraint (7) if  $x$  satisfies the constraint (7).  $\square$

With Lemma 4, we prove that any feasible solution of  $x$  in (11) is a feasible solution of (5)–(8). First, it is easy to check that  $x^{init}$  satisfies the constraints (6)–(7). For each feasible solution of  $x$  in (11), since  $x = x^{init} + [\partial]y = x^{init} + \sum_i [\partial][y_i]$ , by iteratively using Lemma 4, we get that  $x$  satisfies the constraints (6)–(7). Since  $x_e \geq 0$  for all  $e \in E$  is a constraint existing in both linear relaxations,  $x$  is a feasible solution of (5)–(8).

We now show that any feasible solution of (5)–(8) is a feasible solution of (11). Let  $x$  be a feasible solution of (5)–(8). We show that  $x$  is also a feasible solution of (11) by proving that  $x$  can be converted to  $x^{init}$  in (11) via the boundary operator  $\partial$ . First, if there is a diagonal edge

$e = [(u_1, u_2), (v_1, v_2)]$  in  $E$  such that  $x_e > 0$ , then it can be replaced by the horizontal edge  $e_h = [(u_1, u_2), (u_1, v_2)]$  followed by the vertical edge  $e_v = [(u_1, v_2), (v_1, v_2)]$  by using one boundary operation on the 2-simplex  $[(u_1, u_2), (u_1, v_2), (v_1, v_2)]$ . Hence,  $x$  can be converted to a new vector  $x'$ , such that  $x'_e = 0$ ,  $x'_{e_h} = x_{e_h} + x_e$ ,  $x'_{e_v} = x_{e_v} + x_e$ , and all the other entries in  $x'$  are the same as those in  $x$ . It is easy to check that  $x'$  is also a feasible solution of (5)–(8). Therefore, without loss of generality, we assume  $x$  to be a vector such that all the entries corresponding to diagonal edges in  $\mathcal{A}(G_1, G_2)$  are zero.

We then prove that any  $x$  can be converted to  $x^{init}$  in (11) via the boundary operator. Let the source and the sink node of  $x$  in  $\mathcal{A}(G_1, G_2)$  be  $(s_1^1, s_1^2)$  and  $(s_2^1, s_2^2)$ , where  $s_1^i$  is the source node of  $G_i$  and  $s_2^i$  is the sink node of  $G_i$ . When the Eulerian trail is closed (meaning that it is an Eulerian tour) in  $G_i$ , we let  $s_1^i = s_2^i$  be an arbitrary vertex in  $V_i$ .  $x^{init}$  can be seen as a trail (tour) in  $\mathcal{A}(G_1, G_2)$  that starts from  $(s_1^1, s_1^2)$ , walks along an Eulerian trail of  $G_2$  via all the horizontal edges  $P_h$ ,

$$P_h = \{[(s_1^1, s_1^2), (s_1^1, v_1^2)], [(s_1^1, v_1^2), (s_1^1, v_2^2)], \dots, [(s_1^1, v_{i-1}^2), (s_1^1, v_i^2)], [(s_1^1, v_i^2), (s_1^1, s_2^2)]\},$$

and then walks along an Eulerian trail of  $G_1$  via all the vertical edges  $P_v$ ,

$$P_v = \{[(s_1^1, s_2^2), (v_1^1, s_2^2)], [(v_1^1, s_2^2), (v_2^1, s_2^2)], \dots, [(v_{i-1}^1, s_2^2), (v_i^1, s_2^2)], [(v_i^1, s_2^2), (s_2^1, s_2^2)]\},$$

until the sink node  $(s_2^1, s_2^2)$ . Here  $\{s_2^1, v_1^2, v_2^2, \dots, v_{i-1}^2, v_i^2, s_2^2\}$  is an Eulerian trail of  $G_2$  and  $\{s_1^1, v_1^1, v_2^1, \dots, v_{i-1}^1, v_i^1, s_2^1\}$  is an Eulerian trail of  $G_1$ . We use  $P_0 = \{P_h, P_v\}$  to denote the trail from  $(s_1^1, s_1^2)$  to  $(s_2^1, s_2^2)$  that is the concatenation of  $P_h$  and  $P_v$ . It is easy to see that each edge in  $P_0$  is unique.

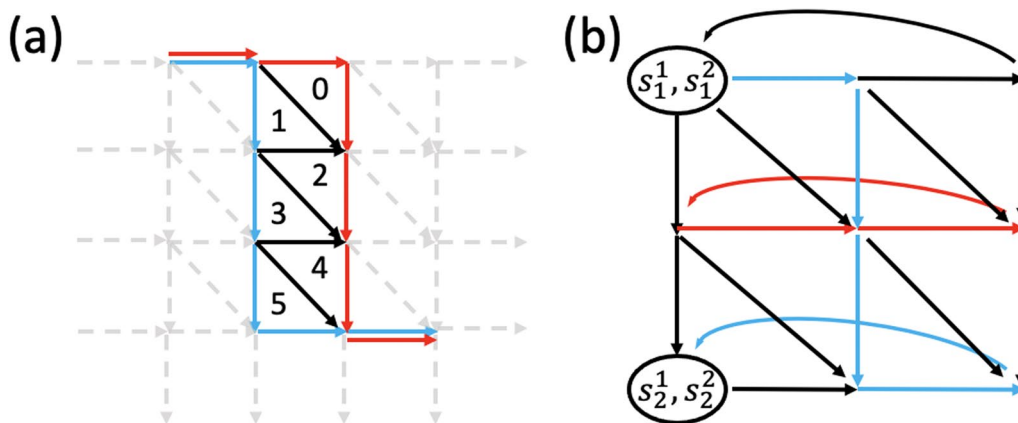
As shown in Qiu and Kingsford [13],  $x$  is a flow of  $\mathcal{A}(G_1, G_2)$  with the additional constraint (7). Therefore, according to the flow decomposition theorem [24, p. 80],  $x$  can be decomposed into a finite set of weighted paths in  $\mathcal{A}(G_1, G_2)$  from  $(s_1^1, s_1^2)$  to  $(s_2^1, s_2^2)$ , which is denoted as  $\{(p_1, w_1^p), \dots, (p_n, w_n^p)\}$ , and a finite set of weight cycles in  $\mathcal{A}(G_1, G_2)$ , which is denoted as  $\{(c_1, w_1^c), \dots, (c_m, w_m^c)\}$ . Each path or cycle only contains horizontal and vertical edges.

For path  $i$ , we use a vector  $x^{p,i}$  to represent  $(p_i, w_i^p)$ :

$$x_e^{p,i} = \begin{cases} w_i^p & \text{if } e \in p_i \\ 0 & \text{otherwise.} \end{cases} \quad (41)$$

By using the boundary operator, each path  $p_i$  can actually be converted to a new trail  $p'_i$  such that each edge in





**Fig. 7 a** An example of converting three vertical edges followed by one horizontal edge (blue line) to one horizontal edge followed by three vertical edges (red line). It can be done by doing boundary operations on 2-simplices labeled from 0 to 5. **b** An example of a cycle path (red line) and its auxiliary trail (blue line)

$p'_i$  is also an edge in  $P_0$ . To prove this, we consider the following two cases:

- If  $p_i$  walks along all the horizontal edges followed by all the vertical edges, then every edge in  $p_i$  is an edge in  $P_0$ . To see that, let  $e$  be an horizontal edge in  $p_i$ , since  $p_i$  starts from  $(s_1^1, s_1^2)$ ,  $e$  has the form  $[(s_1^1, v), (s_1^1, v')]$  where  $[v, v'] \in E_2$ . Since  $P_h$  corresponds to the Eulerian trail of  $G_2$ , for each  $[v, v'] \in E_2$ , we have  $[(s_1^1, v), (s_1^1, v')] \in P_h$ . Therefore  $e \in P_0$ . We can use the same approach to prove  $e \in P_0$  when  $e$  is a vertical edge. Note that in this case, the number of horizontal edges or vertical edges can be zero.
- If not, then we let  $p_i = \{e_1^i, e_2^i, \dots, e_m^i\}$ , and let  $e_t^i$  be the vertical edge with the smallest index  $t$ . There exists an integer  $k (k \geq 1)$  such that  $\{e_t^i, e_{t+1}^i, \dots, e_{t+k-1}^i\}$  are all vertical edges and  $e_{t+k}^i$  is an horizontal edge. We denote each vertical edge  $e_{t+w}^i \in \{e_t^i, e_{t+1}^i, \dots, e_{t+k-1}^i\}$  as  $[(v_w, v_t), (v_{w+1}, v_t)]$  and denote  $e_{t+k}^i$  as  $[(v_k, v_t), (v_k, v_{t+1})]$ . It is easy to see that when  $w = 0, v_w = s_1^1$ . By using the boundary operator, this subpath  $\{e_t^i, e_{t+1}^i, \dots, e_{t+k-1}^i, e_{t+k}^i\}$  can be replaced by another subpath with one horizontal edge  $[(s_1^1, v_t), (s_1^1, v_{t+1})]$  followed by  $k$  vertical edges:

$$\{[(s_1^1, v_{t+1}), (v_1, v_{t+1})], [(v_1, v_{t+1}), (v_2, v_{t+1})], \dots, [(v_{k-1}, v_{t+1}), (v_k, v_{t+1})]\}.$$

Now we have a new path, denoted as  $p_i^1$ , in which the smallest index of the vertical edges becomes  $t + 1$ . Figure 7a shows an example, in which the blue line represents the subpath of  $p_i$  and the red line represents the new subpath in  $p_i^1$ .

To create a new vector that represents  $p_i^1$ , we first create a zero vector  $y^{p,i,1} \in \mathbb{R}^{|T(G_1, G_2)|}$ , and from  $w = 0$  to  $w = k - 1$ , we iteratively update  $y^{p,i,1}$  via the following equations:

$$y_\sigma^{p,i,1} = \begin{cases} y_\sigma^{p,i,1} - w_i^p & \text{if } \sigma = [(v_w, v_t), (v_{w+1}, v_t), (v_{w+1}, v_{t+1})] \\ y_\sigma^{p,i,1} + w_i^p & \text{if } \sigma = [(v_w, v_t), (v_w, v_{t+1}), (v_{w+1}, v_{t+1})] \\ 0 & \text{otherwise.} \end{cases} \tag{42}$$

The vector  $x^{p,i,1} = x^{p,i} + [\partial]y^{p,i,1}$  is the one that represents  $p_i^1$ .

Since the length of  $p_i$  is finite, by doing such a transformation a finite number of times, we can convert  $p_i$  to a new path  $p'_i$  such that  $p'_i$  walks along all the horizontal edges first followed by all the vertical edges, therefore each edge in  $p'_i$  is also an edge in  $P_0$ . We use the vector  $\hat{x}^{p,i}$  to represent  $p'_i$ ,  $\hat{x}^{p,i} = x^{p,i} + [\partial] \sum_{j=1}^q y^{p,i,j}$  where  $q$  is the number of transformations. Apparently,  $\hat{x}_e^{p,i} = 0$  when  $e \notin P_0$ . Let  $y^{p,i} = \sum_{j=1}^q y^{p,i,j}$ , we have  $\hat{x}^{p,i} = x^{p,i} + [\partial]y^{p,i}$ .

For cycle  $i$ , we also use a vector  $x^{c,i}$  to represent  $(c_i, w_i^c)$ ,

$$x_e^{c,i} = \begin{cases} w_i^c & \text{if } e \in c_i \\ 0 & \text{otherwise,} \end{cases} \tag{43}$$

Let  $(v, v')$  be an arbitrary chosen node in  $c_i$ , we construct a trail  $p_{aux}^i$  that passes  $(v, v')$  as follows:

- From  $(s_1^1, s_2^2)$ , walk along  $P_h$  until the node  $(s_1^1, v')$ . It corresponds to a part of an Eulerian trail of  $G_2$ .
- From  $(s_1^1, v')$ , walk along an Eulerian trail of  $G_1$  to  $(s_2^1, v')$ . It must pass the node  $(v, v')$ .
- From  $(s_2^1, v')$ , walk along the remaining part of the Eulerian trail of  $G_2$  to the node  $(s_2^1, s_2^2)$ .



Fig. 7b shows an example, in which the blue line represents  $p_{aux}^i$  and the red line represents  $c_i$ .

We use  $x^{aux,i}$  to denote the vector representing  $p_{aux}^i$ . The combination of  $c_i$  and  $p_{aux}^i$  represented by the vector  $x^{c,i} + x^{aux,i}$  creates a new trail (may have repeated edges) from  $(s_1^1, s_1^2)$  to  $(s_2^1, s_2^2)$ : (1) walk along  $p_{aux}^i$  from  $(s_1^1, s_1^2)$  to  $(v, v')$ , (2) walk along  $c_i$  from  $(v, v')$  to itself, and (3) walk along the remaining part of  $p_{aux}^i$  from  $(v, v')$  to  $(s_2^1, s_2^2)$ . By using the same approach described above, each  $c_i + p_{aux}^i$  or  $p_{aux}^i$  can be converted to a new trail in which each edge is also an edge in  $P_0$ . We use  $\hat{x}^{c,i}$  or  $\hat{x}^{aux,i}$  to represent the new trail accordingly, therefore, we have  $\hat{x}^{c,i} = x^{c,i} + x^{aux,i} + [\partial]y^{c,i}$  and  $\hat{x}^{aux,i} = x^{aux,i} + [\partial]y^{aux,i}$ . Likewise,  $\hat{x}_e^{c,i} = \hat{x}_e^{aux,i} = 0$  when  $e \notin P_0$ .

We define a new vector  $\hat{x}$  such that:

$$\begin{aligned} \hat{x} &= \sum_{i=1}^n \hat{x}^{p,i} + \sum_{j=1}^m \hat{x}^{c,j} - \hat{x}^{aux,j} \\ &= \sum_{i=1}^n x^{p,i} + [\partial]y^{p,i} + \sum_{j=1}^m x^{c,j} + x^{aux,j} + [\partial]y^{c,j} - \left( \sum_{j=1}^m x^{aux,j} + [\partial]y^{aux,j} \right) \\ &= \sum_{i=1}^n x^{p,i} + \sum_{j=1}^m x^{c,j} + [\partial] \left( \sum_{i=1}^n y^{p,i} + \sum_{j=1}^m y^{c,j} - \sum_{j=1}^m y^{aux,j} \right) \\ &= x + [\partial] \left( \sum_{i=1}^n y^{p,i} + \sum_{j=1}^m y^{c,j} - \sum_{j=1}^m y^{aux,j} \right). \end{aligned}$$

Therefore,  $\hat{x}$  is a vector converted from  $x$  via boundary operations.  $\hat{x}$  is equal to  $x^{init}$  because:

1.  $\hat{x}_e = 0$  when  $e \notin P_0$  since  $\hat{x}_e^{p,i} = \hat{x}_e^{c,i} = \hat{x}_e^{aux,i} = 0$  when  $e \notin P_0$  for each  $i$ .
2. As we have proved above, the boundary operator preserves the constraints (6)–(7). Therefore,  $\hat{x}$  satisfies the constraints (6)–(7) since  $x$  is a feasible solution of (5)–(8). Combined with the first point, we have that  $\hat{x}_e = 1$  if  $e \in P_0$  and  $\hat{x}_e = 0$  otherwise, meaning that  $\hat{x} = x^{init}$ .

Hence, for each feasible solution  $x$  of (5)–(8), we have:

$$\begin{aligned} x &= x^{init} - [\partial] \left( \sum_{i=1}^n y^{p,i} + \sum_{j=1}^m y^{c,j} - \sum_{j=1}^m y^{aux,j} \right) \\ &= x^{init} + [\partial] \left( - \sum_{i=1}^n y^{p,i} - \sum_{j=1}^m y^{c,j} + \sum_{j=1}^m y^{aux,j} \right), \end{aligned}$$

meaning that  $x$  is also a feasible solution of (11).

We proved that the feasibility region of  $x$  in (11) is the same as the feasibility region of  $x$  in (5)–(8), and since the objective functions of these two linear relaxations are the same, the optimal solutions of them are equal.

By employing the same approach and taking into account that if all edge weights in a flow network are non-negative integers, the flow decomposition theorem guarantees that the network can be decomposed into a finite set of weighted paths and cycles, each with positive integer weight, we can prove that the ILP in (5)–(8) and the ILP in (11)–(12) are also equivalent.

Based on the proof, we can conclude that the way to index the vertices or edges in the alignment graph, or the 2-simplices in  $T(G_1, G_2)$ , will not affect the equivalence result.

Additionally, different choices of orientations for the 2-simplices in  $T(G_1, G_2)$  will also not affect the equivalence result. This is because for any two sets  $T(G_1, G_2)$  and  $T'(G_1, G_2)$  containing the same 2-simplices with the same indices but different orientations, if  $(x, y)$  is a feasible solution of the ILP in (11)–(12) (or its relaxation) that corresponds to  $T(G_1, G_2)$ , then  $(x, y')$  is a feasible solution of the ILP in (11)–(12) (or its relaxation) that corresponds to  $T'(G_1, G_2)$ , where  $y_i = y'_i$  when  $\sigma_i \in T(G_1, G_2)$  has the same orientation as  $\sigma'_i \in T'(G_1, G_2)$ , and  $y_i = -y'_i$  when  $\sigma_i \in T(G_1, G_2)$  has the opposite orientation to  $\sigma'_i \in T'(G_1, G_2)$ . Therefore, it is acceptable to specify a particular orientation for each 2-simplex when defining  $T(G_1, G_2)$ .

### Appendix 3 The linear relaxation of the ILP in (11)–(12) does not always yield integer solutions

#### Characterizations of the ILP in (11)–(12)

Ebrahimpour Boroojny et al. [1] propose the ILP in (11)–(12), and we show that the  $x$  variables in this ILP have the same feasible region as the  $x$  variables in lower bound ILP. However, Ebrahimpour Boroojny et al. [1] argues that

the linear programming relaxation of the ILP in (11)–(12) always yields integer optimal solutions, and therefore the ILP in (11)–(12) can be solved in polynomial time. We provide a counterexample where the ILP in (11)–(12) yields fractional optimal solutions with fractional variable values. Additionally, we show that the constraint matrix of the LP relaxation of the ILP in (11)–(12) is not totally unimodular given most non-trivial input graphs.

**[ $\partial$ ] is not necessarily totally unimodular**

A linear programming formulation always yields integer solutions if its constraint matrix is totally unimodular, which means that all of its square submatrices have determinants of 0, -1, or 1 [25]. To show that the constraint matrix of the LP relaxation of the ILP in (11)–(12) is not totally unimodular, we first write the LP in standard form.

In a standard form of an LP, all variables are greater than or equal to 0. Since  $y$  vectors in the LP relaxation of the ILP in (11)–(12) can contain negative entries, we decompose it into  $y^+ - y^-$ . Given alignment graph  $\mathcal{A}(G_1, G_2) = (V, E, \delta)$  and  $T(G_1, G_2)$ , we can now write the standard form of the LP in (11)–(12) as

$$\begin{aligned} & \text{minimize} && \sum_{e \in E} x_e \delta(e) \\ & x \in \mathbb{R}^{|E|}, y^+, y^- \in \mathbb{R}^{|T(G_1, G_2)|} && \\ & \text{subject to} && [I, -[\partial], [\partial]][x, y^+, y^-]^T = x^{init} \\ & && x, y^+, y^- \geq 0. \end{aligned} \tag{44}$$

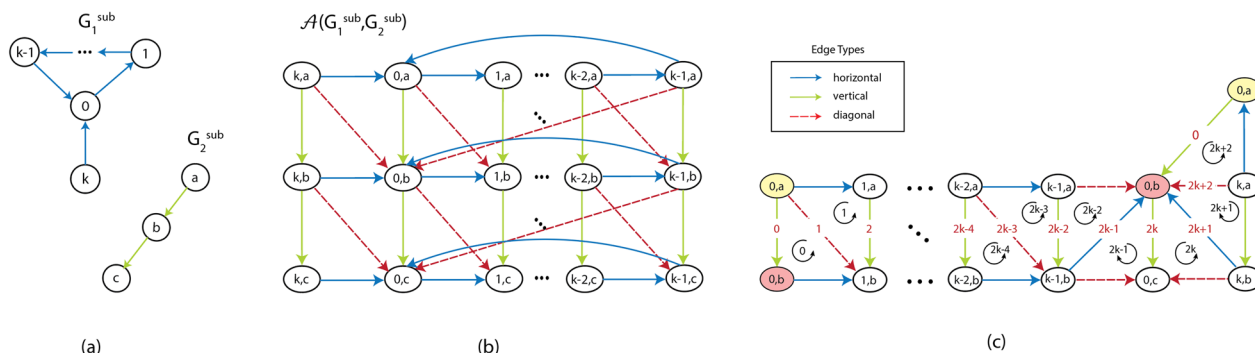
Hence the constraint matrix of the LP relaxation is  $A = [I, -[\partial], [\partial]]$ . According to the characteristics of a totally unimodular matrix [26, p. 280]  $A$  is not totally unimodular if  $[\partial]$  is not totally unimodular. We show that  $[\partial]$

is not TU when the input graphs satisfy the constraints given in the following theorem.

**Theorem 8** *Given two unidirectional, edge-labeled Eulerian graphs  $G_1$  and, the boundary matrix  $G_2$  where  $|E_1| \geq 2$  and  $|E_2| \geq 2$  [ $\partial$ ] constructed from  $\mathcal{A}(G_1, G_2) = (V, E, \delta)$  and  $T(G_1, G_2)$  is not totally unimodular if there is a vertex  $v \in V_1$  or  $V_2$  such that there are at least 3 unique edges in  $E_1$  or  $E_2$  that are incident to  $v$ . Here, unique edges are edges that connect to  $v$  at one end but have different endpoints at the other end.*

*Proof* To prove that the boundary matrix is not TU, we only need to show that it is not TU under one specific chosen orientation for 1- and 2-simplices, as well as one specific chosen set of indices for 1- and 2-simplices. This is because changing the orientations or indices of 1-simplices in  $E$  or 2-simplices in  $T(G_1, G_2)$  corresponds to permuting rows and columns of  $[\partial]$  or multiplying rows and columns of  $[\partial]$  by  $-1$ , which preserves the total unimodularity [26, p. 280].

Without loss of generality, let  $v_0 \in V_1$  be a node that is incident to at least 3 unique edges. Since  $G_1$  is an Eulerian graph,  $v$  must be part of a cycle  $C$  in  $G_1$ . Also, there must exist another node  $v_k$  and an edge between  $v_0$  and  $v_k$  in either direction, such that the edge between  $v_0$  and  $v_k$  is not contained in cycle  $C$  (Fig. 8a). Suppose the number of nodes in the cycle is  $k$  ( $k \geq 3$  due to the unidirectionality constraint), and let the cycle  $C = v_0, v_1, \dots, v_{k-1}$ . Since a specific choice of 1-simplex orientations does not affect the total unimodularity of the boundary matrix, we assume the edge between  $v_0$  and  $v_k$  is  $[v_k, v_0]$  without loss of generality. We use  $G_1^{sub} = (V_1^{sub}, E_1^{sub})$  to denote the subgraph with  $V_1^{sub} = \{v_0, \dots, v_{k-1}, v_k\}$  and  $E_1^{sub} = \{[v_i, v_{i+1}] : i \in \{0, 1, \dots, k-2\}\} \cup \{[v_k, v_0]\}$ .



**Fig. 8** **a** Subgraphs  $G_1^{sub}$  and  $G_2^{sub}$  of input graphs  $G_1$  and  $G_2$ . Dots represent a path from node 1 to  $k - 1$  with middle nodes omitted. **b** The alignment graph  $\mathcal{A}(G_1^{sub}, G_2^{sub})$  with different edges labeled with colors. **c** A subgraph of the alignment graph in **b** with edges and triangles numbered. Dots represent horizontal and diagonal edges omitted. The same vertices that are repeated in **c** are marked with yellow and red filling colors

Since  $|E_2| \geq 2$  and  $G_2$  is a connected graph, there exist two consecutive, directed edges in  $G_2$ . We use  $G_2^{sub} = (V_2^{sub}, E_2^{sub})$  to denote the subgraph of  $G_2$  with  $V_2^{sub} = \{v_a, v_b, v_c\}$  and  $E_2^{sub} = \{[v_a, v_b], [v_b, v_c]\}$ . The alignment graph  $\mathcal{A}(G_1^{sub}, G_2^{sub})$  is formed with  $G_1^{sub}$  and  $G_2^{sub}$  and is a subgraph of  $\mathcal{A}(G_1, G_2)$ , therefore, each subgraph of  $\mathcal{A}(G_1^{sub}, G_2^{sub})$  is also a subgraph of  $\mathcal{A}(G_1, G_2)$ . Similarly, the 2-simplex set  $T(G_1^{sub}, G_2^{sub})$  is a subset of  $T(G_1, G_2)$ .

We extract a sequence of 2-simplices (Fig. 8(c)),  $T_c$ , from  $T(G_1^{sub}, G_2^{sub})$  via following steps:

1. Extract all oriented 2-simplices  $[(v_i, v_a), (v_i, v_b), (v_{i+1}, v_b)]$  and  $[(v_i, v_a), (v_{i+1}, v_a), (v_{i+1}, v_b)]$  for  $0 \leq i \leq k-2$  from  $T(G_1^{sub}, G_2^{sub})$ .

Flip the orientations of  $[(v_i, v_a), (v_{i+1}, v_a), (v_{i+1}, v_b)]$  for all  $0 \leq i \leq k-2$ , obtaining  $[(v_i, v_a), (v_{i+1}, v_b), (v_{i+1}, v_a)]$ . Use  $\sigma_{2i}$  to denote  $[(v_i, v_a), (v_i, v_b), (v_{i+1}, v_b)]$ , and  $\sigma_{2i+1}$  to denote  $[(v_i, v_a), (v_{i+1}, v_b), (v_{i+1}, v_a)]$ .

$\det M_1$

$$\begin{aligned}
 &= \det \begin{bmatrix} -1 & 1 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix} - \det \begin{bmatrix} 1 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \\
 &= (-1)^{2k-2} \times (-1) - 1^{2k+2} = -2.
 \end{aligned}$$

2. Add to the sequence another five oriented 2-simplices from  $T(G_1^{sub}, G_2^{sub})$  in the order as specified:
 
$$\begin{aligned}
 \sigma_{2k-2} &= [(v_{k-1}, v_a), (v_{k-1}, v_b), (v_0, v_b)], \\
 \sigma_{2k-1} &= [(v_{k-1}, v_b), (v_0, v_b), (v_0, v_c)] \\
 \sigma_{2k} &= [(v_k, v_b), (v_0, v_b), (v_0, v_c)] \\
 \sigma_{2k+1} &= [(v_k, v_a), (v_k, v_b), (v_0, v_b)] \quad \text{and finally} \\
 \sigma_{2k+2} &= [(v_k, v_a), (v_0, v_a), (v_0, v_b)].
 \end{aligned}$$

In total, we extract a sequence of  $(2k+3)$  oriented 2-simplices,  $T_c = \{\sigma_0, \sigma_1, \dots, \sigma_{2k+2}\}$ , such that  $\sigma_i$  and  $\sigma_{i+1 \pmod{2k+3}}$  share one edge. The extracted

2-simplices and their orientations as well as all shared edges are shown in Fig. 8(c). We flip the orientations of  $[(v_i, v_a), (v_{i+1}, v_a), (v_{i+1}, v_b)]$  solely to ensure that the submatrix constructed below has a simple form, which makes it easier to compute the determinant.

Based on  $T_c$ , we obtain  $M_1$ , a  $(2k+3) \times (2k+3)$  submatrix of  $[\partial]$  where each row corresponds to a shared edge and each column corresponds to a 2-simplex in  $T_c$ .

The entry values of  $M_1$  are the signed coefficients of each selected 1-simplex from the boundaries of selected 2-simplices.

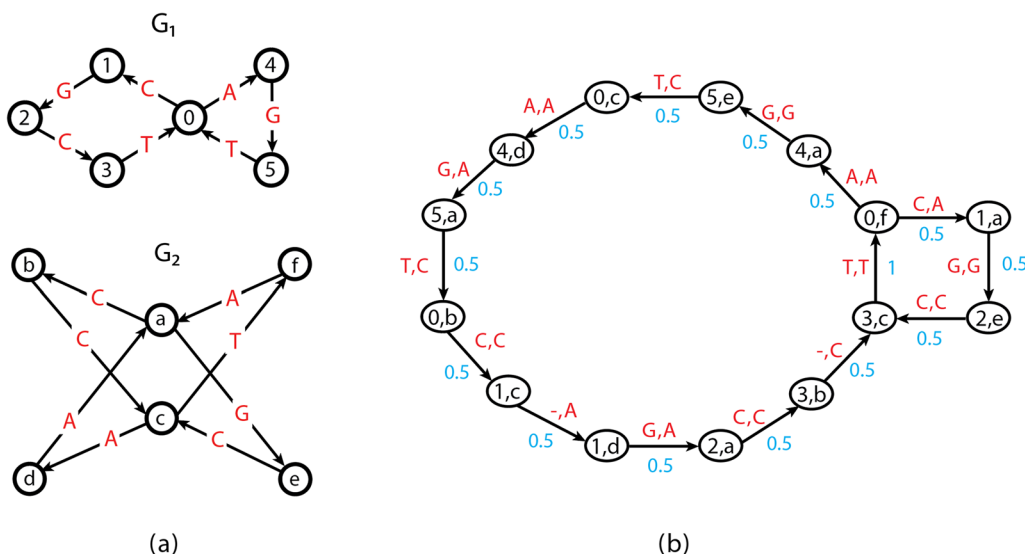
$$M_1 = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ -1 & 1 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & -1 & -1 \end{bmatrix}$$

The determinant of  $M_1$  is:

Since the determinant of  $M_1$  is -2, and  $M_1$  is a submatrix of  $[\partial]$ ,  $[\partial]$  is not totally unimodular.  $\square$

The minimal pair of input graphs that satisfy the conditions in Theorem 8 is a graph with one 3-node cycle and one additional edge incident to the cycle and an acyclic, connected graph with three nodes. In practice, most non-trivial edge-labeled Eulerian graphs satisfy these conditions.

According to the definitions in Dey et al. [25], the subgraph used to construct  $M_1$  in the above proof (Fig. 8(c)) is a Möbius subcomplex, and  $M_1$  is a  $(2k+3)$ -Möbius cycle matrix (MCM). Theorem 8 also



**Fig. 9** An example of a fractional optimal solution to the LP in (11)–(12) and the LP in (5)–(8). **a** A pair of input graphs to the LP in (11)–(12) and the LP in (5)–(8). The letters in red are edge labels. **b** A subgraph of  $\mathcal{A}(G_1, G_2)$  that is induced by alignment edges with non-zero weights (blue font) in an optimal solution to the LPs. The letters in red show the matching between the edge labels or between edge labels and gaps

establishes that there may exist a Möbius subcomplex in an alignment graph, which corrects the false claim made in Lemma 2 in [1].

Theorem 2 in Ebrahimpour Boroojeny et al. [1] attempts to employ a more algebraic approach to demonstrate that  $[\partial]$  is TU by establishing that the alignment graph is a Möbius-free product space. However, the property of being Möbius-free globally does not imply the absence of Möbius subcomplexes locally. As we show in Theorem 8, although the alignment graph  $\mathcal{A}(G_1^{sub}, G_2^{sub})$  is homotopically equivalent to the one-dimensional circle, which is Möbius-free, it still contains a Möbius subcomplex.

**The LP yields optimal fractional solutions**

The fact that  $[\partial]$  is not totally unimodular does not guarantee that the LP in (11)–(12) has a fractional optimal objective value. In this section, we prove that the LP in (11)–(12) does not always yield integer optimal solutions by constructing a specific example with a fractional optimal objective value.

**Theorem 9** *The LP in (5)–(8) and the LP in (11)–(12) do not always yield optimal integer solutions.*

We prove the above theorem by giving an example where the LP in (5)–(8) yields a fractional optimal solution. Since by Theorem 7, two LPs are equivalent,

it follows that the LP in (11)–(12) also yields the same fractional optimal solution.

Construct  $G_1$  and  $G_2$  such that their edges and edge labels are equal to the ones specified in Fig. 9a. Let the edge multi-set of  $\mathcal{A}(G_1, G_2)$  be  $E$ . We assign an edge cost to 0 if the edge matches two equal characters and 1 otherwise. Construct vector  $x^* \in \mathbb{R}^{|E|}$  and set entries corresponding to edges in Fig. 9b to 0.5 except edge  $[(v_3, v_c), (v_0, v_f)]$  to which the corresponding entry is set to 1. Set the rest of the entries of  $x^*$  to 0.

**Lemma 5**  *$x^*$  is an optimal solution to the LP in (5)–(8) constructed with  $\mathcal{A}(G_1, G_2)$  and  $T(G_1, G_2)$ .*

*Proof* We prove the optimality of  $x^*$  via complementary slackness. We first write the LP in (5)–(8) in standard form:

$$\begin{aligned}
 &\text{minimize} && \sum_{e \in E} \delta_e x_e \\
 &\text{subject to} && Ax = b \\
 &&& x_e \geq 0 \quad \text{for all } e \in E.
 \end{aligned}
 \tag{45}$$

Here,  $\delta$  is a vector of size  $|E|$  where each entry is the cost of edge  $e$ . The constraint matrix  $A$  of the primal LP (45) has  $|E|$  columns and  $|V| + |E_1| + |E_2| = m$  rows, where  $V$  is the vertex set of  $\mathcal{A}(G_1, G_2)$ , and  $E_1$  and  $E_2$  are edge multi-sets of the input graphs. The first  $|V|$  rows correspond to the constraints specified in (6). The rest of the rows correspond to the constraints in (7) that enforce the

projected multi-set of edges to be equal to the multi-set of edges in each input graph. Since the input graphs both contain Eulerian tours, the vector  $b$  has size  $m$ , where the first  $|V|$  entries are zeroes and the rest of the entries are 1 s.

We write the dual form of LP (45) as follows:

$$\begin{aligned} & \underset{y \in \mathbb{R}^m}{\text{maximize}} && \sum_{j=1}^m b_j y_j \\ & \text{subject to} && A^\top y \leq \delta. \end{aligned} \tag{46}$$

Let the objective value of LP (45) given a  $x$  as input is  $\text{obj}_x^p$ , and the objective value of LP (46) given a  $y$  as input is  $\text{obj}_y^d$ . To show that  $x^*$  is an optimal solution to the LP in (5)–(8), we need to show that there exists a feasible solution to the dual LP,  $y^*$ , that satisfies the complementary slackness conditions and that  $\text{obj}_{y^*}^d = \text{obj}_{x^*}^p$ .

Since each alignment edge has two endpoints and is projected to at most one edge in each graph, there are at most 4 non-zero entries in each column of  $A$ . The variables in  $y$  of the dual form can be interpreted in three parts. Each of the first  $|V|$  entries of  $y$  can be assigned to each vertex in the alignment graph, and the next  $|E_1|$  entries can be assigned to edges in  $G_1$  and the last  $|E_2|$  entries can be assigned to edges in  $G_2$ . There are  $|E|$  constraints in the dual LP, and the  $e$ -th constraint can be assigned to one edge in the alignment graph has cost  $\delta_e$ . Therefore, each constraint that is assigned to a horizontal or a vertical edge can be written as

$$y_{v_e^{\text{out}}} - y_{v_e^{\text{in}}} + y_{e_i} \leq \delta_e, \tag{47}$$

where  $i = 1$  if  $e$  is a horizontal edge, and  $i = 2$  if  $e$  is a vertical edge.  $y_{v_e^{\text{in}}}$  and  $y_{v_e^{\text{out}}}$  are the  $y$  entries that are assigned to the vertices that are the start and end of edge  $e$ , and  $y_{e_i}$  are the  $y$  entries that assigned to the  $\pi_i(e)$ .

Similarly, each constraint that is assigned to a diagonal edge is

$$y_{v_e^{\text{in}}} - y_{v_e^{\text{out}}} + y_{e_1} + y_{e_2} \leq \delta_e. \tag{48}$$

We can verify that  $x^*$  is a feasible solution of the primal form (45) by checking if constraints (6)–(7) are satisfied. The primal objective value can be computed in a straightforward way, and we can obtain  $\text{obj}_{x^*}^p = 3.5$ .

According to complementary slackness conditions, since  $x_e^* > 0$  for edges shown in Fig. 9b, the corresponding constraints in the dual LP (46) must be tight, meaning

that the equality must hold in these constraints. The rest of the dual constraints could have slacks.

Let the subgraph of  $\mathcal{A}(G_1, G_2)$  shown in Fig. 9b be  $A'$ . Denote the cycle that traverses from  $[(0, f), (4, a)]$  to  $[(3, c), (0, f)]$  be  $C'$  and the 4-node cycle that traverses  $((0, f), (1, a), (2, e), (3, c))$  be  $C''$ . Denote the concatenation of two cycles with  $C$ . The projected cycle from  $C$  to  $G_1$  is

$$C_1 = (v_0, v_4, v_5, v_0, v_4, v_5, v_0, v_1, v_2, v_3, v_0, v_1, v_2, v_3, v_0). \tag{49}$$

The projected cycle from  $C$  to  $G_2$  is

$$C_2 = (v_f, v_a, v_e, v_c, v_d, v_a, v_b, v_c, v_d, v_a, v_b, v_c, v_f, v_a, v_e, v_c, v_f). \tag{50}$$

Sum up all the constraints that are assigned edge  $e$  where  $x_e^* > 0$ . Since these edges form a cycle, we get:

$$\sum_{e \in C} (y_{v_e^{\text{out}}} - y_{v_e^{\text{in}}}) + 2 \left( \sum_{e_1 \in C_1} y_{e_1} + \sum_{e_2 \in C_2} y_{e_2} \right) \tag{51}$$

$$= 0 + 2 \left( \sum_{e_1 \in C_1} y_{e_1} + \sum_{e_2 \in C_2} y_{e_2} \right) \tag{52}$$

$$= \sum_{e \in C} \delta_e = 7, \tag{53}$$

$$\Rightarrow \sum_{e_1 \in C_1} y_{e_1} + \sum_{e_2 \in C_2} y_{e_2} = 3.5. \tag{54}$$

The summed edge cost is 7 as there are 7 edges that are either mismatch edges or vertical edges.

All  $y$  entries that correspond to vertices are free variables and are in every constraint. After fixing the  $y$  variables that satisfy constraint (54), the rest of the  $y$  variables can be set to satisfy the dual constraint. We now obtain  $y^*$  which is a feasible solution to the dual LP.

The only entries in  $y^*$  that could have non-zero dual costs are those that correspond to edges in  $E_1$  and  $E_2$ . Since these corresponding dual costs are all 1,

$$\text{obj}_{y^*}^d = \sum_{e_1 \in C_1} y_{e_1} + \sum_{e_2 \in C_2} y_{e_2} = 3.5 = \text{obj}_{x^*}^p.$$

□

Since the costs of alignment graph edges are all integers, the fact that the LP in (11)–(12) and the LP in (5)–(8) yield fractional optimal objective values means that they must yield fractional solutions and assign fractional values to entries in  $x$ . Theorem 9 follows. Since the LP



in (11)–(12) yields fractional solutions and GTED is always an integer, solving the LP in (11)–(12) does not solve GTED.

#### Acknowledgements

The authors would like to thank the members of the Kingsford Group for their helpful comments throughout this project, in particular Guillaume Marçais. The authors thank Marina L Knittel, Jacob M Gilbert, and Cenk Sahinalp for their insightful discussion on the NP-completeness of CCTED.

#### Author contributions

Y.Q. and C.K. conceived and designed the study. Y.Q., Y.S. and C.K. developed computational methods and proofs for the theorems. Y.Q. and Y.S. prepared the figures. All authors wrote the main manuscript and reviewed and approved the final manuscript.

#### Funding

This work was supported in part by the US National Science Foundation [DBI-1937540, III-2232121], the US National Institutes of Health [R01HG012470] and by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program.

#### Availability of data and materials

The simulated data described in the manuscript can be found at <https://github.com/Kingsford-Group/gtednewilp/data>.

#### Code availability

The GTED ILP implementation can be found at <https://github.com/Kingsford-Group/gtednewilp>.

#### Declarations

##### Ethics approval and consent to participate

Not applicable.

##### Consent for publication

Not applicable.

##### Competing interests

C.K. is a co-founder of Ocean Genomics, Inc.

Received: 21 October 2023 Accepted: 21 March 2024

Published online: 29 April 2024

#### References

- Ebrahimpour Boroojeny A, Shrestha A, Sharifi-Zarchi A, Gallagher SR, Sahinalp SC, Chitsaz H. Graph traversal edit distance and extensions. *J Comput Biol.* 2020;27(3):317–29.
- Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci USA.* 2001;98(17):9748–53.
- Polevikov E, Kolmogorov M. Synteny Paths for Assembly Graphs Comparison. In: Huber, K.T., Gusfield, D. (eds.) 19th International Workshop on Algorithms in Bioinformatics (WABI 2019). Leibniz International Proceedings in Informatics (LIPIcs), vol. 143, pp. 24–12414. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany (2019). <https://doi.org/10.4230/LIPIcs.WABI.2019.24>. <http://drops.dagstuhl.de/opus/volltexte/2019/11054>
- Minkin I, Medvedev P. Scalable pairwise whole-genome homology mapping of long genomes with Bubbz. *IScience.* 2020;23(6): 101224.
- Mangul S, Koslicki D. Reference-free comparison of microbial communities via de Bruijn graphs. In: Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2016;68–77
- Huntsman S, Rezaee A. De Bruijn entropy and string similarity. *arXiv preprint arXiv:1509.02975* (2015)
- Kupferman O, Vardi G. Eulerian paths with regular constraints. In: Faliszewski P, Muscholl A, Niedermeier R, editors. 41st International Symposium on Mathematical Foundations of Computer Science (MFCS 2016), vol. 58. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik; 2016. p. 62–16215.
- Jain C, Zhang H, Gao Y, Aluru S. On the complexity of sequence-to-graph alignment. *J Comput Biol.* 2020;27(4):640–54.
- Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol.* 1970;48(3):443–53.
- Dantzig G, Fulkerson R, Johnson S. Solution of a large-scale traveling-salesman problem. *J Oper Res Soc Am.* 1954;2(4):393–410.
- Dias FH, Williams L, Mumey B, Tomescu AI. Minimum flow decomposition in graphs with cycles using integer linear programming. *arXiv preprint arXiv:2209.00042* (2022)
- Miller CE, Tucker AW, Zemlin RA. Integer programming formulation of traveling salesman problems. *J ACM.* 1960;7(4):326–9.
- Qiu Y, Kingsford C. The effect of genome graph expressiveness on the discrepancy between genome graph distance and string set distance. *Bioinformatics.* 2022;38:404–12.
- Rubner Y, Tomasi C, Guibas LJ. The Earth Mover's distance as a metric for image retrieval. *Int J Comput Vision.* 2000;40(2):99–121.
- Munkres JR. Elements of algebraic topology. Boca Raton: CRC Press; 2018.
- Bradley SP, Hax AC, Magnanti TL. Applied mathematical programming (1977)
- Pele O, Werman M. A linear time histogram metric for improved sift matching. In: Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12–18, 2008, Proceedings, Part III 10, pp. 495–508 (2008). Springer
- Bourque G, Pevzner PA, Tesler G. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res.* 2004;14(4):507–16.
- Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, Diekhans M, Sulovari A, Munson KM, Lewis AP, et al. Segmental duplications and their variation in a complete human genome. *Science.* 2022;376(6588):6965.
- Gurobi Optimization, LLC: Gurobi Optimizer Reference Manual (2023). <https://www.gurobi.com>
- Lefranc M-P. IMG2, the international ImMunoGeneTics information system. *Cold Spring Harbor Protocols.* 2011;2011(6):595–603.
- Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, Khurana E, Waszak S, Korbel JO, Haber JE, et al. Patterns of somatic structural variation in human cancer genomes. *Nature.* 2020;578(7793):112–21.
- Darai-Ramqvist E, Sandlund A, Müller S, Klein G, Imreh S, Kost-Alimova M. Segmental duplications and evolutionary plasticity at tumor chromosome break-prone regions. *Genome Res.* 2008;18(3):370–9.
- Ahujia RK, Magnanti TL, Orlin JB. Network flows: theory, algorithms and applications. New Jersey: Prentice-Hall; 1993.
- Dey TK, Hirani AN, Krishnamoorthy B. Optimal homologous cycles, total unimodularity, and linear programming. *SIAM J Comput.* 2011;40(4):1026–44.
- Schrijver A. Theory of linear and integer programming. Hoboken: Wiley; 1998.

#### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.