Fast linear sum assignment with error-correction and no cost constraints
Sébastien Bougleux, Benoit Gaüzère, David Blumenthal, Luc Brun

To cite this version:

Sébastien Bougleux, Benoit Gaüzère, David Blumenthal, Luc Brun. Fast linear sum assignment with error-correction and no cost constraints. Pattern Recognition Letters, Elsevier, 2018, 10.1016/j.patrec.2018.03.032. hal-02110718
Abstract

We propose an algorithm that efficiently solves the linear sum assignment problem with error-correction and no cost constraints. This problem is encountered for instance in the approximation of the graph edit distance. The fastest currently available solvers for the linear sum assignment problem require the pairwise costs to respect the triangle inequality. Our algorithm is as fast as these algorithms, but manages to drop the cost constraint. The main technical ingredient of our algorithm is a cost-dependent factorization of the node substitutions.

1 Introduction

Finding correspondences between structured or unstructured data is a fundamental problem in data processing and analysis, in particular in computer vision, pattern recognition and machine learning. In various application scenarios, data is represented by attributed graphs, and the correspondence problem hence translates to the task to compute those graphs $G$, from a given collection, that “correspond best” to a given graph $H$.

One way to formalize correspondence between graphs is to say that two attributed graphs $G$ and $H$ correspond well, if and only if they are close to each other w.r.t. some distance measure. A very sensitive and therefore widely used
distance measure is the graph edit distance (GED) \([7, 20, 20]\). It is defined as the minimum cost of an edit path between \(G\) and \(H\). An edit path between \(G\) and \(H\) is a sequence of elementary edit operations that transform \(G\) into \(H\), where the elementary edit operations are removal or insertion of an edge or an isolated node and substitution of a node’s or an edge’s attribute. Each elementary edit operation comes with an associated non-negative edit cost.

Since it is \(NP\)-hard to exactly compute GED \([31]\), algorithms that compute lower and upper bounds are important. Many existing algorithms \([21, 31, 20, 12, 32, 9, 3, 2, 30, 10]\) proceed as follows: In a first step, they lossily transform the problem of computing the graph edit distance between \(G\) and \(H\) into an instance \(C\) of the linear sum assignment problem with error-correction (LSAPE), where \(C\) is a real-valued matrix. LSAPE is similar to the linear sum assignment problem for bipartite graphs (LSAP), but also allows node removals and insertions in addition to node substitutions. In a second step, a LSAPE solver is called to compute an optimal solution \(X^*\) for \(C\). Subsequently, \(X^*\) is interpreted as an edit path between \(G\) and \(H\) and hence gives an upper bound for GED. By using the optimality of \(X^*\), some approaches also allow to derive a lower bound of the GED from the cost of \(X^*\) under \(C\) \([31, 32, 3, 2]\). In addition to that, LSAPE arises as a subproblem in several approaches that refine the upper bound by local or hybrid greedy search strategies \([22, 25, 11]\). This is also the case for approaches based on relaxations of quadratic programming formulations of GED and conditional gradient descent \([3, 10]\).

The runtime performance of the approaches mentioned in the previous paragraph crucially depend on the performance of the LSAPE solver they employ. Existing LSAPE solvers fall in two categories. Solvers of the first kind reduce LSAPE to LSAP \([23, 21, 27, 29, 28]\): Using different strategies, they first transform an instance \(C\) of LSAPE into an instance \(\overline{C}\) of LSAP and then use standard approaches available for LSAP such as the Hungarian Algorithm for computing an optimal solution \(\overline{X}^*\) for \(\overline{C}\). Finally, \(\overline{X}^*\) is transformed into an optimal solution \(X^*\) for the LSAPE instance \(C\). Algorithms of the second kind directly solve LSAPE by adapting existing approaches for LSAP \([13, 14, 4]\). Furthermore, LSAPE solvers can be separated in general methods \([23, 21, 13, 14, 4]\) and cost-constrained methods that are only applicable to those instances of LSAPE that respect the triangle inequality \([27, 29, 28]\).

The fastest currently available LSAPE solver is the algorithm FBP which reduces an instance \(C \in \mathbb{R}^{(m+1)\times(n+1)}\) of LSAPE to an instance \(\overline{C} \in \mathbb{R}^{n\times m}\) of LSAP. However, FBP requires \(C\) to respect the triangle inequality. In this paper, we propose the algorithm FLWC, a fast solver for LSAPE without cost constraints. Like FBP, FLWC reduces \(C\) to a \((n \times m)\)-sized instance of LSAP. Unlike FBP, FLWC does not require \(C\) to respect the triangle inequality. FLWC builds upon the insight that node substitutions whose costs do not respect the triangle inequality can be factorized into removals and insertions. A thorough experimental evaluation shows that, for instances that do respect the triangle inequality, FLWC is as fast as FBP, while, for unconstrained instances, FLWC clearly outperforms all competitors.

The remainder of the paper is organized as follows: In Section \([\_\_\_\_]\) we formally
introduce LSAPE and LSAP. In Section 3, we discuss related work. In Section 4 we present FLWC. In Section 5, we empirically evaluate the performance of FLWC w.r.t. the performance of existing competitors.

2 Preliminaries

2.1 Linear Sum Assignment Problem

The linear sum assignment problem (LSAP) is defined on complete bipartite graphs $K_{n,m} = (U, V, U \times V)$, where $U = \{u_i\}_{i \in I}$ and $V = \{v_j\}_{j \in J}$ are sets of nodes, and $I = \{1, \ldots, n\}$ and $J = \{1, \ldots, m\}$ are index sets. In the following, we will assume w.l.o.g. that $n \leq m$. This can easily be enforced by exchanging the roles of $U$ and $V$, if necessary. An edge set $M \subset U \times V$ is called a maximum matching for $K_{n,m}$ if and only if all nodes in $U$ are incident with exactly one edge in $M$ and all nodes in $V$ are incident with at most one edge in $M$. A maximum matching $M$ can be encoded with a binary matrix $X = (x_{i,j}) \in \{0, 1\}^{n \times m}$ by setting $x_{i,j} = 1$ just in case $(u_i, v_j) \in M$. In the following, we will always represent matchings with their induced binary matrices. The set $\Pi_{n,m}$ of all maximum matchings for $K_{n,m}$ is hence given as follows:

$$\Pi_{n,m} = \left\{ X \in \{0, 1\}^{n \times m} : \right. \\
\quad \forall j \in J, \sum_{i=1}^{n} x_{i,j} \leq 1 \\
\left. \forall i \in I, \sum_{j=1}^{m} x_{i,j} = 1 \right\}$$

(1)

When $n = m$, maximum matchings become perfect matchings, and $\Pi_{n,n}$ becomes the set of all $n \times n$ permutation matrices.

Given a cost matrix $\overline{C} = (c_{i,j}) \in \mathbb{R}^{n \times m}$ for the edges of $K_{n,m}$, LSAP asks to compute a maximum matching for $K_{n,m}$ which minimizes the induced edge costs.

**Problem 1 (LSAP)** Given a cost matrix $\overline{C} \in \mathbb{R}^{n \times m}$, LSAP consists to find a maximum matching $\overline{X}^* \in \Pi_{n,m}$ with $\overline{X}^* = \arg\min_{X \in \Pi_{n,m}} L(\overline{X}, \overline{C})$, where $L(\overline{X}, \overline{C}) = \sum_{i=1}^{n} \sum_{j=1}^{m} c_{i,j} x_{i,j}$.

This is a classical combinatorial optimization problem, also known as the minimum cost maximum (or perfect) matching problem in bipartite graphs. It can be solved in polynomial time and space complexities with several algorithms. For instance, if the cost matrix $\overline{C}$ is balanced, i.e., if $n = m$, it can be solved in $O(n^3)$ time and $O(n^2)$ space with the Kuhn-Munkres Hungarian Algorithm [15, 19]. For the unbalanced case, it can be solved in $O(n^2m)$ time [6]. See [16, 8] for more details.

2.2 Linear Sum Assignment Problem with Error-Correction

Given a maximum matching $\overline{X}$ for $K_{n,m} = (U, V, U \times V)$ and a cost matrix $\overline{C} \in \mathbb{R}^{n \times m}$, a cell $\overline{x}_{i,j}$ of $\overline{X}$ with $\overline{x}_{i,j} = 1$ can be interpreted as a substitution of
the node \( u_i \in \mathcal{U} \) by the node \( v_j \in \mathcal{V} \) with associated substitution cost \( c_{i,j} \). LSAP can hence be viewed as the task to substitute all the nodes of \( \mathcal{U} \) by pairwise distinct nodes of \( \mathcal{V} \) such that the substitution cost is minimized. Note that, under this interpretation, LSAP does not require all nodes in \( \mathcal{V} \) to be taken care of, as, if \( n < m \), there are always nodes in \( \mathcal{V} \) that do not substitute any node in \( \mathcal{U} \).

There are scenarios such as approximation of GED, where one is faced with a slightly different matching problem: First, in addition to node substitutions, one also wants to allow node insertions and node removals. Second, one wants to enforce that all the nodes in \( \mathcal{V} \) are taken care of. The linear sum assignment problem with error-correction (LSAPE) models these settings. To this purpose, the sets \( \mathcal{U} \) and \( \mathcal{V} \) are extended to \( \mathcal{U}_\epsilon = \mathcal{U} \cup \{ \epsilon \} \) and \( \mathcal{V}_\epsilon = \mathcal{V} \cup \{ \epsilon \} \), where \( \epsilon \) is a dummy node. An edge set \( \mathcal{M} = \mathcal{S} \cup \mathcal{R} \cup \mathcal{I} \subset \mathcal{U}_\epsilon \times \mathcal{V}_\epsilon \) is called error-correcting matching for \( K_{n,m,\epsilon} = (\mathcal{U}_\epsilon, \mathcal{V}_\epsilon, \mathcal{U}_\epsilon \times \mathcal{V}_\epsilon) \) if and only if it does not contain the edge \((\epsilon, \epsilon)\) and all nodes in \( \mathcal{U}_\epsilon \) and \( \mathcal{V}_\epsilon \) are incident with exactly one edge in \( \mathcal{M} \). The set \( \Pi_{n,m,\epsilon} \) of all error-correcting matchings for \( K_{n,m,\epsilon} \) is hence given as

\[
\Pi_{n,m,\epsilon} = \left\{ \mathbf{X} = \begin{pmatrix} \mathbf{X}_{\text{sub}} & \mathbf{X}_{\text{rem}} \\ \mathbf{X}_{\text{ins}} & 0 \end{pmatrix} \in \{0, 1\}^{(n+1) \times (m+1)} : \right.
\]

\[
\forall j \in J, \quad x_{i,j} + \sum_{i=1}^{n} x_{i,j} = 1 \\
\forall i \in I, \quad x_{i,\epsilon} + \sum_{j=1}^{m} x_{i,j} = 1 \left. \right\} (2)
\]

where \( \mathbf{X}_{\text{sub}} = (x_{i,j}) \in \mathbb{R}^{n \times m} \) encodes node substitutions, \( \mathbf{X}_{\text{rem}} = (x_{i,\epsilon}) \in \mathbb{R}^{n \times 1} \) encodes node removals, and \( \mathbf{X}_{\text{ins}} = (x_{\epsilon,j}) \in \mathbb{R}^{1 \times m} \) encodes node insertions. Assume now that \( \mathbf{C} \in \mathbb{R}^{(n+1) \times (m+1)} \) is a cost matrix of the form

\[
\mathbf{C} = \begin{pmatrix} \mathbf{C}_{\text{sub}} & \mathbf{C}_{\text{rem}} \\ \mathbf{C}_{\text{ins}} & 0 \end{pmatrix},
\]

where \( \mathbf{C}_{\text{sub}} = (c_{i,j}) \in \mathbb{R}^{n \times m} \) contains the substitution costs, \( \mathbf{C}_{\text{rem}} = (c_{i,\epsilon}) \in \mathbb{R}^{n \times 1} \) contains the removal costs, and \( \mathbf{C}_{\text{ins}} = (c_{\epsilon,j}) \in \mathbb{R}^{1 \times m} \) contains the insertion costs. Then LSAPE asks to compute an error-correcting matching for \( K_{n,m,\epsilon} \) such that the sum of the induced substitution, removal, and insertion costs is minimized.

**Problem 2 (LSAPE)** Given a cost matrix \( \mathbf{C} \in \mathbb{R}^{(n+1) \times (m+1)} \) of form (3), LSAPE consists to find an error-correcting matching \( \mathbf{X}^* \in \Pi_{n,m,\epsilon} \) with \( \mathbf{X}^* \in \arg\min_{\mathbf{X} \in \Pi_{n,m,\epsilon}} L(\mathbf{X}, \mathbf{C}) \), where \( L(\mathbf{X}, \mathbf{C}) = \sum_{i=1}^{n} \sum_{j=1}^{m} c_{i,j} x_{i,j} + \sum_{i=1}^{n} c_{i,\epsilon} x_{i,\epsilon} + \sum_{j=1}^{m} c_{\epsilon,j} x_{\epsilon,j} \).

Table\[\text{I}\] summarizes the notations introduced in this section.
Table 1: Frequently used notations

<table>
<thead>
<tr>
<th>syntax</th>
<th>semantic</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C \in \mathbb{R}^{(n+1) \times (m+1)}$</td>
<td>instance of LSAPE</td>
</tr>
<tr>
<td>$\overline{C} \in \mathbb{R}^{n \times m}$</td>
<td>instance of LSAP</td>
</tr>
<tr>
<td>$X \in \Pi_{n,m,\epsilon}$</td>
<td>error-correcting matching</td>
</tr>
<tr>
<td>$\overline{X} \in \Pi_{n,m}$</td>
<td>maximum matching</td>
</tr>
<tr>
<td>$X^* \in \Pi_{n,m,\epsilon}$</td>
<td>optimal error-correcting matching</td>
</tr>
<tr>
<td>$\overline{X}^* \in \Pi_{n,m}$</td>
<td>optimal maximum matching</td>
</tr>
<tr>
<td>$L(X, C)$</td>
<td>cost of error-correcting matching $X$ w.r.t. $C$</td>
</tr>
<tr>
<td>$L(X, \overline{C})$</td>
<td>cost of maximum matching $X$ w.r.t. $\overline{C}$</td>
</tr>
</tbody>
</table>

3 State of the Art for LSAPE

3.1 Unconstrained Reduction to LSAP

The standard algorithm *extended bipartite matching (EBP)* \([20, 21, 23]\) solves LSAPE by transforming it to LSAP. Given an instance $C$ of LSAPE, EBP constructs an instance $\overline{C} \in \mathbb{R}^{(n+m) \times (m+n)}$ of LSAP as

$$\overline{C} = \begin{pmatrix} C_{\text{sub}} & \omega(1_{m \times m} - I_m) + \text{diag}(c_{\text{rem}}) \\ \omega(1_{n \times n} - I_n) + \text{diag}(c_{\text{ins}}) & 0_{m \times n} \end{pmatrix},$$

where $\omega$ denotes a very large value and the operator $\text{diag}$ maps a vector $(v_i)_{i=1}^k$ to the diagonal matrix $(d_{i,j})_{i,j=1}^k$ with $d_{i,i} = v_i$ and $d_{i,j} = 0$ for all $i \neq j$. EBP then calls a LSAP solver to compute an optimal maximum matching $\overline{X}^* \in \Pi_{n+m,m+n}$ for LSAP. By construction, $\overline{X}^*$ is of the form

$$\overline{X}^* = \begin{pmatrix} \overline{X}_{\text{sub}}^* & \overline{X}_{\text{rem}}^* \\ \overline{X}_{\text{ins}}^* & \overline{X}_{\text{comp}}^* \end{pmatrix},$$

where $\overline{X}_{\text{sub}}^* \in \{0,1\}^{n \times m}$, $\overline{X}_{\text{rem}}^* \in \{0,1\}^{n \times n}$, and $\overline{X}_{\text{ins}}^* \in \{0,1\}^{m \times m}$. The southeast quadrant $\overline{X}_{\text{comp}}^* \in \{0,1\}^{m \times n}$ contains assignments from dummy nodes to dummy nodes which are not needed for encoding node removals or insertions but are anyway computed by LSAP algorithms. $\overline{X}^*$ is then transformed into an optimal error-correcting matching $\overline{X}^* \in \Pi_{n,m,\epsilon}$ for LSAPE with $L(\overline{X}^*, C) = L(\overline{X}^*, \overline{C})$ by setting $X_{\text{sub}}^* = \overline{X}_{\text{sub}}^*$, $x_{\text{rem}}^* = \overline{X}_{\text{rem}}^*1_n$, and $x_{\text{ins}}^* = 1_m^T \overline{X}_{\text{ins}}^*$. The time complexity of EBP is dominated by the complexity of solving the LSAP instance $\overline{C}$. Therefore, EBP runs in $O((n + m)^3)$ time.

3.2 Cost-Constrained Reductions to LSAP

The algorithms *fast bipartite matching (FBP)* \([27]\) and *square fast bipartite matching (SFBP)* \([28, 29]\) build upon more compact reductions of LSAPE to LSAP than EBP. However, both FBP and SFBP are applicable only to those instances $C$ of
LSAPE which respect the following triangle inequalities:

$$\forall (i,j) \in I \times J, \ c_{i,j} \leq c_{i,\epsilon} + c_{\epsilon,j}$$  \hspace{1cm} (5)

In other terms, FBP and SFBP can be used for instances of LSAPE where substituting a node $u_i \in U$ by a node $v_j \in V$ is never more expensive than removing $u_i$ and inserting $v_j$. The following proposition is the key-ingredient of both FBP and SFBP. Recall that we have assumed w.l.o.g. that $n \leq m$.

**Proposition 1 (Cf. [27, 28, 29])** Let $X^* \in \Pi_{n,m,\epsilon}$ be an optimal error-correcting matching for an instance $C \in \mathbb{R}^{(n+1)\times(m+1)}$ of LSAPE which satisfies (5). Then $X$ contains no node removal, i.e., satisfies $x_{\text{rem}} = 0_n$. Note that this implies that $X^*$ contains exactly $m-n$ node insertions and hence that $x_{\text{ins}} = 0_m$, if $n = m$.

Given an instance $C$ of LSAPE that satisfies (5), FBP constructs an instance $C \in \mathbb{R}^{n \times m}$ of LSAP by setting $C = C_{\text{sub}} - c_{\text{rem}} 1^T_n - 1^T_m c_{\text{ins}}$. Subsequently, FBP computes an optimal maximum matching $X^* \in \Pi_{n,m,\epsilon}$ for $C$. Because of Proposition 1 it then holds that the matrix $X^* \in \Pi_{n,m,\epsilon}$ defined by $X^*_{\text{sub}} = X^*$, $x^*_{\text{rem}} = 0_n$, and $x^*_{\text{ins}} = 1^T_m - 1^T_n X^*$ is an optimal error-correcting matching for $C$.

The variant FBP$_0$ of FBP transforms $C$ into a balanced instance $C_0 \in \mathbb{R}^{m \times m}$ of LSAP, by adding the matrix $0_{m-n,n}$ below $C$ defined for FBP. After solving this instance, FBP$_0$ transforms the resulting optimal maximum matching

$$\left( \begin{array}{c} X^* \\ * \end{array} \right)$$

for $C_0$ into an optimal error-correcting matching $X^* \in \Pi_{n,m,\epsilon}$ for $C$, by applying the same transformation rules as FBP on $X^*$.

Like FBP$_0$, SFBP transforms an instance $C$ into a balanced instance $\overline{C} \in \mathbb{R}^{m \times m}$ of LSAP. However, $\overline{C}$ is now defined as follows:

$$\overline{C} = \left( \begin{array}{c} C_{\text{sub}} \\ 1_{m-n} c_{\text{ins}} \end{array} \right)$$

In the next step, SFBP computes an optimal maximum matching

$$X^* = \left( \begin{array}{c} X^*_{\text{sub}} \\ X^*_{\text{ins}} \end{array} \right)$$

for $\overline{C}$. SFBP then constructs the matrix $X^* \in \Pi_{n,m,\epsilon}$ by setting $X^*_{\text{sub}} = X^*_{\text{sub}}$, $x^*_{\text{rem}} = 0_n$, and $x^*_{\text{ins}} = 1^T_n X^*_{\text{ins}}$. Again, Proposition 1 ensures that $X^*$ is indeed an optimal error-correcting matching for LSAPE.

The time complexities of FBP, FBP$_0$, and SFBP are dominated by the complexities of solving the LSAP instances $C$. Therefore, FBP runs in $O(n^2 m)$ time, while FBP$_0$ and SFBP run in $O(m^3)$ time. These are significant improvements over EBP. However, recall that FBP, FBP$_0$, and SFBP can be used only if the cost matrix $C$ respects the triangle inequalities (5).
Adaptations of Classical Algorithms

LSAPE can also be solved directly by adapting algorithms originally designed for LSAP. An adaptation of the Jonker-Volgenant Algorithm is proposed in [14]. An adaption of the Hungarian Algorithm, denoted HNG, in this paper, has been suggested in [4]. Both modifications lead to an overall time complexity of $O(n^2m)$.

4 A Compact Reduction from LSAPE to LSAPE without Cost Constraints

The main contribution of this paper is to show that LSAPE without cost constraints can be reduced to an instance of LSAP of size $n \times m$. The reduction translates into the algorithm FLWC (fast solver for LSAPE without cost constraints), which, like FBP, runs in $O(n^2m)$ time, but, unlike FBP, FBP$_0$, and SFBP, does not assume the costs to respect the triangle inequalities [3]. The following Theorem 1 states the reduction principle. It relies on a cost-dependent factorization of substitutions, removals and insertions. In Section 4.1, we discuss special cases and present FLWC. In Section 4.2, we present the proof of Theorem 1.

Theorem 1 (Reduction Principle) Let $C \in \mathbb{R}^{(n+1) \times (m+1)}$ be an instance of LSAPE and let $X^\star \in \Pi_{n,m}$ be an optimal maximum matching for the instance $C$ of LSAP defined by

$$\forall (i,j) \in I \times J, \quad \tau_{i,j} = \delta_{i,j}\epsilon c_{i,j} + (1 - \delta_{i,j}\epsilon)(c_{i,\epsilon} + c_{\epsilon,j}) - \delta_{n<m}c_{\epsilon,j},$$

(6)

where

$$\delta_{n<m} = \begin{cases} 1 & \text{if } n < m \\ 0 & \text{else} \end{cases} \quad \delta_{i,j}\epsilon = \begin{cases} 1 & \text{if } c_{i,j} \leq c_{i,\epsilon} + c_{\epsilon,j} \\ 0 & \text{else} \end{cases}$$

Furthermore, let the function $f_C : \Pi_{n,m} \to \Pi_{n,m,\epsilon}$ be defined as follows:

$$\forall (i,j) \in I \times J, \quad x_{i,j} = (f_C(X))_{i,j} = \delta_{i,j}\epsilon \tau_{i,j}$$

(7)

$$\forall i \in I, \quad x_{i,\epsilon} = (f_C(X))_{i,\epsilon} = 1 - \sum_{j=1}^{m} \delta_{i,j}\epsilon \tau_{i,j}$$

(8)

$$\forall j \in J, \quad x_{\epsilon,j} = (f_C(X))_{\epsilon,j} = 1 - \sum_{i=1}^{n} \delta_{i,j}\epsilon \tau_{i,j}$$

(9)

Then $X^\star = f_C(X^\star)$ is an optimal error-correcting matching for $C$ with cost $L(X^\star, C) = L(X^\star, \tilde{C}) + \delta_{n<m} \sum_{j=1}^{m} c_{\epsilon,j}$.

Example 1 Assume that $n = 2$, $m = 3$, and consider the instance $C$ of LSAPE and the induced instance $\tilde{C}$ of LSAP:

$$C = \begin{pmatrix} 1 & 2 & 3 & \epsilon \\ 1 & 3 & 5 & 1 & 4 \\ 2 & 8 & 9 & 4 & 4 \\ \epsilon & 2 & 4 & 0 & 0 \end{pmatrix} \quad \text{apply } (6) \quad \tilde{C} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \\ 2 & 4 & 4 \end{pmatrix}$$
(a) $\mathbf{X}^*$  
(b) $\mathbf{X}^*$  
(c) $\mathbf{X}^*$  
(d) $\mathbf{X}^*$

Figure 1: Illustration of the reduction principle stated in Theorem 1

For instance, we have $c_1,1 = \frac{\delta}{C_1,1} + (1 - \frac{\delta}{C_1,1})(2 + 4) - \delta_2 < 3 = 3 - 2 = 1$ and $\bar{\tau}_{2,2} = \frac{\delta}{C_2,2} + (1 - \frac{\delta}{C_2,2})(4 + 4) - \delta_2 < 3 = 4$. Fig. 1(a) shows an optimal maximum matching $\mathbf{X}^* \in \Pi_{n,m}$ for $C$, and Fig. 1(b) shows an optimal error-correcting matching $\mathbf{X}^* = f_C(\mathbf{X}^*) \in \Pi_{n,m,\epsilon}$ for $C$. Note that $f_C$ factorizes the substitution $(u_2, v_2)$ into the removal $(u_2, \epsilon)$ and the insertion $(\epsilon, v_2)$, since $c_2,2 > c_2,\epsilon + c_\epsilon,2$ (Fig. 1(c)). On the other hand, the substitution $(u_1, v_1)$ is not factorized, since $c_1,1 \leq c_1,\epsilon + c_\epsilon,1$ (Fig. 1(d)). Furthermore, we have $L(\mathbf{X}^*, C) = 11$, $L(\mathbf{X}^*, \bar{C}) = 5$, and $\delta_{n < m} \sum_{j=1}^{m \epsilon, j} = 6$. Therefore, it holds that $L(\mathbf{X}^*, C) = L(\mathbf{X}^*, \bar{C}) + \delta_{n < m} \sum_{j=1}^{m \epsilon, j}$, as stated by our reduction principle.

### 4.1 Discussion of Special Cases and Presentation of FLWC

Theorem 1 states that a general instance $C \in \mathbb{R}^{(n+1) \times (m+1)}$ of LSAPE, which is not required to respect the triangle inequalities (5), can be reduced to a $(n \times m)$-sized instance of LSAP. However, there is still room for improvement if $C$ does respect the triangle inequalities.

**Proposition 2** Let $C \in \mathbb{R}^{(n+1) \times (m+1)}$ be an instance of LSAPE that respects the triangle inequalities (5). Then the reduction principle specified in Theorem 1 can be carried out without knowledge of the removal costs $c_{\text{rem}}$. If $n = m$ holds, too, then knowledge of the insertion costs $c_{\text{ins}}$ is not necessary, either.

**Proof 1** The proposition immediately follows from the facts that $\delta_{i,j,\epsilon}^C = 1$ holds for all $(i, j) \in I \times J$ if $C$ respects (6), and that $\delta_{n < m} = 0$ if $n = m$. $\Box$

Proposition 2 is useful, because in many application scenarios for LSAPE, $C$ is not given but has to be computed. Furthermore, $C$ is often known a priori to respect the triangle inequalities, for instance, because its entries contain the distances between elements in a metric space. In such settings, the overall performance of an algorithm that calls a LSAPE solver as a subroutine can improve significantly if only parts of the cost matrix $C$ have to be computed. Table 2 summarizes which parts of $C$ have to be known for our reduction from LSAPE to LSAP. The case $n > m$ can straightforwardly be obtained from the
case \( n < m \) by transposing \( C \). Recall that both FBP and FBP\(_0\) always require the entire cost matrix \( C \) to be known. SFBP requires the same parts of \( C \) as our approach, but reduces LSAPE to a larger instance of LSAP (\( \max\{n, m\} \times \max\{n, m\} \) vs. \( n \times m \)).

Table 2: Required parts of \( C \) for our reduction from LSAPE to LSAP

<table>
<thead>
<tr>
<th>( n = m )</th>
<th>( n &lt; m )</th>
<th>( n &gt; m )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C )</td>
<td>( C, c_{\text{ins}} )</td>
<td>( C, c_{\text{ins}}, c_{\text{rem}} )</td>
</tr>
<tr>
<td>( C, c_{\text{ins}}, c_{\text{rem}} )</td>
<td>( C, c_{\text{ins}}, c_{\text{rem}} )</td>
<td>( C, c_{\text{ins}}, c_{\text{rem}} )</td>
</tr>
</tbody>
</table>

Algorithm 1 shows the algorithm FLWC, which turns our reduction from LSAPE to LSAP into a method for computing an optimal error-correcting matching. FLWC only uses those parts of \( C \) which are really required by the reduction (cf. Table 2).

Algorithm 1: FLWC

input: an instance \( C \in \mathbb{R}^{(n+1)\times(m+1)} \) of LSAPE
output: an optimal error-correcting matching \( X^* \in \Pi_{n,m,\epsilon} \) for \( C \)

1 \text{ if } n > m \text{ then } 
2 \quad C \leftarrow C^T; (n, m) \leftarrow (m, n); 
3 \text{ initialize } \overline{C} \in \mathbb{R}^{n \times m}; 
4 \text{ for } i \in \{1, \ldots, n\} \text{ do } 
5 \quad \text{ for } j \in \{1, \ldots, m\} \text{ do } 
6 \quad \text{ if } C \text{ respects triangle inequalities then } 
7 \quad \quad \text{ if } n = m \text{ then } 
8 \quad \quad \quad \tau_{i,j} \leftarrow c_{i,j}; 
9 \quad \quad \text{ else } 
10 \quad \quad \quad \tau_{i,j} \leftarrow c_{i,j} - c_{i,j}; 
11 \quad \quad \text{ else } 
12 \quad \quad \quad \tau_{i,j} \leftarrow \delta_{i,j} c_{i,j} + (1 - \delta_{i,j})(c_{i,\epsilon} + c_{\epsilon,j}) - \delta_{n<m} c_{i,j}; 
13 \text{ call LSAP solver to compute opt. max. matching } X^* \in \Pi_{n,m} \text{ for } \overline{C}; 
14 \quad X^* \leftarrow f_C(X^*); 
15 \text{ if } C \text{ was transposed in lines 1-2 then } 
16 \quad X^* \leftarrow X^*^T; 
17 \text{ return } X^*; 

If the adaption of the Hungarian Algorithm to unbalanced instances of LSAP is used in line 13, FLWC runs in \( O(\min\{n, m\}^2 \max\{n, m\}) \) time and \( O(nm) \) space. Table 3 compares FLWC’s time and space complexities to the complexities of existing competitors. Note that our reduction principle can also be used for the fast computation of a suboptimal solution for LSAPE. To this end, it suffices
Consider the error-correcting matching

Example 2

This gives us

we have

∑

X

removes

∑

j

|U

have

∑

j

∈U

rem

|V

let


Proof 2

Let

Furthermore, let

s

1

be a maximum matching for

⋆

be the set of indices of those nodes of

U

and

V

whose edges encode all combinations of node removals and insertions, let

Z∗

be the set of of maximum matchings for

Z

, and let the set

YX

be defined as follows:

\[ Y_X = \{ X_{\text{sub}} + Z^* : Z^* \in Z^* \} \]  

(10)

Then

Y

\in \Pi_{n,m}

holds for each

Y

\in Y_X.

Proposition 3

Let

X\in \Pi_{n,m,\epsilon}

be an error-correcting matching. Furthermore, let

Z_X = (U, V, \{ (u_i, v_j) \in U \times V : x_i, x_j = 1 \})

be the bipartite graph between

U

and

V

whose edges encode all combinations of node removals and insertions, let

Z^*

be the set of maximum matchings for

Z_X,

and let

\{ v_j : j \in J \setminus J_{\text{sub}} \}

be the number of substitutions encoded by

X,

and let

Z^*

be a maximum matching for

Z_X.

We observe that

Z_X

can be viewed as the complete bipartite graph between the nodes

U_{\text{rem}} = \{ u_i : i \in I \setminus I_{\text{sub}} \}

and

V_{\text{ins}} = \{ v_j : j \in J \setminus J_{\text{sub}} \}

that are removed and inserted by

X,

and that we have

|U_{\text{rem}}| = n - s \leq m - s = |V_{\text{ins}}|.

These observations imply that we have

\[ \sum_{j \in J} z_{i,j} = 0 \]

for each

i \in I_{\text{sub}}

and

\[ \sum_{j \in J} z_{i,j}^* = 1 \]

for each

i \in I \setminus I_{\text{sub}}.

Similarly, we have

\[ \sum_{i \in I} z_{i,j}^* = 0 \]

for each

j \in J_{\text{sub}}

and

\[ \sum_{i \in I} z_{i,j} \leq 1 \]

for each

j \in J \setminus J_{\text{sub}}.

This gives us

\[ \sum_{i \in I} x_{i,j} + z_{i,j}^* = 1 \]

for each

i \in I

and

\[ \sum_{i \in I} x_{i,j} + z_{i,j} \leq 1 \]

for each

j \in J,

which implies

X_{\text{sub}} + Z^* \in \Pi_{n,m,\epsilon}.

Example 2

Consider the error-correcting matching

X

shown in Fig 2(a). Since

X

removes

u_2

and inserts

v_2

and

v_3,

\[ Z_X \]

contains the edges

(u_2, v_2)

and

(u_3, v_3)

(Fig 2(b)). There are exactly two maximum matchings for

Z_X,

namely

\[ Z_1^* = \]

\[ Z_2^* = \]
{(u_2, v_2)} \text{ and } Z^*_2 = \{(u_2, v_3)\}. This implies that Y_X = \{Y_1, Y_2\} with Y_1 = \{(u_1, v_1), (u_2, v_2)\} (Fig. 2(c)) and Y_2 = \{(u_1, v_1), (u_2, v_3)\} (Fig. 2(d)).

We now introduce the notion of a minimally-sized error-correcting matching. To this purpose, we call two error-correcting matchings $X, X' \in \Pi_{n,m,\epsilon}$ equivalent w. r. t. an instance $C$ of LSAPE (in symbols: $X \sim_C X'$) if and only if, for all $(i,j) \in I \times J$, $x_{i,j} = x'_{i,j}$ or $c_{i,j} = c_{i,\epsilon} + c_{\epsilon,j}$ and $x_{i,j} = x'_{i,\epsilon}x_{\epsilon,j}$ or $x'_{i,j} = x_i x_{\epsilon,j}$. By definition of $\sim_C$, the cost $L$ is invariant on the equivalence classes induced by $\sim_C$.

**Definition 1 (Minimally-Sized Error-Correcting Matching)** Let $C \in \mathbb{R}^{(n+1)\times(m+1)}$ be an instance of LSAPE and $X \in \Pi_{n,m,\epsilon}$ be an error-correcting matching. Then $X$ is called minimally-sized if and only if $|\text{supp}(X)| < |\text{supp}(X')|$ holds for all $X' \in [X]_{\sim_C}$, where $\text{supp}(X)$ is the support of $X$.

In other words, $X$ is minimally-sized just in case it always favours substitution over removal plus insertion, if the costs are the same. By construction, each equivalence class $[X]_{\sim_C}$ contains exactly one minimally-sized error-correcting matching. In particular, there is always a minimally-sized optimal error-correcting matching.

The next step is to characterize a subset $\Pi_{n,m,\epsilon}(C) \subseteq \Pi_{n,m,\epsilon}$ which contains all optimal minimally-sized error-correcting matchings for a given instance $C$ of LSAPE.

**Proposition 4** Let $C \in \mathbb{R}^{(n+1)\times(m+1)}$ be an instance of LSAPE, and let the set $\Pi_{n,m,\epsilon}(C) \subseteq \Pi_{n,m,\epsilon}$ of cost-dependent error-correcting matchings be defined as follows:

$$\Pi_{n,m,\epsilon}(C) = \{ X \in \Pi_{n,m,\epsilon} : \forall (i,j) \in I \times J, \ (1 - \delta^C_{i,j,\epsilon})x_{i,j} = 0, \forall (i,j) \in I \times J, \ \delta^C_{i,j,\epsilon}x_i x_{\epsilon,j} = 0 \}$$

Then $\Pi_{n,m,\epsilon}(C)$ contains all minimally-sized optimal error-correcting matchings for the instance $C$ of LSAPE.

**Proof 3** Assume that there is a minimally-sized optimal error-correcting matching $X^* \in \Pi_{n,m,\epsilon} \setminus \Pi_{n,m,\epsilon}(C)$. Then there is a pair $(i,j) \in I \times J$ such that $(1 - \delta^C_{i,j,\epsilon})x_{i,j}^* = 1$ or $\delta^C_{i,j,\epsilon}x_i^* x_{\epsilon,j}^* = 1$. Assume that we are in the first case, i. e.,

\begin{figure}[h]
\centering
\begin{tikzpicture}
\tikzstyle{every node}=[inner sep=1.5pt,fill,circle,minimum size=6pt]
\node (v1) at (0,0) [label=above:u_1]{}
\node (v2) at (0,-1) [label=above:u_2]{}
\node (v3) at (1,-1) [label=above:v_3]{}
\node (v4) at (1,0) [label=above:v_2]{}
\node (v5) at (1,1) [label=above:v_1]{}
\draw (v1) edge (v2)
\draw (v2) edge (v3)
\draw (v3) edge (v4)
\draw (v4) edge (v5)
\draw (v5) edge (v1)
\end{tikzpicture}
\caption{Illustration of Proposition 3}
\end{figure}
that $\delta_{i,j}^c = 0$ and $x_{i,j}^* = 1$. This implies $c_{i,j} > c_{i,\epsilon} + c_{\epsilon,j}$. Now consider the error-correcting matching $X'$, which, instead of substituting $u_i$ by $v_j$, removes $u_i$ and inserts $v_j$ ($x_i' = x_j^* = 1$). Since $\delta_{i,j}^c = 0$, $X'$ is cheaper than $X^*$. This contradicts $X^*$’s optimality. If we are in the second case, we have $x_{i,\epsilon} + x_{\epsilon,j}^* = 1$ and $c_{i,j} \leq c_{i,\epsilon} + c_{\epsilon,j}$. Since $X^*$ is minimally-sized, we can strengthen the last inequality to $c_{i,j} < c_{i,\epsilon} + c_{\epsilon,j}$. Consider the error-correcting matching $X'$, which, instead of removing $u_i$ and inserting $v_j$, substitutes $u_i$ by $v_j$ ($x_i' = x_j^* = 1$). Again, $X'$ is cheaper than $X^*$, which is a contradiction to $X^*$’s optimality.\]

The following Proposition 5 shows that the transformation function $f_C$ defined in Theorem 4 indeed maps maximum matchings to error-correcting matchings and that it is surjective on $\Pi_{n,m,\epsilon}$.

**Proposition 5** Let $f_C : \Pi_{n,m} \rightarrow \Pi_{n,m,\epsilon}$ be defined as in Theorem 4. Then it holds that $\text{img}(f_C) \subseteq \Pi_{n,m,\epsilon}$ and that $\text{img}(f_C) \supseteq \Pi_{n,m,\epsilon}(C)$.

**Proof 4** Consider a maximum matching $X \in \Pi_{n,m}$ and let $X = f_C(X)$. From (7) and (8), we have $x_{i,\epsilon} + \sum_{j=1}^{m} x_{i,j} = 1$ for each $i \in I$. From (7) and (9), we have $x_{\epsilon,j} + \sum_{i=1}^{n} x_{i,j} = 1$ for each $j \in J$. This implies $X \in \Pi_{n,m,\epsilon}$ and thus $\text{img}(f_C) \subseteq \Pi_{n,m,\epsilon}$.

For showing $\text{img}(f(C)) \supseteq \Pi_{n,m,\epsilon}(C)$, we fix an error-correcting matching $X \in \Pi_{n,m,\epsilon}(C)$. From Proposition 3, there is a set $Y_X$ of maximum matchings representing $X$. Consider a matching $\overline{X} \in Y_X$, i.e., $\overline{X} = X_{\text{sub}} + Z^*$ with $Z^* \in Z_X$. We will show that $X = f_C(\overline{X})$, which proves the proposition. To this end, we first show the following equalities:

$$\forall (i, j) \in I \times J, \quad \delta_{i,j,\epsilon} x_{i,j}^* = 0 \quad (11)$$

Consider a pair $(i, j) \in I \times J$ with $x_{i,j}^* = 1$. From $Z_X \in Z_X$, we know that $x_{i,\epsilon} x_{\epsilon,j} = 1$. As $X \in \Pi_{n,m,\epsilon}(C)$, this implies $\delta_{i,j,\epsilon} = 0$ and hence proves (11).

Now let $(i, j) \in I \times J$. It holds that $f_C(\overline{X})_{i,j} = \delta_{i,j}^c x_{i,j} + \delta_{i,j,\epsilon} x_{i,j}^* = \delta_{i,j}^c x_{i,j} = x_{i,j}$, where the first equality follows from the definitions of $f_C$ and $\overline{X}$, the second equality follows from (11), and the third equality follows from $X \in \Pi_{n,m,\epsilon}(C)$. We have hence shown that $X_{\text{sub}} = f_C(\overline{X})_{\text{sub}}$. Next, we show that $f_C(\overline{X})_{\text{sub}} = X_{\text{ins}}$. Let $j \in J$. We have $f_C(\overline{X})_{\epsilon,j} = 1 - \sum_{i=1}^{n} \delta_{i,j,\epsilon} x_{i,j} - \sum_{i=1}^{n} \delta_{i,j}^c x_{i,j} = 1 - \sum_{i=1}^{n} \delta_{i,j,\epsilon} x_{i,j} - \sum_{i=1}^{n} x_{i,j} = x_{\epsilon,j}$, as required. Again, the first equality follows from the definitions of $f_C$ and $\overline{X}$, the second equality follows from (11), and the third equality follows from $X \in \Pi_{n,m,\epsilon}(C)$. The last equality follows from the fact that $X$ is an error-correcting matching. The argument for showing that $f_C(\overline{X})_{\text{rem}} = X_{\text{rem}}$ is analogous.\]

Now we prove the correctness of our reduction principle.

**Proof of Theorem 1** Let $X \in \Pi_{n,m}$ be a maximum matching. Its cost
Since $X$ is a maximum matching for $\Pi_{n,m}$, we know that $A_i = 1 - \sum_{j=1}^{m} \delta_{i,j,\epsilon} \pi_{i,j} = f_C(X)_{i,\epsilon}$ for each $i \in I$. We now distinguish the cases $\delta_{n<\epsilon} = 1$ and $\delta_{n<\epsilon} = 0$. In the first case, we immediately have $B_j = 1 - \sum_{i=1}^{n} \delta_{i,j,\epsilon} \pi_{i,j} = f_C(X)_{i,\epsilon}$ for each $j \in J$. In the second case, we have $n = m$ and $B_j = f_C(X)_{i,\epsilon}$ holds, too, since, for balanced instances, a maximum matching contains an edge $(u_i, v_j)$ for each $i \in I$. We have thus shown the following equality:

$$\forall X \in \Pi_{n,m}, \quad L(X, C) = L(f_C(X), C) - \delta_{n<\epsilon} \sum_{j=1}^{m} c_{i,j}$$  \hspace{1cm} (12)

Now let $X^\star$ be a minimally-sized optimal error-correcting matching for $C$, $X^\star$ be an optimal maximum matching for $C$, and $X' = f_C(X^\star)$. From [12], we know that $L(X', C) = L(X^\star, C) + \delta_{n<\epsilon} \sum_{j=1}^{m} c_{i,j}$. Therefore, the theorem follows if we can show that $L(X', C) = L(X'^\star, C)$. The $\geq$-direction of the desired equality follows from $X^\star$’s optimality and the fact that, from Proposition [3], we know that $X' \in \Pi_{n,m,\epsilon}$. For showing the $\leq$-direction, we pick an $X' \in \Pi_{n,m}$ with $f_C(X') = X^\star$. Such an $X'$ exists because, from Proposition [4], we know that $X^\star \in \Pi_{n,m,\epsilon}(C)$, and, from Proposition [3], we have $\text{img}(f_C) \supseteq \Pi_{n,m,\epsilon}(C)$. Assume that $L(X', C) > L(X'^\star, C)$. Then [12] implies that $L(X'^\star, C) > L(X', C)$, which contradicts $X^\star$’s optimality. $\square$

5 Experimental Validation

In our experiments, we compared FLWC to all existing competitors mentioned in Table 3. We evaluated the runtime of the different methods (Section 5.1) and their performances when used within approximate approaches for the computation of GED (Section 5.2). All procedures use the same C++ implementation.
of the Hungarian Algorithm as LSAP solver, which is based on the version presented in [16, 8]. HRGε is also based on this version, so that all procedures are comparable.

5.1 Time Comparison

To illustrate the differences between the methods, we recorded their execution time on three types of instances, some of which do and some of which do not respect the triangle inequalities: Machol-Wien instances, flat instances, and random instances. Experiments on further types yielded similar results. Fig. 3 shows the results for Machol-Wien instances and flat instances, Fig. 4 displays...
the results for random instances. For each instance type, the tested methods’
execution time is reported for several values of $n$ and $m$. In the first and
third columns of the figures, $n$ is fixed and $m$ is varied, while in the second
and fourth columns, $m$ is fixed and $n$ is varied. In any case, we have $n \leq m$.
The general methods FLWC, EBP, and HNG, were tested on all instances, the
cost-constrained algorithms FBP, FBP$_0$, and SFBP only on those that respect the
triangle inequalities. Since FBP$_0$ was slower than FBP across all instances, our
plots do not contain curves for FBP$_0$.

Machol-Wien Instances [18]  Machol-Wien instances are defined by $c_{i,j} = i \cdot j \forall (i,j)$ and thus satisfy the triangle inequalities. They are known to be hard
to optimize for classical LSAP solvers such as the Hungarian Algorithm. This
is also the case for the LSAPE solver HNG$_\epsilon$, but not for the other methods. This
is explained by the fact that all methods except for HNG$_\epsilon$ transform an instance
$C$ of LSAPE into an instance $\overline{C}$ of LSAP, which reduces the difficulty of the
problem. The first row of Fig. 3 shows the results of our experiments on Machol-
Wien instances. We observe that FLWC and FBP perform very similarly, provide
the best results in all cases, and are more stable than the other methods. EBP
and SFBP are more time consuming, in particular when $n$ and $m$ increases.

Flat Instances  In a first series of experiments, we considered flat instances
of the form $C = \alpha 1_{n \times m}$, with $\alpha = 10$. These instances satisfy the triangle
inequalities. Moreover, they are easy to solve, which implies that a large part of
the execution time is spent on the initialization step of the Hungarian Algorithm
and the cost transformations. The results for instances of this kind are displayed
in the second row of Fig. 3. As before, FLWC and FBP are more stable and
efficient than the other methods. In a second series of experiments, we varied
the construction of $C$ in order to enforce that $k\%$ of the nodes contained in
the smaller set be removed. This can be achieved by randomly selecting $m - n$
nodes in the larger set and setting their insertion cost to 0. Subsequently, $k\%$ of
the remaining elements are selected and their insertion cost is set to $(\alpha/2) - 1$.
Similarly, the removal costs of $k\%$ of the nodes contained in the smaller set are
set to $(\alpha/2) - 1$. The resulting instances do not satisfy the triangle inequalities,
and thus FBP, FBP$_0$, and SFBP are not tested (they do not compute an optimal
solution). The results are shown in the third and the fourth row of Fig. 3. Both
FLWC and HNG$_\epsilon$ clearly outperform EBP, and FLWC is slightly more stable than
HNG$_\epsilon$.

Random instances  We also carried out experiments on random instances
similar to the ones presented in [28, 29]. These instances are very similar to the
ones that occur in the context of approximation of GED. The cost $c_{i,j}$ encodes
the cost of substituting a node $v_i$ and its set of incident edges in a graph $G$
by a node $v_j$ and its set of incident edges in a graph $H$. Graphs are constructed
locally by assigning node degrees randomly from 1 to $d_{\text{max}} = (3/10)|V_G|$, where
$|V_G|$ is the number of nodes in the graph $G$. Nodes are labeled randomly with
Figure 4: Time comparison for random instances of LSAPE (see text).

an integer value in \( \{1, \ldots, t_{\text{max}}\} \), where \( t_{\text{max}} = \sqrt{n \times m}/10 \). Similarly, edges are labeled with a binary value. Then \( c_{i,j} \) is defined as the cost of substituting the labels (0 if they are the same, or the constant \( c_s \) else) plus the cost of an optimal error-correcting matching between the sets of incident edges. The cost of substituting two edges is defined as before, and the cost of removing or inserting an edge is defined by the constant \( c_{\text{eri}} \). Similarly, removal and insertion costs are defined as the cost of removing the node plus the cost of removing its incident edges, which is given by \( d_i c_{\text{eri}} + c_{\text{nri}} \), where \( d_i \) is the node degree and \( c_{\text{nri}} \) is its removal or insertion cost. The three following set of parameters \((c_s, c_{\text{nri}}, c_{\text{eri}})\) are considered: \((40, 20, 20)\), \((20, 40, 0)\), and \((40, 2, 2)\). The first parameter setting (first row of Fig. 4) satisfies the triangle inequality. The second (second row of Fig. 4) and the third parameter setting (third row of Fig. 4) do not satisfy the triangle inequalities. However, for the second setting, none of the computed optimal error-correcting contains removals and insertions, and so all methods were tested. On the contrary, optimal error-correcting matchings for the third setting indeed contain up to 20% of removals, which is why we did not carry out tests for the cost-constrained methods FBP, FBP0, and SFBP. We again observe that our method FLWC is globally the most stable algorithm and obtains the best results on average.
5.2 Effect on Approximation of GED

As mentioned in the introduction, many methods approximate GED by lossily transforming the problem of its computation into an instance $C$ of LSAPE [20 12 8 21 31 32 9]. Lossy transformations from GED to LSAPE can be used for computing both upper and lower bounds for GED. Given an error-correcting matching for $C$ computed by any LSAPE algorithm, an upper bound for GED is derived by computing the distance associated to the edit path induced by the considered matching. Since this edit path may be suboptimal, the computed distance might be an overestimation of the exact GED, and thus constitutes an upper bound. With this paradigm, each error-correcting matching for $C$ yields a valid upper bound, although the ones induced by optimal matchings are usually tighter. Note that, if the LSAPE instance $C$ constructed by a lossy transformation has only one optimal solution, each optimal LSAPE solver yields the same upper bound for GED. If $C$ has more than one optimal solution, there might be differences in accuracy, since different LSAPE solvers might pick different optimal solutions depending on the organization of the input data. However, these differences in accuracy are arbitrary and hence disappear once the upper bounds are averaged across enough pairs of input graphs. In other words, when it comes to the approximation of GED, the only relevant property of an exact LSAPE solver is its runtime behaviour.

This observation is empirically confirmed by the experiments reported in Table 4, which shows how different methods for solving LSAPE affect the performance of the algorithm BP suggested in [21], which computes an upper bound for GED. Given two graphs $G_1$ and $G_2$ on $n_1$ and $n_2$ nodes, respectively, BP constructs a LSAPE instance $C \in \mathbb{R}^{(n_1+1) \times (n_2+1)}$. For computing the cell $c_{i,j}$ for $(i,j) \in \{1, \ldots, n_1\} \times \{1, \ldots, n_2\}$, BP has to solve another LSAPE instance of size $(\text{deg}_{G_1}(u_i) + 1) \times (\text{deg}_{G_2}(v_j) + 1)$, where $\text{deg}_{G_1}(u_i)$ is the degree of node $u_i$ in $G_1$ and $\text{deg}_{G_2}(v_j)$ is the degree of node $v_j$ in $G_2$. So altogether, BP has to solve $1 + n_1 n_2$ instances of LSAPE. The tests were carried out on the datasets Acyclic and MAO from the ICPR GED contest [1]. In order to ensure that also the cost-constrained LSAPE solvers compute optimal solutions, we defined metric edit costs by setting the cost of substituting nodes and edges to 1 and the cost of deleting and inserting nodes and edges to 3. We see that, as expected, there are no significant differences between the different solvers w.r.t. the tightness of the produced upper bounds. In terms of runtime, our solver FLWC performs best, followed by FBP and HNG.

In contrast to the situation for upper bounds, lossy transformations from GED to LSAPE which aim at the computation of lower bounds [3 2] as well as methods based on conditional gradient descent [5 4] where LSAPE occurs as a subproblem crucially depend on the optimality of the computed error-correcting matching. Therefore, these methods cannot use the existing fast LSAPE solvers FBP, FBP0, and SFBP, unless the triangle inequalities are known to be satisfied. For instance, lower bounds for GED are obtained from the cost $L(X^*, C)$ of an optimal error-correcting matching for the LSAPE instance $C$. If $X^*$ is not optimal, $L(X^*, C)$ is not in general a valid lower bound.
Table 4: Effect of LSAPE methods on algorithm BP suggested in [21] that computes an upper bound for GED on datasets with metric costs.

<table>
<thead>
<tr>
<th>method</th>
<th>time in µs</th>
<th>average UB</th>
<th>time in µs</th>
<th>average UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acyclic</td>
<td></td>
<td></td>
<td>MAO</td>
<td></td>
</tr>
<tr>
<td>cost-constrained methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FBP</td>
<td>1.61</td>
<td>38.86</td>
<td>6.62</td>
<td>107.97</td>
</tr>
<tr>
<td>FBP₀</td>
<td>2.04</td>
<td>38.89</td>
<td>8.10</td>
<td>107.85</td>
</tr>
<tr>
<td>SFBP</td>
<td>2.14</td>
<td>39.28</td>
<td>11.10</td>
<td>107.48</td>
</tr>
<tr>
<td>general methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EBP</td>
<td>4.84</td>
<td>39.14</td>
<td>23.30</td>
<td>107.48</td>
</tr>
<tr>
<td>HNG₀</td>
<td>1.87</td>
<td>38.97</td>
<td>7.46</td>
<td>107.59</td>
</tr>
<tr>
<td>FLWC</td>
<td>1.53</td>
<td>38.86</td>
<td>6.05</td>
<td>108.04</td>
</tr>
</tbody>
</table>

Table 5 shows how different methods for solving LSAPE affect the performance of the algorithm BRANCH suggested in [3, 2], which computes a lower bound for GED. Like BP, given to graphs $G_1$ and $G_2$ on $n_1$ and $n_2$ nodes, respectively, BRANCH has to solve $1 + n_1n_2$ instances of LSAPE. One of them is of size $(n_1 + 1) \times (n_2 + 1)$, the other ones are of size $(\text{deg} G_1(u_i) + 1) \times (\text{deg} G_2(v_j) + 1)$.

The tests were again carried out on the datasets Acyclic and MAO, but this time, edit costs that do not satisfy the triangle inequality were used (cf. [1] for more details on the edit costs, especially equation (1) and setting 3 in Table 2).

First of all, we can see that the classic approach EBP requires more computational time than other optimized approaches, and that FLWC, FBP, and HNG₀ are the fastest methods. Second, we observe that the cost-constrained methods FBP, FBP₀ and SFBP are not able to deal with costs which do not satisfy the triangle inequality: Very often, they compute invalid lower bounds that exceed the exact GED, which we computed using a binary linear programming approach [17]. This is explained by the fact that, if the triangle inequality is not satisfied, optimal error-correcting matchings might well include both insertions and removals. However, those error-correcting matchings are not considered by cost-constrained methods for LSAPE.

Figure 5 shows a case where the cost-constrained methods FBP, FBP₀ and SFBP compute an invalid lower bound. Considering the two graphs $G_1$ and $G_2$ extracted from Acyclic dataset, cost-constrained methods provide a matching $\varphi_{\text{FBP}}$ that favours node substitution over removal plus insertion even if removal plus insertion is cheaper. For instance, $\varphi_{\text{FBP}}$ matches the nodes 7 and 8 of $G_1$ to the nodes 9 and 5 of $G_2$, although it is cheaper to first remove 7 and 8 and then insert 9 and 5, as done by the matching $\varphi_{\text{FLWC}}$ computed by FLWC. In conclusion, cost-constrained methods do not guarantee valid lower bounds for general edit costs, while our algorithm FLWC does.
Table 5: Effect of LSAPE methods on algorithm BRANCH suggested in [3] that computes a lower bound for GED on datasets with non metric costs.

<table>
<thead>
<tr>
<th>method</th>
<th>time in $\mu s$</th>
<th># invalid LB</th>
<th>time in $\mu s$</th>
<th># invalid LB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acyclic</td>
<td>MAO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cost-constrained methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FBP</td>
<td>1.28</td>
<td>4</td>
<td>8.30</td>
<td>416</td>
</tr>
<tr>
<td>FBP$\theta$</td>
<td>1.64</td>
<td>4</td>
<td>10.80</td>
<td>416</td>
</tr>
<tr>
<td>SFBP</td>
<td>1.91</td>
<td>4</td>
<td>12.81</td>
<td>416</td>
</tr>
<tr>
<td>general methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EBP</td>
<td>5.34</td>
<td>0</td>
<td>21.10</td>
<td>0</td>
</tr>
<tr>
<td>HNG$\epsilon$</td>
<td>1.54</td>
<td>0</td>
<td>7.59</td>
<td>0</td>
</tr>
<tr>
<td>FLWC</td>
<td>1.23</td>
<td>0</td>
<td>8.50</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 5: Example of lower bound computation where FBP based methods do not allow to compute a valid lower bound.

6 Conclusions

In this paper, we presented FLWC, a new efficient method for solving the linear sum assignment problem with error correction. The technical backbone of our method is a cost-dependent factorization of substitutions into removals and insertions. FLWC runs in $O(\min\{n, m\}^2 \max\{n, m\})$ time and $O(nm)$ space. Its complexities are hence the same as the ones of the most efficient state of the art methods FBP and HNG$\epsilon$.

The advantage of our method FLWC over FBP is that, unlike FBP, FLWC remains valid for any configuration of the cost matrix and does not require the triangle inequality to holds. Therefore, our method allows to efficiently retrieve a lower
bound for the graph edit distance irrespectively of whether or not the edit costs respect the triangle inequality. This is especially important in settings where the edit costs are deduced from calculus and where one does not have an a priori guarantee that the triangle inequality will be satisfied. Situations of this kind occur, for instance, in some quadratic approximations of the graph edit distance problem [5, 4].

One advantage of FLWC over HNG is that, although both algorithms have the same time complexity, FLWC is slightly faster in practice. In particular, FLWC is more stable than HNG, in the sense that it also allows to quickly solve difficult LSAP instances with whom HNG struggles. A second advantage is that FLWC is much easier to implement: While implementing HNG’s adaptation of the Hungarian Algorithm to LSAP requires a thorough knowledge of matching theory, FLWC can be implemented by slightly transforming the cost matrix and then calling a solver for LSAP. Since LSAP is a very famous combinatorial optimization problem, libraries are available for all major programming languages.

For future work, we are planning to integrate the factorization of the assignment matrix, which is the core of our paper, in a modeling of the graph edit distance problem as a quadratic assignment problem.

References


