Improved local search for graph edit distance
Nicolas Boria, David Blumenthal, Sébastien Bougleux, Luc Brun

To cite this version:

HAL Id: hal-02388317
https://hal.archives-ouvertes.fr/hal-02388317
Submitted on 1 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Improved local search for graph edit distance

Nicolas Boria\textsuperscript{a}, David B. Blumenthal\textsuperscript{b}, Sébastien Bougleux\textsuperscript{a}, Luc Brun\textsuperscript{a}

\textsuperscript{a}Normandie Univ, ENSICAEN, UNICAEN, CNRS, GREYC, 14000 Caen, France
\textsuperscript{b}Free University of Bozen-Bolzano, Faculty of Computer Science, Piazza Dominici 3, 39100 Bolzano, Italy

ABSTRACT

The graph edit distance (GED) measures the dissimilarity between two graphs as the minimal cost of a sequence of elementary operations transforming one graph into another. This measure is fundamental in many areas such as structural pattern recognition or classification. However, exactly computing GED is \textit{NP}-hard. Among different classes of heuristic algorithms that were proposed to compute approximate solutions, local search based algorithms provide the tightest upper bounds for GED. In this paper, we present \textsc{K-REFINE} and \textsc{RANDPOST}. \textsc{K-REFINE} generalizes and improves an existing local search algorithm and performs particularly well on small graphs. \textsc{RANDPOST} is a general warm start framework that stochastically generates promising initial solutions to be used by any local search based GED algorithm. It is particularly efficient on large graphs. An extensive empirical evaluation demonstrates that both \textsc{K-REFINE} and \textsc{RANDPOST} perform excellently in practice.

1. Introduction

In many areas such as pattern recognition or graph classification, computing a graph (dis-)similarity measure is a central issue \cite{1, 2, 3, 4}. One popular approach is to embed the graphs into multi-dimensional vector spaces and then to compare their vector representations \cite{5, 6, 7}. This approach has the advantage that it allows fast computations. The main drawback is that embedding the graphs into vector spaces always comes at the price of information loss.

If local information that would be lost by the embeddings is crucial, it is recommendable to compare the graphs directly in the graph space. The graph edit distance (GED) is one of the most widely used measures for this purpose \cite{8, 9}. GED is defined as the minimum cost of an edit path transforming one graph into another, where an edit path is a sequence of node and edge insertions, deletions, and substitutions. Equivalently, it can be defined as the minimum cost of an edit path induced by a node map that assigns nodes of the source graph to nodes of the target graph \cite{10}.

As exactly computing GED is \textit{NP}-hard \cite{11}, research has mainly focused on heuristics \cite{12}. The development of heuristics was particularly triggered by the algorithms presented in \cite{13} and \cite{11}, which use transformations to the linear sum assignment problem with error correction (LSAPE) \cite{14} — a variant of the linear sum assignment problem (LSAP) where rows and columns may also be inserted and deleted — to compute upper bounds for GED. Further transformations from GED to LSAPE have been proposed in \cite{15, 16, 17, 18, 19, 20, 21, 22}.

LSAPE based heuristics are typically quite fast but yield loose upper bounds on some graphs. Tighter upper bounds for GED can be obtained by algorithms that use variants of local search. Given one or several initial node maps, local search based algorithms explore suitably defined neighborhoods to find improved node maps that induce cheaper edit paths. Several algorithms have been proposed that instantiate this paradigm: \textsc{DEFINE} \cite{11} varies the initial node maps via binary swaps; \textsc{IPFP} \cite{23, 24, 25} computes locally optimal node maps by using a variant of the classical Frank-Wolfe algorithm \cite{26, 27}; \textsc{BP-BEAM} \cite{28} uses beam search to improve the initial node maps; and \textsc{IBP-BEAM} \cite{29} further improves \textsc{BP-BEAM} by iteratively running it with different processing orders.

In this paper, we propose a new local search based algorithm \textsc{K-REFINE} and a warm start framework \textsc{RANDPOST}. \textsc{K-REFINE} generalizes and improves the existing algorithm \textsc{DEFINE} in three respects: Firstly, \textsc{K-REFINE} considers not only binary swaps, but rather swaps of size up to \(K\), where \(K \in \mathbb{N}_{\geq 2}\) is a meta-parameter. Secondly, \textsc{K-REFINE} computes the swap costs more efficiently than \textsc{DEFINE}, which leads to a significant gain in runtime performance. Thirdly, unlike \textsc{DEFINE}, \textsc{K-REFINE} allows the improved node map to contain fewer node substitutions than the original node map, which tightens the produced upper bound.

The framework \textsc{RANDPOST} extends the local search paradigm and hence improves all local search based GED algorithms. In a first step, \textsc{RANDPOST} runs a local search algorithm from a set of initial node maps. Subsequently, \textsc{RANDPOST} stochastically generates a new set of initial node maps from the converged solutions, and then re-runs the local search algorithm. This process iterates until a user-specified number of iterations has been

\textcopyright 2019. This version is made available under the CC-BY-NC-SA 4.0 license: \url{http://creativecommons.org/licenses/by-nc-sa/4.0/}
Journal version available at \url{https://doi.org/10.1016/j.patrec.2019.10.028}
Code available at \url{https://github.com/dbblumenthal/gedlib}
Work supported by Région Normandie (France) under project RIN AGAC
reached. Extensive experiments show that \textsc{K-REFINE} performs extremely well on small to medium sized graphs, while using \textsc{RANDPOST} is particularly effective on larger graphs.

The remainder of this paper is organized as follows: In Section 2, we fix concepts and notations and present the general local search framework. In Section 3, we summarize existing local search algorithms. In Section 4 and Section 5, we present \textsc{K-REFINE} and \textsc{RANDPOST}. In Section 6, \textsc{K-REFINE} and \textsc{RANDPOST} are evaluated empirically. Section 7 concludes the paper. The paper extends the results published in [30], where a preliminary version of \textsc{RANDPOST} was presented.

2. Preliminaries and notations

2.1. Undirected and labeled graphs

An undirected labeled graph \( G \) is a 4-tuple \( G = (V^G, E^G, \ell^G_V, \ell^G_E) \), where \( V^G \) and \( E^G \) are sets of nodes and edges, \( \Sigma_V \) and \( \Sigma_E \) are label alphabets, and \( \ell^G_V : V^G \rightarrow \Sigma_V, \ell^G_E : E^G \rightarrow \Sigma_E \) are labeling functions. The dummy symbol \( \epsilon \) denotes dummy nodes and edges as well as their labels. Throughout the paper, we denote the nodes of a graph \( G \) by \( V^G := \{ u_i \mid i \in [|V^G|] \} \) and the nodes of a graph \( H \) by \( V^H := \{ v_k \mid k \in [|V^H|] \} \), where the index set \([N]\) is defined as \([N] := \{ n \in \mathbb{N} \mid 1 \leq n \leq N \} \), for all \( N \in \mathbb{N} \).

We denote by assignment a pair \((u,v)\) in \((V^G \cup \epsilon) \times (V^H \cup \epsilon)\), and we define a node map \( \pi \) as a set of assignments such that each node in \( V^G \) and \( V^H \) appears in exactly one assignment of \( \pi \). The notation \((u,v) \in \pi\) is considered equivalent to both \( \pi(u) = v \) and \( \pi^{-1}(v) = u \), and \( \Pi(G,H) \) denotes the set of all node maps between \( G \) and \( H \). For edges \( e = (u,u') \in E^G \) and \( f = (v,v') \in E^H \), we introduce the short-hand notations \( \pi(e) := (\pi(u),\pi(u')) \) and \( \pi^{-1}(e) := (\pi^{-1}(v),\pi^{-1}(v')) \).

A node map \( \pi \in \Pi(G,H) \) specifies for all nodes and edges of \( G \) and \( H \) whether they are substituted, deleted, or inserted. Each of these operations has an induced cost. An assignment \((u,v)\) induces a node substitution with cost \( c_V((u,v),\pi) \) if \( u \neq v \) and \( \pi^{-1}(v) \neq u \), a node deletion with cost \( c_V(u,\pi) \) if \( u \neq v \) and \( \pi^{-1}(v) = u \), and a node insertion with cost \( c_V(v,\pi) \) if \( u = v \) and \( \pi^{-1}(v) \neq u \). Similarly, a pair of assignments \((u,v), (u',v')\) induces an edge substitution with cost \( c_E((u',v'),(v',\pi)) \) if \( u \neq u' \) and \( v \neq v' \), an edge deletion with cost \( c_E((u,u'),\pi) \) if \( u = u' \) and \( \pi^{-1}(v) \neq u \), and an edge insertion with cost \( c_E((v,v'),\pi) \) if \( u = u' \) and \( \pi^{-1}(v') \neq u \). The edit cost functions \( c_V \) and \( c_E \) are defined in terms of the labeling functions \( \ell^G_V, \ell^H_V, \ell^G_E, \ell^H_E \), i.e., nodes and edges with the same labels induce the same edit costs.

Any node map \( \pi \in \Pi(G,H) \) hence induces an edit path \( P_\pi \) between \( G \) and \( H \). The cost \( c(P_\pi) \) of \( P_\pi \) is given as follows:

\[
c(P_\pi) = \sum_{u \in V^G} c_V(u,\pi(u)) + \sum_{u \in V^G} c_V(u,\pi(v)) + \sum_{v \in V^H} c_V(v,\pi^{-1}(v)) + \sum_{e \in E^G} c_E(e,\pi(e)) + \sum_{e \in E^G} c_E(e,\pi^{-1}(e)) + \sum_{f \in E^H} c_E(f,\pi^{-1}(f)) + \sum_{f \in E^H} c_E(f,\pi(f))
\]

We can now formally define GED.

**Definition 1 (GED).** The graph edit distance (GED) between two graphs \( G \) and \( H \) is defined as \( \text{GED}(G,H) := \min_{\pi \in \Pi(G,H)} c(P_\pi) \).

2.2. Upper bounding GED via local search

By Definition 1, each node map \( \pi \in \Pi(G,H) \) induces an upper bound \( UB := c(P_\pi) \) for \( \text{GED}(G,H) \). Hence, a straightforward application of the local search paradigm to the problem of upper bounding GED works as follows: Given an initial node map \( \pi \in \Pi(G,H) \), run a local search algorithm to obtain an improved node map \( \pi' \) with \( c(P_{\pi'}) \leq c(P_\pi) \).

With this approach, the quality of the obtained node map \( \pi' \) clearly depends a lot on the initial node map \( \pi \). In order to reduce this dependency, it was suggested in [31] to generate \( \kappa \) different initial solutions, run the local search algorithm on each of them (possibly in parallel), and return the best among the \( \kappa \) computed local optima. In order to reduce the computing time when parallelization is available, it was suggested in [30] to run in parallel more local searches than the number of desired local optima and to stop the whole process once the number local searches that have converged has reached the number of desired local optima. In this context, the framework runs with two parameters: \( \kappa \) represents the number of initial solutions, and \( 0 < \rho \leq 1 \) is defined such that \( [\rho \cdot \kappa] \) represents the number of desired computed local optima.

3. Existing local search algorithms for GED

3.1. The algorithm \textsc{REFINE}

Given an initial node map \( \pi \in \Pi(G,H) \), the algorithm \textsc{REFINE} [11] proceeds as follows: Let \( ((u_s,v_s))_{s=1}^{n_{ini}} \) be an arbitrary ordering of the initial node map \( \pi \), and \( \Gamma_s := (V^G_s \cup V^H_s, A_s) \) be an auxiliary directed bipartite graph, where \( \Gamma_s := (V^G_s \cup V^H_s) \) for \( \pi \) and \( (V^G_s \cup V^H_s) \cup \{ (s,s') \mid (s,s') \in [|[[\pi]]| \setminus \{s \} \} \) in other words, \( \Gamma_s \) contains a forward arc for each assignment contained in \( \pi \), and backward arcs between nodes in \( V^G \) and \( V^H \) that are not assigned to each other by \( \pi \). A directed cycle \( C \subseteq \Gamma_s \) in \( \Gamma_s \) with \( |C| = 4 \) is called swap.

For each swap \( C = \{(u_s,v_s),(v_s, u_{s'}) \} \cup \{(u_{s'}, v_{s'}),(u_s, v_s)\} \), \textsc{REFINE} checks if the swapped node map \( \pi' := (\pi \setminus \{(u_s,v_s),(u_{s'},v_{s'})\}) \cup \{(u_s,v_{s'}),(u_{s'},v_s)\} \) induces a smaller upper bound than \( \pi \). If, at the end of the for-loop, a node map has been found that improves the upper bound, \( \pi \) is updated to the node map that yields the largest improvement and the process iterates. Otherwise, the current node map \( \pi \) is returned.

3.2. The algorithm \textsc{IPFP}

The algorithm \textsc{IPFP} [27] is a variant of the Frank-Wolfe algorithm [26] for cases where an integer solution is required. Its adaptation to GED suggested in [23, 24, 25] implicitly constructs a matrix \( D \in \mathbb{R}^{(|V^G|+1) \times (|V^H|+1) \times (|V^G|+1) \times (|V^H|+1)} \) such that \( \min_{x \in \Pi(G,H)} \text{vec}(X)^T D \text{vec}(X) = \text{GED}(G,H) \), where \( \Pi(G,H) \subseteq [0,1]^{(|V^G|+1) \times (|V^H|+1)} \) contains all node maps between \( G \) and \( H \) in matrix form and \( \text{vec}() \) is a vectorization operator.

Let the cost function \( c : [0,1]^{(|V^G|+1) \times (|V^H|+1)} \rightarrow \mathbb{R} \) be defined as \( c(X) := \text{vec}(X)^T D \text{vec}(X) \). Starting from an initial node map \( X_0 \in \Pi(G,H) \) with induced upper bound \( UB := c(X_0) \), the
algorithm initializes $X' := X_0$ and converges to a, possibly fractional, local minimum by repeating the following five steps:

1. Populate LSAPE instance $C_k := D \text{vec}(X_k)$.
2. Compute $B_{k+1} \in \arg \min_{B_{k+1}} \text{vec}(C_k)^T \text{vec}(B)$.
3. Set $X' := \arg \min_{X \in \{X, B_{k+1}\}} c(X)$.
4. Compute $\alpha_{k+1} := \min_{\alpha \in [0, 1]} c(X_k + \alpha \cdot (B_{k+1} - X_k))$.
5. Set $X_{k+1} := X_k + \alpha_{k+1}(B_{k+1} - X_k)$.

BP-BEAM iterates until $(c(X_k) - \text{vec}(C_k))^T \text{vec}(B_{k+1})) / c(X_k)$ is smaller than a threshold $\epsilon$ or a maximal number of iterations $I$ has been reached. Subsequently, the local optimum $X_{k+1}$ is projected to the closest integral solution $\tilde{X}$, and the best encountered node map $X' := \arg \min_{X \in \{X, \tilde{X}\}} c(X)$ is returned.

3.2.1. The algorithm BP-BEAM

Given an initial node map $\pi \in \Pi(G, H)$ and a constant $B \in \mathbb{N}_{\geq 1}$, the algorithm BP-BEAM [28] starts by producing a random ordering $((u_s, v_s))_{s = 1}^{\pi}$ of the initial node map $\pi$. BP-BEAM now constructs an improved node map $\pi'$ by partially traversing an implicitly constructed tree $T$ via beam search with beam size $B$. The nodes of $T$ are tuples $(\pi'', s)$, where $\pi'' \in \Pi(G, H)$ is an ordered node map and $s \in [|\pi|]$ is the depth of the tree node in $T$. Tree nodes $(\pi'', s)$ with $s = |\pi|$ are leaves, and the children of an inner node $(\pi'', s)$ are $\{(\text{SWAP}((\pi'', s, s'), s + 1) | s' \in [s, \ldots, |\pi|]\}$. Here, $\text{SWAP}((\pi'', s, s'), s)$ is the ordered node map obtained from $\pi''$ by swapping the assignments $(u_s, v_s)$ and $(u_{s'}, v_{s'})$.

At initialization, BP-BEAM sets the output node map $\pi'$ to the initial node map $\pi$. Furthermore, BP-BEAM maintains a priority queue $q$ of tree nodes which is initialized as $q := ((\pi, 1))$ and sorted w.r.t. non-decreasing induced edit cost of the contained node maps. As long as $q$ is non-empty, BP-BEAM extracts the top node $(\pi'', s)$ from $q$ and updates the output node map $\pi'$ if $c(P_{\pi'}) < c(P_{\pi''})$. If $s < |\pi|$, i.e., if the extracted tree node is no leaf, BP-BEAM adds all of its children to $q$ and subsequently discards all but the first $B$ tree nodes contained in $q$. Once $q$ is empty, the cheapest encountered node map $\pi'$ is returned.

3.3. The algorithm IBP-BEAM

Since the size of the priority queue $q$ is restricted to $B$, which parts of the search tree $T$ are visited by BP-BEAM crucially depends on the ordering of the initial node map $\pi$. Therefore, BP-BEAM can be improved by considering not one but several initial orderings. The algorithm IBP-BEAM suggested in [29] does exactly this. That is, given a constant number of iterations $I \in \mathbb{N}_{\geq 1}$, IBP-BEAM runs BP-BEAM with $I$ different randomly created orderings of the initial node map $\pi$, and then returns the cheapest node map $\pi'$ encountered in one of the iterations.

4. The local search algorithm K-REFINE

In this section, we extend and improve the algorithm REFINE [11] in three ways. Firstly, instead of considering only binary swaps, we make K-REFINE consider all $K'$-swaps for all $K' \in [K] \setminus \{1\}$, where $K \in \mathbb{N}_{\geq 2}$ is a constant (Section 4.1). Secondly, we show that for computing the induced cost $c(P_{\pi'})$ of a node map $\pi'$ obtained from $\pi$ via a $K'$-swap $C$, it suffices to consider the nodes and edges that are incident with $C$ (Section 4.2). This observation yields an improved implementation of K-REFINE, which is much more efficient than the naïve implementation suggested in [11]. Thirdly, we suggest to include the dummy assignment $(\epsilon, \epsilon)$ into the initial node map $\pi$ before enumerating the swaps (Section 4.3). This modification allows the number of node substitutions to decrease and hence improves the quality of the obtained upper bound.

4.1. Generalization to swaps of size larger than two

Algorithm 1 gives an overview of the algorithm K-REFINE, which generalizes REFINE to swaps of size larger than two. Given graphs $G$ and $H$, an initial node map $\pi \in \Pi(G, H)$, and a maximal swap size $K \in \mathbb{N}_{\geq 2}$, K-REFINE starts by initializing the current swap size $K'$, the best swap $C^*$, and the best swap cost $\Delta^* \geq 1$. Algorithm 1 first enumerates the set $C_{\pi,K'} := \{C \subseteq A_\pi | C \text{ is cycle of length } K' \text{ in } G_\pi\}$ of all $K'$-swaps of $\pi$ (line 4). The auxiliary directed bipartite graph $G_\pi := (V_\pi^G \cup V_\pi^H, A_\pi)$ is defined as in Section 3.1 above, i.e., we have $V_\pi^G := \{u_s | s \in [|\pi|]\}$, $V_\pi^H := \{v_s | s \in [|\pi|]\}$, and $A_\pi := \pi \cup \{(v_s, u_s') \in V_\pi^G \times V_\pi^H | (s, s') \in [|\pi|] \times [|\pi|], s \neq s'\}$, where $(u_s, v_s)'_{s = 1}^{|\pi|}$ is an arbitrary ordering of $\pi$.

For each $K'$-swap $C \in C_{\pi,K'}$, let $F(C) := C \cap \pi$ and $B(C) := \{(u, v) \in (V_\pi^G \cup \{\epsilon\}) \times (V_\pi^H \cup \{\epsilon\}) | (v, u) \in C \setminus \pi\}$ be the set of node assignments corresponding to forward and backwards arcs contained in $C$, respectively. K-REFINE computes the swap cost

$$\text{SWAP-COST}(\pi, C) := c(P_{\pi'}) - c(P_{\text{SWAP}(\pi, C)}).$$
where $SWAP(\pi, C) := (\pi \setminus F(C)) \cup \{(u, v) \in B(C) \mid (u, v) \neq (e, e)\}$ is the node map obtained from $\pi$ by carrying out the swap encoded by $C$ (line 5). If $C$ yields an improvement, $K$-REFINE updates the best swap cost $C^*$ and the best swap cost $\Delta^*$ (lines 6 to 7).

Once all $K$'s swaps $C \in C_{\pi, K}$ have been visited, $K$-REFINE checks whether one of them yields an improvement w.r.t. the current node map $\pi$ (line 8). If this is the case, $K$-REFINE updates $\pi$ (lines 9 to 11) and resets the current swap size to $K' := 2$ (line 12). Otherwise, $K'$ is incremented (line 14). Subsequently, $K$-REFINE resets the best swap and the best swap cost to $C^* := \emptyset$ and $\Delta^* := 0$, respectively (line 15). Upon termination of the main loop, $K$-REFINE returns the current node map $\pi$ (line 16).

Assume that $SWAP-COST(\pi, C)$ can be computed in $O(\alpha)$ time (cf. Section 4.2 for details). Furthermore, let $f \in \mathbb{N}$ be the number of times $K$-REFINE finds an improved node map in line 8. Note that, if the edit costs are integral, it holds that $f \leq |C(\pi)|$. Thus, $K$-REFINE’s overall runtime complexity is $O(I(|V_G| + |V_H|)^K)$.

**Proposition 1.** For each node map $\pi \in \Pi(G, H)$ and each $K' \in \mathbb{N}_{\geq 2}$, it holds that $|C_{\pi, K'}| = \binom{|\mathcal{K}|}{K'}(K' - 1)!$.

**Proof.** The proposition immediately follows from the definition of $C_{\pi, K'}$. Details are omitted due to space constraints.

4.2. Efficient computation of swap costs

Given a node map $\pi \in \Pi(G, H)$ and a $K$'-swap $C \in C_{\pi, K'}$, let $\pi' := SWAP(\pi, C)$ be the node map obtained from $\pi$ by swapping the forward and backward arcs contained in $C$. Assume that $c(\pi_C)$ has already been computed. By line 16, the swap cost $SWAP-COST(\pi, C)$ can be computed naïvely by computing the induced costs $c(\pi_C)$ of the swapped node map and then considering the difference between $c(\pi_C)$ and $c(\pi_C')$. By definition of $c(\pi_C)$, this requires $O(|E_G|, |E_H|)$ time. Since $SWAP-COST(\pi, C)$ has to be computed in every iteration of $K$-REFINE’s inner loop, it is highly desirable to implement $SWAP-COST(\pi, C)$ more efficiently. The following Proposition 2 provides the key ingredient of a more efficient implementation.

**Proposition 2.** Let $\pi \in \Pi(G, H)$ be a node map, $K' \in \mathbb{N}_{\geq 2}$ be a constant, and $C \in C_{\pi, K'}$ be a $K$'-swap. Then the following equation holds:

$$
\Delta = \sum_{e \in E_G^C} c_G(e, \pi'(e)) + \sum_{e \in E_H^C} c_H(e, \pi(e)) + \sum_{f \in E_H^C} c_H(e, f) + \sum_{e \in E_G^C} c_G(e, \pi'(e)) - \sum_{e \in E_G^C} c_G(e, \pi(e)) - \sum_{f \in E_H^C, \pi'(f)} c_H(e, f).
$$

**Proof.** By construction of $V_G^C$ and $V_H^C$, we have $\pi'(u) = \pi'(u)$, for all $u \in V_G \setminus V_H^C$, and $\pi'(v) = \pi'(v)$, for all $v \in V_H \setminus V_H^C$. Similarly, $\pi(e) = \pi'(e)$ and $\pi'(f) = \pi'(f)$ holds for all $e \in E_G \setminus E_G^C$ and all $f \in E_H \setminus E_H^C$. This proves the proposition.

**Proposition 2** implies that, for computing $SWAP-COST(\pi, C)$, only the nodes and edges contained in $V_G^C, V_H^C, E_G^C$, and $E_H^C$ must be considered. Since $K'$ is constant, $|V_G^C|, |V_H^C| \leq K', |E_G^C| \leq K' \max \deg(G)$, and $|E_H^C| \leq K' \max \deg(H)$, $SWAP-COST(\pi, C)$ can hence be computed in $O(\max(|E_G|, |E_H|))$ time. This is a significant improvement w.r.t. the naïvely computing $SWAP-COST(\pi, C)$ in $O(\max(|E_G|, |E_H|))$ time.

4.3. Improvement of upper bound via dummy assignment

For each node map $\pi \in \Pi(G, H)$, let $S(\pi) := ||(u, v) \in \pi \mid u \neq e \land v \neq e||$ denote the number of node substitutions contained in $\pi$. Now assume that $K$-REFINE as specified in Algorithm 1 is run from an initial node map $\pi \in \Pi(G, H)$ that does not contain the dummy assignment $(e, e)$. Since $\pi$ and $\pi \cup \{(e, e)\}$ induce the same edit path, the application of our algorithm is likely to hold in most implementations of $K$-REFINE. The following Proposition 3 shows that, under this assumption, the search space of $K$-REFINE is restricted in the sense that it includes only node maps $\pi' \in \Pi(G, H)$ with $S(\pi') \geq S(\pi)$. This has a negative effect on the quality of the upper bound produced by $K$-REFINE, as some potentially promising node maps are excluded a priori.

**Proposition 3.** Let $\pi \in \Pi(G, H)$ be a node map that satisfies $(e, e) \not\in \pi$ and $\pi \in \Pi(G, H)$ be the improved node map obtained from $\pi$ by running $K$-REFINE as specified in Algorithm 1. Then it holds that $S(\pi') \geq S(\pi)$.

**Proof.** Let $\pi \in \Pi(G, H)$ be a node map that satisfies $(e, e) \not\in \pi$, $K' \in \mathbb{N}_{\geq 2}$ be a constant, and $C \in C_{\pi, K'}$ be a $K$'-swap. By definition of $B(C)$, we have $(e, e) \not\in SWAP(\pi, C)$. Therefore, the proposition follows by induction on the number of times $K$-REFINE finds an improved node map in line 8, if we can show that $S(SWAP(\pi, C)) \geq S(\pi)$. To show this inequality, we define $S'_G(\pi, C) := \||(u, v) \in F(C) \mid u = e \land v = e||$ and $S'_H(\pi, C) := \||(u, v) \in B(C) \mid u = e \land v = e||$. It is easy to see that we have $S(SWAP(\pi, C)) = S(\pi) + S'_G(\pi, C) - S'_f(\pi, C)$. Since $(e, e) \not\in \pi$, we additionally know that $S'_f(\pi, C) = 0$. We hence obtain $S(SWAP(\pi, C)) = S(\pi) + S'_G(\pi, C) \geq S(\pi)$, as required.

The proof of Proposition 3 tells us how we have to modify $K$-REFINE in order to ensure that node maps with fewer node substitutions than the initial node map are contained in its search space: We simply have to update the current node map $\pi$ as $\pi := \pi \cup \{(e, e)\}$ before enumerating all $K$'-swaps $C \in C_{\pi, K'}$ in line 4 of Algorithm 1. This modification is particularly important if the edit costs are non-metric, i.e., if it can happen that deleting plus inserting is cheaper than substituting.
Algorithm 2: The framework RANDPOST.

Input: Graphs $G$ and $H$, constants $\kappa \in \mathbb{N}_{\geq 1}$, $L \in \mathbb{N}$, $\rho \in (0, 1]$, and $\eta \in [0, 1]$, local search algorithm $\text{ALG}$, initial node map set $S_0 \subseteq \Pi(G, H)$ with $|S_0| = \kappa$, lower bound $LB$ for GED($G, H$).

Output: Upper bound $UB$ for GED($G, H$).

1. $S'_0 := \text{ALG}(S_0, \rho)$; // run local search on initial node maps
2. $UB := \min_{\pi \in S_0} c(P_{\pi})$; // set first upper bound
3. $M := 0_{(|V^G|+1) \times (|V^H|+1)}$; // initialize scores matrix
4. for $r \in [L]$ do
   5. $M := \text{UPD-SCORES}(M, S'_{r-1}, \eta, LB, UB)$; // update scores
   6. $S_r := \text{GEN-NODE-MAPS}(M, \kappa)$; // generate node maps
   7. $S'_r := \text{ALG}(S_r, \rho)$; // run local search on new node maps
   8. $UB := \min \{UB, \min_{\pi \in S'_r} c(P_{\pi})\}$; // update upper bound
5. return $UB$; // return upper bound

5. The framework RANDPOST

In this section, we present RANDPOST, a framework that can be used to improve any local search based algorithm for upper bounding GED. Intuitively, RANDPOST iteratively runs a given local search algorithm. In each iteration, previously computed locally optimal node maps are combined stochastically to obtain new promising initial node maps to be used in the next iteration.

Algorithm 2 provides an overview of the framework. Given a set of initial node maps $S_0 \subseteq \Pi(G, H)$ with $|S_0| = \kappa$, a constant $\rho \in (0, 1]$, and a local search algorithm ALG, RANDPOST computes a set $S'_0 \subseteq \Pi(G, H)$ of improved node maps with $|S'_0| = \lfloor \rho \cdot \kappa \rfloor$ by (parallelly) running ALG on all initial node maps and terminating once $\lfloor \rho \cdot \kappa \rfloor$ runs have converged (line 1). Subsequently, the upper bound $UB$ is initialized as the cost of the cheapest induced edit path encountered so far (line 2). Note that, up to this point, RANDPOST is equivalent to the local search framework with multi-start described in Section 2.2 above.

RANDPOST now initializes a matrix $M \in \mathbb{R}^{(|V^G|+1) \times (|V^H|+1)} := 0_{(|V^G|+1) \times (|V^H|+1)}$ that contains scores $m_{i,k}$ for each possible node assignment $(u_i, v_k) \in (V^G \cup \{e\}) \times (V^H \cup \{e\})$ (line 3). The score for each substitution $(u_i, v_k) \in V^G \times V^H$ is represented by the value $m_{i,k}$, while the scores for the deletion $(u_i, e)$ and the insertion $(e, v_k)$ are represented by the values $m_{i,|V^H|+1}$ and $m_{|V^G|+1,k}$, respectively. Throughout the algorithm, M is maintained in such a way that $m_{i,k}$ is large just in case the corresponding node assignment appears in many cheap locally optimal node maps.

After initializing $M$, RANDPOST carries out $L$ iterations of its main for-loop, where $L \in \mathbb{N}$ is a meta-parameter (lines 4 to 8). Inside the $i$th iteration, RANDPOST starts by updating the scores matrix $M$ by calling UPD-SCORES$(M, S'_{r-1}, \eta, LB, UB)$, where $M$ is the current scores matrix, $S'_{r-1} \subseteq \Pi(G, H)$ is the set of improved node maps obtained from the previous iteration, $\eta \in [0, 1]$ is a meta-parameter used to give greater weight to cheap node maps, $LB$ is a previously computed lower bound for GED($G, H$), and $UB$ is the current upper bound (line 5). Let the matrix $M' \in \mathbb{R}^{(|V^G|+1) \times (|V^H|+1)}$ be defined as $M' := \text{UPD-SCORES}(M, S'_{r-1}, \eta, LB, UB)$. Then $M'$ is given as

$$M' := M + \sum_{\pi \in S'_{r-1}} (1 - \eta) \frac{UB - LB}{c(P_{\pi}) - LB} X' \cdot$$

where $X' \in [0, 1]^{(|V^G|+1) \times (|V^H|+1)}$ is the matrix representation of the improved node map $\pi' \in S'_{r-1}$, i.e., for all $u_i \in V^G$ and all $v_k \in V^H$, we have $X'_{i,k} = 1$ just in case $(u_i, v_k) \in \pi'$, $X'_{|V^G|+1} = 1$ just in case $(u_i, e) \in \pi'$, and $X'_{|V^G|+1,k} = 1$ just in case $(e, v_k) \in \pi'$. If $\eta = 0$, $m_{i,k}$ represents the number of converged local optima that contain the corresponding assignment. If $\eta > 0$, assignments that appear in node maps with lower costs receive higher scores.

Once $M$ has been updated, RANDPOST creates a new $\kappa$-sized set $S_r \subseteq \Pi(G, H)$ of initial node maps by calling GEN-NODE-MAPS$(M, \kappa)$ (line 6). GEN-NODE-MAPS$(M, \kappa)$ works as follows: For each of the first $|V^G|$ rows $M_i$ of $M$, RANDPOST draws a column $k \in [|V^H| + 1]$ from the distribution encoded by $M_i$. If $k = |V^H| + 1$, the node deletion $(u_i, e)$ is added to the node map $\pi$ that is being constructed. Otherwise, the substitution $(u_i, v_k) \in \pi$ is added to $\pi$, the score $m_{i,k}$ is temporarily set to 0 for all $j \in [|V^G|] \setminus \{i\}$, and the column $k$ is marked as covered. Once all nodes of $G$ have been processed, node insertions $(e, v_k)$ are added to $\pi$ for all uncovered columns $k \in [|V^H|]$. This process is repeated until $\kappa$ different node maps have been created.

After creating the set $S_r$ of new initial node maps, RANDPOST constructs a new $[\rho \cdot \kappa]$-sized set $S'_r \subseteq \Pi(G, H)$ of improved node maps by (parallelly) running the local search algorithm ALG on the initial node maps contained in $S_r$ and terminating once $[\rho \cdot \kappa]$ runs have converged (line 7). Subsequently, the upper bound is updated as the minimum of the current upper bound and the cost of the cheapest edit path induced by one of the newly computed improved node maps (line 8). Finally, RANDPOST returns the best encountered upper bound (line 9).

6. Empirical evaluation

In order to empirically evaluate K-REFINE and RANDPOST, extensive tests were conducted on four standard datasets from the IAM Database Repository [32, 33]: MUTA, PROTEIN, GREC, and FP (cf. Table 1). For all datasets, we tested on the metric edit costs suggested in [33]. For MUTA, we additionally defined non-metric edit costs by setting the costs of node and edge deletions and insertions to 1, and setting the costs of node and edge substitutions to 3 (the resulting dataset is denoted as MUTA-N). For each dataset, subsets of 50 graphs were chosen randomly, and upper bounds for GED were computed for each pair of graphs in the subsets, as well as for each graph and a shuffled copy of itself. In the following, $d$, $\bar{d}$, and $t$ denote the average upper bound, the average upper bound between graphs and their shuffled copies, and the average runtime in seconds, respectively.
Note that the test metric \( \hat{d} \) gives us a hint to how close to optimality each algorithm is, as the optimal value, namely 0, is known. All methods were implemented using the GEDLIB library and were run in 20 parallel threads.\(^1\)

We tested two versions 2-REFINE and 3-REFINE of our local search algorithm K-REFINE, which use swaps of maximum size two and three, respectively. We compared them to the existing local search algorithms REFINE, IPFP, BP-BEAM, and IBP-BEAM. As suggested in [28] and [29], we set the beam size employed by BP-BEAM and IBP-BEAM to 5 and the number of iterations employed by IBP-BEAM to 20. IPFP was run with convergence threshold set to \( 10^{-3} \) and maximum number of iterations set to 100, as proposed in [25]. In order to evaluate RANDPOST, we ran each algorithm with \( \kappa = 40 \) initial solutions. Indeed, experiments reported in [31] show that on all tested datasets, using more initial solutions does not bring a significant decrease of the estimated GED. We also varied the pair of meta-parameters \((L, \rho)\) on the set \{0, 1\}, \{(0, 0.5), (3, 0.25), (7, 0.125)\}. Recall that \( L \) is the number of RANDPOST loops and \( \rho \) is defined such that each iteration produces exactly \([\rho \cdot \kappa]\) locally optimal node maps. Therefore, our setup ensures that each configuration produces exactly 40 local optima. For each algorithm and each dataset, we conducted pre-tests where we varied the penalty parameter \( \eta \) on the set \{\( n/10 \mid n \in \mathbb{N}_{\leq 10} \}\}, and then picked the value of \( \eta \) for the main experiments that yielded the best average upper bound across all RANDPOST configurations.

6.1. K-REFINE vs. REFINE

In a first series of experiments, we compared the versions 2-REFINE and 3-REFINE of our improved and generalized local search algorithm K-REFINE to the baseline algorithm REFINE. All algorithms were run without RANDPOST, i.e., with \((L, \rho) = (0, 1)\), and the tests were carried out on a computer using an Intel Xeon E5-2620 v4 2.10GHz CPU. Table 2 shows the results. By comparing \( t(2\text{-REFINE}) \) and \( t(\text{REFINE}) \), we see that efficiently computing the swap costs as suggested in Section 4.2 indeed significantly improves the runtime performance. Unsurprisingly, the speed-up is especially large on the datasets PROTEIN and MUTA containing the larger graphs. Comparing \( d(2\text{-REFINE}) \) and \( d(\text{REFINE}) \) shows that the inclusion of the dummy assignment proposed in Section 4.3 slightly improves the quality of the produced upper bound. As expected, the per- centual improvement is largest on the dataset MUTA-N with non-metric edit costs. Finally, we observe that running K-REFINE with swaps of size three slightly improves the upper bounds on all datasets, but significantly increases the runtime of the algorithm.

6.2. Behaviour of RANDPOST framework

In a second series of experiments, we evaluated the behaviour of RANDPOST by running each algorithm with four different pairs of meta-parameters \((L, \rho) \in \{(0, 1), (1, 0.5), (3, 0.25), (7, 0.125)\}\). We remind that the case \((L, \rho) = (0, 1)\) amounts to a basic multi-start framework with no RANDPOST loop. An additional subset of 10 graphs having exactly 70 nodes was extracted from MUTA and is denoted by MUTA-L. The tests were run on a computer using an Intel(R) Xeon E5-2640 v4 2.4GHz CPU. Figure 1 visualizes the results. Since 2-REFINE was always faster and more accurate than the baseline REFINE (cf. Section 6.1), we do not show plots for REFINE. Table 3 provides detailed numerical

---

\(^1\)Sources and datasets: https://github.com/dbblumenthal/gedlib/.
data for the MUTA-L subset, which turned out to be the subset with the highest variability in distances and computing times.

Figure 1 indicates that, on the datasets FP, GREC, and PROTEIN containing small graphs, near-optimality is reached by most algorithms when run with \( L \geq 1 \) number of RANDPOST loops. In these contexts, our algorithm 2-REFINE with RANDPOST configuration \((L, \rho) = (1, 0.5)\) provides the best tradeoff between runtime and accuracy, as it reaches the same accuracy as best algorithms, and, in terms of runtime, outperforms all algorithms except for BP-BEAM by approximately one order of magnitude. The only faster algorithm BP-BEAM computes much more expensive node maps, even in the RANDPOST settings with higher number of loops. We also note that our algorithms 2-REFINE and 3-REFINE are already among the best local search algorithms when run in a simple multi-start setting without RANDPOST (i.e., when \( L = 0 \)), both in terms of distance and computing time.

On the datasets MUTA and MUTA-L containing larger graphs, the behavior of the RANDPOST framework appears clearly and independently of the local search algorithm it is applied to. In all cases, a higher number of RANDPOST loops — and lower number of computed solutions per loop — leads to a higher computation time (the computation time is approximately doubled whenever the number of loops is doubled), and to a lower average distance. In other words, the framework RANDPOST provides a very useful algorithmic tool in situations where some time can be dedicated to compute tight upper bounds on big graphs.

### 7. Conclusions

In this paper, we proposed K-REFINE, an improved and generalized version of the local search based GED algorithm REFINE, and suggested the general framework RANDPOST, which stochastically generates promising initial solutions to tighten the upper bounds produced by all local search algorithms. Both K-REFINE and RANDPOST perform excellently in practice: On small graphs, K-REFINE is among the algorithms computing the tightest upper bounds and, in terms of runtime, clearly outperforms all existing algorithms that yield similar accuracy. On larger graphs, K-REFINE provides a very good tradeoff between runtime and accuracy, as it is only slightly less accurate but much faster than the most accurate algorithms. The framework RANDPOST is particularly effective on larger graphs, where it significantly improves the upper bounds of all local search algorithms.

## References


### Table 3. Detailed experimental results on MUTA-L.

<table>
<thead>
<tr>
<th>((L, \rho))</th>
<th>2-REFINE</th>
<th>3-REFINE</th>
<th>IPFP</th>
<th>BP-BEAM</th>
<th>IBP-BEAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>((0, 1))</td>
<td>104.57 5.42 2.85 (10^{-1})</td>
<td>101.82 5.16 2.18</td>
<td>93.81 2.94 3.34</td>
<td>136.21 13.25 2.95 (10^{-2})</td>
<td>116.92 9.35 8.46 (10^{-1})</td>
</tr>
<tr>
<td>((1, 0.5))</td>
<td>98.61 3.81 6.96 (10^{-1})</td>
<td>94.34 3.32 4.51</td>
<td>85.05 1.13 6.21</td>
<td>113.18 7.41 5.81 (10^{-2})</td>
<td>106.39 6.35 1.71</td>
</tr>
<tr>
<td>((3, 0.25))</td>
<td>94.69 3.13 7.61 (10^{-1})</td>
<td>90.64 2.28 6.74</td>
<td>82.34 0.29 1.01 (10^{1})</td>
<td>111.12 6.57 1.11 (10^{-1})</td>
<td>104.64 5.65 2.61</td>
</tr>
<tr>
<td>((7, 0.125))</td>
<td>92.63 2.45 1.22</td>
<td>88.49 1.97 1.16 (10^{1})</td>
<td>81.36 0.29 1.81 (10^{1})</td>
<td>110.32 6.59 2.35 (10^{-1})</td>
<td>102.84 5.16 5.19</td>
</tr>
</tbody>
</table>


