Approximating Edit Distance Within Constant Factor in Truly Sub-Quadratic Time

Diptarka Chakraborty
Computer Science Institute of Charles University
Malostranské náměstí 25, 118 00
Praha 1, Czech Republic
Email: diptarka@iuuk.mff.cuni.cz

Debarati Das
Computer Science Institute of Charles University
Malostranské náměstí 25, 118 00
Praha 1, Czech Republic
Email: debarati710@gmail.com

Elazar Goldenberg
The Academic College Of Tel Aviv-Yaffo
School of Computer Science
Tel Aviv-Yaffo, Israel
Email: elazargo@mta.ac.il

Michal Koucký
Computer Science Institute of Charles University
Malostranské náměstí 25, 118 00
Praha 1, Czech Republic
Email: koucky@iuuk.mff.cuni.cz

Michael Saks
Department of Mathematics
Rutgers University
Piscataway, NJ, USA
Email: msaks30@gmail.com

Abstract—Edit distance is a measure of similarity of two strings based on the minimum number of character insertions, deletions, and substitutions required to transform one string into the other. The edit distance can be computed exactly using a dynamic programming algorithm that runs in quadratic time. Andoni, Krauthgamer and Onak (2010) gave a nearly linear time algorithm that approximates edit distance within approximation factor poly(log n).

In this paper, we provide an algorithm with running time \( \tilde{O}(n^{2-2/7}) \) that approximates the edit distance within a constant factor.

Keywords.Edit distance; Approximation algorithm; Sub-quadratic time algorithm; Randomized algorithm;

I. INTRODUCTION

Exact computation of edit distance. The edit distance (aka Levenshtein distance) [1] between strings \( x, y \), denoted by \( d_{edit}(x, y) \), is the minimum number of character insertions, deletions, and substitutions needed to convert \( x \) into \( y \). It is a widely used distance measure between strings that finds applications in fields such as computational biology, pattern recognition, text processing, and information retrieval. The problems of efficiently computing \( d_{edit}(x, y) \), and of constructing an optimal alignment (sequence of operations that converts \( x \) to \( y \)), are of significant interest.

Edit distance can be evaluated exactly in quadratic time via dynamic programming (Wagner and Fischer [2]). Landau et al. [3] gave an algorithm that finds an optimal alignment in time \( O(n + d_{edit}(x, y)^2) \), improving on a previous \( O(n^2 + d_{edit}(x, y)) \) algorithm of Ukkonen [4]. Masek and Paterson [5] obtained the first (slightly) sub-quadratic \( O(n^2 / \log n) \) time algorithm, and the current asymptotically fastest algorithm (Grabowski [6]) runs in time \( O(n^2 \log \log n / \log \log n) \).

The research leading to these results has received funding from the European Research Council (FP7/2007-2013)/ERC Grant Agreement no. 616787.

Supported in part by Simons Foundation under award 332622

Backurs and Indyk [7] showed that a truly sub-quadratic algorithm \( O(n^{2-\delta}) \) for some \( \delta > 0 \) would imply a \( 2^{(1-\gamma)n} \) time algorithm for CNF-satisfiability, contradicting the Strong Exponential Time Hypothesis (SETH). Abboud et al. [8] showed that even shaving an arbitrarily large polylog factor from \( n^2 \) would have the plausible, but apparently hard-to-prove, consequence that NEXP does not have non-uniform NC\(^1\) circuits. For further “barrier” results, see [9], [10].

Approximation algorithms. There is a long line of work on approximating edit distance. The exact \( O(n + k^2) \) time algorithm (where \( k \) is the edit distance of the input) of Landau et al. [3] yields a linear time \( \sqrt{n} \)-factor approximation. This approximation factor was improved, first to \( n^{3/7} \) [11], then to \( n^{1/3+o(1)} \) [12] and later to \( 2^{O(\sqrt{\log n})} \) [13], all with slightly superlinear runtime. Batu et al. [14] provided an \( O(n^{1-\alpha}) \)-approximation algorithm with runtime \( O(n^{\max\{\delta, 20\alpha - 1\}}) \). The strongest result of this type is the \( (\log n)^{O(1/\epsilon)} \) factor approximation (for every \( \epsilon > 0 \)) with running time \( n^{1+\epsilon} \) of Andoni et al. [15]. Abboud and Backurs [16] showed that a truly sub-quadratic deterministic time \( 1 + o(1) \)-factor approximation algorithm for edit distance would imply new circuit lower bounds.

Independent of our work, Boroujeni et al. [17] obtained a truly sub-quadratic quantum algorithm that provides a constant factor approximation. Their latest results [18] are a \( (3 + \epsilon) \) factor with runtime \( O(n^{2-\frac{3}{21}}/\epsilon^O(1)) \) and a faster \( O(n^{1.708}) \)-time with a larger constant factor approximation.

Andoni and Nguyen [19] found a randomized algorithm that approximates Ulam distance of two permutations of \( \{1, \ldots, n\} \) (edit distance with only insertions and deletions) within a (large) constant factor in time \( O(\sqrt{n} + n/k) \), where \( k \) is the Ulam distance of the input; this was improved by Naumovitz et al. [20] to a \( (1+\epsilon) \)-factor approximation (for any \( \epsilon > 0 \)) with similar runtime.

Our results. We present the first truly sub-quadratic time
Theorem 1. There is a randomized algorithm ED-UB that on input strings $x, y$ of length $n$ over any alphabet $\Sigma$ outputs an upper bound on $d_{ed}(x,y)$ in time $O(n^{12/7})$ that, with probability at least $1-n^{-5}$, is at most a fixed constant multiple of $d_{ed}(x,y)$.

If the output is $U$, then the algorithm has implicitly found an alignment of cost at most $U$. The algorithm can be modified to explicitly output such an alignment.

The approximation factor proved in this preliminary version is 1680, can be greatly improved by tweaking parameters. We believe, but have not proved, that with sufficient care the algorithm can be modified (with no significant increase in runtime) to get $(3 + \epsilon)$ approximation.

Theorem 1 follows from:

Theorem 2. For every $\theta \in [n^{-1/5}, 1]$, there is a randomized algorithm GAP-UB$\theta$ that on input strings $x, y$ of length $n$ outputs $u = GAP-UB_\theta(x,y)$ such that: (1) $d_{ed}(x,y) \leq u$ and (2) on any input with $d_{ed}(x,y) \leq \theta n$, $u \leq 840\theta n$ with probability at least $1-n^{-5}$. The runtime of GAP-UB$\theta$ is $O(n^{12/7})$.

The name GAP-UB$\theta$ reflects this that this is a "gap algorithm", which distinguishes inputs with $d_{ed}(x,y) \leq \theta n$ (where the output is at most $840\theta n$), and those with $d_{ed}(x,y) > 840\theta n$ (where the output is greater than $840\theta n$).

Theorem 1 follows via a routine construction of ED-UB from GAP-UB$\theta$, presented in Section V. The rest of the paper is devoted to proving Theorem 2.

The framework of the algorithm. We use a standard two-dimensional representation of edit distance. Visualize $x$ as lying on a horizontal axis and $y$ as lying on a vertical axis, with horizontal coordinate $i \in \{1, \ldots, n\}$ corresponding to $x_i$ and vertical component $j$ corresponding to $y_j$. The width $\mu(I)$ of interval $I \subseteq \{0, 1, \ldots, n\}$ is $\max(I) - \min(I) + 1$. Also, $x_I$ denotes the substring of $x$ indexed by $I - \{\min(I)\}$. (Note: $x_{\min(I)}$ is not part of $x_I$, e.g., $x_I = x_{I[0,\ldots,n]}$.

This convention is motivated by Proposition 3.)

We refer to $I$ as an $x$-interval to indicate that it indexes a substring of $x$, and $J$ as a $y$-interval to indicate that it indexes a substring of $y$. A box is a set $I \times J$ where $I$ is a $x$-interval and $J$ is a $y$-interval. $I \times J$ corresponds to the substring pair $(x_I, y_J)$. $I \times J$ is a $w$-box if $\mu(I) = \mu(J) = w$. We often abbreviate $d_{ed}(x_I, y_J)$ by $d_{ed}(I, J)$.

A decomposition of an $x$-interval $I$ is a sequence $I_1, \ldots, I_{\ell}$ of subintervals with $\min(I_j) = \min(I)$, $\max(I_j) = \max(I)$ for $j \in [\ell - 1]$, $\max(I_{\ell}) = \min(I_{\ell+1})$. Associated to $x, y$ is a directed graph $G_{x,y}$ with edge costs called a grid graph with vertex set $\{0, \ldots, n\} \times \{0, \ldots, n\}$ and all edges of the form $(i - 1, j) \to (i, j)$ (H-steps), $(i, j - 1) \to (i, j)$ (V-steps) and $(i - 1, j - 1) \to (i, j)$ (D-steps). Every H-step or V-step costs 1, and D-steps cost 1 if $x_i \neq y_j$ and 0 otherwise. There is a 1-1 correspondence that maps a path from $(0, 0)$ to $(n, n)$ to an alignment from $x$ to $y$, i.e., a set of character deletions, insertions and substitutions that changes $x$ to $y$, where an H-step $(i-1, j) \to (i, j)$ means "delete $x_i$", a V-step $(i, j - 1) \to (i, j)$ means "insert $y_j$ between $x_i$ and $x_{i+1}$" and a D-step $(i-1, j-1) \to (i, j)$ means replace $x_i$ by $y_j$, unless they are already equal. We have:

Proposition 3. The cost of an alignment, cost$(\tau)$, is the sum of edge costs of its associated path $\tau$, and $d_{ed}(x,y)$ is equal to cost$(G_{x,y})$, the min cost of an alignment path from $(0,0)$ to $(n,n)$.

For $I, J \subseteq \{0, \ldots, n\}$, $G_{x,y}(I \times J) \cong G_{x_I,y_J}$ is the grid graph induced on $I \times J$, and $d_{ed}(I, J) = \text{cost}(G_{x,y}(I \times J))$.

The natural high-level idea that GAP-UB$\theta$ appears (explicitly or implicitly) in previous work. The algorithm has two phases. First, the covering phase identifies a set $R$ of certified boxes which are pairs $(I \times J, \kappa)$, where $\kappa$ is an upper bound on the normalized edit distance $\Delta_{ed}(x_I, y_J) = \frac{d_{ed}(x_I, y_J)}{\mu(I)}$. The runtime of $\Delta_{ed}(I, J)$ is more convenient than $d_{ed}(I, J)$ for the covering phase.) Second, the min-cost path phase, takes input $R$ and uses a straightforward customized variant of dynamic programming to find an upper bound $U(R)$ on $d_{ed}(x,y)$ in time quasilinear in $|R|$. The central issue is to ensure that the covering phase outputs $R$ that is sufficiently informative so that $U(R) \leq c \cdot d_{ed}(x,y)$ for constant $c$, while running in sub-quadratic time.

Simplifying assumptions. The input strings $x, y$ have equal length $n$. (It is easy to reduce to this case: pad the shorter string to the length of the longer using a new symbol. The edit distance of the new pair is between the original edit distance and twice the original edit distance. This factor 2 increase in approximation factor can be avoided by generalizing our algorithm to the case $|x| \neq |y|$, but we won’t do this here.) We assume $n$ is a power of 2 (by padding both strings with a new symbol, which leaves edit distance unchanged). We assume that $\theta$ is a (negative) integral power of 2. The algorithm involves integer parameters $w_1, w_2, d$, all of which are chosen to be powers of 2.

Organization of the paper. Section II is a detailed overview of the covering phase algorithm and its analysis. Section III presents the pseudo-code and analysis for the covering phase. Section IV presents the min-cost path phase algorithm. Section V summarizes the full algorithm and discusses improvements in runtime via recursion.

II. COVERING ALGORITHM: DETAILED OVERVIEW

We give a detailed overview of the covering phase and its time analysis and proof of correctness, ignoring minor technical details. The pseudo-code in Section III corresponds to the overview, with technical differences mainly to improve runtime. We will illustrate the sub-quadratic time analysis
with the sample input parameter \( \theta = n^{-1/50} \) and algorithm parameters \( w_1 = n^{1/10} \), \( w_2 = n^{3/10} \) and \( d = n^{1/5} \).

The covering phase outputs a set \( \mathcal{R} \) of certified boxes. The goal is that \( \mathcal{R} \) includes an adequate approximating sequence for some min-cost path \( \tau \) in \( G_{x,y} \), which is a sequence \( \sigma \) of certified boxes \( (I_1 \times J_1, \kappa_1), \ldots, (I_\ell \times J_\ell, \kappa_\ell) \) that satisfies:

1. \( I_1, \ldots, I_\ell \) is a decomposition of \( \{0, \ldots, n\} \).
2. \( I_i \times J_i \) is an adequate cover of \( \tau_i \), where \( \tau_i = \tau_1 \) denotes the minimal subpath of \( \tau \) whose projection to the \( x \)-axis is \( I_i \), and adequate cover means that the \( \ell \)-interval (vertical) distance from the start vertex (resp. final vertex) of \( \tau_i \) and the lower left (resp. upper right) corner of \( I_i \times J_i \) is at most a constant multiple of \( \text{cost}(\tau_i) + \theta \).
3. The sequence \( \sigma \) is adequately bounded, i.e., \( \sum_i \mu(I_i)\kappa_i \leq c(\text{cost}(\tau) + \theta n) \), for a constant \( c \).

This is a slight oversimplification of Definition 3 of \((k, \zeta)\)-approximation of \( \tau \) by a sequence of certified boxes.

The intuition for the second condition is that \( \tau_i \) is "almost" a path between the lower left and upper right corners of \( I_i \times J_i \). Now \( \tau_i \) might have a vertical extent \( J' \) that is much larger than its horizontal extent \( I_i \), in which case it is impossible to place a square \( I_i \times J_i \) with corners close to both endpoints of \( \tau_i \). However, \( \tau_i \) has a very high cost (at least \( |\mu(J') - \mu(I_i)| \)). The closeness required is adjusted based on \( \text{cost}(\tau_i) \), with relaxed requirements if \( \text{cost}(\tau_i) \) is large.

The output of the min-cost path phase should satisfy the requirements of \( \text{GAP-UB}_\theta \). Lemma 17 shows that if the min-cost path phase receives \( \mathcal{R} \) that contains a \((k, \theta)\)-approximating sequence to some min-cost path \( \tau \), then it will output an upper bound to \( d_{\text{edit}}(x, y) \) that is at most \( k^\ell (d_{\text{edit}}(x, y) + \theta n) \) for some \( k \). So that on input \( x, y \) with \( d_{\text{edit}}(x, y) \leq \kappa n \), the output is at most \( 2k^\ell \kappa n \), satisfying the requirements of \( \text{GAP-UB}_\theta \). This formalizes the intution that an adequate approximating sequence captures enough information to deduce a good bound on \( \text{cost}(\tau) \).

Once and for all, we fix a min-cost path \( \tau \). Our task for the covering phase is that, with high probability, \( \mathcal{R} \) includes an adequate approximating sequence for \( \tau \).

A \( \tau \)-match for an \( x \)-interval \( I \) is a \( y \)-interval \( J \) such that \( I \times J \) is an adequate cover of \( \tau \). It is easy to show (Proposition 7) that this implies \( d_{\text{edit}}(I, J) \leq (\text{cost}(\tau) + \theta \mu(I)) \).

A box \( I \times J \) is said to be \( \tau \)-compatible if \( J \) is a \( \tau \)-match for \( I \) and a box sequence is \( \tau \)-compatible if every box is \( \tau \)-compatible. A \( \tau \)-compatible certified box sequence whose distance upper bounds are (on average) within a constant factor of the actual cost, satisfies the requirements for an adequate approximating sequence. Our cover algorithm will ensure that \( \mathcal{R} \) contains such a sequence.

A natural decomposition is \( \mathcal{I}_{w_1} \), with all parts of width \( w_1 \) (think of \( w_1 \) as a power of 2 that is roughly \( n^{1/10} \)) so \( \ell = \lceil w_1 \rceil \) and \( I_j = \{ (j-1)w_1, \ldots, jw_1 \} \). The naive approach to building \( \mathcal{R} \) is to include certified boxes for enough choices of \( J \) to guarantee a \( \tau \)-match for each \( I_j \).

An interval of width \( w_1 \) is \( \delta \)-aligned if its upper and lower endpoints are both multiples of \( \delta w_1 \) (which we require to be an integral power of 2). We restrict attention to \( x \)-intervals in \( \mathcal{I}_{w_1} \), called \( x \)-candidates and \( \theta \)-aligned \( y \)-intervals of width \( w_1 \) called \( y \)-candidates. It can be shown (see Proposition 8) that an \( x \)-interval \( I \) always has a \( \tau \)-match \( J \) that is \( \theta \)-aligned. (In this overview we will fix \( \delta \) to \( \theta \); the actual algorithm has \( O(\log n) \) iterations during which the value of \( \delta \) varies, giving improvements in runtime that are unimportant in this overview.) For each \( x \)-candidate \( I \), designate one such \( \tau \)-match as the canonical \( \tau \)-match, \( J^\tau(I) \) for \( I \), and \( I \times J^\tau(I) \) is the canonical \( \tau \)-compatible box for \( I \).

In the exhaustive approach, for each \((x \text{-candidate}, y \text{-candidate)} \text{-pair} \) \((I, J)\), its edit distance is computed in time \( O(w_1^2) \), and the certified box \((I \times J, \Delta_{\text{edit}}(I, J)) \) is included. There are \( \frac{n}{w_1} \frac{n}{\theta w_1} \) boxes, so the time for all edit distance computations is \( O\left(\frac{n^2}{\theta w_1^2}\right) \), which is worse than quadratic. (The factor \( \frac{1}{\theta} \) can be avoided by standard techniques, but this is not significant to the quest for a sub-quadratic algorithm, so we defer this until the next section.) Note that \( |\mathcal{R}| \) is \( \frac{\log n}{\theta w_1^2} \) (which is \( n^{1.82} \) for our sample parameters) so at least the min-cost path phase (which runs in time quasi-linear in \( \mathcal{R} \)) is truly sub-quadratic.

Two natural goals that will improve the runtime are: (1) Reduce the amortized time per box needed to certify boxes significantly below \( (w_1)^2 \) and (2) Reduce the total number of certified boxes created significantly below \( \frac{n^2}{\theta w_1^2} \). Neither goal is always achievable, and our covering algorithm combines them. In independent work [17], [18], versions of these two goals are combined, where the second goal is accomplished via Grover search, thus yielding a constant factor sub-quadratic time quantum approximation algorithm.

Reducing amortized time for certifying boxes: the dense case algorithm. We aim to reduce the amortized time per certified box to be much smaller than \( (w_1)^2 \). We divide our search for certified boxes into iterations \( i \in \{0, \ldots, \log n\} \). For iteration \( i \), with \( c_i = 2^{-i} \), our goal is that for all candidate pairs \((I, J)\) with \( \Delta_{\text{edit}}(I, J) \leq c_i \), we include the certified box \((I \times J, c_i)\) for a fixed constant \( c \). If we succeed, then for each \( I \) and its canonical \( \tau \)-match \( J^\tau(I) \), and for the largest index \( i \) for which \( \Delta_{\text{edit}}(I, J^\tau(I)) \leq c_i \), iteration \( i \) will certify \((I \times J^\tau(I), \kappa_i)\) with \( \kappa_i \leq c_i \leq 2c \Delta_{\text{edit}}(I, J^\tau(I)) \), as needed.

For a string \( z \) of size \( w_1 \), let \( H(z, \rho) \) be the set of \( x \)-candidates \( I \) with \( \Delta_{\text{edit}}(z, x_I) \leq \rho \) and \( V(z, \rho) \) be the set of \( y \)-candidates \( J \) with \( \Delta_{\text{edit}}(z, y_J) \leq \rho \). In iteration \( i \), for each \( x \)-candidate \( I \), we will specify a set \( Q_i(I) \) of \( y \)-candidates that includes \( V(x_I, c_i) \) and is contained in \( V(x_I, 5c_i) \). The set of certified boxes \((I \times J, 5c_i)\) for all \( x \)-candidates \( I \) and \( J \in Q_i(I) \) satisfies the goal of iteration \( i \).

Iteration \( i \) proceeds in rounds. In each round we select an \( x \)-candidate \( I \), called the pivot, for which \( Q_i(I) \) has not yet been specified. Compute \( \Delta_{\text{edit}}(x_I, y_J) \) for all \( y \)-candidates
and $\Delta_{edist}(x_I, x_J)$ for all $x$-candidates $I'$; these determine $\mathcal{H}(x_I, x_J)$ and $\mathcal{V}(x_I, x_J)$ for any $\rho$. For all $I' \in \mathcal{H}(x_I, 2\epsilon_I)$, set $Q_i(I') = \mathcal{V}(x_I, 3\epsilon_I)$. By the triangle inequality, for each $I' \in \mathcal{H}(x_I, 2\epsilon_I)$, $\mathcal{V}(x_I, 3\epsilon_I)$ includes $\mathcal{V}(x_I, \epsilon_I)$ and is contained in $\mathcal{V}(x_I, 5\epsilon_I)$ so we can certify all the boxes with upper bound $5\epsilon_I$. Mark intervals in $\mathcal{H}(x_I, 2\epsilon_I)$ as fulfilled and proceed to the next round, choosing a new pivot from among the unfulfilled $x$-candidates.

The number of certified boxes produced in a round is $|\mathcal{H}(x_I, 2\epsilon_I)| \times |\mathcal{V}(x_I, 3\epsilon_I)|$. If this is much larger than $O\left(\frac{n^2w_I}{\theta}\right)$, the number of edit distance computations, then we have significantly reduced amortized time per certified box. (For example, in the trivial case $i = 0$, every candidate box will be certified in a single round.) But in worst case, there are $\frac{n}{w_I}$ rounds each requiring $\Omega\left(\frac{n^2w_I}{\theta}\right)$ time, for an unacceptable total time $\Theta(n^2/\theta)$.

Here is a situation where the number of rounds is much less than $\frac{n}{w_I}$. Since any two pivots are necessarily greater than $2\epsilon_I$ apart, the sets $\mathcal{V}(x_I, \epsilon_I)$ for distinct pivots are disjoint. Now for some parameter $d$ ($\text{think of } d = n^{1/5}$) an $x$-candidate is $d$-dense for $\epsilon_I$ if $|\mathcal{V}(x_I, \epsilon_I)| \geq d$, i.e., $x_I$ is $\epsilon_I$-close in edit distance to at least $d$ $y$-candidates; it is $d$-sparse otherwise. If we manage to select a $d$-dense pivot $I$ in each round, then the number of rounds is $O\left(\frac{n}{w_I}\right)$ and the overall time will be $\Theta\left(\frac{n^2w_I}{\theta}\right)$. For the sample parameters this is $\Theta(n^{1.84})$. But there’s no reason to expect that we’ll only choose dense pivots; indeed there need not be any dense pivot.

Let’s modify the process a bit. When choosing potential pivot $I$, first test whether or not it is (approximately) $d$-dense. This can be done with high probability, by randomly sampling $\widetilde{\Theta}\left(\frac{n}{w_I}\right)$ $y$-candidates and finding the fraction of the sample that are within $\epsilon_I$ of $x_I$. If this fraction is less than $\frac{d/n \cdot d}{n}$ then $I$ is declared sparse and abandoned as a pivot; otherwise $I$ is declared dense, and used as a pivot. With high probability, all $d$-dense intervals that are tested are declared dense, and all tested intervals that are not $d/4$-dense are declared sparse, so we assume this is the case. Then all pivots are processed (as above) in time $O\left(\frac{n^2w_I}{\theta}\right)$ (under sample parameters: $O(n^{1.84})$). We pay $\widetilde{O}\left(\frac{n^2w_I}{\theta}\right)(w_I)^2$ to test each potential pivot (at most $\frac{n}{w_I}$ of them) so the overall time to test potential pivots is $\widetilde{O}\left(\frac{n^2w_I}{\theta}\right)$ (with sample parameters: $\widetilde{O}(n^{1.82})$).

Each iteration $i$ (with different $\epsilon_I$) splits $x$-candidates into two sets, $S_i$ of intervals that are declared sparse, and all of the rest for which we have found the desired set $Q_i(I)$. With high probability every interval in $S_i$ is indeed $d$-sparse, but a sparse interval need not belong to $S_i$, since it may belong to $\mathcal{H}(x_I, 2\epsilon_I)$ for some selected pivot $I$.

For every $x$-candidate $I \notin S_i$ we have met the goal for the iteration. If $S_i$ is very small for all iterations, then the set of certified boxes will suffice for the min-cost path algorithm to output a good approximation.

But if $S_i$ is not small, another approach is needed.

**Reducing the number of candidates explored: the diagonal extension algorithm.** For each $x$-candidate $I$, although it suffices to certify the single box $(I, J''(I))$ with a good upper bound, since $\tau$ is unknown, the exhaustive and dense case approaches both include certified boxes for all $y$-candidates $J$. The potential savings in the dense case approach comes from certifying many boxes simultaneously using a relatively small number of edit distance computations.

Here’s another approach: for each $x$-candidate $I$ try to quickly identify a relatively small subset $\mathcal{Y}(I)$ of $y$-candidates that is guaranteed to include $J''(I)$. If we succeed, then the number of boxes we certify is significantly reduced, and even paying quadratic time per certified box, we will have a sub-quadratic algorithm.

We need the notion of diagonal extension of a box. The main diagonal of box $I \times J$ is the segment joining the lower left and upper right corners. The square box $I' \times J'$ is a diagonal extension of a square subbox $I \times J$ if the main diagonal of $I \times J$ is a subsegment of the main diagonal of $I' \times J'$. (See Definition 2.) Given square box $I \times J$ and $I' \in \mathcal{J}(I \times J)$ the diagonal extension of $I \times J$ with respect to $I'$ is the unique diagonal extension of $I \times J$ having $x$-interval $I'$. The key observation (Proposition 9) is: if $I \times J$ is an adequate cover of $\tau_I$ then any diagonal extension $I' \times J'$ is an adequate cover of $\tau_{I'}$.

Now let $w_1, w_2$ be two numbers with $w_1 | w_2$ and $w_2 | n$. (Think of $w_1 = n^{1/10}$ and $w_2 = n^{3/10}$.) We use the decomposition $I_{w_2} = \{0, \ldots, n\}$ into intervals of width $w_2$. The set of $y$-candidates consists $\theta$-aligned vertical intervals of width $w_2$ and has size $\frac{n}{w_2}$. To identify a small set of potential matches for $I' \in I_{w_2}$, we will identify a set (of size much smaller than $\frac{n^2}{w_2}$) of $w_1$-boxes $B(I')$ having $x$-interval in $I_{w_2}(I')$ (the decomposition of $I'$ into width $w_1$ intervals). For each box in $B(I')$ we determine the diagonal extension $I' \times J'$ with respect to $I'$, compute $\kappa = \Delta_{edist}(I', J')$ and certify $(I' \times J', \kappa)$. Our hope is that $B(I')$ includes a $\tau$-compatible $w_1$-box $I'' \times J''(I'')$, then the observation above implies that its diagonal extension provides an adequate cover for $\tau_{I''}$.

Here’s how to build $B(I')$: Randomly select a polylog($n$) size set $\mathcal{H}(I')$ of $w_1$-intervals from $I_{w_2}(I')$. For each $I'' \in \mathcal{H}(I')$ compute $\Delta_{edist}(I'', J'')$ for each $y$-candidate $J''$, and let $\mathcal{J}(I'')$ consist of the $d$ candidates $J''$ with smallest edit distance to $I''$. Here $d$ is a parameter; think of $d = n^{1/5}$ as before. $B(I')$ consists of all $I'' \times J''$ where $I'' \in \mathcal{H}(I')$ and $J'' \in \mathcal{J}(I'')$.

To bound runtime: Each $I' \in I_{w_2}$ requires $\widetilde{O}(\frac{n^2 w_2}{\theta})$ width-$w_1$ $\Delta_{edist}()$ computations, taking time $\widetilde{O}(\frac{n^2w_2}{\theta})$. Diagonal extension step requires $\widetilde{O}(d)$ width-$w_2$ $\Delta_{edist}()$ computations, for time $\widetilde{O}(dw_2^2)$. Summing over $\frac{n}{w_2}$ choices for $I'$ gives time $\widetilde{O}(n^2 w_2^2 + ndw_2^2)$ (with sample parameters: $\widetilde{O}(n^{1.82})$).
Why should \(B(I')\) include a box that is an adequate approximation to \(\tau_I\)? The intuition behind the choice of \(B(I')\) is that an adequate cover for \(\tau_I\) should typically be among the cheapest boxes of the form \(I' \times J'\), and if \(I' \times J'\) is cheap then for a randomly chosen \(w_1\)-subinterval \(I''\), we should also have \(I'' \times J'(I'')\) is among the cheapest boxes for \(I''\).

Clearly this intuition is faulty: \(I'\) may have many inexpensive matches \(J'\) such that \(I' \times J'\) is far from \(\tau_I\), which may all be much cheaper than the match we are looking for. In this bad situation, there are many \(y\)-intervals \(J'\) such that \(\Delta_{ed}(I', J')\) is smaller than the match we are looking, and this is reminiscent of the \textit{good} situation for the dense case algorithm, where we hope that \(I'\) has lots of close matches. This suggests combining the two approaches, and leads to our full covering algorithm.

The full covering algorithm. This is now easy to describe. The parameters \(w_1, w_2, d\) are as above. We iterate over \(i \in \{0, \ldots, \log n\}\) with \(\epsilon_i = 2^{-i}\). In iteration \(i\), we first run the dense case algorithm, and let \(S_i\) be the set of intervals declared sparse. Then run the diagonal extension algorithm described earlier (with small modifications): For each \(w_2\)-interval \(I'\), select \(\mathcal{H}(I') = \mathcal{H}_i(I')\) to consist of \(\theta(\log^2 n)\) independent random selections from \(S_i\). For each \(I'' \in \mathcal{H}_i(I')\), find the set of vertical candidates \(J''\) for which \(\Delta_{ed}(I'', J'') \leq \epsilon_i\). Since \(I''\) is (almost certainly) \(d\)-sparse, the number of such \(J''\) is at most \(d\). Proceeding as in the diagonal extension algorithm, we produce a set \(\mathcal{P}_i(I')\) of \(O(d)\) certified \(w_2\)-boxes with \(x\)-interval \(I'\). Let \(R_D\) (resp. \(R_E\)) be the set of all certified boxes produced by the dense case algorithm, resp. diagonal extension iterations. The output is \(R = R_D \cup R_E\). (See Figure 1 for an illustration of the output \(R\).

The runtime is the sum of the runtimes of the dense case and diagonal extension algorithms, as analyzed above. Later, we will give a more precise runtime analysis for the pseudocode.

To finish this extended overview, we sketch the argument that \(R\) satisfies the covering phase requirements.

Claim 4. Let \(I'\) be an interval in the \(w_2\)-decomposition. Either (1) the output of the dense case algorithm includes a sequence of certified \(w_1\)-boxes that adequately approximates the subpath \(\tau_I\), or (2) with high probability the output of the sparse case algorithm includes a single \(w_2\)-box that adequately approximates \(\tau_I\).

(This claim is formalized in Claim 14.) Stitching together the subpaths for all \(I'\) implies that \(R\) will contain a certified box sequence that adequately approximates \(\tau\).

To prove the claim, we establish a sufficient condition for each of the two conclusion and show that if the sufficient condition for the second conclusion fails, then the sufficient condition for the first holds.

Let \(I''\) denote the \(w_1\)-decomposition \(I_{w_1}(I')\) of \(I'\). Every interval \(I'' \in I''\) has a \(\theta\)-aligned \(\tau\)-match \(J''(I'')\). It will be shown (see Proposition 8), that \(\Delta_{ed}(I'', J''(I'')) \leq 2\frac{\text{cost}(I', J')(\gamma(I'R))}{\mu(I'R)} + \theta\). Let \(u(I'')\) denote this upper bound. Consider the first alternative in the claim. During the dense case iteration \(i = 0\), every interval is declared dense, and \((I'' \times J''(I''), 5)\) is in \(R_D\) for all \(I''\). To get an adequate approximation, we try to show that later iterations provide much better upper bounds on these boxes, i.e., \((I'' \times J''(I''), \gamma(I'')) \in R_D\) for a small enough value of \(\gamma(I'')\). By definition of adequate approximation, it is enough that \(\sum_{I'' \in I''} \gamma(I'') \leq c \sum_{I'' \in I''} u(I'')\), for some \(c\). Let \(l(I'') = \tau_i(I'')\) be the last (largest) iteration for which \(\epsilon_i(I'') \geq u(I'')\) and \(I'' \notin S_i(I'')\) (which is well defined since \(S_0 = \emptyset\)). Let \(b(I'') = \epsilon_i(I'')\). Since \(b(I'') \geq u(I'') \geq \Delta_{ed}(I'', J''(I''))\), the box \((I'' \times J''(I''), 5b(I''))\) is certified. The collection \(\{(I'' \times J''(I''), 5b(I''))\}\) is a sequence of certified boxes that satisfies the first two conditions for an adequate approximation of \(\tau\). The third condition will follow if:

\[
\sum_{I'' \in I''} 5b(I'') \leq c \sum_{I'' \in I''} u(I'')
\]  

so this is sufficient to imply the first condition of the claim.

Next consider what we need for the second alternative to hold. Let \(S_i(I')\) be the set of intervals declared sparse in iteration \(i\). An interval \(I'' \in S_i(I')\) is a \textit{winner} (for iteration \(i\)) if \(\Delta_{ed}(I'', J''(I'')) \leq \epsilon_i\), and \(W_i(I')\) is the set of winners. In iteration \(i\) of the diagonal extension.

---

Figure 1. Illustration of the Covering Algorithm: Green boxes are low cost boxes in dense \(w_1\)-strips, while the pink ones are in sparse \(w_1\)-strips. The blue line corresponds to the path \(\tau\) that we are trying to cover. In each \(w_2\)-strip, \(\tau\) is covered by either a collection of many \(w_1\)-boxes or it is covered by a diagonal extension of a low cost \(w_1\)-box. The various boxes might overlap vertically which is not shown in the picture.
algorithm, we sample \(\theta (\log^2 n)\) elements of \(S_i(I')\). If for at least one iteration \(i\) our sample includes a winner \(I''\) then the second condition of the claim will hold: \(I'' \times J^*(I'')\) is extended diagonally to a \(w_2\)-box, and by the diagonal extension property, the extension is an adequate cover of \(\tau_{I''}\), which we will certify with its exact edit distance.

Thus for the second alternative to fail with nonnegligible probability:

\[
\text{For all } i, |W_i(I')| < |S_i(I') - W_i(I')|,
\]

We argue that if (2) holds, then the success condition (1) holds. Multiply (2) by \(\epsilon_i\) and sum on \(i\) to get:

\[
\sum_{I'' \in \mathcal{I}_w(I')} \epsilon_i < \sum_{I'' \in \mathcal{I}_w(I')} \sum_{\epsilon_i \in S_i(I') - W_i(I')} \epsilon_i. \tag{3}
\]

For \(I'' \in \mathcal{I}_w(I')\), consider the iterations \(i\) for which \(I'' \in W_i(I')\) and those for which \(I'' \in S_i(I') - W_i(I')\). First of all if \(\epsilon_i \geq u(I'')\) and \(I'' \in S_i(I')\) then since \(\Delta_{\text{edit}}(I', J^*(I'')) \leq u(I'') \leq \epsilon_i\) we conclude \(I'' \in W_i(I')\). So \(I'' \in S_i(I') - W_i(I')\) implies that \(\epsilon_i < u(I'')\), so the inner sum of the right side of (3) is at most \(2u(I'')\) (by summing a geometric series).

Furthermore, for \(i\) with \(u(I'') \leq \epsilon_i < b(I'')\), \(I'' \in S_i\) by the choice of \(t(I'')\). Either \(b(I'')/2 \leq u(I'')\) or \(u(I'') < b(I'')/2\). The latter implies \(I'' \in W_i(I'' + 1(I'))\), and then \(b(I'')/2\) is upper bounded by the inner sum on the left of (3). Therefore:

\[
\sum_{I''} b(I'') \leq \sum_{I''} \left( 2u(I'') + \sum_{\epsilon_i \in W_i(I')} 2\epsilon_i \right) < \sum_{I''} \left( 2u(I'') + 2 \sum_{\epsilon_i \in S_i(I') - W_i(I')} \epsilon_i \right) \leq 6 \sum_{I''} u(I''),
\]

as required for (1).

This completes the overview of the covering algorithm.

### III. COVERING ALGORITHM: PSEUDO-CODE AND ANALYSIS

The pseudo-code consists of \textit{CoveringAlgorithm} which calls procedures \textit{DenseStripRemoval} (the dense case algorithm) and \textit{SparseStripExtensionSampling} (the diagonal extension algorithm). These are abbreviated, respectively by CA, DSR and SSES. The technical differences between the pseudo-code and the informal description, are mainly to improve runtime analysis.

#### A. Pseudo-code

The parameters of CA are as described in the overview: \(x, y\) are input strings of length \(n\), \(\theta\) comes from \textit{GAP-UB}_\(\theta\), \(w_1 < w_2 < n\) and \(d < n\) are integral powers of 2, as are the auxiliary input parameters. The output is a set \(R\) of certified boxes. The algorithm uses local constants \(c_0 \geq 0\) and \(c_1 \geq 120\), where the former one is needed for Proposition 11.

We use a subroutine \textit{SMALL-ED} which takes strings \(z_1, z_2\) of length \(w\) and parameter \(\kappa\) and outputs \(\infty\) if \(\Delta_{\text{edit}}(z_1, z_2) > \kappa\) and otherwise outputs \(\Delta_{\text{edit}}(z_1, z_2)\). The algorithm of [4] implements \textit{SMALL-ED} in time \(O(\kappa w^2)\).

One technical difference from the overview, is that the pseudo-code saves time by restricting the search for certified boxes to a portion of the grid close to the main diagonal. Recall that \textit{GAP-UB}_\(\theta\) has two requirements, that the output upper bounds \(d_{\text{edit}}(x, y)\) (which will be guaranteed by the requirement that \(R\) contains no falsely certified boxes), and that if \(d_{\text{edit}}(x, y) \leq \theta_n\), the output is at most \(\theta_n\) for some constant \(c\). We therefore design our algorithm assuming \(d_{\text{edit}}(x, y) \leq \theta_n\), in which case every min-cost \(G_{x,y}\)-path \(r\) consists entirely of points within \(\frac{n}{2}\) steps from the main diagonal, i.e. \(|i - j| \leq \frac{n}{2}\). So we restrict our search for certified boxes as follows: set \(m = \frac{1}{d}\theta_n\), and consider the \(\frac{n}{m}\) overlapping equally spaced boxes of width \(8m = 20n\) lying along the main diagonal. Together these boxes cover all points within \(\theta_n\) of the main diagonal.

The algorithm of the overview is executed separately on each of these \(n/m\) boxes. Within each of these executions, we iterate over \(i \in \{0, \ldots, \log \frac{1}{2} \frac{n}{m}\}\) (rather than \(\{0, \ldots, \log n\}\) as in the overview). In each iteration we apply the dense case algorithm and the diagonal extension algorithm as in the overview. The output is the union over all \(n/m\) boxes and all iterations, of the boxes produced.

In the procedures DSR and SSES, the input \(G\) is an induced grid graph corresponding to a box \(I_G \times J_G\), as described in the "framework" part of Section I. The procedure DSR on input \(G\), sets \(T\) to be the \(w_1\)-decomposition of \(I_G\) (the \(x\)-candidates) and \(B\) to be the set of \(\frac{1}{\theta}\)-aligned \(y\)-candidates. As in the overview, the dense case algorithm produces a set of certified boxes (called \(R_1\) in the pseudo-code) and a set \(S\) of intervals declared sparse. SSES is invoked if \(S \neq \emptyset\) and iterates over all \(x\)-intervals \(I'\) in the decomposition \(I_{\mathcal{L}_w}(I_G)\). The algorithm skips \(I'\) if \(S\) contains no subset of \(I'\), and otherwise selects a sample \(H\) of \(\theta (\log^2 n)\) subintervals of \(I'\) from \(S\). For each sample interval \(I''\) it finds the vertical candidates \(J''\) for which \(\Delta_{\text{edit}}(I'', J'') \leq \epsilon_i\), does a diagonal extension to \(I'\) and certifies each box with an exact edit distance computation.

There are a few parameter changes from the overview that provide some improvement in the time analysis: During each iteration \(i\), rather than take our vertical candidates to be from a \(\theta\)-aligned grid, we can afford a coarser grid that is \(\epsilon_i/8\)-aligned. Also, the local parameter \(d\) in DSR and SSES
is set to \(d/\epsilon_i\) during iteration \(i\).

There is one counterintuitive quirk in SSES: each certified box is replicated \(O(\log n)\) times with higher distance bounds. This is permissible (increasing the distance bound cannot decertify a box), but seems silly (why add the same box with a higher distance bound?). This is just a convenient technical device to ensure that the second phase min-cost path algorithm gives a good approximation.

Algorithm 1 CA(\(x, y, n, w_1, w_2, d, \theta\))

Covering Algorithm

Input: Strings \(x, y\) of length \(n\), \(w_1, w_2, d \in [n]\), \(w_1 < w_2 < \theta n/4\), and \(\theta \in [0, 1]\). \(n, w_1, w_2, \theta\) are powers of 2.

Output: A set \(R\) of certified boxes in \(G\).

1. Initialization: \(R = G_x, y\) \(R_D = R_E = \emptyset\).
2. Let \(m = \Theta(\frac{n}{2})\).
3. for \(k = 0, \ldots, \frac{m}{4}\) do
4. Let \(I = J = \{km, km + 1, \ldots, (k + 8)m\}\).
5. for \(i = \log(1/\theta) + \ldots, 0\) do
6. Set \(\epsilon_i = 2^{-i}\).
7. Invoke \(DSR(G(I \times J), n, w_1, d, \epsilon_i, \epsilon_i)\) to get \(S\) and \(R_1\).
8. if \(S \neq \emptyset\) then
9. Invoke \(SSES(G(I \times J), S, n, w_1, w_2, d, \epsilon_i, \epsilon_i, \theta)\) to get \(R_2\).
10. else
11. \(R_2 = \emptyset\).
12. end if
13. end for
14. end for
15. Output \(R = R_D \cup R_E\).

For the analysis we must prove that \(R\) contains an "adequate approximation" of some min-cost alignment path \(\tau\). To state this precisely, we start with definitions and observations that formalize intuitive notions from the overview.

Cost and normalized cost. The cost of a path \(\tau\), \(\text{cost}(\tau)\), from \((u_1, u_2)\) to \((v_1, v_2)\) in a grid-graph (see Section I), is the sum of the edge costs, and the normalized cost is \(\text{ncost}(\tau) = \frac{\text{cost}(\tau)}{\text{dist}(u_1, u_2)}\). \(\text{cost}(G(I \times J))\) is the cost of a subgraph \(G(I \times J)\), and \(\text{ncost}(G(I \times J))\) is a cost of a path from the lower left to the upper right corner. The normalized cost is \(\text{ncost}(G(I \times J)) = \frac{\text{cost}(G(I \times J))}{\text{dim}(I \times J)}\).

We note the following simple fact without proof:

Proposition 5. For \(I, J, J' \subseteq \{0, \ldots, n\}\), \(d_{ed}(x, y) = d_{ed}(x, y)\) \(|J|\Delta J'\), where \(\Delta\) denotes symmetric difference.

Projections and subpaths. The horizontal projection of a path \(\tau = (i_1, j_1), \ldots, (i_\ell, j_\ell)\) is the set of \(i_1, i_\ell, \ldots, i_\ell\). We say that \(\tau\) crosses box \(I \times J\) if the vertices of \(\tau\) belong to \(I \times J\) and its horizontal projection is \(I\). If the horizontal projection of \(\tau\) contains \(I', \tau_I'\) denotes the (unique) minimal subpath of \(\tau\) whose projection is \(I'\).

Algorithm 2 DSR(\(G, n, w, d, \delta, \epsilon\))

Dense Strip Removal

Input: \(G = G_{x, y}(I_G \times J_G)\) for some \(I_G, J_G \subseteq \{0, 1, \ldots, n\}\), \(w, d \in [n]\), the endpoints of \(I_G\) and \(J_G\) are multiples of \(w\) and \(\epsilon, \delta \in [0, 1]\).

Output: Set \(S\) which is a subset of the \(w\)-decomposition of \(I_G\) and a set \(R\) of \(\delta\)-aligned certified \(w\)-boxes all with distance bound \(5\epsilon_i\).

1. Initialization: \(S = R = \emptyset\). \(T = \{I_G\}\)
2. \(B\), the set of \(y\)-candidates, is the set of width \(w\) \(\delta\)-aligned subintervals of \(J_G\) (having endpoints a multiple of \(\delta w\)).
3. while \(T\) is non-empty do
4. Pick \(I \in T\)
5. Sample \(c_0|B|\frac{1}{2} \log n\) intervals \(J \in B\) uniformly at random and for each test if \(\Delta_{ed}(x, y) \leq \epsilon\).
6. if for at most \(\frac{1}{2} \log n\) sampled \(J\)’s, \(\text{SMALL-ED}(x, y, \epsilon) < \infty\) then
7. \(S = S \cup \{I\}, T = T - \{I\}\) (Is declared sparse)
8. else
9. \((I\) is declared dense and used as a pivot)
10. Compute:
11. \(Y = \{J \in B; \text{SMALL-ED}(x, y, 3\epsilon) < \infty\}\).
12. \(X = \{I \in T; \text{SMALL-ED}(x, y, 3\epsilon) < \infty\}\).
13. Add \((I', J', 5\epsilon)\) to \(R\) for all pairs \((I', J') \in X \times Y\).
14. \(T = T - X\).
15. end if
16. end while
17. Output \(S\) and \(R\).

985
Algorithm 3 SSES($G, S, n, w_1, w_2, d, \delta, \epsilon, \theta$)
SparseStripExtensionSampling

Input: $G = G_{S,w}(I_G,J_G)$ with $I_G, J_G \subseteq \{0,1,\ldots,n\}$, $w_1,w_2,d,n$ are powers of 2, with $w_1,w_2,d < n$ and $w_1 < w_2$. Endpoints of $I_G$ and $J_G$ are multiples of $w_2$.
$S$ is a subset of the $w_1$-decomposition of $I_G$ and $\delta,\epsilon,\theta$ are non-positive integral powers of 2.
Output: A set $R$ of certified $w_2$-boxes in $G$.

1: Initialization: $R = \emptyset$.
2: $B$, the set of $y$-candidates, is the set of width $w$ $\delta$-aligned subintervals of $J_G$ (endpoints are multiples of $\delta w$.)
3: for $I' \in I_{w_2}(I_G)$ do
4: if $S$ includes a subset of $I'$ then
5: Select $c_1 \log^2 n$ intervals $I_1 \in S$ independently and uniformly at random from $I_{w_1}(I') \cap S$, to obtain $H$.
6: for each $I_i \in H$ and each $J \in B$ do
7: if SMALL-ED($x_f,y_f,\epsilon$) $< \infty$ then
8: Let $J'$ be such that $I' \times J'$ is the diagonal extension of $I \times J$ in $I' \times J$.
9: Let $p = \text{SMALL-ED}(x_f,y_f,3\epsilon)$
10: if $p < \infty$ then
11: For $k = 0,\ldots,\log n$, add $(I',J',p + \theta + 2^{-k})$ to $R$.
12: end if
13: end if
14: end for
15: end if
16: end for
17: Output $R$.

2) If $J''$ is any vertical interval, then $I' \times J''$ $(1 - \delta - |J' \Delta J''|/\mu(I'))$ covers $\tau$.

The routine proof is in the full version.

\textbf{\textit{\delta-aligned boxes}} A $y$-interval $J$ of width $w$ is $\delta$-aligned for $\delta \in (0,1)$ if its endpoints are multiples of $\delta w$ (which we require to be an integer). The easy proof of the following is in the full version:

\textbf{Proposition 8.} Let $\tau$ be a path that crosses $I \times J$. Suppose that $I' \subseteq I$ has width $w$, and $\mu(J) \geq w$.

1) There is an interval $J_1$ with $\mu(J_1) = \mu(I')$ so that $\text{ncost}(I' \times J_1) \leq 2\text{ncost}(\tau_{I'}I)$ and $I' \times J_1$ $(1 - \text{ncost}(\tau_{I'}I))$-covers $\tau$.
2) There is a $\delta$-aligned interval $J' \subseteq J$ of width $w$ so that $\text{ncost}(I' \times J') \leq 2\text{ncost}(\tau_{I'}I) + \delta$ and $I' \times J'$ $(1 - \text{ncost}(\tau_{I'}I) - \delta)$-covers $\tau$.

($J_1, J'$ are "$\tau$-matches" for $I'$, in the sense of the overview.)

\textbf{Definition 2.} 1) The main diagonal of a box is the segment joining the lower left and upper right corners.
2) For a square box $I' \times J'$, and $I' \subseteq I$, the true diagonal extension of $I' \times J'$ to $I$ is the square box $I \times J$ whose main diagonal contains the main diagonal of $I' \times J'$.

3) For a $w$-box $I' \times J'$ contained in strip $I \times J$, the adjusted diagonal extension of $I' \times J'$ within $I \times J$ is the box $I \times J'$ obtained from the true diagonal extension of $I' \times J'$ to $I$ by the minimal vertical shift so that it is a subset of $I \times J$. (The adjusted diagonal extension is the true diagonal extension if the true diagonal extension is contained in $I \times J$; otherwise it's lower edge is $\min(J)$ or its upper edge is $\max(J)$.)

\textbf{Proposition 9.} Suppose path $\tau$ crosses $I \times J$ and $\text{ncost}(\tau_I) \leq \epsilon$. Let $w = \mu(I)$. Let $I' \times J'$ be a $w'$-box that $(1 - \delta)-$covers $\tau'$. Then the adjusted diagonal extension $I \times J'$ of $I' \times J'$ within $I \times J$ $(1 - (\epsilon + \frac{\delta w}{w'}))$-covers $\tau$ and satisfies $\text{ncost}(I \times J') \leq 3\epsilon + 2\delta \frac{w}{w'}$.

The straightforward proof appears in the full version.

\textbf{Definition 3.} Let $G$ be a grid graph on $I \times J$. Let $\zeta, \epsilon \in [0,1]$. Let $\tau$ be a path that crosses $G$. A sequence of certified boxes $\sigma = \{(I_1 \times J_1, \epsilon_1),(I_2 \times J_2, \epsilon_2),\ldots,(I_n \times J_n, \epsilon_n)\}$ $(k,\zeta)$-approximates $\tau$ provided:
1) $I_1,\ldots,I_n$ is a decomposition of $I$.
2) For each $i \in [n]$, $I_i \times J_i$ $(1 - \epsilon_i)$-covers $\tau$.
3) $\sum_{i \in [n]} \epsilon_i \mu(I_i) \leq (k \cdot \text{ncost}(\tau) + \zeta)\mu(I)$.

\textbf{Proposition 10.} Suppose path $\tau$ crosses $I \times J$ and $I_1,\ldots,I_m$ is a decomposition of $I$, and for $i \in [m], \sigma_i$ is a certified box sequence that $(k,\zeta)$-approximates $\tau_{I_i}$. Then $\sigma_1,\ldots,\sigma_m$ $(k,\zeta)$-approximates $\tau$.

The routine proof appears in the full version.

\textbf{(d,\delta,\epsilon)-dense and -sparse.} Fix a box $I \times J$. An interval $I' \subseteq I$ of width $w$ is $(d,\delta,\epsilon)$-sparse (wrt $I \times J$) for integer $d$ and $\epsilon, \delta \in (0,1]$ if there are at most $d$ $\delta$-aligned $w$-boxes in $I' \times J$ of ncost at most $\epsilon$, and is $(d,\delta,\epsilon)$-dense otherwise.

The \textbf{sets $S_i$ and $S_i(I')$.} For fixed $k$ in the outer loop of CA, the set $S$ created in iteration $i$ of CA is denoted by $S_i$. For any interval $I'$, $S_i(I')$ is the set of subintervals of $I'$ belonging to $S_i$.

\textbf{Successful Sampling.} The algorithm uses random sampling in two places, in the $i$ loop inside CA and within the conditional on $S$ containing a set from $I_{w_1}(I')$ in SSSE. We now specify what we need from the random sampling.

\textbf{Definition 4.} A run of the algorithm has successful sampling provided that for every $k \in \{0,\ldots,4/\theta\}$ and $i \in \{0,\ldots,\log \frac{1}{\theta} \}$ in the nested CA loops:

- For every $w_1$ interval $I$ with endpoints a multiple of $w_1$, if $I$ is $(d,\frac{\delta}{w_1},\epsilon_1)$-dense interval (in terms of global parameters), DSR does not assign $I$ to $S$ and if $I$ is $(\frac{d}{w_1},\frac{\delta}{w_1},\epsilon_1)$-sparse, DSR places $I$ in $S$. 

986
• On all calls to SSES, for every $w_2$ interval $I$ with endpoints a multiple of $w_2$, if $|W_i(I)|$ has size at least
\[
|S_i(I) - W_i(I)|/32 \text{ then the sample } \mathcal{H} \text{ selected contains an element of } W_i(I). \]
(Here $S_i(I)$ and $W_i(I)$ are that defined in the proof of Claim 14, whose definitions don’t depend on the randomness used to select $\mathcal{H}$.)

The proof of the following (via standard tail bounds) appears in the full version:

**Proposition 11.** For large enough $n$, a run of CA has successful sampling with probability at least $1 - n^{-7}$.

We assume that coins are fixed in a way that gives successful sampling.

**B. Properties of the covering algorithm**

The main property of CA to be proved is:

**Theorem 12.** Let $x, y$ be strings of length $n$, $1/n \leq \theta \leq 1$ be a real. Let $w_1, w_2, d$ satisfy $w_1 \leq \theta w_2, w_2 \leq \frac{\theta n}{4}$ and $1 \leq d \leq \frac{\theta n}{w_2}$. Assume $n, w_1, w_2, d, \theta$ are powers of $2$. Let $\mathcal{R}$ be the set of weight boxes obtained by running CA $(x, y, n, w_1, w_2, d, \theta)$ with $c_1 > 120$. Then (1) Every $(I \times J, \epsilon) \in \mathcal{R}$ is correctly certified, i.e., $\Delta_{edit}(x_I, y_J) \leq \epsilon$, and (2) In a run that satisfies successful sampling, for every path $\tau$ from the source to the sink in $G = G_{x,y}$ of cost at most $\theta$ there is a subset of $\mathcal{R}$ that $(45, 150)$-approximates $\tau$.

**Proof:** All boxes output are correctly certified: Each box in $\mathcal{R}_{E}$ comes from SSES which only certifies boxes with atleast their exact edit distance. For $(I \times J, \epsilon) \in \mathcal{R}_{D}$, there must be an $I'$ such that $\Delta_{edit}(x_{I'}, y_{I'}) \leq \frac{\epsilon}{2} \cdot \epsilon$ and $\Delta_{edit}(x_{I'} - x_I, y_{I'} - y_I) \leq \frac{\epsilon}{2} \cdot \epsilon$ and so $\Delta_{edit}(x_{I'}, y_{I'}) \leq \epsilon$.

It remains to establish (2). Fix a source-sink path $\tau$ of normalized cost $\kappa$. By Proposition 10 it is enough to show that for each $I' \in \mathcal{R}_{w_2}$, $\mathcal{R}$ contains a box sequence that $(45, 150)$-approximates $\tau_{I'}$. So we fix $I' \in \mathcal{R}_{w_2}$.

The main loop (on $k$) of CA processes $G$ in overlapping boxes. Since $ncost(\tau) \leq \theta$, one of these boxes, which we’ll call $I \times J$, must contain $\tau_{I'}$. (See the full version for a proof.)

We note:

**Claim 13.** Let $i \in \{0, \ldots, \log 1/\theta\}$. Suppose $I'' \in \mathcal{R}_{w_1}(I)$ and $J'' \in J$ is $\epsilon_i/8$-aligned. If $I'' \not\in S_i$ and $\text{cost}(I'' \times J'') \leq \epsilon_i$ then $(I'' \times J'', 5\epsilon_i) \in \mathcal{R}_D$.

**Proof:** If $I'' \not\in S_i$ then in the call to DSR($G(I \times J)$, $n, w_1, d/\epsilon_i, \epsilon_i/8, \epsilon_i$) there is an iteration of the main loop where the selected interval $I$ from $\mathcal{T}$ is declared dense and $\Delta_{edit}(x_I, x_{I''}) \leq 2\epsilon_i$. Since $\Delta_{edit}(x_{I''}, y_{J''}) \leq \epsilon_i$, $\Delta_{edit}(x_I, y_J) \leq 3\epsilon_i$ and so $I'' \in \mathcal{X}$ and $J'' \in \mathcal{Y}$. Thus, DSR certifies $(I'' \times J'', 5\epsilon_i)$, which is added to $\mathcal{R}_D$.

The theorem follows from:

**Claim 14.** For an interval $I' \in \mathcal{I}_{w_2}$, assuming successful sampling $\mathcal{R}_{E}$ or $\mathcal{R}_{D}$ contains a $(45, 150)$-approximation of $\tau_{I'}$.

The proof is similar to that of Claim 4, with adjustments for some technicalities.

**Proof:** Let $\tau' = \tau_{I'}$ and $\kappa = \text{ncost}(\tau_{I'})$. Let $\mathcal{T}' = \mathcal{T}_{w_1}(I')$. For $I'' \in \mathcal{T}'$, let $\kappa_{I''} = \text{ncost}(\tau_{I''})$. By Proposition 8, for all $I'' \in \mathcal{T}'$ and $\epsilon_i \geq \kappa_{I''}$ there is an $\epsilon_i/8$-aligned vertical interval $J''(I'')$, such that $\text{ncost}(I'' \times J''(I'')) \leq 2\kappa_{I''} + \epsilon_i/8$ and $I'' \times J''(I'')(1 - \kappa_{I''} - \epsilon_i/8)$-covers $\tau_{I''}$.

Let $s(I'')$ be the largest integer such that $\epsilon(s(I'')) \geq 3\kappa_{I''} + \kappa + \theta$. Let $t(I'') \leq s(I'')$ be the largest integer such that $I'' \not\in S_t(I'')$. (Since $\theta n/w_1 \geq d$, $S_0 = \emptyset$, so $t(I'')$ is well-defined.) Let $a(I'') = \epsilon(s(I''))$ (this plays a similar role to $u(I'')$ in Section II) and $b(I'') = \epsilon(t(I''))$.

For all $\epsilon_i \in [a(I''), b(I'')]$, $\text{ncost}(I'' \times J''(I'')) \leq \epsilon_i$ and $I'' \times J''(I'')(1 - \epsilon_i)$-covers $\tau'$. By the definition of $b(I'')$ and Claim 13, $\mathcal{R}_D$ contains the certified box $I'' \times J''_{s(I'')}$. So $\mathcal{R}_D$ contains a $(45, 150)$-approximation of $\tau'$ provided that:

\[
\sum_{I'' \in \mathcal{T}'} 5b(I'') \leq \frac{45}{8} \sum_{I'' \in \mathcal{T}'} a(I'') \tag{4}
\]

since $a(I'') \leq 2(3\kappa_{I''} + \kappa + \theta)$.

Next we determine a sufficient condition that $\mathcal{R}_E$ contain a box sequence (consisting of a single box) that $(5, 4\theta)$-approximates $\tau'$. Let $S_i(I') = S_i \cap \mathcal{T}'$. Interval $I'' \in S_i(I')$ is a winner for iteration $i$ if $\epsilon_i \geq a(I'')$. This set of winners is denoted by $W_i(I')$. It suffices that during iteration $i$, the set of $c_1 \log^2 n$ samples taken in SSSES includes a winner $I''$; then since $\Delta_{edit}(I'', J''(I'')) \leq \epsilon_i$, the $(\epsilon_i)$-diagonal extension $I'' \times J''(I'')$ of $I'' \times J''(I'')$ will be certified. By Proposition 9, $I' \times J$ has normalized cost at most $3\kappa + 2\epsilon_i w_1/w_2 \leq 3\kappa + 2\theta \leq 3\epsilon_i$ and it $(1 - (\kappa + \theta))$-covers $\tau'$. If $\kappa = 0$ then $I' \times J, \text{ncost}(I' \times J) + \theta + 2^{-\log n}$ is in $\mathcal{R}_E$ by the behavior of SSES and it $(5, 4\theta)$-approximates $\tau'$. Otherwise $\kappa \geq 1/n$, so set $k = \lceil \log 1/\kappa \rceil$. Thus, $k \leq \log n$ and $2^{-k} \in [\kappa, 2\kappa)$. Then $(I' \times J, \text{ncost}(I' \times J) + \theta + 2^{-k})$ is in $\mathcal{R}_E$ and it $(5, 4\theta)$-approximates $\tau'$.

Under successful sampling if $|W_i(I')| \geq \frac{1}{32}|S_i(I') - W_i(I')|$, at least one interval from $W_i(I')$ will be included in our $c_1 \log^2 n$ samples during SSSES and $\mathcal{R}_E$ will contain a $(5, 4\theta)$-approximation of $\tau'$ as above. So suppose this fails:

For all $i, |W_i(I')| \leq \frac{1}{32}|S_i(I') - W_i(I')|$. \tag{5}

We show that this implies (4). Multiplying (5) by $\epsilon_i$ and summing on $i$ yields:

\[
\sum_{I'' \in \mathcal{T}'} \sum_{I'' \in W_i(I')} \epsilon_i \leq \frac{1}{32} \sum_{I'' \in \mathcal{T}'} \epsilon_i \sum_{I'' \in S_i(I') - W_i(I')} \epsilon_i \tag{6}
\]

$I'' \in S_i(I') - W_i(I')$ implies $\epsilon_i < a(I'')$. Summing the geometric series:

\[
\sum_{i : I'' \in S_i(I') - W_i(I')} \epsilon_i \leq 2a(I''). \tag{7}
\]
Either \(a(I'') = b(I'')\) or \(a(I'') < b(I'')\). If the latter, then \(I'' \in W_i(I')\) for \(\epsilon_i = b(I'')/2\). So:

\[
\sum_{I'} b(I'') \leq \sum_{I'} \left( a(I'') + \frac{1}{16} \sum_{i : I'' \in W_i(I')} 2\epsilon_i \right) < \sum_{I'} \left( a(I'') + \frac{1}{16} \sum_{i : I'' \in S_i(I') - W_i(I')} \epsilon_i \right) \leq \frac{9}{8} \sum_{I'} a(I'')
\]

which implies Equation 4. (The second inequality follows from (6) and the last inequality from (7).)

\[\boxed{}
\]

**C. Time complexity of CA**

We write \(t(w, \epsilon)\) for the time of \textsc{Small-Ed}(\(z_1, z_2, \epsilon\)) on strings of length \(w\). We assume \(t(w, \epsilon) \geq w\), and that for \(n \geq 1\), there is a constant \(c(k)\) such that for all \(\epsilon \in [0, 1]\) and all \(w > 1\), \(t(w, k\epsilon) \in \text{time}(w, c(k))\). As mentioned earlier, by [4], we can use \(t(w, \epsilon) = O(w^{2\epsilon})\).

**Theorem 15.** Let \(n\) be a sufficiently large power of 2 and \(\theta \in [1/n, 1]\) be a power of 2. Let \(x, y\) be strings of length \(n\). Let \(\log n \leq w_1 \leq w_2 \leq \theta n/4\), \(1 \leq d \leq n\) be powers of 2, where \(w_1, w_2\) are powers of \(2\), and \(w_1/w_2 \leq \theta\). The size of the set \(R\) output by \(CA\) is \(O((w_1^{2\epsilon} + w_2^{2\epsilon})n)\) and in any run that satisfies successful sampling, \(CA\) runs in time:

\[
O \left( |R| + \sum_{k = \log 1/\theta, \ldots, n} \left( \frac{\theta n^2 \log^2 n}{d \epsilon w_1^2}, t(w_1, \epsilon) + \frac{\theta n^2 \log^2 n}{d \epsilon w_2^2} , t(w_2, \epsilon) \right) \right).
\]

**Proof:** To bound \(|R|\) note that for each choice of \(k, i\) in the outer and inner loops of \(CA\), the set bounds the number of boxes certified by \(DSR\). The call to \(SSES\) constructs at most one diagonal extension for each such candidate box, and each diagonal extension gives rise to at most \(O(\log n)\) certified boxes. Thus, for each \((k, i)\) there are \(O((w_1^{2\epsilon} + w_2^{2\epsilon})n)\) certified boxes. Summing the geometric series over \(i\), we get that \(\min(\epsilon_i) = \theta\), and summing over \(O(1/\theta)\) values of \(k\) gives the required bound on \(|R|\).

The steps in the algorithm that actually construct certified boxes (13 of \(DSR\), 11 of \(SSES\), 13 of \(CA\)) cost \(O(1)\) per box giving the first term in the time bound.

We next bound the other contributions to runtime. The outer loop of \(CA\) has \(1 + \log \frac{\theta}{\delta}\) iterations on \(k\)'s. The inner loop has \(1 + \log \frac{\theta}{\delta}\) iterations on \(i\). Each iteration invokes \(DSR\) and \(SSES\) on \(I \times J\) with \(I\) and \(J\) of width at most \(4\theta n\).

We bound the time of a call to \(DSR\). To distinguish between local variables of \(DSR\) and global variables of \(CA\), we denote local input variables as \(G, \bar{n}, \bar{w}, \bar{d}, \bar{\delta}, \bar{\epsilon}\). For \(B\) and \(T\) as in \(DSR\), \(|B| \leq \mu(I_{G_{\bar{\delta}}})\). Since \(\mu(I_{G_{\bar{\delta}}}) = \mu(I_{G_{\delta}})\), the main while loop of \(DSR\) repeatedly picks intervals \(I \in T\) and samples \(c_0|B| \log^2 \frac{\theta}{\delta} \leq c_0 \mu(I_{G_{\bar{\delta}}}) \log^2 \frac{\theta}{\delta} \delta \bar{w} \)

vertical intervals \(J\) and tests whether \(\Delta_{\bar{\delta}}(x_1, y_1) \leq \epsilon\). Each such test takes time \(t(\bar{w}, \epsilon)\). This is done at most once for each of the \(\mu(I_{G_{\bar{\delta}}})\) horizontal candidates for a total time of \(O(\mu(I_{G_{\bar{\delta}}}) \log^2 \frac{\theta}{\delta} \bar{w} \delta)\). We next bound the cost of processing a pivot \(\bar{T}\). This requires testing \(\Delta_{\bar{\delta}}(x_I, y_I) \leq 3\epsilon\) for \(J \in B\) and \(\Delta_{\bar{\delta}}(x_I, y_I) \leq 2\epsilon\) for \(I \in T\). Each test costs \(O(t(\bar{w}, \epsilon))\) (by our assumption on \(t(\bar{w}, \epsilon)\)) and since \(|T| \leq |B| = \mu(I_{G_{\bar{\delta}}})\), \(I\) is processed in time \(O(\mu(I_{G_{\bar{\delta}}}) t(\bar{w}, \epsilon))\). This is multiplied by the number of intervals declared dense, which we now upper bound. If \(I\) is declared dense then at the end of processing \(I\), \(X\) is removed from \(T\). This ensures \(\Delta_{\bar{\delta}}(I, I') > 2\epsilon\) for any two intervals \(I, I'\) declared dense. By the triangle inequality the sets \(B(I) = \{J \in B : \Delta_{\bar{\delta}}(x_I, y_I) \leq \epsilon\}\) are disjoint for different pivots. By successful sampling, for each pivot \(I\), \(|B(I)| \geq \frac{4}{\delta}\) and thus at most \(|B|/|d/\delta| = 4\mu(I_{G_{\bar{\delta}}})\) intervals are declared dense, so all intervals declared dense are processed in time \(O(\mu(I_{G_{\bar{\delta}}})^2 t(\bar{w}, \epsilon))\).

The time for dense/sparse classification of intervals and for processing intervals declared dense is at most \(O(\mu(I_{G_{\bar{\delta}}})^2 \log^2 \frac{\theta}{\delta} \bar{w} \delta)\). During iteration \(i\) of the inner loop of \(CA\), the local variables of \(DSR\) are set as \(\bar{n} = n\), \(\mu(I_{G_{\bar{\delta}}}) \leq 4\theta n\), \(\bar{w} = w_1\), \(d = d/\epsilon_i\), \(\delta = \epsilon_i/8\). Substituting these parameters yields time \(O(\bar{n}^2 \bar{w}^2 n) t(w_1, \epsilon_i)\). Multiplying by the \(O(1/\theta)\) iterations on \(k\) gives the first summand of the theorem.

Next we turn to \(SSES\). The local input variables \(n, w_1, w_2, S, \theta\) are set to their global values so we denote them without . The other local input variables are denoted as \(\bar{G}, \bar{d}, \bar{\delta}, \bar{\epsilon}\). The local variable \(B\) has size \(\mu(I_{G_{\bar{\delta}}})\). By successful sampling, we assume that on every call, every interval in \(S\) is \((d, \bar{\delta}, \bar{\epsilon})\)-sparse. The outer loop enumerates the \(\mu(I_{G_{\bar{\delta}}})/w_2\) intervals \(I'\) of \(I_2(I_{G_{\bar{\delta}}})\). We select \(H\) to be \(\epsilon_i \log n\) random subsets from subsets of \(I'\) belonging to \(S\). For each \(I \in H\) and \(J \in B\), we call \textsc{Small-Ed}(\(x_1, y_J, \epsilon\)) taking time \(t(w_1, \epsilon_i)\). The total time of all tests is \(O(\mu(I_{G_{\bar{\delta}}}) \log n^2 n t(w_1, \epsilon_i))\). Using \(d = d/\epsilon_i\), \(\delta = \epsilon_i/8\) and \(\bar{\epsilon} = \epsilon_i\) from the ith call to \(SSES\) gives \(O(\bar{n}^2 \bar{w}^2 \log n) t(w_1, \epsilon_i)\). Multiplying by the \(O(1/\theta)\) iterations on \(k\) gives the second summand in the theorem.

Assuming successful sampling, all intervals in the set \(S\) passed from \(DSR\) to \(SSES\) are \((d, \bar{\delta}, \bar{\epsilon})\)-sparse. Therefore, for each sampled \(I\), at most \(d\) intervals \(J\) are within \(\bar{\epsilon}\) of \(I\). For each of these we do a diagonal extension of \(I \times J\) to a \(w_2\)-box \(I' \times J'\), and call \textsc{Small-Ed}(\(x_1, y_J, \delta, \epsilon\)) at cost \(O(t(w_2, \epsilon_i))\) for each call. The number of such calls is \(O(\nu(I_{G_{\bar{\delta}}}) \log n)\). Using the parameter \(d = d/\epsilon_i\) in the ith call of the inner iteration of \(CA\), we get a cost of
$O\left(\frac{\theta d \log^2 n}{\epsilon \epsilon' w_2}\right)t(w_2, \epsilon')$ and multiplying by the $O(1/\theta)$ gives the third summand in the theorem.

Choosing the parameters to minimize the maximum term in the time bound, subject to the restrictions of the theorem and using $t(w, \epsilon) = O(\epsilon w^2)$ we have:

**Corollary 16.** For all sufficiently large $n$, and for $\theta \geq n^{-1/3}$ (both powers of 2) choosing $w_1, w_2,$ and $d$ to be the largest powers of 2 satisfying: $w_1 \leq \theta^{-2/7}(n)^{3/7},$ $w_2 \leq \theta^{1/7}(n)^{3/7},$ and $d \leq \theta^{3/7}(n)^{2/7}$, with probability at least $1 - n^{-1/7},$ CA runs in time $O(n^{12/7} \theta^{4/7}),$ and outputs the set $R$ of size at most $\tilde{O}(n^{12/7} \theta^{4/7}).$

**IV. MIN-COST PATHS IN SHORTCUT GRAPHS**

We now describe the second phase of our algorithm, which uses the set $R$ output by CA to upper bound $d_{\text{edit}}(x, y).$ A shortcut graph on vertex set $\{0, \ldots, n\} \times \{0, \ldots, n\}$ consists of the $H$ and $V$ edges of cost 1, together with an arbitrary collection of shortcut edges $(i, j) \rightarrow (i', j')$ where $i < i'$ and $j < j'$, also denoted by $e_{i,j}$ where $I = \{i, \ldots, i'\}$ and $J = \{j, \ldots, j'\}$, along with their costs. A certified graph (for $x, y$) is a shortcut graph where every shortcut edge $e_{i,j}$ has cost at least $\min_d(x, y)$. The min-cost path from $(0, 0)$ to $(n, n)$ in a certified graph upper bounds $d_{\text{edit}}(x, y)$. The second phase algorithm uses $R$ to construct a certified graph, and computes the min-cost path to upper bound on $d_{\text{edit}}(x, y)$.

A certified box $(I \times J, \kappa)$ corresponds to the $e_{i,j}$ with cost $\kappa \mu(I)$. In the certified graph we use non-normalized costs. However, the certified graph built from $R$ in this way may not have a path of cost $O(d_{\text{edit}}(x, y) + \theta n)$. We need a modified version of $(I \times J, \kappa)$. If $\kappa \geq 1/2$ we add no shortcut. Otherwise $(I \times J, \kappa)$ converts to the edge $e_{i,j}$ with cost $\kappa \mu(I)$ where $J'$ is obtained by shrinking $J$: $\min(J') = \min(J) + \ell$ and $\max(J') = \max(J') - \ell$ where $\ell = \lfloor \kappa \mu(I) \rfloor$. By Proposition 5, this is a certified edge. Call the resulting graph $\tilde{G}$. The following straightforward claim is proved in the full version:

**Lemma 17.** Let $\tau$ be a path from source to sink in $G_{x,y}$. If $R$ contains a sequence $\sigma$ that $(k, \ell)$-approximates $\tau$ then there is a source-sink path $\tau'$ in $G$ that consists of the shortcuts corresponding to $\sigma$ together with some $H$ and $V$ edges with cost $\tilde{G}(\tau') \leq 5(k \cdot \text{cost}_{G_{x,y}}(\tau) + \theta n)$.

**Computing the min-cost.** We present an $O(n + m \log(mn))$ algorithm to find a min cost source-sink path in a shortcut graph $G$ with $m$ shortcuts. It’s easier to switch to the max-benefit problem: Let $H$ be the same graph with cost $c_e$ of $e = (i, j) \rightarrow (i', j')$ replaced by benefit $b_e = (i' - i) + (j' - j) - c_e$, (so $H$ and $V$ edges have benefit 0). The min-cost path of $\tilde{G}$ is $2n$ minus the max-benefit path of $\tilde{H}$. To compute the max-benefit path of $\tilde{H}$, we use a binary tree data structure with leaves $\{1, \ldots, n\}$, where each node $v$ stores a number $b_v$, and a collection of lists $L_1, \ldots, L_n$, where $L_i$ stores pairs $(e, q(e))$ where the head of $e$ has $x$-coordinate $i$ and $q(e)$ is the max benefit of a path that ends with $e$.

We proceed in $n - 1$ rounds. Let the set $A_i$ consist of all the shortcuts whose tail has $x$-coordinate $i$. The preconditions for round $i$ are: (1) for each leaf $j$, the stored value $b_j$ is the max benefit path to $(i, j)$ that includes a shortcut whose head has $y$-coordinate $j$ (or 0 if there is no such path), (2) for each internal node $v$, $b_v = \max(b_j : j$ is a leaf in the subtree of $v)$. and (3) for every edge $e = (i', j') \rightarrow (i'', j'')$ with $i' < i$, the value $q(e)$ has been computed and $(e, q(e))$ is in list $L_{i'}$. During round $i$, for each shortcut $e = (i, j) \rightarrow (i', j')$ in $A_i$, $q(e)$ equals the max of $b_{i'} + b_j$ over tree leaves $v$ with $v \leq j$. This can be computed in $O(\log n)$ time as $b_{i'} + b_j$, over $(j)$ union the set of left children of vertices on the root-to-$j$ path that are not themselves on the path. Add $(e, q(e))$ to list $L_{i'}$. After processing $A_i$, update the binary tree: for each $(e, q(e)) \in L_{i+1}$, let $j$ be the $y$-coordinate of the head of $e$ and for all vertices $v$ on the root-to-$j$ path, replace $b_v$ by $\max(b_v, q(e))$. The tree then satisfies the precondition for round $i + 1$. The output of the algorithm is $b_n$ at the end of round $n - 1$. It takes $O(n)$ time to set up the data structure, $O(m \log n)$ time to sort the shortcuts, and $O(\log n)$ processing time per shortcut (computing $q(e)$ and later updating the data structure).

**V. SUMMING UP AND SPEEDING UP**

To summarize, the algorithm GAP-UB runs CoveringAlgorithm of Section III, converts the output into a shortcut graph, and runs the min-cost path algorithm of Section IV. By Corollary 16, and the quasilinear runtime (in the number of shortcuts) of the min-cost path algorithm, the algorithm GAP-UB runs in time $O(n^{12/7} \theta^{4/7})$. The construction of the main algorithm ED-UB from GAP-UB is standard:

**Proof of Theorem 1 from Theorem 2:** Given GAP-UB, we construct ED-UB: Run the aforementioned exact algorithm of [3] with runtime $O(n + k^2)$ time on instances of edit distance $k$, for $O(n + n^{2-2/5})$ time. If it terminates then it outputs the exact edit distance. Otherwise, the failure to terminate implies $d_{\text{edit}}(x, y) \leq n^{4/5}$. Now run GAP-UB on $(x, y)$ for $\theta_j = (1/2)^j$ for $j = \{0, \ldots, \log n\}$ and output the minimum of all upper bounds obtained. Let $j$ be the largest index with $\theta_j n \geq d_{\text{edit}}(x, y)$ (such an index exists since $j = 0$ works). The output is at most $840\theta_j n \leq 1680d_{\text{edit}}(x, y)$. We run at most $O(\log n)$ iterations, each with runtime $O(n^{2-2/7})$.

**Speeding up the algorithm.** The runtime of ED-UB is dominated by the cost of SMALL-ED$(z_1, z_2, \epsilon)$ on pairs of strings of length $w \in \{w_1, w_2\}$. We use Ukkonen’s algorithm [4] with $t(w, \epsilon) = O(\epsilon w^2)$. In the full paper we describe a revised algorithm ED-UB1, replacing the Ukkonen’s algorithm with ED-UB. This worsens the approximation factor (roughly multiplying it by the approximation factor of ED-UB) but improves runtime. The internal parameters
$w_1, w_2, d$ are adjusted to maximize savings. One can iterate this process any constant number of times to get faster algorithms with worse (but still constant) approximation factors. Because of the dependence of the analysis on $\theta$, we do not get a faster edit distance algorithm for all $\theta \in [0, 1]$ but only for $\theta$ close to 1. (This may be an artifact of our analysis rather than an inherent limitation.)

**Theorem 18.** For $\epsilon > 0$, there are constants $c \geq 1$ and $\beta \in (0, 1)$ and an algorithm with runtime $O(n \frac{d_{\text{edit}}(x, y)}{\epsilon^2 \log d_{\text{edit}}(x, y) + n^{1-\beta}})$ with probability at least $1 - n^{-1}/n$.

**Acknowledgements.** We thank the FOCS 2018 committee and especially several anonymous referees for helpful comments and suggestions. Michael Saks thanks C. Seshadhri for insightful discussions on the edit distance problem.

**REFERENCES**


