Simple Low-Overhead Communication-Efficient String Reconciliation and Edit Distance

Michael T. Goodrich[1][0000−0002−8943−191X], Gonzalo Navarro[2][0000−0002−2286−741X], and Claire A. To[1][0009−0008−9102−2219]

[1] University of California, Irvine, Irvine CA, USA
[2] Dept. of Computer Science, University of Chile, Santiago, Chile

Abstract. Suppose two parties, Alice and Bob, hold long character strings, X and Y, respectively, and they are interested in determining how similar X and Y are. Moreover, they want to exchange the strings with cost proportional to their degree of dissimilarity. Such problems arise, for example, in database and file system synchronization operations, as well as in DNA sequence comparisons. Since the strings are long, we are interested in methods that are communication-efficient and have low overhead in terms of the computations that Alice and Bob must perform, when the strings are similar enough. In this paper, we provide simple low-overhead communication-efficient algorithms for such string reconciliation and edit distance problems. In the general case, we show how to determine the edit distance k between X and Y using only $O(k^2 \log n)$ bits of communication and optimal $O(n)$ time overhead, with high probability. For specialized cases, such as typical English text or DNA sequences, where we can make additional well-justified assumptions about the distribution of the input strings, we show how to achieve possibly better bounds, such as $O(k \log^2 n)$ bits of communication.

Keywords: distributed algorithms · randomized algorithms · edit distance · hash functions · communication complexity · string reconciliation

1 Introduction

A fundamental challenge in distributed computing lies in managing distributed data across communication networks. A problem arising frequently is that of checking or maintaining consistency across records in replicated databases, synchronizing files in peer-to-peer systems, and repairing or updating databases across the network. Thus, there is a need for efficient distributed protocols for reconciling character strings, which can, e.g., represent files, Web pages, or DNA sequences, avoiding the communication overhead of retransmitting the whole strings. As a concrete scenario, sequencing labs repeatedly generate genomes, and their clients then download them over the Internet to update genome libraries. In such scenarios, where clients already hold other genomes of the same species, which are known to be very similar to the new ones, clients and servers would like to minimize communication complexities as well as computational overheads.

In some cases, version control systems can keep track of which the differences are, or both parties can share a common string, so that they can easily exchange only the needed edits to reconcile the strings. The scenarios we have described, however, may lack such shared string or edit information, so the parties *know* that the differences between their strings are small, but do not know *which* they are. Still, they want to find those differences, and reconcile the strings, within a communication cost that depends only on the amount of differences.

1.1 Problem Statement

Let $X = x_1x_2 \ldots x_l$ and $Y = y_1y_2 \ldots y_n$ be two strings held by two parties, Alice and Bob, respectively, where each $x_i, y_j \in \Sigma$, for $i = 1, 2, \ldots, l$ and $j = 1, 2, \ldots, n$, and Σ is an alphabet whose size may or may not depend on l or n. In the case where $l = n$, the *Hamming distance*, $\mathrm{hd}(X, Y)$, between X and Y is the number of indices, i, where $x_i \neq y_i$, that is, the number of character substitutions needed to transform X into Y [11, 27]. A related notion that does not assume $l = n$ is the ***edit distance***, also known as *Levenshtein distance*, $L(X, Y)$, between X and Y: it is the minimum number character insertions, deletions, or substitutions needed to transform X into Y (or vice versa).

In this paper, we are interested in distributed string reconciliation algorithms with optimal (i.e., linear) computational overhead and small communication complexity for computing the edit distance between the two strings, X and Y, that are respectively held by Alice and Bob. This turns out to be equivalent to the problem of Alice and Bob exchanging their strings X and Y with communication cost based on their communicating a minimal set of edits between X and Y. We assume that Alice and Bob can communicate over some channel but they do not share any other computational resources, such as CPUs or memory. The ***communication complexity*** for a communication protocol for Alice and Bob is the number of bits that are communicated between them over the course of their distributed protocol, where we assume initially that neither party has knowledge of the other party's string.

In the ***general distributed edit distance*** problem, we are interested in a protocol for Alice and Bob to determine the edit distance, $k = L(X, Y)$, between X and Y using a small number of bits of communication (as a function of k). Alternatively, in ***specialized*** string reconciliation problems, we make additional assumptions regarding X and Y and their edit distance. For example, we develop efficient distributed protocols for the case where X and Y come from well-motivated real-world domains, such as English text and DNA sequences, which support assumptions regarding X and Y that we can exploit. Further, for practical reasons, in this paper we are interested in simple algorithms that have low communication complexity and computational overhead for Alice and Bob, for both the general and specialized versions of the string reconciliation problem.

Our results will hold ***with high probability*** (whp), meaning that they occur with probability $1 - 1/n^c$, for some constant $c \geq 1$. W.l.o.g. let us assume $l \leq n$.

1.2 Our Contributions

For the general distributed edit distance problem, we present a simple algorithm that achieves a communication complexity of $O(k^2 \log n)$ bits and optimal $O(n)$ time overhead, where k is the edit distance, with a probability of success over $1 - 1/n$. Our method is a communication-efficient adaptation of a dynamic-programming table "sliding" algorithm of Landau, Myers, and Schmidt [23] combined with Ukkonen's technique [36]. We avoid performing full sliding steps, however, as in these classical algorithms, since such steps would be communication inefficient. Instead, we perform what we call "fast sliding" steps by combining rolling hash functions and noisy binary searching.

We also show how to solve string reconciliation in the specialized context, as described. For such scenarios, we provide an algorithm that uses $O(n)$ time and a communication complexity of

$$O(k((d_X + d_Y + \log n) \log |\Sigma| + (\log \eta) \log n))$$

bits, succeeding with probability $1 - 1/\eta$, where k is an upper bound on the edit distance between X and Y known a priori, and d_X and d_Y are the lengths of the longest repeated substrings in X and Y, which are expected to be small for real-world applications and are $O(\log n)$ on widely used statistical models. Under such models, the communication complexity is $O(k \log^2 n)$ bits if $|\Sigma| = O(n^c)$ and the probability of error is $O(1/n^c)$ for any constant c. With $O(n \log k)$ time overhead we obtain the same communication cost, now with k being the actual edit distance, without need to know it a priori. Our method involves a non-trivial use of the invertible Bloom lookup table (IBLT) data structure [13, 15, 18].

2 Related Work

There has been considerable prior work on the general distributed edit distance problem, but we are not aware of any previous methods that are simple and communication-efficient while also having low computational overhead. For example, Orlitsky [28] gives a communication-optimal method that achieves $O(k \log n)$ communication complexity, where k is an upper bound on the edit distance between two strings of size $\Theta(n)$, but requires computational overhead that is $n^{O(k)}$, i.e., exponential in k. The computational overhead for such protocols has been subsequently improved [2, 8, 19, 20], but the best methods still require $O(n \operatorname{polylog}(n))$ overhead and are still fairly complicated. For instance, Chakraborty, Goldenberg, and Koucký [8] present a protocol with communication efficiency of $O(k^2 \log n)$ bits and $O(n \operatorname{polylog}(n))$ overhead by embedding strings in a Hamming space using random walks and doing their communication in this Hamming space. Belazzougui and Zhang [3] achieve a communication complexity of $O(k(\log^2 k + \log n))$ bits and $O(n \operatorname{polylog}(n))$ overhead assuming $k < n^{1/c}$ for a large constant c, but their method is quite complex and their bounds have large constant factors. In general, the previous methods for general distributed edit distance [2, 3, 8, 19, 20] all require

$O(n \operatorname{polylog}(n))$ overhead, or worse, which in practice is unacceptable for the long strings that arise in various real-world settings.

For specialized string reconciliation, Agarwal, Chauhan, and Trachtenberg [1] present distributed methods for reconciling typical real-world strings. Their method has communication complexity that scales linearly in the edit distance between the reconciling strings, as is also true for our methods. Their reconstruction methods are based on using masks and shingling to encode the strings and enumerating Eulerian paths to perform the reconciliation. As they admit, however, the number of such paths (and, hence, the computational burden of such algorithms) can be exponentially large, hence, they do not achieve low overhead. Like our approach, Kontorovich and Trachtenberg [22] reduce string reconciliation to the set reconciliation problem, but their approach still has suboptimal performance. Song and Trachtenberg [33] revisit the approach using masks and shingles with an improved Eulerian-path reconstruction method, which they show emprically improves over the prior work by Agarwal, Chauhan, and Trachtenberg [1], but the asymptotic overhead of their method is unfortunately still not efficient.

3 A General Algorithm for Edit Distance

In this section, we provide a general communication-efficient distributed algorithm for Alice and Bob to compute the edit distance of their respective strings, $X = x_1x_2 \dots x_l$ and $Y = y_1y_2 \dots y_n$, using $O(k^2 \log n)$ bits of communication and optimal $O(n)$ time overhead, where k is the edit distance between X and Y. Note that in this case we can also assume, without loss of generality, that l is $\Theta(n)$ and $k < \sqrt{n/\log_{|\Sigma|} n}$, since if this is not the case, Alice and Bob can simply exchange X and Y with each other, with $O(n \log |\Sigma|) \subseteq O(k^2 \log n)$ bits of communication, and each can privately compute $L(X, Y)$: Our algorithm can start with the $O(k^2 \log n)$-bits plan and switch to this second option right after having communicated $O(n \log |\Sigma|)$ bits. We first recall the classic sliding algorithm, then present our distributed version using $O(k^2 \log^2 n)$ bits of communication, and finally show how to reduce it to $O(k^2 \log n)$.

3.1 The Classic Sliding Algorithm

We first review a classic "sliding" algorithm to compute edit distance by Landau, Myers, and Schmidt [23], combined with Ukkonen's technique [36]. Let us begin describing the classic dynamic programming algorithm for computing edit distance [39]. Let $L(i, j)$ denote the edit distance for the prefixes $X_i = x_1x_2 \dots x_i$ of X and $Y_j = y_1y_2 \dots y_j$ of Y. Then, we have the boundary cases $L(i, 0) = i$ and $L(0, j) = j$, and, in general, the recurrence:

$$L(i,j) = \min \begin{cases} L(i-1,j)+1 & \text{(deletion)} \\ L(i,j-1)+1 & \text{(insertion)} \\ L(i-1,j-1)+\delta(x_i,y_j), & \end{cases}$$

where $\delta(x_i, y_j) = 0$ if $x_i = y_j$ (match) and $\delta(x_i, y_j) = 1$ if $x_i \neq y_j$ (substitution). This set of equations defines an $l \times n$ matrix, L, where $L(l, n)$ is the edit distance between X and Y, and the sequence of values that lead to this value can be viewed as a path that starts at $L(0, 0)$ and follows horizontally rightward, vertically downward, or down-right diagonal edges between adjacent cells. If $L(l, n) \leq k$, this path will be restricted to fall between the $j = i - k$ and $j = i + k$ diagonals of $L(i, j)$ entries [37], which we call the ***central band*** of L.

As noted by Landau *et al.* [23], most of the edges in the central band of L go between cells that have the same values and only $O(k^2)$ of these go between cells whose values differ, and these differ by 1. The key idea in their algorithm is to use suffix trees [24, 40] to quickly "slide" along diagonal paths where values in L are not changing.

For a diagonal, d, and row, i, let $\text{Slide}(d, i)$ denote the furthest row, $\rho \geq i$, that can be reached from row i on diagonal d with edit cost 0. That is,

$$\text{Slide}(d, i) = \max\{\rho \mid i \leq \rho \leq \min\{l, n - d\} \text{ and } x_{i+1} \ldots x_\rho = y_{i+d+1} \ldots y_{\rho+d}\}.$$

Note that $\text{Slide}(d, i)$ does not compare x_i and y_i. Also, note that $\text{Slide}(0, 0)$ is the length of the longest common prefix of X and Y. Further, if $x_{i+1} \neq y_{i+d+1}$, then $\text{Slide}(d, i) = i$, since the matching substrings of X and Y in the definition of $\text{Slide}(d, i)$ in this case are empty.

In addition, define $M^h(d)$ to be the maximum row index of an entry on diagonal d that has value h, and define the ***h-wave*** to be the set

$$M^h = \{M^h(-h), M^h(-h+1), \ldots, M^h(h-1), M^h(h)\}.$$

To handle boundary cases, we set $M^h(d) = \infty$ if $M^h(d)$ does not actually exist.

At a high level, the algorithm computes the h-waves, for $h = 0, 1, 2, \ldots$, until a wave h' is computed for which $M^{h'}(n - l) = l$, in which case the algorithm outputs h' as the edit distance. The "inner loop" formula for computing the h-wave from the $(h-1)$-wave, for $h > 0$, is the following [23]:

$$M^h(d) = \text{Slide}(d, i_{\max}(d, h)),$$

where

$$i_{\max}(d, h) = \max \begin{cases} M^{h-1}(d+1) + 1 & \text{if } d < h - 1 \\ M^{h-1}(d) + 1 & \text{if } -h < d < h \\ M^{h-1}(d-1) & \text{if } d > -h + 1. \end{cases}$$

Since there are only $O(k)$ entries from the matrix L in each h-wave that we need to consider and only k h-waves that we need to consider, this implies that we just need to perform $O(k^2)$ slide computations to determine the edit distance.

A problem is that we do not know k in advance. This is solved by Ukkonen [36] as follows: we first compute the 0-wave, for which we need only the diagonal 0; we then compute the 1-wave, for which we need only the diagonals -1, 0, and 1. In general, we need only the diagonals $-h$ to h to compute the h-wave. We

will then compute $O(k^2)$ entries $M^h(d)$ in order to find the edit distance $k = h'$. By computing each entry in $O(1)$ time using the suffix trees of X and Y, it is possible to compute the edit distance in time $O(k^2)$.

3.2 Our Fast-Sliding Algorithm

Our algorithm, which we call the "fast-sliding algorithm," is a communication-efficient adaptation of the sliding algorithm; see also, e.g., [8]. In the distributed setting, we cannot afford to use suffix trees as done by Landau *et al.*, as exchanging them would be too costly. Our method instead uses a Karp–Rabin rolling hash function, rh (see Appendix B). In particular, to compute Slide(d, i), in the general case, we need to find the maximum index, ρ, such that

$$x_{i+1} \dots x_\rho = y_{i+d+1} \dots y_{\rho+d}.$$

We do this by a doubling-binary search [4] starting from the index $i + 1$ in X and index $i + d + 1$ in Y, where we communicate and compare $\mathrm{rh}(x_{i+1} \dots x_t)$ and $\mathrm{rh}(y_{i+d+1} \dots y_{t+d})$, for $t > i + 1$.

We choose the parameters for rh in this case so that the hash values have at least $3 \log n$ bits, so that with probability at least $1 - n^{-3}$ each test during a doubling-binary search will successfully determine whether $x_{i+1} \dots x_t = y_{i+d+1} \dots y_{t+d}$ just by comparing the respective rolling-hash values. Since each doubling-binary search requires $O(\log n)$ such tests, this implies that the entire algorithm has a communication complexity of $O(k^2 \log^2 n)$ bits and succeeds whp, by a union bound.

3.3 Improving the Communication Complexity

We can improve the communication complexity or our fast-sliding algorithm by replacing the doubling-binary search with a communication-efficient "noisy" doubling-binary search, similar to a technique used by Viola [38]. In particular, we expand rh to have $24 \log n$ bits, but rather than using all $24 \log n$ bits of the rh function with each test in our doubling-binary search, we instead divide these bits into groups of 4 bits each and just use the corresponding sets of bits in each comparison of a doubling-binary search. That is, for the first comparison in such a doubling-binary search, we use the first 4 bits, in the second we use the second 4 bits, and so on, so each comparison succeeds with probability at least $15/16$ and the result of each comparison is independent of any other. The resulting version may require some backtracking, but uses only $6 \log n$ tests to obtain a success probability of $1 - n^{-3}$, see, e.g., [12, 14, 16]. Thus, we have the following.

Theorem 1. *Let X and Y be strings of length l and n, respectively, with $l \leq n$, from a finite alphabet, stored with separate parties. Then we can compute the edit distance k between X and Y with an overhead of $O(n)$ computational operations for the two parties and a communication complexity of $O(k^2 \log n)$ bits whp.*

Note that Alice and Bob not only discover the edit distance; they can reconstruct the path in their matrix L that leads to such distance. This allows each party reconstruct the other string, with $O(n)$ additional time and space [25].

4 A Specialized Algorithm based on IBLT Chunking

In many applications, the strings stored by Alice and Bob come from input distributions that follow widely accepted statistical models [10, 32], as is the case of DNA and English text. Those models imply that all strings over some short length are unique whp.[3] In such cases, and if we know a priori an upper bound k on the edit distance, we are able to possibly improve the communication complexity of the fast-sliding algorithm of the previous section, to $O(k \log^2 n)$ bits. Concretely, we devise an efficient string reconciliation algorithm that is based on ***content-defined chunking*** [9, 41]. We refer to this as our ***IBLT-chunking*** algorithm, since it also uses an invertible Bloom lookup table (IBLT) [18]. This data structure is also used by Eppstein *et al.* [15] to efficiently compute Hamming distance, but our usage is different. We first give some background on IBLTs and then describe our algorithm.

4.1 Invertible Bloom Lookup Table (IBLT)

IBLT [13,18] is a space-efficient probabilistic data structure, based on Bloom filters, that supports various operations implemented by the XOR primitive. It can be used to solve the *set reconciliation* problem [15], which is the task of synchronizing two sets stored at two nodes across a communication link. IBLTs allow the parties to efficiently identify the items that are unique to each set so that both parties can update their local copies and obtain the other set as well. The main step in this process is to compute a *set difference*, finding the set of elements that one party possesses but the other does not.

Suppose we are given a set $S = \{s_1, s_2, \ldots, s_n\}$, which we want to store probabilistically in an IBLT table, T, with m cells. We assume that we have an ***encoding function***, encode, which encodes each element, s_i, into an associated unique key, e_i. We also have λ hash functions that map any key to λ distinct locations, and we let HashToIndices(e_i, λ, m) denote the set of λ locations determined by these hash functions for the key e_i. Each IBLT cell, $T[j]$, stores a `(keysum, hashsum)` pair such that `keysum` contains the bitwise-XOR (denoted $\oplus$) of all encoded keys e_i that are mapped to $T[j]$ by one of T's hash functions; `hashsum` contains the XOR of all fingerprints or hashes $H(e_i)$, where $H(e_i)$ is a b-bit secondary hash of e_i, such as a checksum, or cryptographic hash, used to verify or detect errors in the decoding process.

The IBLT supports both insertion and deletion of elements using XOR operations. Inserting an element corresponds to XOR-ing its encoded value into the table (at the λ locations), while deleting it corresponds to *the same* operations. To encode an entire set, S, we simply insert each of its elements.

For a set reconciliation protocol, Alice and Bob build respective tables, T_A and T_B, by inserting all the elements of their sets, X and Y, respectively, into

[3] Note that this is not the case in some applications, for example those where repetitive sequences are handled [26].

Algorithm 1 IBLT Subtract ($T = T_A \oplus T_B$)

```
for i = 0 to m − 1 do
    T[i].keysum ← T_A[i].keysum ⊕ T_B[i].keysum
    T[i].hashsum ← T_A[i].hashsum ⊕ T_B[i].hashsum
```

Algorithm 2 IBLT Decode T

```
while ∃ i, s.t. H(T[i].keysum) = T[i].hashsum do
    s ← decode(T[i].keysum)
    Output s
    delete(T, s)
for i = 0 to m − 1 do
    if T[j].keysum ≠ 0 OR T[j].hashsum ≠ 0 then
        return FAIL
return SUCCESS
```

initially zeroed IBLTs of the same size using the same hash functions, and exchange the tables. From these two IBLTs, T_A and T_B, both Alice and Bob can compute an IBLT representing the symmetric difference between X and Y. This simple method is shown in Algorithm 1. The fact that this algorithm computes an IBLT representation of the symmetric difference of X and Y follows immediately from the fact that $x \oplus x = 0$ for any value x.

Listing Set Entries If an IBLT is not "too" full, we can list out its entries. The listing of entries or decoding is a destructive $O(m)$-time procedure that recovers all keys in the IBLT or reports that the IBLT is too full to achieve this successfully. Say that a cell $T[i]$ is ***pure*** if

$$H(T[i].\texttt{keysum}) = T[i].\texttt{hashsum},$$

which clearly holds for a cell holding exactly one key but could admittedly also hold if there is a checksum collision. Note that if a cell is pure, we can recover its key by extracting the value from $T[i].\texttt{keysum}$ and then deleting the element from the table, which is also referred to as ***peeling*** they key from the corresponding λ hashed cells. This may make other cells pure, which can eventually allow us to recover all the values stored in the IBLT. Algorithm 2 shows this process.

The following results are proved in Appendix A.

Lemma 1. *The probability that a cell containing $t \geq 2$ keys produces a false positive on the purity test is 2^{-b}.*

Corollary 1. *In an IBLT of m cells, the probability that the purity test yields a false positive for any cell is at most $m2^{-b}$.*

In our setting, we want to recover a symmetric difference of size at most d with high probability, even if d is a constant. We use an IBLT of size $m = c \cdot d \log n$ for a $c \geq 1$, and λ being $\Theta(\log n)$, so that by an analysis similar to that of Goodrich,

Kitagawa, and Mitzenmacher [17], when the current number of elements in the IBLT is at most d, we can successfully list all entries of T whp:

Theorem 2. *Let T be an IBLT with $m = cd \log n$ cells and $\lambda = (c/2) \log n$ hash functions, for $c \geq 4$. If T stores at most d elements, then it can be successfully decoded by the peeling algorithm with probability at least $1 - 1/n^{c/2-1}$.*

To sum up, suppose two sets, X and Y, of size $O(n)$, are held by Alice and Bob respectively, which are assumed to differ by at most d elements. Each party independently constructs an IBLT, T_A and T_B of $m = O(d \log n)$ cells, and inserts all elements into the table. Even if either T_A and T_B is too full to decode individually, there will be pure cells in $T_A \oplus T_B$, i.e., containing only one key. That is, with properly chosen parameters, Alice and Bob can reconcile their sets whp, in $O(n)$ time and $O(d\beta \log n)$ bits of communication, where $\beta > 0$ is the size of the keys, by exchanging their tables T_A and T_B and performing the decoding algorithm on $T_A \oplus T_B$.

4.2 Our Algorithm

Content-defined chunking (CDC) [9, 41] divides a string, $X = x_1 x_2 \ldots x_n$, into a set, $\mathcal{C}$, of contiguous substrings, called ***chunks***, X_1, X_2, ..., X_r, so that $X = X_1 || X_2 || \cdots || X_r$, where $||$ denotes concatenation. This is achieved by defining a minimum-size parameter, s, and a windowed rolling hash function, h, with a window size w, and subdividing the string, X, by a simple algorithm that computes h for each substring of X of length w, starting from the beginning of X. At each position where $h(x_i \ldots x_{i+w-1}) = 0$, if the length of the current substring being processed is at least s, then we cut X at the index $i + w - 1$ to form the next chunk.

In the edit distance algorithm of this section, we desire that the chunks in $\mathcal{C}$ are unique and not too long, which we can characterize in terms of a parameter d_X: the length of the longest repeated string in X. Note that d_X is computed in linear time using suffix trees. The following observation should be obvious.

Lemma 2. *Let $\mathcal{C} = \{X_1, X_2, \ldots, X_r\}$ be a set of chunks for the string $X = X_1 || X_2 || \cdots || X_r$. If $|X_i| > d_X$ for $i = 1, 2, \ldots, r$, then each chunk in $\mathcal{C}$ is unique.*

Szpankowski [34] analyzes the value of d_X for strings X that can be modeled as pseudo-random sequences, such as those determined by Bernoulli or Markov processes, and shows that in such cases d_X is $O(\log |X|)$ whp. Note that d_X is always at least $\log_{|\Sigma|} |X|$. Appendix C discusses practical values.

Before running the string reconciliation protocol, Alice and Bob first compute and exchange d_X and d_Y, so that both use the same minimum chunk size, $s = \max\{d_X, d_Y\} + 1$, and the same CDC window size, $w = s$, for their respective chunking processes. We also choose a confidence bound, $0 < \eta \leq n^c$, for some constant $c \geq 1$, where we desire our protocol to succeed with probability at least $1 - 1/\eta$ (i.e., whp).

Lemma 3. *With high probability, the maximum chunk size in $\mathcal{C}(X)$ (resp., $\mathcal{C}(Y)$) is $O(s + \log|X|)$ (resp., $O(s + \log|Y|)$).*

Proof: W.l.o.g., let us focus on Alice's set of chunks, $\mathcal{C}(X)$. By definition, every substring of X with length at least s is guaranteed to be unique. Suppose that the uniformly random rolling hash function h outputs b bits. The probability of a cut at any window w is $p = Pr[h(x_i \dots x_{i+w-1}) = 0] = 1/2^b$. Suppose a chunk of length $l > w$ starts at position i. Then, this implies that every chunk window of size w with starting positions in range $[i, i + l - w]$ did not produce a cut. There are $t = l - w + 1$ distinct windows. The probability no cut occurs in t distinct consecutive windows is $(1-p)^t \leq e^{-pt}$. Let $t = (c/p)\log|X|$ for a constant $c \geq 2$, then $Pr[l > w + t] \leq e^{-pt} = e^{-c\log|X|} = 1/|X|^c$. By a union bound over all $|X|$ possible starting positions, the probability any chunk exceeds $w + t$ is at most $1/|X|^{c-1}$. Since $w = s$ and $t = O(\log|X|)$ for constant hash output size b (which makes p constant), with high success probability, at least $1 - 1/|X|^{c-1}$ where $c \geq 2$, the maximum chunk size is $O(s + \log|X|)$. ■

Let us further assume that we have an upper-bound estimate, $k \geq 1$, for the edit distance between the strings, X and Y, held respectively by Alice and Bob (we later lift this assumption). Then we create an IBLT of appropriate size to allow listing its elements with probability at least $1 - 1/\eta$, for our desired confidence parameter, $0 < \eta \leq n^c$, if there are most k elements, i.e., an IBLT of $O(k \log \eta)$ cells, with each cell storing a `keysum` and `hashsum` of $\Theta(\log n)$ bits. Alice then stores a suitably defined set, $S(X)$, in her IBLT of this chosen size such that, given a string, X, and a set of chunks, $\mathcal{C}(X) = \{X_1, X_2, \dots, X_r\}$, for X, where $X = X_1||X_2||\cdots||X_r$, her IBLT is defined with respect to the set $S(X)$ defined for her chunks. In particular, she encodes $X = x_1 x_2 \dots x_n$ as a set of pairs of checksum hashes of consecutive chunks,

$$S(X) = \{H(X_i)||H(X_{i+1}) \mid X_i \in \mathcal{C}(X), i = 1, \dots, r-1\},$$

where $H(X')$ is a checksum hash with $\Theta(\log n)$ bits for the chunk X', so that, whp, all the checksum hashes are unique, given that all the chunks are unique. Likewise, Bob encodes a similar set of pairs of chunk checksums, $S(Y)$, and stores these in a similarly-defined IBLT of the same size. Alice and Bob then exchange their IBLTs and perform the peeling algorithm to find the set difference.

For each differing chunk pair that either Alice or Bob decodes, they can determine if it is one of their chunk pairs or a chunk pair from the other party. Further, after Alice and Bob have respectively decoded the set-difference IBLT and determined which chunk pairs belong to X and which belong to Y, Alice and Bob then each send the other their corresponding chunk pairs. That is, for each chunk-checksum pair, $H(X_i)||H(X_{i+1})$, of hers that Alice decodes from the IBLT as being a part of the symmetric difference, Alice sends Bob the pair (X_i, X_{i+1}). Likewise, Bob sends a similar set of chunk pairs to Alice.

Lemma 4. *Let k be an upper bound on the edit distance between two strings, X and Y, each of length $\Theta(n)$, defined over an alphabet Σ, and let d_X and d_Y*

be the length of their longest repeated substring, respectively. Then our IBLT-chunking algorithm finds the set of $O(k)$ differing chunk pairs between X and Y with probability at least $1 - 1/\eta$, for a desired confidence bound, $0 < \eta \leq n^c$, for a constant $c \geq 1$, using $O(k((d_X + d_Y + \log n) \log |\Sigma| + (\log \eta) \log n))$ bits of communication.

Proof: By the above discussion and Lemma 3, each chunk will have size $O((d_X + d_Y + \log n) \log |\Sigma|)$ bits whp and the IBLT needs to have $O(k \log \eta)$ cells of $O(\log n)$ bits each in order to decode with probability at least $1 - 1/\eta$, assuming the symmetric difference has size $O(k)$.

To see that this is so note that, since any edit between X and Y will change at most $w + 1$ consecutive rolling hashes, and since $s = w$, those positions span at most two consecutive chunks, so at most two consecutive chunks can differ for each edit. Therefore, each edit of a string can change at most $O(1)$ chunks when doing CDC chunking (i.e., chunking quickly resynchronizes after an edit). Thus, if k is an upper bound on the edit distance between X and Y, then at most $O(k)$ chunks are different between $\mathcal{C}(X)$ and $\mathcal{C}(Y)$; hence, at most $O(k)$ chunk-checksum pairs are different between $S(X)$ and $S(Y)$ whp. Thus, after the decoding of the IBLT occurs, Alice and Bob each send the other lists of $O(k)$ chunk pairs, each of at most $O((d_X + d_Y + \log n) \log |\Sigma|)$ bits. ■

In addition, the information exchanged is sufficient for each party to reconstruct the other party's string. W.l.o.g., let us focus on Bob, who wishes to reconstruct X by "linking" all of Alice's chunk pairs. Also, with probability at least $1 - 1/\eta$, we can assume all the elements in the symmetric difference were decoded correctly and that all the chunk checksums are unique. For instance, the chunks in $\mathcal{C}(X)$ (resp., $\mathcal{C}(Y)$) are unique by construction; hence, the chunk-checksum pairs in $S(X)$ and $S(Y)$ are also respectively unique whp. Further, assume for simplicity that Alice always sends Bob her first chunk pair (X_1, X_2), so Bob can determine if their first chunk pairs match or not. In either case, he can determine the first two chunks in $\mathcal{C}(X)$. Also, note that Bob has received from Alice each chunk pair, (X_j, X_{j+1}), such that $H(X_j)||H(X_{j+1})$ is a chunk-checksum pair in the symmetric difference of $S(X)$ and $S(Y)$ decoded from the IBLT. Suppose (X_{i-1}, X_i) is the last pair of chunks in $\mathcal{C}(X)$ that Bob has discovered; hence, he is interested in discovering the next pair, (X_i, X_{i+1}).

- Case 1. There is a chunk pair, $H(X_i)||H(X')$, in the IBLT symmetric difference. Since chunks are unique, this implies $(X_i, X_{i+1}) = (X_i, X')$. Moreover, in this case, Alice has sent Bob the pair, (X_i, X'); hence, Bob can determine (X_i, X_{i+1}).
- Case 2. There is no chunk pair, $H(X_i)||H(X')$, in the IBLT symmetric difference. Then there is a chunk-checksum pair, $H(X_i)||H(X')$, that is also in $S(Y)$; i.e., $(X_i, X') = (Y_j, Y_{j+1})$, for some $j > 1$. Since chunks are unique, this implies $(X_i, X_{i+1}) = (Y_j, Y_{j+1})$; hence, Bob can determine (X_i, X_{i+1}) by looking up his chunk, Y_{j+1}, given that he has already determined that X_i matches Y_j.

Corollary 2. *The exchanged information is sufficient for Bob (resp., Alice) to reconstruct X (resp., Y) with probability at least $1 - 1/\eta$.*

Therefore, we have the following result, which yields $O(k \log^2 n)$-bit communication complexity in the discussed models where d_X and d_Y are $O(\log n)$.

Theorem 3. *Let k be an a-priori-known upper bound on the edit distance between two strings, X and Y, of length $\Theta(n)$, each defined over an alphabet Σ, held by two separate parties, Alice and Bob, and let d_X and d_Y be the lengths of their longest common substrings. Then our IBLT-chunking algorithm allows Alice and Bob to determine the other party's string in $O(n)$ time using*

$$O(k((d_X + d_Y + \log n) \log |\Sigma| + (\log \eta) \log n))$$

bits of communication with probability at least $1 - 1/\eta$.

Once both Alice and Bob have the chunks of both strings, they can compute their edit distance in $O(n+k^2)$ additional time [23]. Thus, when $k = O(\sqrt{n})$ and d_X and d_Y are $O(\log n)$, we can also compute the precise edit distance within this cost, which may outperform our general fast-sliding algorithm.

To conclude, we note that we can obtain a similar result *without* knowing an *a priori* bound k on the edit distance, by slightly increasing the time overhead.

Corollary 3. *The same result of Theorem 3 can be obtained if k is the actual edit distance between X and Y, a priori unknown, with $O(n \log k)$ time overhead.*

Proof: Apply Theorem 3 with upper bound $k' = 1, 2, 4, 8, \ldots$ until it succeeds. With probability at least $1 - 1/\eta$, the try with $k \leq k' < 2k$ will succeed, and each party will obtain the other's string. The total communication cost is as in Theorem 3, replacing k by $\sum_{j=0}^{\lceil \log_2 k \rceil} 2^j = O(k)$, and the time overhead is $O(n \log k)$ because we process the chunks $O(\log k)$ times, once per value of k'. ■

5 Conclusions

We have introduced simple low-overhead communication efficient algorithms for string reconciliation between two parties that separately store two strings of size $\Theta(n)$. Our general method, the fast-sliding algorithm, runs with linear time complexity for the parties and exchange $O(k^2 \log n)$ bits whp, where k is the edit distance between both strings. We also gave an IBLT-based method based on content-defined chunking that also has linear time complexity for the parties and, for typical real-world strings whose longest repeated substrings are of length $O(\log n)$, exchange $O(k \log^2 n)$ bits whp.

Directions for future work include exploring efficient low-overhead ways of performing string reconciliation for compressed strings, particularly in the scenario of highly repetitive string collections, which do not follow typical statistical models but can be sharply compressed with dictionary methods. We will also consider other similarity models beyond edit distance, like allowing substring reversals and block moves.

References

1. S. Agarwal, V. Chauhan, and A. Trachtenberg, "Bandwidth efficient string reconciliation using puzzles," *IEEE Transactions on Parallel and Distributed Systems*, vol. 17, no. 11, pp. 1217–1225, 2006.
2. D. Belazzougui, "Efficient deterministic single round document exchange for edit distance," *arXiv preprint arXiv:1511.09229*, 2015.
3. D. Belazzougui and Q. Zhang, "Edit distance: Sketching, streaming, and document exchange," in *IEEE Symp. on Foundations of Computer Science (FOCS)*, 2016, pp. 51–60.
4. J. L. Bentley and A. C.-C. Yao, "An almost optimal algorithm for unbounded searching," *Information Processing Letters*, vol. 5, no. 3, pp. 82–87, 1976.
5. T. Brants and A. Franz, "Web 1t 5-gram version 1," Web Download, LDC2006T13, Linguistic Data Consortium, Philadelphia, 2006, accessed via Linguistic Data Consortium.
6. C. Buck, K. Heafield, and B. Ooyen, "N-gram counts and language models from the common crawl," in *Proceedings of the Language Resources and Evaluation Conf. 2014*, May 2014, pp. 3579–3584.
7. Y. Bussi, R. Kapon, and Z. Reich, "Large-scale k-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy," *PloS one*, vol. 16, no. 10, p. e0258693, 2021.
8. D. Chakraborty, E. Goldenberg, and M. Koucký, "Streaming algorithms for embedding and computing edit distance in the low distance regime," in *ACM Symp. on Theory of Computing (STOC)*, 2016, pp. 712–725.
9. B. Chapuis, B. Garbinato, and P. Andritsos, "Throughput: A key performance measure of content-defined chunking algorithms," in *IEEE 36th Int. Conf. on Distributed Computing Systems Workshops (ICDCSW)*, 2016, pp. 7–12.
10. T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.
11. M. Crochemore, C. Hancart, and T. Lecroq, *Algorithms on Strings*. Cambridge University Press, 2007.
12. D. Dereniowski, A. Łukasiewicz, and P. Uznański, "Noisy (binary) searching: Simple, fast and correct," in *42nd Symp. on Theoretical Aspects of Computer Science (STACS)*, 2025, pp. 29–1.
13. D. Eppstein and M. T. Goodrich, "Straggler identification in round-trip data streams via newton's identities and invertible Bloom filters," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 2, pp. 297–306, 2010.
14. D. Eppstein, M. T. Goodrich, and V. Sridhar, "Computational geometry with probabilistically noisy primitive operations," *arXiv preprint arXiv:2501.07707*, 2025.
15. D. Eppstein, M. T. Goodrich, F. Uyeda, and G. Varghese, "What's the difference? efficient set reconciliation without prior context," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, p. 218–229, Aug. 2011.
16. U. Feige, P. Raghavan, D. Peleg, and E. Upfal, "Computing with noisy information," *SIAM J. on Computing*, vol. 23, no. 5, pp. 1001–1018, 1994.
17. M. T. Goodrich, R. Kitagawa, and M. Mitzenmacher, "Parallel peeling of invertible Bloom lookup tables in a constant number of rounds," in *Int. Conf. on Current Trends in Theory and Practice of Computer Science*. Springer, 2025, pp. 70–84.
18. M. T. Goodrich and M. Mitzenmacher, "Invertible Bloom lookup tables," 2015. [Online]. Available: https://arxiv.org/abs/1101.2245

19. U. Irmak, S. Mihaylov, and T. Suel, "Improved single-round protocols for remote file synchronization," in *24th Joint Conf. of the IEEE Computer and Communications Societies*, vol. 3, 2005, pp. 1665–1676.
20. H. Jowhari, "Efficient communication protocols for deciding edit distance," in *European Symp. on Algorithms (ESA)*, 2012, pp. 648–658.
21. R. M. Karp and M. O. Rabin, "Efficient randomized pattern-matching algorithms," *IBM Journal of Research and Development*, vol. 31, no. 2, pp. 249–260, 1987.
22. A. Kontorovich and A. Trachtenberg, "String reconciliation with unknown edit distance," in *2012 IEEE Int. Symp. on Information Theory Proceedings*, 2012, pp. 2751–2755.
23. G. M. Landau, E. W. Myers, and J. P. Schmidt, "Incremental string comparison," *SIAM J. on Computing*, vol. 27, no. 2, pp. 557–582, 1998.
24. E. McCreight, "A space-economical suffix tree construction algorithm," *Journal of the ACM*, vol. 23, no. 2, pp. 262–272, 1976.
25. E. W. Myers, "An $O(ND)$ difference algorithm and its variations," vol. 1, pp. 251–266, 1986.
26. G. Navarro, "Indexing highly repetitive string collections, part I: Repetitiveness measures," *ACM Computing Surveys*, vol. 54, no. 2, p. article 29, 2021.
27. ——, "A guided tour to approximate string matching," *ACM Computing Surveys*, vol. 33, no. 1, pp. 31–88, 2001.
28. A. Orlitsky, "Interactive communication: Balanced distributions, correlated files, and average-case complexity," in *IEEE Symp. on Foundations of Computer Science (FOCS)*, 1991, pp. 228–238.
29. A. J. Ponsero, M. Miller, and B. L. Hurwitz, "Comparison of k-mer-based de novo comparative metagenomic tools and approaches," *Microbiome research reports*, vol. 2, no. 4, p. 27, 2023.
30. A. Rahman, I. Hallgrímsdóttir, M. Eisen, and L. Pachter, "Association mapping from sequencing reads using k-mers," *Elife*, vol. 7, p. e32920, 2018.
31. A. Shajii, D. Yorukoglu, Y. William Yu, and B. Berger, "Fast genotyping of known snps through approximate k-mer matching," *Bioinformatics*, vol. 32, no. 17, pp. i538–i544, 2016.
32. C. E. Shannon, "A mathematical theory of communication," *Bell Systems Technical Journal*, vol. 27, pp. 398–403, 1948.
33. B. Song and A. Trachtenberg, "Scalable string reconciliation by recursive content-dependent shingling," in *57th Allerton Conf. on Communication, Control, and Computing (Allerton)*, 2019, pp. 623–630.
34. W. Szpankowski, "A generalized suffix tree and its (un)expected asymptotic behaviors," *SIAM J. on Computing*, vol. 22, no. 6, pp. 1176–1198, 1993.
35. Trustees of Indiana University, "How to recognize plagiarism: Tutorials and tests," https://plagiarism.tedfrick.me/IUcriteria.html, 2025.
36. E. Ukkonen, "Algorithms for approximate string matching," *Information and Control*, vol. 64, no. 1-3, pp. 100–118, 1985.
37. ——, "Approximate string-matching over suffix trees," in *Combinatorial Pattern Matching*, A. Apostolico, M. Crochemore, Z. Galil, and U. Manber, Eds. Springer, 1993, pp. 228–242.
38. E. Viola, "The communication complexity of addition," *Combinatorica*, vol. 35, no. 6, pp. 703–747, 2015.
39. R. A. Wagner and M. J. Fischer, "The string-to-string correction problem," *Journal of the ACM*, vol. 21, no. 1, pp. 168–173, 1974.
40. P. Weiner, "Linear pattern matching algorithm," in *Proc. 14th Annual IEEE Symposium on Switching and Automata Theory*, 1973, pp. 1–11.

41. W. Xia, X. Zou, H. Jiang, Y. Zhou, C. Liu, D. Feng, Y. Hua, Y. Hu, and Y. Zhang, "The design of fast content-defined chunking for data deduplication based storage systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 9, pp. 2017–2031, 2020.
42. J. Zentgraf and S. Rahmann, "Swiftly identifying strongly unique k-mers," in *24th Workshop on Algorithms in Bioinformatics (WABI)*, ser. LIPIcs, G. Kucherov and V. Miele, Eds., vol. 312, 2024, pp. 15:1–15:15.

A IBLT Proofs

Proof: ***(of Lemma 1)*** Let H be a uniform random hash function that produces b-bit hash values. Suppose a cell contains $t \geq 2$ keys, with $\texttt{keysum} = x_1 \oplus x_2 \oplus \ldots \oplus x_t$ and $\texttt{hashsum} = H(x_1) \oplus H(x_2) \oplus \ldots \oplus H(x_t)$. Define a random variable $Z = H(\texttt{keysum}) \oplus \texttt{hashsum} = H(x_1 \oplus \ldots \oplus x_t) \oplus H(x_1) \oplus \ldots \oplus H(x_t)$. Because H is uniform and random, $H(x_1 \oplus \cdots \oplus x_t)$ is independent of $H(x_1), \ldots, H(x_t)$ and each value is uniformly distributed over $\{0,1\}^b$. Thus, the probability of failure is 2^{-b}. ■

Proof: ***(of Corollary 1)*** Apply the union bound to the failure probability from Lemma 1. ■

Proof: ***(of Theorem 2)*** Let X_i be a random variable that is 1 if the i-th element has no pure cell and is 0 otherwise. Since there are at most d elements in T and it has $m = cd \log n$ cells and $\lambda = (c/2) \log n$ hash functions, at least half the cells in the IBLT are empty. Thus, any one of the λ hash functions will map an element to what would otherwise be an empty cell with probability at least $1/2$. Since the λ hash functions are independent, this implies that

$$\Pr(X_i = 1) \leq \frac{1}{2^\lambda} = \frac{1}{n^{c/2}}.$$

Thus, by a union bound, the probability that every element in T has a pure cell is at least $1 - 1/n^{c/2-1}$. ■

B Hash Functions

In this paper, we assume that hash functions are pseudo-random and can be considered as random functions for the sake of analysis. One particular type of hash function we use is a ***rolling hash function***, which can be computed for substrings of a string, $X = x_1 x_2 \ldots x_l$. In particular, the well-known Rabin-Karp string pattern matching algorithm [21] uses the following rolling hash function, for hashing the substring, $x_i x_{i+1} \ldots x_j$, in which we view each x_i as an integer:

$$\mathrm{rh}(x_i x_{i+1} \ldots x_j) = x_i b^{j-i} + x_{i+1} b^{j-i-1} + \cdots + x_j b^0.$$

Here, all arithmetic is done on a suitably chosen finite field (e.g., modulo some prime p) and b is a randomly chosen generator for that field. We also use this

rolling hash function, but rather than using it only for fixed-length substrings, as in the Rabin-Karp algorithm, we use it in some cases for fixed-length substrings and other times for arbitrarily long substrings. One useful property of this rolling hash function is that we can compute the rolling hash for every prefix of a length-l string, X, in $O(l)$ time using the following inductive formula for each prefix:

$$\text{rh}(x_1 x_2 \ldots x_i) = \text{rh}(x_1 x_2 \ldots x_{i-1})b + x_i.$$

Then, if we store these hash values for every prefix of X, we can compute the value of $\text{rh}(x_i x_{i+1} \ldots x_j)$ for any substring, $x_i x_{i+1} \ldots x_j$, in constant time as follows:

$$\text{rh}(x_i x_{i+1} \ldots x_j) = \text{rh}(x_1 x_2 \ldots x_j) - \text{rh}(x_1 x_2 \ldots x_{i-1})b^{j-i+1}.$$

Alternatively, if we are interested in computing rolling hash values of substrings that have a fixed-length, r, then we can use the following formula to compute the next hash value from the previous one in constant time, assuming we have stored the value of b^r:

$$\text{rh}(x_i x_{i+1} \ldots x_{i+r-1}) = \text{rh}(x_{i-1} x_i \ldots x_{i+r-2})b + x_{i+r-1} - x_{i-1}b^r.$$

C Lengths of Unique q-grams in Practice

In practice, many studies involving DNA strings select q in the range of 20-40 base pairs to achieve q-grams (aka q-mers) that are mostly unique [29]. This range balances the high probability of representing unique genomic sequences while reducing potential sequencing errors [30]. For example, $q = 32$ captures 85.7% of unique sequences in the human genome [31], $q = 25$ suffices for majority of q-mers to be unique in mammalian genomes, many of which are already unique for 19-mers [42], and $q = 21$ distinguishes many eukaryote, bacteria, and archea species [7]. For natural language, there is no formal source guaranteeing an exact number of words or characters needed for uniqueness. However, analyses of large corpora [5, 6] show that 5-grams are sufficient for most such sequences of words to be unique, supporting the use of 4-6 word sequences in practical NLP applications. For example, the Indiana University test for word-for-word plagiarism is a matching sequence of seven or more words between two documents [35]. In both applications, the choice of q-gram length is related to the depth, $d(T_X)$, in practice.