Reductions from Tree Reconstruction to String Reconstruction

Thomas Jacob Maranzatto

University of Illinois Chicago (Ph.D.)
University of Maryland College Park

July 2024

Outline

- Some Background/ Definitions
- Combinatorics of the Deletion Channel
- Tree Reconstruction Reductions

Deletion Channel Models

- A trace of a string s is obtained by sending the string through a deletion channel that removes each bit of s independently with probability q. Each trace is generated independently of other traces.

Deletion Channel Models

- A trace of a string s is obtained by sending the string through a deletion channel that removes each bit of s independently with probability q. Each trace is generated independently of other traces.
- For any tree R, in the *tree edit distance* (TED) model when a node w is removed, the children of w become the children of w's parent. A trace is obtained by removing each node with probability q (the order of removal does not matter).

TED Illustration

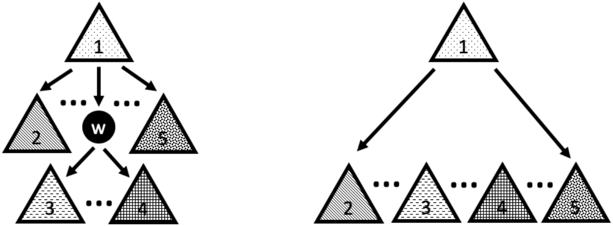


Figure: The generic picture before (left) and after (right) node w is removed from tree R in the TED model. The subtrees 3 and 4 are inserted as children of 1 when w is removed.

Reconstruction Problems

- Given an arbitrary string s of length n, the *string trace reconstruction problem* is to recover s with probability at least $1 - \delta$ using as few traces as possible

Reconstruction Problems

- Given an arbitrary string s of length n, the *string trace reconstruction problem* is to recover s with probability at least $1 - \delta$ using as few traces as possible
- Let $T(n, q, \delta)$ be the number of traces needed to reconstruct an arbitrary string s on n bits with deletion rate q and confidence δ

Reconstruction Problems

- Given an arbitrary string s of length n, the *string trace reconstruction problem* is to recover s with probability at least $1 - \delta$ using as few traces as possible
- Let $T(n, q, \delta)$ be the number of traces needed to reconstruct an arbitrary string s on n bits with deletion rate q and confidence δ
- Given an arbitrary n node tree R, the *tree trace reconstruction problem* is to reconstruct R with probability at least $1 - \delta$ using as few TED traces as possible.

Reconstruction Problems

- Given an arbitrary string s of length n, the *string trace reconstruction problem* is to recover s with probability at least $1 - \delta$ using as few traces as possible
- Let $T(n, q, \delta)$ be the number of traces needed to reconstruct an arbitrary string s on n bits with deletion rate q and confidence δ
- Given an arbitrary n node tree R, the *tree trace reconstruction problem* is to reconstruct R with probability at least $1 - \delta$ using as few TED traces as possible.
- We relate the sample complexity of the tree reconstruction problem to $T(n, q, \delta)$

Best known bounds

- Chase (2019) proved $T(n, q, \delta) > \tilde{\Omega}(n^{3/2})$

Best known bounds

- Chase (2019) proved $T(n, q, \delta) > \tilde{\Omega}(n^{3/2})$
- In a separate paper, Chase (2020) proved that $T(n, q, \delta) < \exp(\tilde{O}(n^{1/5}))$

Best known bounds

- Chase (2019) proved $T(n, q, \delta) > \tilde{\Omega}(n^{3/2})$
- In a separate paper, Chase (2020) proved that $T(n, q, \delta) < \exp(\tilde{O}(n^{1/5}))$
- These are the best generic upper and lower bounds, a huge gap!

Best known bounds

- Chase (2019) proved $T(n, q, \delta) > \tilde{\Omega}(n^{3/2})$
- In a separate paper, Chase (2020) proved that $T(n, q, \delta) < \exp(\tilde{O}(n^{1/5}))$
- These are the best generic upper and lower bounds, a huge gap!
- Davies, Racz, and Rashtchian (2020) introduced the TED model, and proved upper bounds for spider graphs and complete k-ary trees

Best known bounds

- Chase (2019) proved $T(n, q, \delta) > \tilde{\Omega}(n^{3/2})$
- In a separate paper, Chase (2020) proved that $T(n, q, \delta) < \exp(\tilde{O}(n^{1/5}))$
- These are the best generic upper and lower bounds, a huge gap!
- Davies, Racz, and Rashtchian (2020) introduced the TED model, and proved upper bounds for spider graphs and complete k-ary trees
- The spider upper bounds generalized some techniques from De et al. (2017), Nazarov and Peres (2017).

Functions Applied to Traces

Definition

If $s \in \{\pm 1\}^n$, then $\mathcal{D}_s^k$ is the distribution of traces generated from the deletion channel applied to s when conditioned to have length k.

Theorem

Let $s \in \{\pm 1\}^n$ be any string. Then for every $0 \leq k \leq n$ and for any function $h : \{\pm 1\}^k \to \mathbb{R}$,

$$\mathbb{E}_{\mathcal{D}_s^k}[h] = \frac{1}{\binom{n}{k}} \sum_{I \in \{0,1\}^k} \left(\binom{n-k}{k - \|I\|} \cdot \mathbb{E}_{\mathcal{D}_{s^{k\to}}^{k-\|I\|}} [h_I^s] \right)$$

Proof Idea

- Write the expected value in terms of string densities
- Split the sum into parts corresponding to fixed bit values
- Recursively fix bit values until h is saturated

Combinatorics of Infinite Deletion Channel

Theorem

Let $s \in \{0,1\}^\infty$, and suppose the 1's occur at indices $\mathcal{I} \subset \mathbb{N}$. Consider the generating function for s, $f(s;x) := \sum_{i \in \mathcal{I}} x^i$. Then under the deletion channel with rate q,

$$\mathbb{P}[j\text{'th bit of the trace is a } 1] = \frac{1}{j!} p^{j+1} \frac{\partial^j f(s;\cdot)}{\partial x^j}\bigg|_{(1-p)}$$

Proof

Proof.

Observe $\frac{\partial^j f(t;x)}{\partial x^j} = \sum_{i\in\mathcal{I}} \frac{(i)!}{(i-j)!} x^{i-j}$.
Also, $\mathbb{P}[j$'th bit of the trace is a $1] = \sum_{i\in\mathcal{I}} \binom{i}{j} p^{j+1}(1-p)^{i-j}$ by inspecting each 1 in s and noting the probability it ends up at position j, and using the fact that no bit $\mathcal{I} \ni i < j+1$ contributes anything to the sum, and by convention $i < j \implies \binom{i}{j} = 0$. Therefore,

$$\mathbb{P}_{\mathcal{D}_t}[A_j] = \frac{1}{j!} p^{j+1} \sum_{i\in\mathcal{I}} \frac{i!}{(i-j)!} (1-p)^{i-j} = \frac{1}{j!} p^{j+1} \frac{\partial^j f(k;\cdot)}{\partial x^j}(1-p)$$

Note: This argument also applies to k-mer probabilities with s replaced by $s \bigoplus r$ $\square$

TED Lower Bound Reduction

Definition

In the TED deletion channel, when vertex v is removed, contract the edge between v and its parent. Each vertex is removed with probability p.

Theorem

Let $q \in (0,1)$ and $\delta > 0$ be constants. Then,

TED Lower Bound Reduction

Definition

In the TED deletion channel, when vertex v is removed, contract the edge between v and its parent. Each vertex is removed with probability p.

Theorem

Let $q \in (0, 1)$ and $\delta > 0$ be constants. Then,

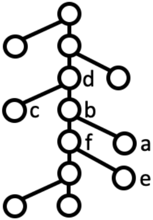
At least $\Omega(T(n, q^2, \delta))$ TED traces are needed to distinguish arbitrary unlabelled trees with probability at least $1 - \delta$.

Proof Idea for Theorem 3

Consider string $s = 010110$. We construct an *unlabelled* tree based on s:

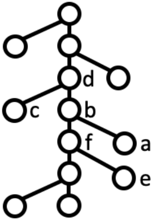
Proof Idea for Theorem 3

Consider string $s = 010110$. We construct an *unlabelled* tree based on s:



Proof Idea for Theorem 3

Consider string $s = 010110$. We construct an *unlabelled* tree based on s:



Deletion of nodes a, e or c are OK, have to be careful with nodes b, d, f.

Upper Bounds on Tree Families

Theorem

Let $q, \delta > 0$*. Given an ordered tree* R *with degree at least* $\log_{\frac{1}{q}}(nT(n,q,\delta))$ *on* n *nodes, , we can reconstruct* R *using* $T(n,q,\delta)$ *traces with probability at least* $1-\delta$*.*

Theorem

For any ordered tree R *on* n *nodes, and* $q, \delta > 0$*, if the leaves of* R *have label 0 and internal nodes have label 1, under the TED deletion channel a.a.s. we can reconstruct* R *and its labelling using* $T(n,q,\delta)$ *traces.*

Proof Idea

- For both theorems, on observing a TED trace from tree R, do a pre-order traversal, and write down two binary strings corresponding to leaf positions
- Show that the ordering of bits in these strings is preserved with high probability
- Argue that the distribution of these binary strings are equal to the string deletion channel applied to corresponding pre-order traversals of R.
- Apply an optimal string reconstruction algorithm to recover the leaf positions, and thus R itself.

Future Work

- Apply previous two theorems to problems in reconstruction or channel coding
- Find a generic upper bound for tree reconstruction with no bit labels. Is $T(n, q, \delta)$ sufficient in the TED channel?