Suffix Trees

Main Goal: Given text file $T$, process to create structure that supports pattern substring queries in time $O(|P|)$ for pattern $P$. All strings are over fixed alphabet $\Sigma$ (finite). Construction to take time linear in $|T|$, and requires linear space.

Example of suffix tree

$$T = aabaabb$$

(Add new symbol to right end so that $T$ becomes $aabaabb\$$. Then no suffix is a prefix of some other suffix.)

Each suffix corresponds to a full rooted path whose edge labels combine to form the suffix.
Storage is reduced by encoding each edge label by the pair of integers that delimit its position in $T$.

\[
S = abcde \\
\text{label} = (3, 5) \leftrightarrow cde
\]

Any pattern present in $T$ is a prefix of some suffix of $T$ and defines a unique (partial) path from the root of the suffix tree,

**Construction** (Outline)

$T_i = \text{suffix starting at position } i$.

$\tau_0$ denotes an empty tree.

For $j = 1$ to $\text{length}(T)$

$\tau_i = \text{insert}(\tau_{i-1}, T_i)$
Set $i = \text{position in } T \text{ of current suffix being inserted.}$
Set $m = \text{position in } T \text{ of mismatch at which path for } T_i \text{ in } \tau_{i-1} \text{ must diverge (where new edge sprouts to form } \tau_i).$

**Observations**

$m$ never decreases from one insertion to the next.

**Proof:**

Case I. (Previously inserted edge sprouted from the root) $m$ was $i - 1$ and its next value is $\geq i.$
Case II. (Previously inserted edge sprouted from non-root node) (Define $\text{Tail}(cw) = w \text{ for any single character } c)$
\(c\alpha = \text{path label (combined edge labels) to node } v\). The mismatch at \(m\) between \(T_{j-1}\) and \(T_{i-1}\) that causes \(v\) likewise appears between \(T_i\) and \(T_j\) causing \(v'\). If there’s a continuation path in \(\tau\) beyond \(v'\) matching \(T_i\) (past \(\alpha\)), then \(m\) advances, otherwise \(m\) stays fixed (and \(\tau\) sprouts a new edge at \(v'\)).

**Suffix link**
Link from \(v\) to \(v'\) is the *suffix link* at \(v\)
\[
\operatorname{Tail}(\text{path_label}(v)) = \text{path_label}(v')
\]

**Observe** Tree-depth\((v')\) (number of tree edges)
on path to \( v' \)
\[
\geq \text{Tree-depth}(v) - 1,
\]
since each node on path to \( v \) (other than the root) has an image on the path to \( v' \). Suffix links point to the images.

**Insertion and suffix link placement (details)**

Loop Invariant: All suffix links have been placed from existing non-root nodes, except from node created by the current insertion.

Let \( m_i \) be the mismatch location when insertion of \( T_i \) is completed. To insert \( T_{i+1} \), trace forward from \( w' \) – reached by following the suf-
fix link from \( w \) (or trace forward from \( v' \) if suffix link from \( v \) already present) – to reach the character location \( L \) (say) on the path that \( T_{i+1} \) traces, corresponding to \( m_i \). Each edge or partial edge below \( w' \) requires only constant time to trace (each appropriate edge is chosen so that the first character of its edge label is consistent with \( T_{i+1} \)). (If \( w \) is the root – or \( v \) is the root, then a suffix link is not utilized, and tracing forward proceeds from the root.) Then perform character comparisons between \( T_{i+1} \) and \( \tau \) labels starting at position \( j = m_i \) (and location \( L \) in \( \tau \)), advancing \( j \) until a mismatch occurs and \( T_{i+1} \) then diverges from the existing path, sprouting a new edge. A node \( x \) just above location \( L \) is now present. A suffix link from \( v \) to \( x \) gets placed, and \( j \) is now the new value for \( m \).
Analysis
Since $m$ never decreases, there are at most $O(|T|)$ character comparisons between characters of $T$ and characters in path labels of $\tau$ as it grows – ignoring those involved in the forward edge tracing steps at constant cost per edge – since a comparison can result in a mismatch most once per suffix insertion.) Next, consider forward edge tracing steps in $\tau$. The final node from which a suffix sprouts a new edge in $\tau$ is at the root (when $\$ \$ \$ \$ \$ \$ \$ \$ gets inserted). The initial node from which a suffix sprouts is also the root. Since $\text{tree-depth}(w') \geq \text{tree-depth}(v) - 2$, (so that the initial depth reduction with each insertion is $\leq 2$), and each forward edge trace increases tree-depth by one, the number of forward edge traces is bounded by $2|T|$. 