Analyzing the sensitivity of pq-gram distance with p and q

Nitish Srivastava
Indian Institute of Technology
Kanpur, India
nitishs@cse.iitk.ac.in

Varunesh Mishra
Indian Institute of Technology
Kanpur, India
varunesh@cse.iitk.ac.in

Arnab Bhattacharya
Indian Institute of Technology
Kanpur, India
arnabb@cse.iitk.ac.in

ABSTRACT

The pq-gram distance is a recently proposed approach for approximate matching of hierarchical data. It works by dividing a tree into small subtrees of a fixed shape and uses the number of common subtrees as a measure of similarity between two trees. The distance is efficiently computable and being parametrized by p and q, has the ability to be tuned to assign importance to different factors when comparing two trees. Higher values of p and q lead to more emphasis on structural rigidity. However, the resulting index structure becomes larger. Choosing the correct values of p and q is a matter of trade-off, the exact nature of which needs to be analyzed. A better understanding of the dependence of pq-gram distance on its parameters is helpful for domain experts to determine the correct parameter values to be used. We address this issue both analytically and experimentally by working with random trees. Our experiments and analyses provide deeper insight into the working of pq-gram distance for different models of tree corpora such as data clustered around random seeds, hierarchically clustered data and edit distance separated data. These models closely relate to real-world datasets. We analyze the sensitivity of pq-gram distance with respect to corpus parameters such as cluster radius, tree sparsity, fan-out and height. Our results show that pq-gram distance offers a high resolution power in a region close to a given tree, which is desirable for nearest neighbour queries.

1. INTRODUCTION

The pq-gram distance is a recent approach for approximate matching of hierarchical data proposed by Augsten et al. [1]. It is a distance measure between ordered, labeled trees. It is efficiently computable (($O(n \log n)$ time and $O(n)$ space). Intuitively, the pq-grams of a tree are all its subtrees of a specific shape. Two trees are said to be similar if they have more pq-grams in common. Apart from computational efficiency, pq-gram distance has an additional advantage. By adjusting the two parameters p and q, which specify the shape of the pq-grams, the distance measure can be tuned. In this paper, we study the nature of this tuning. The values of p and q must usually be determined by a domain expert who understands the underlying semantics of the data and can assess how important various factors are in determining the distance between two trees. Increasing the values of p and q makes the distance more *rigid*, i.e. more importance is being given to the structure of the tree as compared to the data. Decreasing them makes it insensitive to structure. As an extreme case, the values $p = 1$ and $q = 1$ would result in a $1, 1$-gram profile which would be just an unordered list of all parent-child pairs in the tree. Though some structural information can be retrieved using common parent relations, the structure of the tree becomes ambiguous beyond the immediate parent if labels of two different nodes are same. Increasing p relative to q implies that more importance is being given to the ancestors than to the children, i.e., two nodes are being considered same only when they share p common ancestors. Higher parameter values also lead to a larger index structure. This makes it important to determine the best p and q values. Our experiments and analyses give insight into this problem and demonstrate the effect of different values of p and q on different models of tree corpora: edit-separated trees, randomly clustered trees and hierarchically clustered trees. These classes capture various features that are relevant in real-world structured data.

VLDB '10, September 13-17, 2010, Singapore

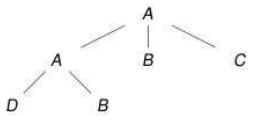
2. RELATED WORK

Augsten et al. [1] [2] proved the pseudo-metric nature of the pq-gram distance measure. Their work demonstrated the effectiveness of this measure in terms of efficiency ($O(n \log n)$ time and $O(n)$ space). The sensitivity of the distance with respect to p and q had been analyzed by experiments on leaf and non-leaf deletions in [1]. The authors showed that the sensitivity to leaf changes depends only on q and structural sensitivity is emphasized with higher values of p. For non-leaf deletions, the pq-gram distance is larger than for leaf deletions. We extend the scope and depth of this analysis by our experiments with different edit operations over stochastic models of tree corpora which model relevant characteristics of real world tree databases. Augsten et al. [3] showed that the pq-gram distance is a lower bound of the fanout weighted tree edit distance (the cost of editing a node is proportional to its fanout). To the best of our knowledge, there is no other work that analyzes pq-gram distance.

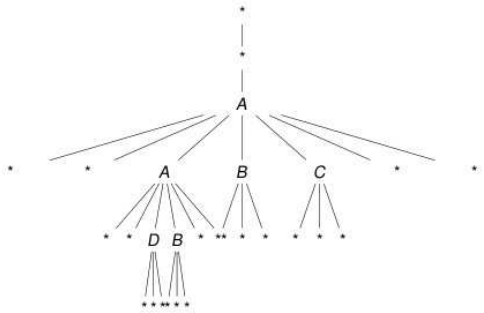
3. PRELIMINARIES

We present the notation and definitions required for analysis. These follow from those used by Augsten et al. [1]

Definition 1. pq-Extended-Tree. Let T be a tree, and $p > 0$ and $q > 0$ be two integers. The pq-extended tree, $T^{p,q}$, is constructed from T by adding $p - 1$ ancestors to the root node, inserting $q - 1$ children before the first and after the last child of each non-leaf



(a) Tree T



(b) Extended Tree $T^{3,3}$

Figure 1: A sample tree T along with its corresponding extended tree

node, and adding q children to each leaf of T. All newly inserted nodes are dummy nodes (denoted by $*$) that do not occur in T.

Definition 2. pq-Gram. Let T be a tree, $T^{p,q}$ the respective extended tree, $p > 0$, $q > 0$. A subtree of $T^{p,q}$ is a pq-gram G of T iff

1. G has q leaf nodes and p non-leaf nodes,
2. all leaf nodes of G are children of a single node $a \in N(G)$ with fan-out q, called the anchor node
3. the leaf nodes of G are consecutive siblings in $T^{p,q}$.

Definition 3. Label Tuple. Let G be a pq-gram with the nodes $N(G) = \{v_1, \ldots, v_p, v_{p+1}, \ldots, v_{p+q}\}$, where v_i is the i-th node in preorder. The tuple $\lambda^*(G) = (\lambda(v_1), \ldots, \lambda(v_p), \lambda(v_{p+1}), \ldots, \lambda(v_{p+q}))$ is called the label tuple of G.

where $\lambda(v)$ refers to the label of node v. Subsequently, if the distinction is clear from the context, we use the term pq-gram for both, the pq-gram itself and its representation as a label tuple.

Definition 4. pq-Gram Index. Let P be the set of all pq-grams of a tree T, $p > 0$, $q > 0$. The pq-gram index, $\mathcal{I}^{p,q}(T)$, of T is defined as the bag of label tuples of all pq-grams of T, i.e., $\mathcal{I}^{p,q}(T) = \uplus_{G \in P} \lambda^*(G)$.

For the tree T shown in Figure 1, $\mathcal{I}^{3,3}(T)$ is ,

```
(*,*,A,*,*,A)    (*,A,A,D,B,*)
(*,*,A,*,A,B)    (*,A,A,B,*,*)
(*,*,A,A,B,C)    (*,A,B,*,*,*)
(*,*,A,B,C,*)    (*,A,C,*,*,*)
(*,*,A,C,*,*)    (A,A,D,*,*,*)
(*,A,A,*,*,D)    (A,A,B,*,*,*)
(*,A,A,*,D,B)
```

Definition 5. pq-Gram distance. Let T_1 and T_2 be trees, $\mathcal{I}_1 = \mathcal{I}^{p,q}(T_1)$, $\mathcal{I}_2 = \mathcal{I}^{p,q}(T_2)$, $p > 0, q > 0$. The pq-gram distance, $d^{p,q}(T_1, T_2)$, between the trees T_1 and T_2 is defined as the symmetric difference between the respective profiles:

$$d^{p,q}(T_1, T_2) = |\mathcal{I}_1 \uplus \mathcal{I}_2| - 2|\mathcal{I}_1 \cap \mathcal{I}_2|$$

It is the number of pq-grams that differ between $\mathcal{I}_1$ and $\mathcal{I}_2$. The pq-gram distance is a pseudo-metric.

Definition 6. Normalized pq-gram distance. Let T_1 and T_2 be trees, $\mathcal{I}_1 = \mathcal{I}^{p,q}(T_1)$, $\mathcal{I}_2 = \mathcal{I}^{p,q}(T_2)$, $p > 0, q > 0$. The normalized pq-gram distance, $dist^{p,q}_{norm}(T_1, T_2)$, between the trees T_1 and T_2 is defined as :

$$d^{p,q}_{norm}(T_1, T_2) = \frac{d^{p,q}(T_1, T_2)}{|\mathcal{I}_1 \uplus \mathcal{I}_2| - |\mathcal{I}_1 \cap \mathcal{I}_2|}$$

This normalization preserves the pseudo-metric properties of pq-distance [1].

4. ANALYSIS OF EDIT OPERATIONS ON PQ-GRAMS

In order to study the nature of dependence of pq-gram distance on its parameters, we study its effect on suitably generated databases of random trees. Our model of random trees captures a general class of trees that is commonly encountered in domains involving structured data.

4.1 Generating random trees

A random tree T of height h is grown by starting with a root node and adding child nodes iteratively. Each node has a probability P_0 of bearing child nodes. This is modeled by a binomial distribution ($B(P_0)$), where success is interpreted as having children. The number of child nodes is chosen from a uniform distribution $U[1 \ldots N]$. Labels are assigned uniformly randomly from a label set Σ. This process is described in Algorithm 1. This is a general model that closely relates to real-world trees. Most trees that are encountered in domains involving structured data can be modeled using this scheme. In most real trees, each node may have different fan-out with a maximum value fixed. A uniform distribution of the type $U[1 \ldots N]$ is a natural choice to model this. Also each node may not have children. This is modeled using a binomially distributed variable.

Algorithm 1 Generating a random tree.

```
RANDOM-TREE (Σ, N, P_0, h)
T ← root
for i = 0 to h − 1 do
    for each leaf l at height i do
        has_children ∼ B(P_0)
        if has_children == true then
            no_of_children ∼ U[1 . . . N]
            Add no_of_children nodes to T as children of l
        end if
    end for
end for
Assign a label to each node of T by sampling uniformly randomly from the set of labels Σ
return T
```

4.2 Computing tree statistics

Let n_h, l_h and m_h denote the number of nodes, leaves and internal nodes respectively in a random tree T^H at height h ($h \leq H$).

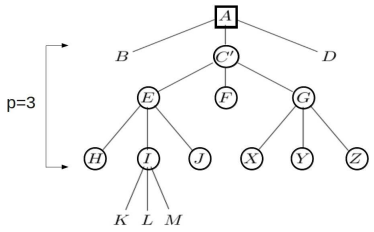


Figure 2: *Renaming of node C to C'*: The pq-grams with anchor node A have their q parts affected. The circled nodes have their p parts affected (p=3).

Then,

$$E[n_h] = \left(P_0 \frac{N+1}{2}\right)^h = \alpha^h \quad (1)$$

$$E[l_h] = \begin{cases} E[n_h](1-P_0) & h < H \\ E[n_h] & h = H \end{cases} \quad (2)$$

$$E[m_h] = \begin{cases} E[n_h](P_0) & h < H \\ 0 & h = H \end{cases} \quad (3)$$

See appendix for the proof. The above statistics can be summarized as exponential growth in the number of nodes with P_0 and N acting as controlling factors. We use these now to analyze the effect of edit operations.

4.3 Rename Operation

The first step towards understanding the effect of p and q is to see how pq-grams of a random tree are affected on performing edit operations on them. A single edit operation on a random tree affects a large number of pq-grams. This number depends on parameters p and q and also on the place in the tree where these operations are performed. Augsten et al. have demonstrated in [1] that the pq-gram distance weighs leaf deletions less than non-leaf deletions. We analyze the exact nature of this behavior. We find the expected pq-gram distance between a random tree T_1 and tree T_2, which is obtained on performing 1 edit operation on T_1.

Let a node C of T^H be renamed. This is illustrated in Figure 2. We find the expected pq-gram distance in terms of tree parameters for each of the following cases (details in appendix). Here $\alpha = \frac{P_0(N+1)}{2}$.

1. C is a leaf node

$$d^{p,q}_{norm}(T_1, T_2) = \frac{2(q+1)}{\alpha^{H-1}(qP_0 + 2 - P_0 + P_0N)} \quad (4)$$

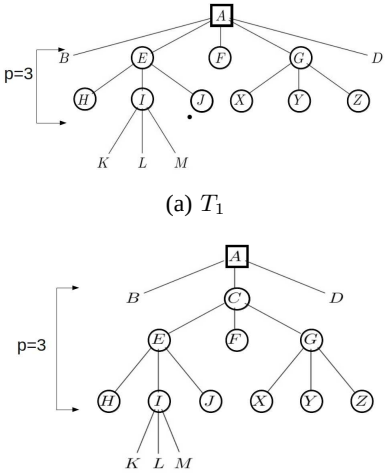
2. C is at height h such that $h + p < H$

$$d^{p,q}_{norm}(T_1, T_2) = \frac{2q + P_0(N-1)}{q(P_0\alpha^{H-p} + 1) + P_0(N-1)\alpha^{H-p}} \quad (5)$$

3. C is at height h such that $h + p \geq H$

$$d^{p,q}_{norm}(T_1, T_2) = 2\alpha^{-h} \frac{q + P_0N}{q(P_0 + \alpha^{-h}) + 2 - P_0 + 2P_0N} \quad (6)$$

The above equation describe the behavior of pq-gram distance in terms of p, q and the tree parameters. Several insights can be drawn from these which can be useful for a database designer to decide the values of p and q to be used for the pq-gram distance. Eq. (4) is independent of p showing that p does not affect leaf edit operations. Hence if the domain of application involves trees with a large number of leaf nodes and most edit operations are expected to be on these leaves, then the value of p does not matter much. The q value should be tuned so that given a tree, other trees which are close to it and of interest while processing queries are not too far from it. Eq. (4) gives a relation which describes how exactly the distance would be affected on changing q. Eq. (5) shows the dependence on p and q explicitly for the most general case in a large tree contaning sizeable number of non-leaf nodes. It clearly shows that dependence on p is exponential while that on q is inverse linear. In this sense the dependence on p is stronger than on q. This fact accounts for the sharper variation of intra to inter cluster distances with variation in p that we study later. The dependence on p is missing in Eq. (6) since the edit operation is to close to the root and p-part of the corresponding pq-grams includes dummy nodes. Variation in p only affects the number of dummy nodes in the p-part of corresponding pq-grams which do not lead to difference in number of affected pq-grams. The formulation of pq-gram distance in terms of the tree parameters is useful for a user of pq-gram distance to tune p and q according to the relevant database.



(a) T_1

(b) T_2

Figure 3: *Insertion of internal node C as parent of E, F and G*: pq-grams with anchor at the circled nodes have their p-parts changed. In T_1 all nodes at depth $p-1$ or less from the parent of the inserted node are affected. In T_2 new pq-grams corresponding to anchor node at C are created.

4.4 Insert Operation

Insertion of leaf nodes involves much smaller changes in the pq-gram profile than insertion of internal nodes.

1. Insertion of leaf node l in T_1 to give T_2

$$d^{p,q}_{norm}(T_1, T_2) = \frac{4q}{\alpha^{H-1}(qP_0 + 2 - P_0 + P_0N)} \quad (7)$$

2. Insertion of internal node at height h, $h + p < H$

$$d^{p,q}_{norm}(T_1, T_2) = \frac{2q + P_0(N-1)}{q(P_0\alpha^{H-p} + 1) + 2\alpha^{H-p}(1 + P_0 N)} \quad (8)$$

3. Insertion of internal node at height h, $h + p \geq H$

$$d^{p,q}_{norm}(T_1, T_2) = 2\alpha^{-h}\frac{q + P_0 N}{q(P_0 + \alpha^{-h}) + P_0(N+1)} \quad (9)$$

These equations describe how pq-distance changes with p, q and tree parameters. The nature of the equations is very similar to rename operation. For example Eq. (7) and (9) are independent of p, each representing border cases involving dummy nodes. Eq. (8) represents the most common case. This equation shows a *mixing* of p and q in the form of $\alpha^{H-p}qP_0$ which was also present for the case of a rename operation. The coupling is not very strong since the exponential in p converges quickly. The inverse linear dependence on q is also retained from the rename case. The difference is only in terms of the tree parameters N and P_0.

4.5 Delete Operation

The deletion operation is inverse of the insert operation. If a delete operation on T_1 gives T_2, then by a unique insert operation T_2 can be converted to T_1. Since pq-gram distance is symmetric, $d^{pq}(T_1, T_2) = d^{pq}(T_2, T_1)$. Thus the functional dependencies for delete operation are exactly the same as in the previous case of insert operation.

4.6 Bounds for n edit operations

The above analysis gives functional dependence of pq-gram distance on p and q for single edit operations. These results are difficult to obtain for general n edit operations. However the distance can be trivially upper bounded using n times the affected number of pq-grams for single edit operation. That does not change the nature of the dependence on p, q or the tree parameters. Hence, the above inferences and results generalize to n edit operations.

5. EDIT-SEPARATED TREES

Trees which make up most natural tree corpora are not distributed uniformly randomly over the set of all possible trees with a given height. Therefore, a simplistic analysis of pq-gram distance over a set of random trees would not be useful in these cases. One way of modeling a natural database of structured data is to think of it as a finite set of clusters where the seed of each cluster is a random variable chosen from some probability distribution. In our experiments we assume two such distributions: a random distribution and a hierarchical distribution over the seeds. Before analyzing pq-gram distance with respect to these, we study the dependence of pq-gram distance on edit-separated trees ,i.e., trees which are separated by a fixed number of tree edit operations. This analysis is helpful for studying clustered corpora later.

In these experiments we analyze the dependence of pq-gram distance on p and q for different parameters of the random trees. Figure 4 shows the effect of varying p on the average normalized pq-gram distance between a random tree T_1 and another tree T_2 which is 20 edit operations away from T_1 for different tree heights. The edit operations are random (insert, delete, rename) and are applied to random nodes in the tree. Algorithm 2 describes this procedure. The default values of parameters are $p = 3$, $q = 3$, $h = 6$, $|\Sigma| = 100$, $N = 5$, $P_0 = 0.7$. Each experiment is averaged over 50 runs.

Algorithm 2 Performing random edit operations.

```
RANDOM-EDIT (T, Σ, n)
for i = 1 to n do
    Choose a node u randomly from T
    Choose an edit operation E randomly from {INS, DEL, REN}
    if E == INS then
        Let u have children {v_1, ..., v_t}
        Choose 2 numbers i and j randomly from U[1 ..., t]
        Create a node p and assign a label randomly from Σ
        Add {v_i, ..., v_j} as children of p
        Add p to T as child of u
    else if E == DEL then
        Let u have children {v_1, ..., v_t}
        Delete u from T and add {v_1, ..., v_t} as children of parent[u]
    else
        Assign a random label chosen from Σ to u
    end if
end for
return T
```

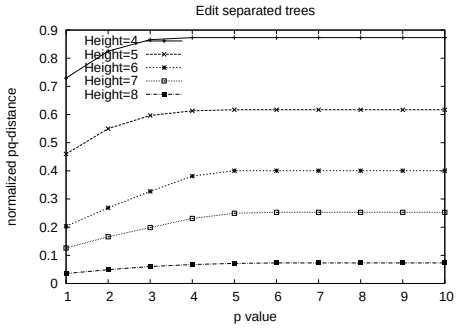


Figure 4: Effect of varying p on edit-separated trees for different tree heights.

For trees of height h, the corresponding plot attains a fixed value at $p = h$. This is to be expected as increasing p beyond the height of the tree cannot affect the pq-gram distance Figure 5 shows the effect of varying q on the average pq-gram distance between T_1 and T_2. The plots are almost perfectly linear showing the pq gram distance seems to grow only by a factor as q is increased. Figure 6 describes the effect of varying p on the average pq-gram distance between T_1 and T_2 as the distance between T_1 and T_2 increases. As the number of edit operations separating the two trees are doubled, the distance increases almost linearly. This illustrates a significant property that pq gram distance can be applied to databases containing very different trees. The distance has the ability to resolve closely in a region close to a given tree while all trees which are far away are almost at the same distance from it. In most application domains, only trees close to a given tree are of interest. A distance measure need not distinguish between two trees which are both far away from a given tree but it is important to preserve the distances to closer trees. pq-gram distance demonstrates this property. In the results shown in Figures 6 and 7, the number of edit-operations need to be doubled to obtain a constant increase in pq-gram distance. This implies that index structures based on pruning using pq-gram distance are very well suited for supporting nearest neighbour queries.

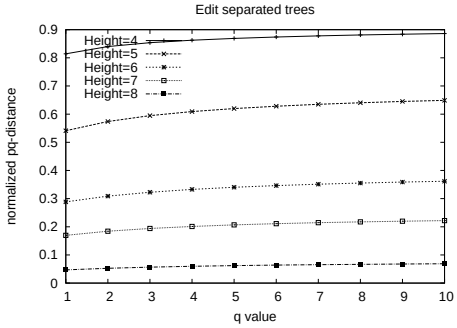


Figure 5: Effect of varying q on edit-separated trees for different tree heights.

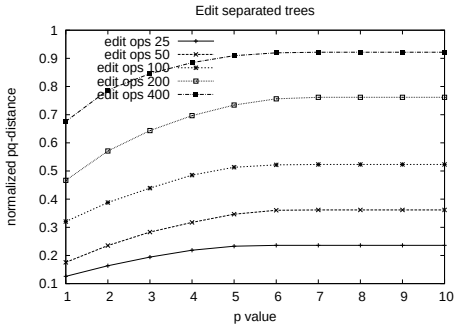


Figure 6: Effect of varying p on edit-separated trees with different number of edit operations.

6. MODELING TREE CORPORA

6.1 Clusters with random seeds

In this model, the database consists of m clusters $\{c_1, \ldots, c_m\}$. The seeds $\{s_1, \ldots, s_m\}$ are randomly chosen trees of height h generated by RANDOM-TREE. To generate an element of c_i, n random edit operations are performed on s_i. The resulting tree is added to c_i. This is repeated k times to generate a cluster of size k. The process is described in Algorithm 3. The database model thus has 4 parameters m, h, n and k, apart from the random tree parameters (Σ, N, P_0). These are the characteristics of the database. Most natural databases consisting of structured data have a number of base templates and all data points are small modifications of this template. This model is apt for such databases.

Algorithm 3 Generating randomly clustered trees.

```
1: RANDOM-CLUSTER (Σ, N, P_0, m, h, n, k)
2: for i = 1 to m do
3:     s[i] ← RANDOM-TREE (Σ, N, P_0, h)
4:     c[i] ← φ
5:     for j = 1 to k do
6:         T ← RANDOM-EDIT (s[i], Σ, n)
7:         c[i] ← c[i] ∪ T
8:     end for
9: end for
10: return {c_1, c_2, ..., c_m}
```

6.2 Hierarchical clusters

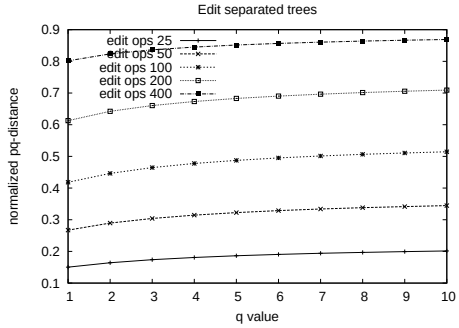


Figure 7: Effect of varying q on edit-separated trees with different number of edit operations

Construction of such a corpus is similar to the previous model except that the seeds are not chosen randomly. To choose the seeds, a random tree T_0 of height h is generated using RANDOM-TREE. This tree acts as the root of a hierarchy $\mathcal{T}$. The process is described in Algorithm 4. The database model thus has 5 parameters e, m, h, n and k apart from the tree parameters (Σ, N, P_0). These parameters control the nature of the database. This model applies to data consisting of clusters where the base templates are not random but related at a higher level.

We study the ability of the pq-gram distance to distinguish between clusters. The aim is to see how this ability changes with variations in p and q. A good way to measure this ability is to observe the variation in intra-cluster and inter-cluster pq-gram distances with change in parameters p and q. In our experiments, the largest intra-cluster distance in the first cluster (c_1) and the smallest inter cluster distance between any tree in c_1 to any tree outside are found. The ratio of these distances averaged over 50 runs is plotted for different p and q values. The evolution of this ratio shows the sensitivity of pq-gram distance with respect to its parameters. We observe this ratio for different database parameters. Default parameter values are $p = 3$, $q = 3$, $h = 6$, $|\Sigma| = 100$, $N = 5$, $P_0 = 0.7$, $m = 4$, $n = 5$, $k = 25$, $e = 10$. Each experiment is averaged over 50 runs.

Algorithm 4 Generating hierarchically clustered trees.

```
1: H-CLUSTER (Σ, N, P_0, e, m, h, n, k)
2: T_0 ← RANDOM-TREE (Σ, N, P_0, h)
3: for i = 1 to m do
4:     s[i] ← RANDOM-EDIT (T_0, Σ, e)
5:     c[i] ← φ
6:     for j = 1 to k do
7:         T ← RANDOM-EDIT (s[i], Σ, n)
8:         c[i] ← c[i] ∪ T
9:     end for
10: end for
11: return {c_1, c_2, ..., c_m}
```

7. SENSITIVITY TO CLUSTER RADIUS

Figures 8 and 9 show the result of varying p and q for different intra-cluster edit operations for random and hierarchical clusters respectively. The effect of increasing the number of edit operations made for creating each cluster (n) is to increase the "radius" of each cluster. As n increases, the clusters become bigger. Therefore the maximum intra-cluster distance increases. This pushes the clusters

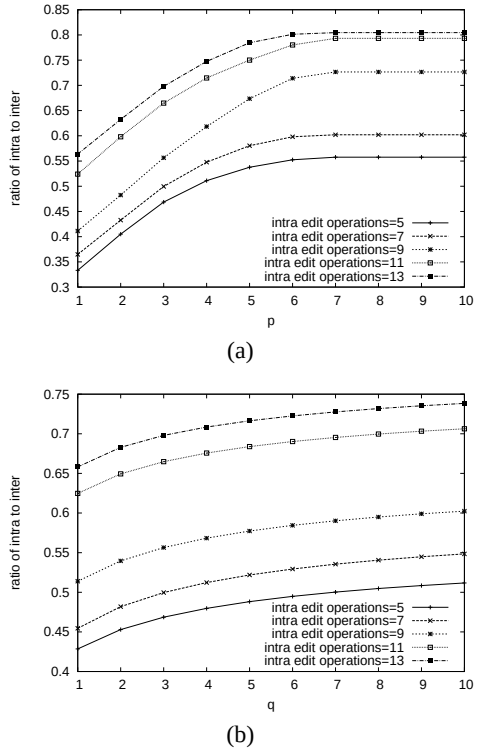


Figure 8: Randomly clustered trees with change in intra-cluster edit operations (n).

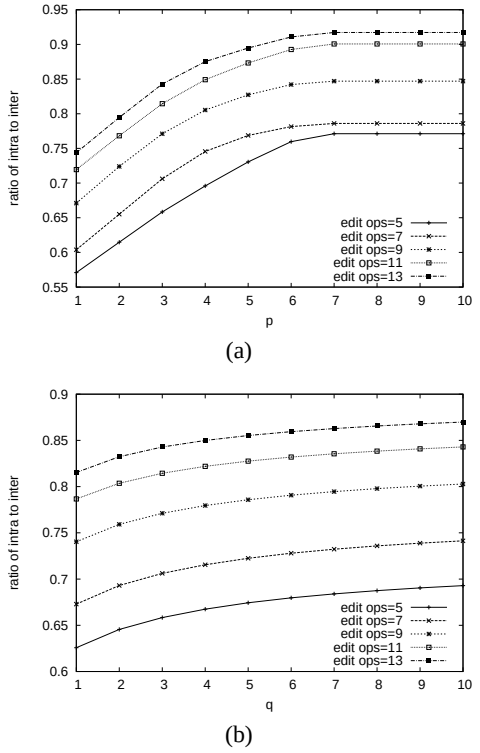


Figure 9: Hierarchically clustered trees with change in intra-cluster edit operations (n).

closer and the minimum inter-cluster distance decreases. Their ratio thus increases but in all cases converges to a linear asymptote. For varying p, it converges to a constant. This is to be expected as increasing p beyond the default height (h=6) does not affect the pq-gram distance. The corresponding plots for q do not converge as abruptly but attain a low constant slope for the observed q values. These observations are common to both models. However, they differ in that the ratio of intra to inter cluster distance is much higher for hierarchical clusters. This is to be expected as the trees are not clustered randomly and the clusters are closer to each other. Thus the inter cluster distance is lower, leading to a higher average ratio.

This experiment gives insight into the sensitivity of the intra to inter cluster distance ratio with respect to p and q. The range of deviation of the ratio is more for varying p than for q. For example, in the case of hierarchical clusters in figure 9 the range of deviation of the ratio for 9 edit operations is 0.05 for q variation while it is about 0.17 for p. The deviation is even higher in the case of random clusters(0.30 for p variation with 9 edit operations). This emphasizes the fact that the ability of pq-gram distance to differentiate between clusters is more sensitive to p variation than to q. This same trait was observed for edit-separated trees and has been analytically obtained for single edit operations in section 4. This observation validates our analytical results.

8. SENSITIVITY TO TREE HEIGHT

Figures 10 and 11 show the result of varying p and q for different tree heights. Increasing the value of h blows up the tree-space exponentially. Taller random trees are much further apart from each other than random trees of smaller height in terms of edit distance. Since the number of edit operations n performed are same, the *effective* radius of each cluster is smaller for taller trees. By *effective* we mean that the radius is to be normalized with the size of the tree. The ratio to which the plots converge decreases with height. This is expected since taller trees will have larger inter-cluster distance than shorter trees. The ratio is almost constant and equal to 1 for trees of height 3 and 4, showing that the clusters are quite close to one another. The ratio decreases dramatically as height increases because the clusters become more concentrated. There is not much difference between the plots for the two models, showing that the exponential increase in tree-space subsumes the fact the clusters are closer to each other in case of hierarchical clusters.

Another major inference that can be drawn from this analysis is the negligible dependence of the ratio on q. The plots in both the corpus models are almost flat indicating a very feeble dependence on q. This follows from the analytical results described earlier. The effect is more emphasized for height variation.

9. SENSITIVITY TO FAN-OUT

Figures 12 and 13 show the result of varying p and q for different values of N. N is the maximum fan-out of a node in the random tree model. The fan-out of a node is a uniformly distributed random variable $U[1 \ldots N]$. Using Eq. (1), N increases the size of the tree-space polynomially ($O(N^h)$, where h is the height of the tree which is kept fixed). This accounts for the decrease in the value to which the ratio converges, following the same reasoning as in the sensitivity analysis for tree height. The decrease is not as sharp since the size increases polynomially, and not exponentially (as with height). The spacing between consecutive plots increases with N. However, the increase is much less pronounced than the increase in spacing for height variation (Figure 13).

The stronger dependence on p than on q is again observed in this

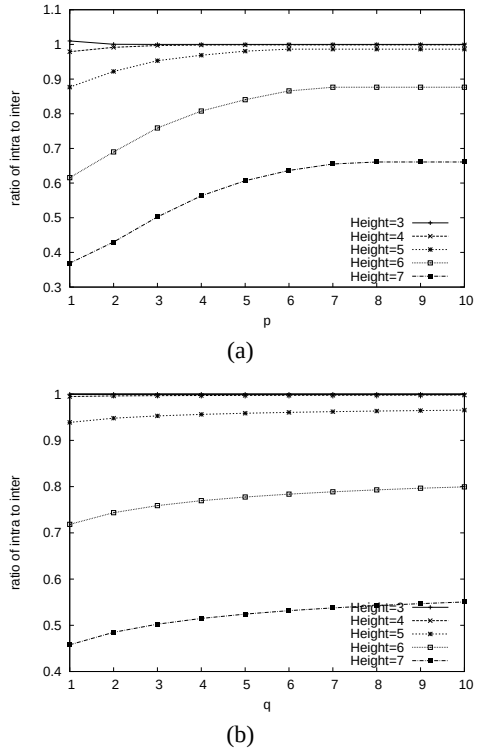


Figure 10: Randomly clustered trees with change in tree height (h).

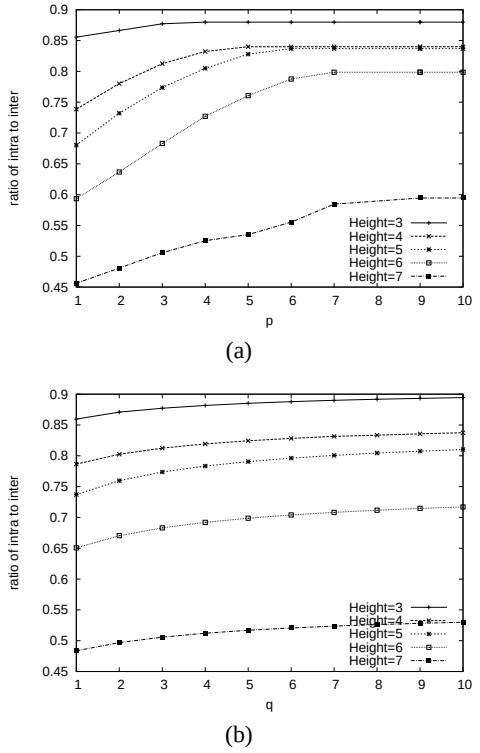


Figure 11: Hierarchically clustered trees with change in tree height (h).

case. A distinguishing characteristic for sensitivity with respect to fan-out is the wide range of intra to inter cluster distance ratios obtained. For example, the ratio ranges from a minimum of 0.28 for p=1 at N=7 to 1.00 for N=3 for almost all values of p at q=3 (default). Though the size of the tree increases polynomially in N, the ratio takes a wider range of values compared to that for variation in tree height, where the size grew exponentially. A wider range of ratios is obtained in all plots for variation in N. This shows that pq-gram distance is quite sensitive to fan-out and can differentiate between clusters with a higher resolution if N is larger.

10. SENSITIVITY TO SPARSITY

Figures 14 and 15 show the result of varying p and q for different values of P_0. P_0 is the probability with which a node in the tree bears children. Hence, P_0 controls the sparsity or *fatness* of the trees. A smaller P_0 leads to a shorter and thinner tree. P_0=1 corresponds to a full N-ary tree. This feature is crucial as it determines the nature of the tree with respect to its organization, i.e. the same data can be organized as a fat tree with leaf nodes only at the lowest level or as a sparser tree with leaves inside as well. The ratio of distances increases rapidly with p but the convergence values are not too different from each other for different values of P_0.

This illustrates a unique property of pq-gram distance. The ratio is almost independent of the value of P_0. This follows from the structure of the pq-grams. Recall that a pq-gram consists of a chain of $p-1$ immediate ancestors and q contiguous children. Thus a pq-gram is a tree which looks like a thin strand followed by a large fan-out at the tail. Due to the linear nature of the p-part, the structure does not depend on the sparsity of the tree from which this is extracted as long as the fan-out parameter is held constant, which is the case here (N is fixed). In a corpus of sparse trees, the number of such strands will be smaller compared to a denser tree but the structure of the pq-grams will not be much different if the fan-out and height are same. By taking the normalized pq-gram distance, the effect of smaller number of pq-grams is counteracted. Thus the distance measure becomes almost independent of the sparsity parameter.

11. CONCLUSIONS

In this paper, we analyzed the pq-gram distance, which has been shown to be useful for approximate matching of ordered labeled trees. We studied how the pq-gram distance changes with p and q when an edit operation is performed on a tree. We also investigated the sensitivity of clustering random trees using the pq-gram distance on p, q along with the various cluster parameters.

12. ACKNOWLEDGMENTS

The authors would like to thank the author of the original pq-gram papers, Prof. Nikolaus Augsten, for his suggestions, fruitful discussions and code for implementation of pq-gram distance.

13. REFERENCES

[1] N. Augsten, M. Böhlen, and J. Gamper. Approximate matching of hierarchical data using pq-grams. In *VLDB*, pages 301–312, 2005.

[2] N. Augsten, M. Böhlen, and J. Gamper. An incrementally maintainable index for approximate lookups in hierarchical data. In *VLDB*, pages 247–258, 2006.

[3] N. Augsten, M. H. Böhlen, and J. Gamper. The *q*-gram distance between ordered labeled trees. *ACM Trans. Database Syst.*, 35(1), 2010.

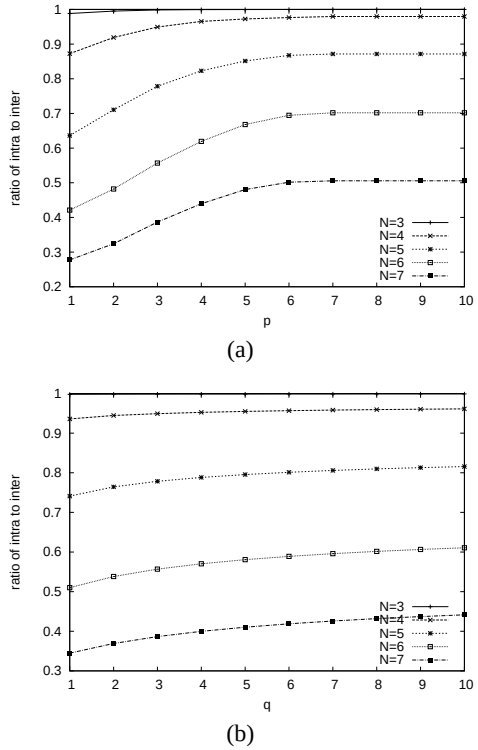


Figure 12: Randomly clustered trees with change in maximum fan-out (N).

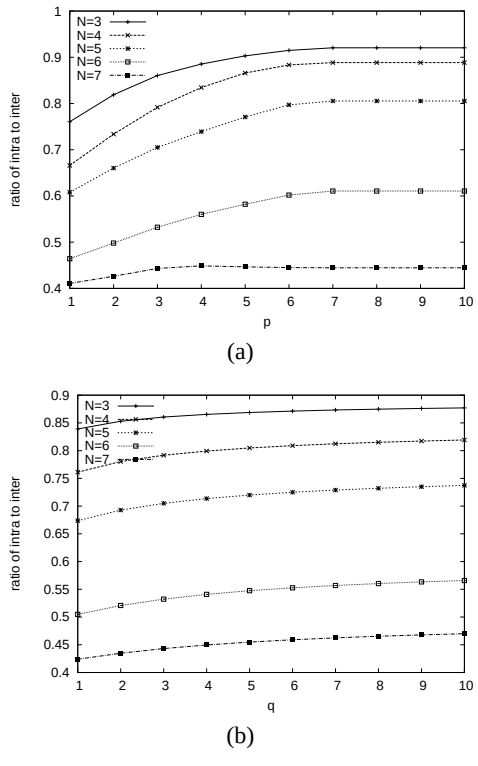


Figure 13: Hierarchically clustered trees with change in maximum fan-out (N).

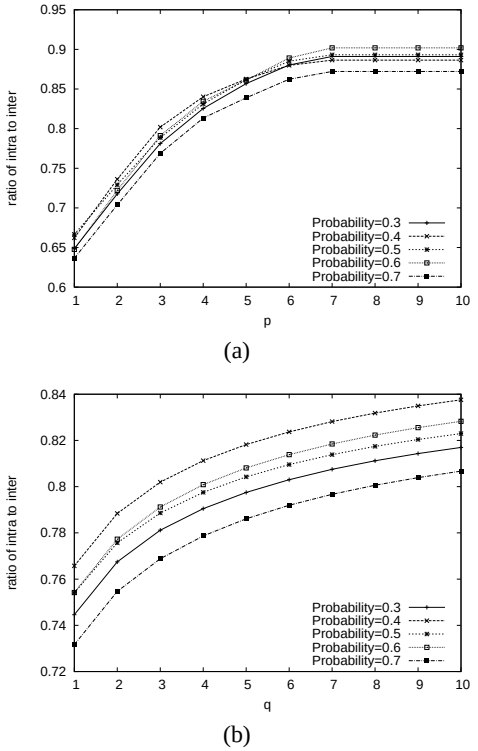


Figure 14: Randomly clustered trees with change in probability of bearing children (P_0).

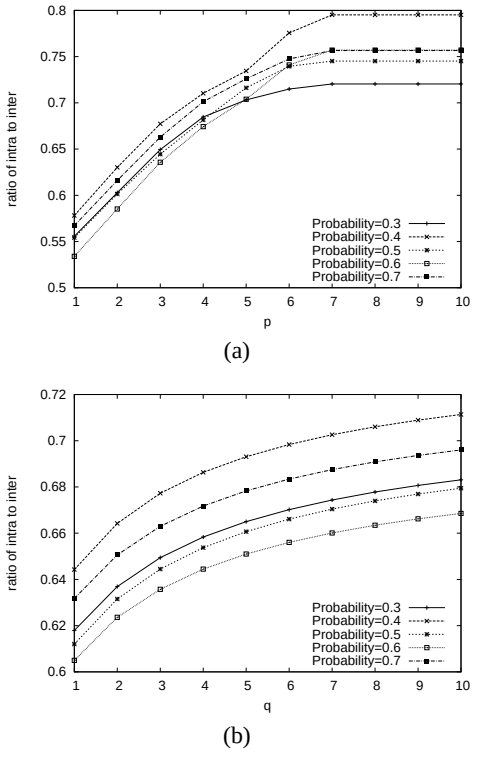


Figure 15: Hierarchically clustered trees with change in probability of bearing children (P_0).

APPENDIX

In the appendix, we derive the expressions of how an edit operation on a tree T_1 affects its distance to the new tree T_2 thus formed. The distance is expressed in terms of p, q, and the tree parameters: (i) binomial probability of a node having children, P_0, (ii) maximum fan-out, N, (iii) height of the tree, h. In the equations derived in the subsequent sections, we use the following notations:

$$\begin{aligned}\alpha &= \frac{P_0(N+1)}{2}\\ \beta &= \frac{P_0(N+1)}{2}+q-1\\ \Delta^h &= \left(\frac{\alpha^{h+1}-1}{\alpha-1}\right)\end{aligned}$$

A. TREE STATISTICS

n_h be a random variable which denotes the number of nodes in a random tree at height h.
$b_{h,i}$ be a random variable drawn from a binomial distribution with probability of success p_0. Success means that the ith node at height h has children.
$v_{h,i}$ is a random variable drawn from a uniform distribution over $\{1,2,\ldots,n\}$. It denotes the number of children of the ith node at height h, if it has any.

$$\begin{aligned}n_{h+1} &= \sum_{i=1}^{n_h} b_{h,i}v_{h,i}\\ E[n_{h+1}] &= \sum_{k=0}^{\infty}\left(P(n_h=k)E\left[\sum_{i=1}^{k} b_{h,i}v_{h,i}\right]\right)\\ &= \sum_{k=0}^{\infty}\left(P(n_h=k)\sum_{i=1}^{k}E[b_{h,i}]E[v_{h,i}]\right)\\ &= \sum_{k=0}^{\infty}\left(P(n_h=k)\sum_{i=1}^{k}P_0\frac{N+1}{2}\right)\\ &= \sum_{k=0}^{\infty}\left(P(n_h=k)kp_0\frac{N+1}{2}\right)\\ &= P_0\frac{N+1}{2}\sum_{k=0}^{\infty}kP(n_h=k)\\ &= \alpha E[n_h]=\alpha^2E[n_{h-1}]=\ldots\\ &= \alpha^{h+1}E[n_0]=\alpha^{h+1}\end{aligned}$$

The expected number of nodes in a tree of height h be Δ^h. Then,

$$\begin{aligned}\Delta^h &= \sum_{i=0}^{h}E[n_i]=\sum_{i=0}^{h}\alpha^i=\frac{\alpha^{h+1}-1}{\alpha-1}\\ |\mathcal{I}^{p,q}| &= 2l+qi-1 \text{ (As shown by Augsten et al. [1])}\end{aligned}$$

where l is the number of leaves, i is the number of non-leaf nodes in the tree.

$$\begin{aligned}&= 2\left((1-P_0)\Delta^{h-1}+\alpha^h\right)+q\left(P_0\Delta^{h-1}\right)-1\\ &= \Delta^{h-1}(2-P_0(2-q))+2\alpha^h-1\end{aligned}$$

B. RENAME OPERATION

Let a node C of T^H be renamed. The following cases arise

1. C is a leaf node
2. C is at height h such that $h+p<H$
3. C is at height h such that $h+p\geq H$

B.1 Rename of leaf node

Affected pq-grams contain node C as anchor node or as a node in the q part of pq-grams with its parent as anchor node. Number of pq-grams containing C as anchor is 1 and number of pq-gram containing C in q part is q. Total number of pq-gram affected is $q+1$. Thus the corresponding pq-gram indices differ in $q+1$ places (assuming that the label set is large enough so that the new label is different from its previous label).

$$d^{p,q}(T_1,T_2) = |\mathcal{I}_1 \uplus \mathcal{I}_2| - 2|\mathcal{I}_1 \upharpoonright \mathcal{I}_2|$$

$\mathcal{I}_1$ and $\mathcal{I}_2$ differ for $q+1$ label tuples. Each common pq-gram occurs twice in $\mathcal{I}_1 \uplus \mathcal{I}_2$ and once in $\mathcal{I}_1 \upharpoonright \mathcal{I}_2$. Thus common pq-grams do not contribute anything to the distance. Each affected pq-gram occurs once in $\mathcal{I}_\infty$ and the corresponding changed pq-gram occurs once in $\mathcal{I}_\in$. They do not occur in $\mathcal{I}_1 \upharpoonright \mathcal{I}_2$. Hence each affected pq-gram contributes 2 to the distance. The total distance is hence twice the number of affected pq-grams. This holds for all rename operations.

$$\begin{aligned}d^{p,q}(T_1,T_2) &= 2(q+1)\\ d^{p,q}_{norm}(T_1,T_2) &= \frac{d^{p,q}(T_1,T_2)}{\mathcal{I}^{p,q}+\frac{d^{p,q}(T_1,T_2)}{2}}\\ &= \frac{2(q+1)}{\Delta^{H-1}(2-P_0(2-q))+2\alpha^H+q}\end{aligned}$$

Simplifying under the assumption that α is large enough,

$$\begin{aligned}d^{p,q}_{norm}(T_1,T_2) &= \frac{2(q+1)}{\Delta^{H-1}(2-P_0(2-q))+2\alpha^H+q}\\ &\approx \frac{2(q+1)}{\alpha^{H-1}(2-P_0(2-q))+2\alpha^H+q}\\ &= \frac{2(q+1)}{q(1+P_0\alpha^{H-1})+\alpha^{H-1}(2-2P_0)+2\alpha^H}\\ &\approx \frac{2(q+1)}{q(P_0\alpha^{H-1})+\alpha^{H-1}(2-2P_0)+2\alpha^H}\\ &= \frac{2(q+1)}{\alpha^{H-1}(qP_0+(2-2P_0)+2\alpha)}\\ &= \frac{2(q+1)}{\alpha^{H-1}(qP_0+2-P_0+P_0N)}\end{aligned}$$

B.2 Rename of node at height h ($h+p<H$)

For one rename operation, affected pq-grams will contain the affected node in q part or p part.
The number of affected pq-grams when node is in q part are q
The number of affected pq-grams when node is in p part are

$$\sum_{j=0}^{p-1}\left(\sum_{i=1}^{m'_j}(v_{ij}+q-1)+l'_j\right) \quad (10)$$

where
m'_j is the number of non leaf nodes at height $j+h$ which anchor pq-grams containing C in their p part
l'_j is the number of leaf nodes at level $j+h$ which anchor pq-grams containing C in their p part
v_{ij} is the number of children of the i-th non-leaf node at height $j+h$ which anchors pq-grams containing C in their p part.
A change in label of node C affects the p parts for anchor nodes in a tree of height $p-1$ rooted at C. At each level j below the root of this affected tree, there are m'_j anchor nodes which bear children and l'_j anchor nodes which are leaves. Each internal node has v_{ij} children and thus anchors $v_{ij}+q-1$ pq-grams. A leaf node anchors 1 pq-gram. Hence Eq. (10) follows. Total number of affected pq-grams

$$= q + \sum_{j=0}^{p-1}\left(\alpha^j\left(\frac{P_0\,(N+1)}{2}+q-1\right)+\alpha^j\,(1-P_0)\right)$$
$$= q + \sum_{j=0}^{p-1}\alpha^j\beta + \alpha^j(1-P_0)$$
$$= q + \sum_{j=0}^{p-1}\alpha^j(\beta+1-P_0)$$
$$= q + \frac{\alpha^p-1}{\alpha-1}(\beta+1-P_0) = q+\Delta^{p-1}(\beta+1-P_0)$$

These are the number of affected pq-grams but it is possible that the pq-gram distance is not exactly twice of this because some affected label tuples in $\mathcal{I}_1^{p,q}$ may match with some other label tuples of $\mathcal{I}_2^{p,q}$. Therefore the expected distance can be written as

$$d^{p,q}(T_1,T_2) = 2\left(q(1-|\Sigma|^{-q}) + \Delta^{p-1}(\beta+1-P_0)(1-|\Sigma|^{-q})\right)$$

$|\Sigma|$ is usually very large, therefore $|\Sigma|^{-q}\to 0$

$$d^{p,q}(T_1,T_2) = 2\left(q+\Delta^{p-1}(\beta+1-P_0)\right)$$

$$d^{p,q}(T_1,T_2) = 2\left(q+\Delta^{p-1}(\beta+1-P_0)\right)$$

$$\begin{aligned} d^{p,q}_{norm}(T_1,T_2) &= \frac{d^{p,q}(T_1,T_2)}{2\mathcal{I}^{p,q}-\left(\mathcal{I}^{p,q}-\frac{d^{p,q}(T_1,T_2)}{2}\right)} \\ &= \frac{d^{p,q}(T_1,T_2)}{\mathcal{I}^{p,q}+\frac{d^{p,q}(T_1,T_2)}{2}} \end{aligned}$$

Simplifying under the assumption that α is large enough,

$$= \frac{2\left(q+\Delta^{p-1}(\beta+1-P_0)\right)}{\Delta^{H-1}(2-P_0(2-q))+2\alpha^H-1+\frac{2\left(q+\Delta^{p-1}(\beta+1-P_0)\right)}{2}}$$
$$\approx \frac{2\left(q+\alpha^{p-1}(P_0(N-1)/2+q)\right)}{\alpha^{H-1}(2-P_0(2-q))+2\alpha^H-1+q+\alpha^{p-1}(P_0(N-1)/2+q)}$$
$$\approx \frac{2q(1+\alpha^{p-1})+\alpha^{p-1}P_0(N-1)}{q(\alpha^{H-1}P_0+\alpha^{p-1}+1)+2\alpha^{H-1}(1-P_0+\alpha)+\alpha^{p-1}(P_0(N-1)/2)}$$
$$\approx \frac{\alpha^{p-1}(2q+P_0(N-1))}{q(\alpha^{H-1}P_0+\alpha^{p-1})+2\alpha^{H-1}(1-P_0+\alpha)+\alpha^{p-1}(P_0(N-1)/2)}$$
$$\approx \frac{2q+(P_0(N-1))}{q(\alpha^{H-p}P_0+1)+2\alpha^{H-p}(1-P_0+\alpha)+P_0(N-1)/2)}$$
$$\approx \alpha^{p-H}\frac{2q+(P_0(N-1))}{q(P_0+\alpha^{p-H})+2(1-P_0)+2\alpha+P_0(N-1)\alpha^{p-H}/2}$$
$$\approx \alpha^{p-H}\frac{2q+(P_0(N-1))}{q(P_0+\alpha^{p-H})+2-P_0+P_0N+P_0(N-1)\alpha^{p-H}/2}$$
$$= \frac{2q+(P_0(N-1))}{q(P_0\alpha^{H-p}+1)+(2-P_0+P_0N)\alpha^{H-p}+P_0(N-1)/2}$$
$$\approx \frac{2q+P_0(N-1)}{q(P_0\alpha^{H-p}+1)+P_0(N-1)\alpha^{H-p}}$$

B.3 Rename of node at height h ($h+p \geq H$)

Let $h+k=H$ where $k\leq p$
Affected number of pq-grams here are

$$q+\sum_{j=0}^{k}(\sum_{i=1}^{n_j}(v_{ij}+q-1)+l_j)+\alpha^k \tag{11}$$

$$\begin{aligned} d^{p,q}(T_1,T_2) &= 2(q+\sum_{j=0}^{k-1}(\alpha^j\beta+\alpha^j(1-P_0))+\alpha^k) \\ &= 2(q+\Delta^{k-1}(\beta+1-P_0)+\alpha^k) \\ d^{p,q}_{norm}(T_1,T_2) &= \frac{d^{p,q}(T_1,T_2)}{\mathcal{I}^{p,q}+\frac{d^{p,q}(T_1,T_2)}{2}} \end{aligned}$$

Simplifying under the assumption that α is large enough,

$$= \frac{2(q+\Delta^{k-1}(\beta+1-P_0)+\alpha^k)}{\Delta^{H-1}(2-P_0(2-q))+2\alpha^H-1+q+\Delta^{k-1}(\beta+1-P_0)+\alpha^k}$$
$$\approx \frac{2(q+\alpha^{k-1}(\beta+1-P_0)+\alpha^k)}{\alpha^{H-1}(2-P_0(2-q))+2\alpha^H-1+q+\alpha^{k-1}(\beta+1-P_0)+\alpha^k}$$
$$= \frac{2q+2\alpha^{k-1}(P_0N+q)}{\alpha^{H-1}(2-P_0(2-q))+2\alpha^H-1+q+\alpha^{k-1}(P_0N+q)}$$
$$\approx \frac{2q(1+\alpha^{k-1})+2\alpha^{k-1}(P_0N)}{q(1+P_0\alpha^{H-1}+\alpha^{k-1})+2\alpha^{H-1}(1-P_0)+2\alpha^H+\alpha^{k-1}P_0N}$$
$$\approx \frac{2\alpha^{k-1}(q+P_0N)}{q(P_0\alpha^{H-1}+\alpha^{k-1})+2\alpha^{H-1}(1-P_0)+2\alpha^H+\alpha^{k-1}P_0N}$$
$$= \frac{2(q+P_0N)}{q(P_0\alpha^{H-k}+1)+2\alpha^{H-k}(1-P_0)+2\alpha^{H-k+1}+P_0N}$$
$$\approx 2\alpha^{k-H}\frac{q+P_0N}{q(P_0+\alpha^{k-H})+2-P_0+2P_0N}$$

C. INSERT OPERATION

C.1 Insertion of leaf node

Insertion of leaf nodes involves much smaller changes in the pq-gram profile than insertion of internal nodes. Let leaf node l be inserted in T_1 to give T_2. This does not affect the p-part of any pq-gram in $\mathcal{I}_1^{p,q}$. One additional pq gram is created with l as the anchor node. The pq-grams with the parent of l as anchor node are affected since the q part of some of them changes. The additional node l must appear in the q-part of q pq-grams in $\mathcal{I}_2^{p,q}$ with anchor node as parent of l. With very high probablity $(1-|\Sigma|^{-q})$, these q pq-grams are all different from the ones previously present there. pq-grams in which the left and right siblings of l occured adjacently will now be absent in $\mathcal{I}_2$. There were $q-1$ of such pq-grams.

$$\begin{aligned} d^{p,q}(T_1,T_2) &= 4q \\ d^{p,q}_{norm}(T_1,T_2) &= \frac{d^{p,q}(T_1,T_2)}{\mathcal{I}^{p,q}+\frac{d^{p,q}(T_1,T_2)}{2}} \\ &= \frac{4q}{\Delta^{H-1}(2-P_0(2-q))+2\alpha^H+2q} \end{aligned}$$

Simplifying,

$$\begin{aligned} d^{p,q}_{norm}(T_1,T_2) &= \frac{4q}{\Delta^{H-1}(2-P_0(2-q))+2\alpha^H+q} \\ &\approx \frac{4q}{\alpha^{H-1}(2-P_0(2-q))+2\alpha^H+2q} \\ &= \frac{4q}{q(2+\alpha^{H-1}P_0)+\alpha^{H-1}(2-2P_0)+2\alpha^H} \\ &= \frac{4q}{q\alpha^{H-1}P_0+\alpha^{H-1}(2-2P_0)+2\alpha^H} \\ &\approx \frac{4q}{\alpha^{H-1}(qP_0+(2-2P_0)+2\alpha)} \\ &= \frac{4q}{\alpha^{H-1}(qP_0+2-P_0+P_0N)} \end{aligned}$$

C.2 Insertion of internal node at h ($h + p < H$)

Let internal node m be inserted in place of $n_1, n_2, \ldots n_k$ as child of node t in T_1 to give T_2. To account for the affected pq-grams we first look at the pq-grams with t as anchor node. All pq -grams in which the q part contains any of $n_1, n_2, \ldots n_k$ are absent in $\mathcal{I}_2$. There are $q + k - 1$ such pq-grams. $\mathcal{I}_2$ contains q new pq-grams in which m occurs in the q part. Thus the q parts contribute $q + k - 1 + q$ changed pq-grams.

For the nodes $n_1, n_2, \ldots n_k$, all ancestors have shifted one level up and m has become parent. Also for any pq-gram which had t in the p-part, m must be inserted just after t and all ancestors before t shifted one level up. So with high probablity, the new sequence of ancestors will be different from the old one for each of these affected pq-grams. The number of such affcetd pq-grams will be total number of anchor nodes which are depth less than p from t times the numbr of pq-grams each such anchor node contributes which is

$$\sum_{i=1}^{p-1}\left(\sum_{j=1}^{n_i}(v_{ij}+q-1)+l_i\right)$$

In addition, there are $q + k - 1$ pq-grams with m as anchor node. Therefore

$$\begin{aligned} d^{p,q}(T_1,T_2) &= 2q+k-1+2\sum_{i=1}^{p-1}\left(\sum_{j=1}^{n_i}(v_{ij}+q-1)+l_i\right) \\ &+ q+k-1 \end{aligned}$$

$$\begin{aligned} & d^{p,q}(T_1,T_2) \\ &= 3q+2k-2+2\textstyle\sum_{i=1}^{p-1}\left(\sum_{j=1}^{n_i}(v_{ij}+q-1)+l_i\right) \\ &= 3q+2k-1+2\textstyle\sum_{i=1}^{p-1}\left(\alpha^i(\beta+1-P_0)\right) \\ &= 3q+2k-1+2\Delta^{p-1}(\beta+1-P_0) \\ & d^{p,q}_{norm}(T_1,T_2) = \frac{d^{p,q}(T_1,T_2)}{\mathcal{I}^{p,q}+\frac{d^{p,q}(T_1,T_2)}{2}} \\ &= \frac{3q+2k-1+2\Delta^{p-1}(\beta+1-P_0)}{\Delta^{H-1}(2-P_0(2-q))+2\alpha^H-1+\frac{3}{2}q+k-\frac{1}{2}+\Delta^{p-1}(\beta+1-P_0)} \\ &= \frac{q(3+2\Delta^{p-1})+2k-1+\Delta^{p-1}P_0(N-1)}{q(P_0\Delta^{H-1}+\Delta^{p-1}+\frac{3}{2})+2\Delta^{H-1}(1-P_0)+2\alpha^H-1+k-\frac{1}{2}+\Delta^{p-1}P_0(N-1)/2} \end{aligned}$$

Simplifying,

$$\begin{aligned} &\approx \frac{q(3+2\alpha^{p-1})+2k-1+\alpha^{p-1}P_0(N-1)}{q(P_0\alpha^{H-1}+\alpha^{p-1}+\frac{3}{2})+2\alpha^{H-1}(1-P_0)+2\alpha^H-1+k-\frac{1}{2}+\alpha^{p-1}P_0(N-1)/2} \\ &\approx \frac{2q\alpha^{p-1}+\alpha^{p-1}P_0(N-1)}{q(P_0\alpha^{H-1}+\alpha^{p-1})+2\alpha^{H-1}(1-P_0)+2\alpha^H+\alpha^{p-1}P_0(N-1)/2} \\ &= \alpha^{p-1}\frac{2q+P_0(N-1)}{q(P_0\alpha^{H-1}+\alpha^{p-1})+2\alpha^{H-1}(1-P_0)+2\alpha^H+\alpha^{p-1}P_0(N-1)/2} \\ &= \frac{2q+P_0(N-1)}{q(P_0\alpha^{H-p}+1)+2\alpha^{H-p}(1-P_0)+2\alpha^{H-p+1}+P_0(N-1)/2} \\ &= \frac{2q+P_0(N-1)}{q(P_0\alpha^{H-p}+1)+2\alpha^{H-p}(1+P_0N)+P_0(N-1)/2} \\ &\approx \frac{2q+P_0(N-1)}{q(P_0\alpha^{H-p}+1)+2\alpha^{H-p}(1+P_0N)} \end{aligned}$$

C.3 Insertion of internal node at h ($h + p \geq H$)

let $h + h' = H$ where $h' \leq p$

$$\begin{aligned} & d^{p,q}(T_1,T_2) \\ &= 3q+2k-2+2\textstyle\sum_{i=1}^{h'-1}\left(\sum_{j=1}^{n_i}(v_{ij}+q-1)+l_i\right)+2\alpha^{h'} \\ &= 3q+2k-2+2\Delta^{h'-1}\beta+2\alpha^{h'} \\ & d^{p,q}_{norm}(T_1,T_2) = \frac{d^{p,q}(T_1,T_2)}{\mathcal{I}^{p,q}+\frac{d^{p,q}(T_1,T_2)}{2}} \\ &= \frac{3q+2k-2+2\Delta^{h'-1}\beta+2\alpha^{h'}}{\Delta^{H-1}(2-P_0(2-q))+2\alpha^H-1+\frac{3}{2}q+k-1+\Delta^{h'-1}\beta+\alpha^{h'}} \\ &= \frac{q(3+2\Delta^{k-1})+2k-2+\Delta^{k-1}P_0(N-1)+2\alpha^k}{q(P_0\Delta^{H-1}+\Delta^{h'-1}+\frac{3}{2})+2\Delta^{H-1}(1-P_0)+2\alpha^H+k-2+\Delta^{h'-1}P_0(N-1)/2+\alpha^{h'}} \end{aligned}$$

Simplifying under the assumption that $|\Sigma|$ is large and constant k is small

$$\begin{aligned} &\approx \frac{2q\Delta^{h'-1}+\Delta^{k-1}P_0(N-1)+2\alpha^{h'}}{q(P_0\Delta^{H-1}+\Delta^{h'-1})+2\Delta^{H-1}(1-P_0)+2\alpha^H+\Delta^{h'-1}P_0(N-1)/2+\alpha^{h'}} \\ &\approx \frac{2q\alpha^{h'-1}+\alpha^{h'-1}P_0(N-1)+2\alpha^{h'}}{q(P_0\alpha^{H-1}+\alpha^{h'-1})+2\alpha^{H-1}(1-P_0)+2\alpha^H+\alpha^{h'-1}P_0(N-1)/2+\alpha^{h'}} \\ &= \alpha^{h'-1}\frac{2q+2P_0N}{q(P_0\alpha^{H-1}+\alpha^{h'-1})+2\alpha^{H-1}(1-P_0)+2\alpha^H+\alpha^{h'-1}P_0(N-1)/2+\alpha^{h'}} \\ &= \frac{2q+2P_0N}{q(P_0\alpha^{H-h'}+1)+2\alpha^{H-h'}(1-P_0)+2\alpha^{H-h'+1}+P_0N} \\ &\approx \frac{2q+2P_0N}{q(P_0\alpha^{H-h'}+1)+2\alpha^{H-h'}(1-P_0)+2\alpha^{H-h'+1}} \\ &= \alpha^{-h}\frac{2q+2P_0N}{q(P_0+\alpha^{-h})+2(1-P_0)+2\alpha} \\ &= 2\alpha^{-h}\frac{q+P_0N}{q(P_0+\alpha^{-h})+P_0(N+1)} \end{aligned}$$