# Beyond Highway Dimension: Small Distance Labels Using Tree Skeletons[1]

Seminar Advanced Algorithms and Data Structures - Report

Andrei Herasimau

## 1 Introduction

One of the most common queries in a graph setting is a shortest path query. The key concept in this paper, the *skeleton dimension*, helps answer these queries by giving information about the structure of a graph. It is related to the notion of *hub sets*. These are sets which contain so-called *transit nodes* which appear on shortest paths between many node pairs. As the authors of [1] mention, the concept of transit nodes started as an empirical observation more than a concrete discovery. It makes intuitive sense that some nodes in a real world graph route more traffic than others (e.g. train stops that aren't popular destinations themselves, but many routes to other destinations go through them). Out of this fairly simple observation emerged a new family of algorithms called *transit node routing algorithms*(TNR-algorithms). The core idea of these algorithms is to pre-process the input graph in a way that would allow for efficient queries later on. This is done by giving each node a *distance label*, and then answering shortest path queries by reading these labels.

In general, the asymptotic lower bounds on the size of distance labels are not very optimistic. Even for sparse graphs, the lower bound is in $\Omega(\sqrt{n})$, where $n$ is the number of nodes. For a general graph the lower bound is linear. Even though in a theoretical setting these bounds don't look too terrible, one can imagine that in a practical application even $\Omega(\sqrt{n})$ can already be too much.

TNR-algorithms try to reduce the size of distance labels by essentially "pruning" unnecessary entries from the distance labels by storing only transit nodes, thereby side-stepping the lower bounds mentioned in the paragraph above. So, in order to bound the size of distance labels produced by a TNR-algorithm, one has to bound the size of hub sets.

In an attempt to analyze the efficiency of TNR-algorithms, a parameter called *highway dimension* was introduced in [2]. The skeleton dimension is an improvement on this parameter, as the exact computation of the highway dimension, which is based on a problem called "hitting set problem", is known to be NP-hard. In order to better introduce the skeleton dimension, we first define a couple of key concepts.

**Definition 1.1**: The *tree skeleton* $T^*$ of $T$ is defined as the subtree of $\tilde{T}$ (geometric realization of $T$) with nodes whose reach is at least half its distance from the root, i.e. $\{v \in V(\tilde{T}) | Reach_{\tilde{T}}(v) \geq \frac{1}{2} d_{\tilde{T}}(u, v)\}$, where $u$ is the root of $T$ and *Reach* is the distance to the furthest node away from $v$.

We note that the $\frac{1}{2}$ in the above definition is chosen arbitrarily by the authors of the paper. In general, one could define a skeleton with any parameter $\alpha > 0$, as we will see later.

**Definition 1.2**: *Tree width* is defined as the maximum number of nodes at any distance from root $u$, i.e. $Width(T) := max_{r>0} |Cut_r(d_{\tilde{T}}(u, v))|$, where $Cut_r(d_{\tilde{T}}(u, v))$ is the set of nodes $v \in V(\tilde{T})$ with $d_{\tilde{T}}(u, v) = r$.

**Definition 1.3**: *Skeleton dimension* $k$ of a graph $G$ is the maximum width of the skeleton of a shortest path tree, i.e. $k = max_{u \in V(G)} Width(T_u^*)$, where $T_u^*$ is the shortest path tree of $u$, obtained as the union of all shortest paths from $u$ to every other node in $G$.

As one can notice, the skeleton dimension gives us a sort of lower bound on the amount of branches of shortest path trees that we can prune in a graph (more pruning is essentially reducing the width of the tree skeleton). Taking the union of all shortest paths, it is natural to assume that transit nodes will appear on a lot of these paths. One can therefore have a dramatic decrease in the amount of information to be stored at every node. Consider the simple example in Figure 1.

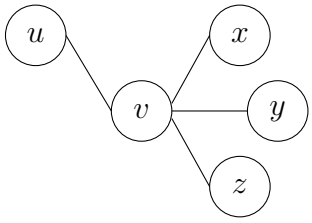


Figure 1

We now look at node $u$. Storing the lengths of all paths $\{u, v, x\}, \{u, v, y\}, \{u, v, z\}$ is superfluous, as node $v$ appears on all of them. Node $v$ is in this a case a *transit node*, and the *hub set* of node $u$ is $S(u) = \{v\}$. It is also worth noting that the hub sets for all nodes except $v$ in this example are the same. Thus, if we want to answer a shortest path query for $u$ and $x$, we look at $S(u) \cap S(x) = \{v\}$ and know that we have to go through $v$ to reach $x$, and return the distance from $u$ to $x$ as the distance from $u$ to $v$ plus the distance of $v$ to $x$. We can see that merely defining hub sets naturally leads to a reduction in the amount of information we need to store.

# 2 Comparing Highway Dimension and Skeleton Dimension

We recall that the highway dimension of a graph $G$ is defined as a (relatively) small integer $h$, such that for any ball with radius $r$, the set $S$ of vertices of $G$ with $|S| \leq h$ covers all shortest paths greater than $r$ within this ball.

## 2.1 Geometric Highway Dimension as an Upper Bound for Skeleton Dimension

In this section we present a claim that shows that in the worst case, the skeleton dimension and highway dimension of a graph are equal. It is important to note that in the proof of the following claim, the authors show the geometric highway dimension as an upper bound. However, in road networks the continuous and discrete versions ("continuous" and "discrete" here refer to the length function) of the highway dimension are expected to coincide almost exactly, as noted in [1]. The proof of the lemma is not so important, but the result is.

**Claim 1.** *If the geometric realization $\tilde{G}$ of a graph $G$ has highway dimension $\tilde{h}$, then $G$ has skeleton dimension $k \leq \tilde{h}$.*

## 2.2 Example of a Gap between Skeleton Dimension and Highway Dimension

In order to illustrate that there can be a significant gap between skeleton and highway dimension, the authors give an example of a family of graphs with an exponential gap between the two. This setting is the so-called Manhattan-type road network. It is a square grid of size $2^L \times 2^L$ with edge lengths, for which nodes whose coordinates are multiples of high powers of 2 have slightly lower transit times.
They also define the length function in a way that guarantees uniqueness of shortest paths by introducing small length perturbations for vertical and horizontal edges.
The result of all this is the following propsition.

**Proposition 2.** *For any $L > 0$, grid $G_L$ has highway dimension $\Omega(\sqrt{n})$ and skeleton dimension $O(\log n)$, where $n = 2^{2L}$ is the number of node in $G_L$.*

The authors note however that there exist length functions, for which the skeleton dimension of the grid is as high as $\Theta(\sqrt{n})$. So, as one can see already, the skeleton dimension of a graph heavily depends on the length function used to measure distances in the graph. This is perhaps not that surprising if one recalls Definition 1.1 and Definition 1.3. The nodes that compose the tree skeleton directly depend on the distance function, which in turn is defined in terms of the length fucntion. Skeleton dimension is defined in terms of the tree skeleton, and therefore also depends on the length function. We will see more examples that illustrate this dependence later on.

We finish off this section with some results from practical experiments. The authors computed the skeleton dimension of a New York travel-time graph and it turned out to be $k = 73$, with average skeleton width of 30. The highway dimension for the same graph is $h \geq 172$.

# 3 Hub Labeling using Tree Skeletons

In this section we will introduce one of the key results of the paper, namely, we shall use the skeleton dimension to obtain an asymptotic bound on the size of hub sets of a graph. We will first present the result and then give an outline of the proof.

**Theorem 1.** *With probability at least* $1 - \frac{8}{n}$*, all nodes* $u \in V$ *satisfy the following bound on hub set size:*

$$S(u) \leq 2\sum_{r=0}^{19} |Cut_u^{*(r)}| + 24 \ln n + 2 \max_{\gamma_i} \sum_{i=1,2,\ldots,\lfloor 16 \ln D \rfloor} l_i \ln \gamma_i.$$

With the help of the skeleton dimension and the concavity of the logarithm function we can derive the following useful corollaries from Theorem 1.

**Corollary 2.** *With probability at least* $1 - O(\frac{1}{n})$*, the hub set size of every node is bounded by:*

$O(k \log D \max\{1, \frac{\log n}{\log D}\})$*, where* $D$ *is the diameter of the graph.*

**Corollary 3.** *With probability at least* $1 - O(\frac{1}{n})$*, the hub set size of every node is bounded by:*

$$O(k \log \log k \log D).$$

Corollary 3 is obtained by substituting $\log n = \log k \cdot \log D$ into the bound in Corollary 2.

As was observed in the example with Figure 1. limiting the hub set size equals to limiting the size of distance labels. With the results above, the authors show a clear connection between hub set size and skeleton dimension. Moreover, they imply that constructing hub sets with the help of tree skeletons with high probability will bypass the pessimistic lower bounds seen in the introduction of this report, which is a good thing to know for anyone that would like to use a TNR-algorithm.
We now proceed to give an intuitive overview of the proof for Theorem 1, as the proof itself is very technical and long. In the overview, we will work with an integer weighted graph, for which we will emulate the geometric realization of the graph by inserting additional nodes with degree 2 into the graph and sub-dividing every edge with length $\ell(v, w)$ into $12\ell(v, w)$ small fragments of size $\frac{1}{12}$ each (constants chosen for ease of analysis). After this process, we can treat the graph as unweighted.

# Proof Overview for Theorem 1.

We first introduce an important definition that will come up throughout the proof.

**Definition 1.4:** $P_u(v, w)$ denotes the unique path between nodes $v$ and $w$ in the tree $T_u$. Also, $P_u(u, v) = P_u(v)$.

## 3.1 Constructing Hub Sets

We construct the hub sets in a randomized fashion by first introducing the concept of the *central sub-path*.

**Definition 1.5:** A *central sub-path* $P'_u(v) \subseteq P_u(v)$ is a sub-path of $P_u(v)$ that contains the middle $\frac{d_u(v)}{6}$ edges.

The hub set of a node $v$ is defined as $S(u) := \{\eta_u(v) | v \in V, v \neq u\}$, where $\eta_u(v) := \arg\min_{e \in P'_u(v)} \rho(e)$ and $\rho(e)$ is a real value in the interval $[0, 1]$ that is assigned to every $e \in E$ uniformly and independently at random. This is done to simulate the uniqueness of shortest paths. Intuitively one can view this as "choosing arbitrary transit nodes" in the central sub-path.

## 3.2 Bounding Average Hub Set Size

We present the following lemma that will give us a bound on the hub set size. The term $\hat{k}(u)$ is directly related to the skeleton dimension per the inequality $\hat{k}(u) \leq O(k \log D)$.

**Lemma 2.** *The expected hub set size of a node $u \in V$ satisfies the bound:*

$$\mathbb{E}[|S(u)|] \leq 16\hat{k}(u).$$

The idea behind the proof for this lemma is as follows. For every node $y$ in a given tree skeleton, we look at a distinct node $x_y$ with $d_u(x_y) = \lfloor \frac{7}{8} d_u(y) \rfloor$ (constant chosen with respect to another lemma from the paper that is omitted). We now look at the path $P_u(x_y, y)$, and more specifically at the probability that one of it's extreme edges is part of the hub set of the two nodes. This is given by an indicator random variable and bounded as:

$$\frac{2}{|P_u(x_y, y)|} \leq \frac{2}{d_u(y) - \frac{7}{8} d_u(y)} = \frac{16}{d_u(y)}.$$

Summing over all $y$ in the tree skeleton and applying linearity of expectation results in the claim of the lemma, where $\hat{k}$ is shorthand for the above-mentioned sum.

A direct application of Markov's inequality to the bound of Lemma 2, as well as the equation $\hat{k}(u) \leq O(k \log D)$, results in the following corollary.

**Corollary 1.** *The average hub set size satisfies*
$\frac{1}{n} \sum_{u \in V} |S(u)| = O(\frac{1}{n} \sum_{u \in V} \hat{k}(u)) \leq O(k \log D)$*, with probability at least 1/2 w.r.t. choice of random values $\rho$.*

With this we see that the hub set size can indeed be bounded with the help of the skeleton dimension.

## 3.3 Concentration Bound on Hub Set Size

Now that we have a bound on the expected size, we now want to tighten this bound, i.e. show that the minimum and maximum size don't lie too far away from the expectation (an observant reader may have noticed that the bound in Corollary 3 is $O(\hat{k}(u) \log \log k)$). Obtaining a concentration bound on hub set size is much more tricky. This is achieved in three steps:

(i) Partitioning $T_u^*$ into Layers

(ii) Bounding the size of $X^-(\eta)$

(iii) Bounding the size of $X^+(\eta)$

Combining the bounds above will give us Theorem 1. We first, however define some random variables.

$$|S(u)| = \sum_{\eta \in E(T_u^*)} X(\eta),$$

where $X(\eta) \in \{0, 1\}$ is an indicator variable for the event "$\eta \in S(u)$". $X^+(\eta)$ and $X^-(\eta)$ is a decomposition of $X(\eta)$ into contributions from paths located towards the root and away from the root respectively (with respect to $\eta$).

**Partitioning $T_u^*$ into Layers**

We partition the edges of $T_u^*$ as follows. We divide our tree skeleton into layers $L^{*[i]}$ that contain edges located between certain radii. These radii are given as an increasing sequence starting at 0. We then define $L^{**[i]} := L^{*[i]} \cap E(T_u^{**})$. These are essentially edges of the tree skeleton of the tree skeleton of the original graph. We can use this partition to show the following lemma.

**Lemma 3.** *For all $i \geq 1$, edge set $L^{**[i]}$ admits a partition into paths, $L^{**[i]} = \cap_{j=1}^{l_i} P^{[i,j]}$, s.t. $l_i < 2 \min_{r \in [r^{[i+1]}, r^{[i+2]}]} |Cut_u^{*(r)}| \leq 2k$, each $P^{[i,j]}$ is a descending path, and all internal nodes of all such paths have degree exactly 2 in $T_u^{**}$.*

We have now constructed a sequence of partitions within partitions and shown that at the lowest level, we obtain a bound that depends on the skeleton dimension. What's left is to now "propagate" this bound upwards and show that the whole thing ($|S(u)|$), can also be bounded using the skeleton dimension.

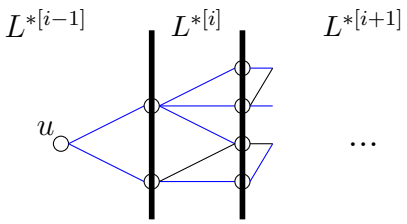


Figure 2. A visualization of the partitions $L^{*[i]}$.
The edges in $L^{**[i]}$ are colored blue. The thick vertical lines are the radii.

**Bounding the Sum of $X^+(\eta)$**
In order to bound this sum, the authors consider an arbitrary edge $\eta$ which doesn't belong to layers 0 or 1 in the tree partition and use the decomposition into paths from Lemma 3. From this, it follows that there exists exactly one path from the partition that $\eta$ is a part of. They then observe that in order for the claim to hold, two conditions must be satisfied:

(i) *Prefix minimum condition*: $\eta = \arg\min_{e \in P^{[i,j]} \cap P_u(\eta^-)} \rho(e)$.

(ii) $\rho(\eta) < \min \rho(Q^{[i,j]})$, where $Q^{[i,j]}$ is a descending path in $T_u$ that extends to $P^{[i,j]}$.

The proof continues with bounding the sum in the following way. We consider the set of all edges the satisfy (ii) and then introduce an indicator random variable for the event that an edge from this set satisfies (i) and then consider the sum of these variables. As a result, the following bound is obtained:

$$\sum_{\eta \in E(T_u^{**})} X^+(\eta) \leq \sum_{r=0}^{19} |Cut_u^{*(r)}| + 6c \ln n + \max_{(\gamma_i)} \sum_{i \leq i_{max}} l_i \ln \gamma_i.$$

**Bounding the Sum of $X^-(\eta)$**
This bound is the same as the bound for $X^+(\eta)$ with some technical differences that are omitted here.

**Combining the Bounds** We now see that by combining the bounds for $X^+(\eta)$ and $X^-(\eta)$, we get:

$$|S(u)| \leq 2 \sum_{r=0}^{19} |Cut_u^{*(r)}| + 12c \ln n + 2 \max_{(\gamma_i)} \sum_{i \leq i_{max}} l_i \ln \gamma_i.$$

By applying Lemma 3, using the bound $i_{max} \leq 16 \ln D$ and setting $c = 2$, we obtain the result of Theorem 1.

# 4 Computing Skeleton Dimension and Additional Results

One of the main results in the paper is the ease with which one can compute the skeleton dimension relative to the NP-hard computation of the highway dimension. Obtaining a discrete tree skeleton can be done with a scan of the vertices in reverse topological order. Its width $k$ can then be computed by a scan of the vertices of the tree skeleton and storing edges that are incident to vertices in $Cut_r(T)$ in a priority queue. This can be done in $O(n \log \log k)$[3]. The skeleton of a graph can be obtained with an all pairs shortest path computation in $O(nm + n^2 \log \log n)$[3].
Obtaining a hub set can be done in $O(n \log C \log(n \log C))$. As a result, we can compute distance labels in expected time $O(nm + n^2 \log C(\log n + \log \log C))$. The authors note however that with shared randomness, the distance labels can be computed independently(e.g. in parallel), we can obtain the labels in $O(mn \log C(\log n + \log \log C))$.

Some interesting results include that using different length fucntions in the same graph can lead to different results. In the New York graph mentioned in Section 2, measuring distance between nodes as travel time, geographical distance and hop count yielded different skeleton dimensions (73, 66 and 56 respectively).
Another result is that one can choose an arbitrary cutoff point in Definition 1.3, i.e. instead of $\frac{1}{2}$, one can choose an arbitrary $\alpha > 0$, resulting in a whole new family of related parameters with $\alpha < \beta \implies k_\beta < k_\alpha$, whose behavior can then be studied w.r.t. $\alpha$ and $\beta$.

# 5 Conclusion

The take-home message here is that despite the very technical proofs presented in this paper, the final result is very beneficial for practical purposes. The skeleton dimension is an example of something that started as an empirical observation, and in the end turned into a "side-stepping" of lower bounds, something that can help apply distance labeling to larger networks (e.g. Google Maps, large road networks, etc.). In a sense, it is a bridge between theory and practice. It shows that theoretical results do not simply abide in an ivory tower separate from practice. They can actually, if done right, help solve very practical problems.
On the other hand, the paper also shows that practical concerns can be used to push theoretical knowledge. If we were satisfied with the decent-looking

bounds on distance label size, research probably never would have gone into skeleton dimension and related works.

# 6 References

[1] Adrian Kosowski and Laurent Viennot. Beyond Highway Dimension: Small Distance Labels Using Tree Skeletons. SODA 2017 - 28th ACM-SIAM Symposium on Discrete Algorithms, Jan 2017, Barcelona, Spain.

[2] Ittai Abraham, Daniel Delling, Amos Fiat, Andrew Goldberg, and Renato Werneck. Highway dimension and provably efficient shortest path algorithms. Technical report, September 2013.

[3] Mikkel Thorup. Integer priority queues with decrease key in constant time and the single source shortest paths problem. Journal of Computer and System Sciences, 69(3):330  353, 2004.