Practical Construction of Metric t-Spanners *

Gonzalo Navarro † Rodrigo Paredes †

Abstract

Let $G(V, A)$ be a connected graph with a nonnegative cost function $d : A \rightarrow \mathbb{R}^+$. Let $d_G(u, v)$ be the cost of the cheapest path between $u, v \in V$. A t-spanner of G is a subgraph $G'(V, E)$, $E \subseteq A$, such that $\forall\ u, v \in V,\ d_{G'}(u, v) \leq t \cdot d_G(u, v),\ t > 1$. We focus on the metric space context, which means that $A = V \times V$, d is a metric, and $t \leq 2$. Several algorithms to build t-spanners are known, but they do not seem to apply well to our case. We present four practical algorithms to build t-spanners with empirical time costs of the form $C_t \cdot n^{2+\frac{0.1...0.2}{t-1}}$ and number of edges of the form $C_e \cdot n^{1+\frac{0.1...0.2}{t-1}}$. These algorithms are useful on general graphs as well.

1 Introduction

Let G be a connected graph $G(V, A)$ with a nonnegative cost function $d(e)$ assigned to its edges $e \in A$. The shortest path among every pair of vertices $u, v \in V$ is the one minimizing the sum of the cost of the edges traversed, $d_G(u, v)$. This can be computed with Floyd's algorithm or with $|V|$ iterations of Dijkstra's algorithm considering each vertex as the origin node [18]. A t-spanner it is a subgraph $G'(V, E)$, with $E \subseteq A$, which permits to compute paths with *stretch* t, that is, ensuring that $\forall\ u, v \in V,\ d_{G'}(u, v) \leq t \cdot d_G(u, v)$ [13]. We call this the t-condition.

In this work we are interested in using t-spanners as tools for searching metric spaces [6]. A metric space is a set of objects $\mathbb{X}$ and a distance function d defined among objects, which satisfies the metric properties (positiveness, reflexivity, symmetry, triangle inequality). Given a finite subset $\mathbb{U} \subseteq \mathbb{X}$, of size n, the goal is to build a data structure over $\mathbb{U}$ such that later, given a query object $q \in \mathbb{X}$, one can find the elements of $\mathbb{U}$ close to q with as few distance computations as possible.

*This work has been supported in part by the Millenium Nucleus Center for Web Research, Grant P01-029-F, Mideplan, Chile and MECESUP Project UCH0109 (Chile).

†Center for Web Research, Dept. of Computer Science, University of Chile. Blanco Encalada 2120, Santiago, Chile. {gnavarro,raparede}@dcc.uchile.cl

One of the best existing algorithms to search metric spaces is AESA [17]. AESA precomputes and stores the matrix of $n(n-1)/2$ distances among elements of $\mathbb{U}$. This huge space requirement makes it unsuitable for most applications, however.

This matrix can be seen as a complete graph $G(V, A)$ where the set of vertices $V = \mathbb{U}$ corresponds to the objects of the metric space, and the set of edges A corresponds to the $n(n-1)/2$ distances among these objects. A t-spanner G' of G would represent all these distances using a small number of edges $E, E \subseteq A$, and still would be able to approximate all the distances with a maximum error t, that is:

$$d(u, v) \quad \leq \quad d_{G'}(u, v) \quad \leq \quad t \cdot d(u, v) \tag{1.1}$$

In most metric spaces the distance histogram follows a distribution that becomes concentrated as the dimension increases [6]. This means that in practice we are interested in the range $t \in (1, 2]$.

We pursue this line in [12], where we focus on the search process but not on t-spanner construction. In that paper we show that the search algorithm is competitive against current approaches, e.g., we need 1.09 times the time cost of AESA using only 3.83% of its space requirement, in a metric space of documents; and 1.5 times the time cost of AESA using only 3.21% of its space requirement, in a metric space of strings. We also show that t-spanners provide better space-time tradeoffs than classical alternatives such as pivot-based indexes.

There are metric spaces where computing the distance evaluation among the objects is highly expensive. For instance, in the metric space of documents under the cosine distance [3], in order to compute the distance numerous disk accesses and million of basics arithmetics operations are required. In this case the distance evaluation could take hundredths of seconds, which is really expensive even compared against the operations introduced by the graph. In particular, the cost of the distance evaluation absorbs the cost of the shortest paths computation using Dijkstra's algorithm.

Hence our interest in this paper is in building t-spanners over metric spaces which work well in practice. Few algorithms exist apart from the basic $O(mn^2)$ technique ($m = |E|$), which inserts the edges needed

one by one and recomputes all the shortest paths to every edge inserted.

Four t-spanner construction algorithms are presented in this paper, with the goals of decreasing CPU and memory cost and of producing t-spanners of good quality, i.e., with few edges. Our four algorithms are:

1. An optimized basic algorithm, where we limit the propagation of edge insertions.
2. A massive edge insertion algorithm, where we amortize the cost of recomputing distances across many edge insertions.
3. An incremental algorithm, where nodes are added one by one to a correct t-spanner.
4. A recursive algorithm applying a divide and conquer technique.

Table 1 shows the complexities obtained. We obtain empirical time costs of the form $C_t \cdot n^{2.24}$ and number of edges of the form $C_e \cdot n^{1.13}$. This shows that good quality t-spanners can be built in reasonable time (just the minimum spanning tree computation needs $O(n^2)$ time). We take no particular advantage of the metric properties of the edge weights, so our algorithms can be used on general graphs too. As far as we know, there has not been previous work on comparing, in practice, t-spanner construction algorithms on metric spaces.

	CPU time	Memory	Distance evaluations
Basic	$O(mn^2)$	$O(n^2)$	$O(n^2)$
Basic optimized	$O(mk^2)$	$O(n^2)$	$O(n^2)$
Massive edge insertion	$O(nm \log m)$	$O(m)$	$O(nm)$
Incremental	$O(nm \log m)$	$O(m)$	$O(n^2)$
Recursive	$O(nm \log m)$	$O(m)$	$O(n^2)$

Table 1: t-Spanner algorithm complexities comparison. The value k refers to the number of nodes that have to be checked when updating distances due to a new inserted edge.

2 Previous Work

Several studies on general graph t-spanners have been undertaken [8, 13, 14]. Most of them resort to the basic $O(mn^2)$ time construction approach detailed in the next section, where $n = |V|$ and $m = |E|$ refer to the resulting t-spanner. It was shown in [1, 2] that this technique produces t-spanners with $n^{1+O(\frac{1}{t-1})}$ edges on general graphs of n nodes.

More sophisticated algorithms have been proposed in [7], producing t-spanners with guaranteed $O(n^{1+(2+\varepsilon)(1+\log_n m)/t})$ edges in worst case time $O(mn^{(2+\varepsilon)(1+\log_n m)/t})$, where in this case m refers to the original graph. In a metric space $m = \Theta(n^2)$, which means worst case time $O(n^5)$. Additionally, the algorithms in [7] work for $t \in [2, \log n]$, unsuitable for our application. (Some of these algorithms could be adapted to work heuristically for smaller t, but to the best of our knowledge, this has not been done so far.) Other recent algorithms [16] work only for $t = 1, 3, 5, \ldots$ also unsuitable for us. Parallel algorithms have been pursued in [11], but they do not give new sequential algorithms.

As it regards to Euclidean t-spanners, i.e., the subclass of metric t-spanners where the objects are points in a D-dimensional space with Euclidean distance, much better results exist [8, 1, 2, 10, 9, 15], showing that one can build t-spanners with $O(n)$ edges in $O(n \log^{D-1} n)$ time. These results, unfortunately, make heavy use of coordinate information and cannot be extended to general metric spaces.

Other related results refer to probabilistic approximations of metric spaces using tree metrics [4, 5]. The idea is to build a set of trees such that their union makes up a t-spanner with high probability. However, the t values are of the form $O(\log n \log \log n)$.

Hence the need to find practical algorithms that allow building appropriate t-spanners for metric spaces, that is, with $t \leq 2$, for complete graphs, and taking advantage of the triangle inequality.

3 Basic t-Spanner Construction Algorithm

The intuitive idea to solve this problem is iterative. We begin with an initial t-spanner that contains all the vertices and no edges, and calculate the distance estimations among all vertex pairs. These are all infinite at step zero, except for the distances between a node and itself ($d(u, u) = 0$). The edges are then inserted until all the distance estimations fulfill the t-condition.

The edges are considered in ascending cost order, so we start by sorting them. Using smaller-cost edges first is in agreement with the geometric idea of inserting edges between near neighbors and making up paths from low cost edges in order to use few edges overall.

Hence the algorithm uses two matrices. The first, *real*, contains the true distance between all the objects, and the second, *estim*, contains the distance estimations obtained with the t-spanner under construction. The t-spanner is stored in an adjacency list.

The insertion criterion is that an edge is added to the set E only when its current estimation does not

satisfy the t-condition. After inserting the edge, it is necessary to update *all* the distance estimations. The update mechanism is similar to the distance calculation mechanism of Floyd's algorithm, but considering that edges, not nodes, are inserted into the set. Figure 1 depicts the basic t-spanner construction algorithm.

```
t-Spanner0 (Stretch t, Vertices U)

real ← real distance matrix
estim ← estimated distance matrix
t-Spanner ← ∅ // t-spanner edge structure

for e = (e_u, e_v) ∈ real chosen in increasing
                          cost order do
  if estim(e) > t · real(e)
    // e is not well t-estimated
    t-Spanner ← t-Spanner ∪ {e}
    for v_i, v_j ∈ U
      d_1 ← estim(v_i, e_u) + estim(v_j, e_v)
      d_2 ← estim(v_j, e_u) + estim(v_i, e_v)
      estim(v_i, v_j) ← min(estim(v_i, v_j),
                             min(d_1, d_2)+real(e))
```

Figure 1: Basic t-spanner construction algorithm (t-Spanner 0).

This algorithm makes $O(n^2)$ distance evaluations, like AESA [17]; $O(mn^2)$ CPU time (recall that $n = |V|$ and $m = |E|$); and $O(n^2 + m) = O(n^2)$ memory. Its main deficiencies are excessive edge insertion cost and too high memory requirements.

4 Optimized Basic Algorithm

Like the basic algorithm (Section 3), this algorithm considers the use of *real* and *estim* matrices, choosing the edges in increasing weight order. The optimization focuses on the distance estimation update mechanism.

The main idea is to control the propagation of the computation, that is, only updating the distance estimations that are affected by the insertion of a new edge. Figure 2 shows the insertion of a new edge. In the first update we must modify only the edge that was inserted, between nodes a_1 and a_2. The computation then propagates to the neighbors of the a_i nodes, namely the nodes $\{b_1, b_2, b_3\}$; then to the nodes $\{c_1, c_2\}$ and finally d_1. The propagation stops when a node does not improve its current estimation or when it does not have further neighbors.

In order to control the propagation, the algorithm uses two sets, *ok* and *check*.

- *ok*: The nodes that already have updated their shortest path estimations due to the inserted edge.
- *check*: The adjacency of *ok*, $check = adyacency(ok) - ok = \{u \in \mathbb{U}, \exists v \in ok, (u, v) \in E\} - ok$. These are the nodes that we still need to update.

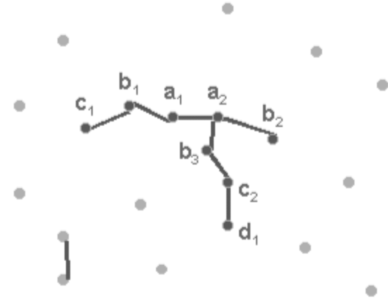


Figure 2: Propagation of distance estimations.

Note that it is necessary to propagate the computation only to the nodes that improve their estimation to a_1 or a_2. The complete algorithm reviews all the edges of the graph. For each edge, it iterates until no further propagation is necessary. Figure 3 depicts the optimized basic algorithm.

```
t-Spanner1 (Stretch t, Vertices U)

real ← real distance matrix
estim ← estimated distance matrix
t-Spanner ← ∅ // t-spanner edge structure

for e = (e_u, e_v) ∈ real chosen in increasing
                          cost order do
  if estim(e) > t · real(e)
    // e is not well t-estimated
    t-Spanner ← t-Spanner ∪ {e}
    ok ← {e_u, e_v}
    check ← adjacency(ok) − ok
    for c ∈ check
      if (estim(c, e_v) + real(e) ≤ estim(c, e_u) or
         (estim(c, e_u) + real(e) ≤ estim(c, e_v)
        for o ∈ ok
          d_1 ← estim(c, e_u) + estim(o, e_v)
          d_2 ← estim(c, e_v) + estim(o, e_u)
          estim(c,o) ← min(estim(c,o),
                           min(d_1,d_2) + real(e))
        check ← check ∪ (adjacency(c) - ok)
      ok ← ok ∪ {c}
      check ← check − {c}
```

Figure 3: Optimized basic algorithm (t-Spanner 1).

This algorithm takes $O(n^2)$ distances evaluations. In terms of CPU time it takes $O(mk^2)$, where k is the number of neighbors to check when inserting an edge. In the worst case this becomes $O(mn^2)$ just like the basic algorithm, but the average is much better. From the point of view of the memory it still takes $O(n^2 + m) = O(n^2)$. This algorithm reduces the CPU time used, but even so this is still very high, and the memory requirements are still too high.

A good feature of this algorithm is that, just like the basic algorithm, it produces good-quality t-spanners (few edges). So we have used its results to predict the expected number of edges per node in order to speed up other algorithms that rely on massive edge insertion. We call $E_{t-\mathrm{Spanner1}}(n, d, t)$ the expected number of edges in a metric space of n objects, distance function d, and stretch t. In Section 8 we show some estimations obtained, see Table 2.

5 Massive Edges Insertion Algorithm

This algorithm tries to reduce both the CPU processing time and memory requirements. To reduce the CPU time, the algorithm updates the distance estimations only after performing many edge insertions, using an $O(m \log n)$-time Dijkstra's algorithm to update distances. To reduce the memory requirement, it computes the distances between objects on the fly.

Since we insert edges less carefully than before, the resulting t-spanner is necessarily of lower quality. Our effort is in minimizing this effect.

The algorithm has three stages. In the first one, it builds the t-spanner backbone by inserting whole minimum spanning trees (MSTs), and determines the global wrongly t-estimated edge list (*pending*); in the second one, it refines the t-spanner by adding more edges to improve the wrongly t-estimated edges; and in the third one, it inserts all the remaining "hard" edges.

This algorithm uses two heuristic values:

H_1 determines the expected number of edges per node, and it is obtained from the t-Spanner 1 edge model: $H_1 = |E_{t-\mathrm{Spanner}\ 1}(n, d, t)|/n$. With H_1 we will define thresholds to determine whether or not to insert the remaining edges (those still wrongly t-estimated) of the current node. Note that $|E_{t-\mathrm{Spanner}\ 1}(n, d, t)|$ is an optimistic predictor of the resulting t-spanner estimated size using the massive edges insertion algorithm, so we can use $H_1 = |E_{t-\mathrm{Spanner}\ 1}(n, d, t)|/n$ as a lower bound of the number of edges per node.

H_2 is used to determine the *pending* list size and will give a criterion to determine when to insert an additional MST. The maximum *pending* list size is $H_2 = 1.2 \cdot |E|$, where E refers to the t-spanner under construction. We made preliminar experiments in order to fix this value. With values lower than 1.2 the algorithm takes more processing time without improving the number of edges, and with higher values the algorithm inserts more edges than necessary and needs more memory to build the t-spanner.

The algorithm stages are:

1. We insert successive MSTs to the t-spanner. The first MST follows the basics Prim algorithm [18], but the next MSTs are built using Prim over the edges that have not been inserted yet.

   We traverse the nodes sequentially, building the list of pending edges (wrongly t-estimated). At the same time, we insert successive MSTs and remove pending edges accordingly. Additionally, when the current node has no more than $H_1/2$ pending edges, we just insert them (since we only need a small set of edges in order to correct the distance estimations of this node). The insertion of MSTs continues as long as there are more than H_2 pending edges (note that H_2 depends on the current t-spanner size $|E|$).

   This stage continues until we review all the nodes. The output is the t-spanner backbone (*t-Spanner*) and the gobal list of pending edges (*pending*).

2. In the second stage we reduce the *pending* list. For this sake, we traverse the list of nodes with pending edges ($pendingNodes$), from more to less pending edges. For each such node, we check which edges have to improve their t-estimation and which do not (edges originally in the pending list may have become well t-estimated along the process). From the still wrongly t-estimated edges, we insert a set of the smaller cost edges of size $H_1/4$ and proceed to the next node (we need to insert more edges in order to improve the distance estimation; with values lower than $H_1/4$ the algorithm takes more processing time without improving the number of edges and with higher values the algorithm inserts more edges than necessary).

   This allows us to review in the first place the nodes that require more attention, without concentrating all the efforts in the same node.

   The process considers two special cases. The first one is that we have inserted more than n edges, in which case we regenerate and re-sort the $pendingNodes$ list and restart the process. The second one is that the pending list of the current node is so small that we simply insert its elements.

The output condition of the second stage is that the *pending* list size is smaller than $n/2$ (we made preliminar experiments in order to fix this value, and we obtained the best results with $n/2$).

3. We insert the *pending* list to the t-spanner.

Figure 4 depicts the massive edges insertion algorithm. This algorithm takes $O(nm)$ distance evaluations, $O(nm \log m)$ CPU time (since we run Dijkstra's algorithm once per node), and $O(n+m) = O(m)$ memory. It is easy to see that the space requirement is $O(m)$: the *pending* list is never larger than $O(m)$ because at each iteration of stage 1 it grows at most by n, and as soon as it becomes larger than $1.2 \cdot m$ (H_2) we take out edges from it by adding a new MST, until it becomes short enough. The CPU time comes from running Dijkstra's algorithm once per node at stage 1. At stage 2 we insert edges in groups of $O(m/n)$, running Dijkstra's algorithm after each insertion, until we have inserted $|pending| - n/2 = O(m)$ edges overall. This accounts for other n times we run Dijkstra's algorithm. Hence the $O(nm \log m)$ complexity.

This algorithm reduces both CPU time and memory requirements, but the amount of distance evaluations is very high ($O(nm) \geq O(n^2)$).

6 Incremental Node Insertion Algorithm

This version reduces the amount of distance evaluations to just $n(n-1)/2$, while preserving the amortized update cost idea.

This algorithm, unlike the previous ones, makes a local analysis of nodes and edges, that is, it makes decisions before having knowledge of the whole edge set. We insert the nodes one by one, not the edges. The invariant is that for nodes $1 \ldots i-1$ we have a well formed t-spanner, and we want to insert the i-th node to the growing t-spanner. Since the insertion process only locally analyzes the edge set, the resulting t-spanner is suboptimal.

For each new node i, the algorithm makes two operations: the first is to connect the node to the growing t-spanner using the cheapest edge (towards a node $< i$); the second one is to verify that the distance estimations satisfy the t-condition, adding some edges to node i until the invariant is restored. We repeat this process until we insert the whole node set.

We also use the H_1 heuristic, with the difference that we recompute H_1 at every iteration (since the t-spanner size changes). We fixed that the number of edges to insert at a time should be $\delta = H_1/(5 \cdot i)$ in order to reduce the processing time and the amount of edges inserted to the t-spanner. Inserting more edges at a time obtains lower processing time but the size of the t-spanner is increased; inserting less edges at a time, increases the processing times whitout decreasing the t-spanner size.

For the distance verification we use an incremental Dijkstra's algorithm with limited propagation, that is, the first time, Dijkstra's algorithm takes an array with precomputed distances initialized at $t \cdot d(u_i, u_j) + \varepsilon$, with $\varepsilon > 0$, $j \in [1, i-1]$. This is because, if a distance to node i is not well t-estimated, we do not really need to know how bad estimated it is. For the next iterations, Dijkstra's algorithm reuses the previously computed array, because there is no need to propagate distances from nodes whose estimation has not improved.

Figure 5 depicts the incremental node insertion algorithm. This algorithm takes $O(n^2)$ distance evaluations, $O(nm \log m)$ CPU time, and $O(n+m) = O(m)$ memory. The CPU time comes from the fact that every node runs Dijkstra's algorithm $n/\delta = O(1)$ times.

```
t-Spanner3 (Stretch t, Vertices 𝕌)

t-Spanner ← ∅ // t-spanner edge structure

for i ∈ [1, n]
  // incremental H1
  δ ← |E_{t-Spanner1}(i, d, t)| / (i · 5)
  k ← argmin_{j∈[1,i−1]} {d(node_i, node_j)}
  // inserting the cheapest edge
  t-Spanner ← t-Spanner ∪ {(node_i, node_k)}
  // defining the propagation limit
  distances ← {(node_j, t · d(node_i, node_j) + ε)/
                j ∈ [1, i − 1]}
  while node_i has wrongly t-estimated edges
    // incremental Dijkstra
    distances ← Dijkstra(t-Spanner, u,
                         distances)
    pending_i ← {(node_i, node_j), j < i/
          distance(node_j) > t · d(node_i, node_j)}
    smallest ← δ cheapest edges in pending_i
    t-Spanner ← t-Spanner ∪ smallest
```

Figure 5: Incremental node insertion algorithm (t-Spanner 3).

7 Recursive Algorithm

The incremental algorithm is an efficient approach to construct t-spanners, but it does not consider spatial proximity (or remoteness) among the objects. A way to solve this is that the set in which the t-spanner is incrementally built is made up of near objects. Following this principle, we present a solution that recursively divides the object set into two compact

```
t-Spanner2 (Stretch t, Vertices 𝕌)

t-Spanner ← MST // t-spanner edge structure, initially has the first MST
pending ← ∅ // global pending egde list
H_1 ← |E_{t−Spanner1}(n, d, t)| / n

Stage 1:  generating t-Spanner and pending
for u ∈ 𝕌
  if |pending| > 1.2 · |t-Spanner| // using H_2
    t-Spanner ← t-Spanner ∪ MST // built over the non-inserted edges
  distances ← Dijkstra(t-Spanner, u) // distances(v) = d_{t-Spanner}(u, v)
  for v ∈ 𝕌
    if distance(v) ≤ t · d(u, v)  pending ← pending − {(u, v)}
    else pending ← pending ∪ {(u, v)}
  if |pending(u)| ≤ H_1/2
    t-Spanner ← t-Spanner ∪ pending(u)
    pending ← pending − pending(u)

Stage 2:  Reducing pending
while |pending| > n/2
  pendingNodes ← nodes sorted in decreasing number of pending edges
  for u ∈ pendingNodes
    if more than n edges have been inserted break // special case 1
    if |pending(u)| < H_1/4 // special case 2
      t-Spanner ← t-Spanner ∪ pending(u)
      pending ← pending − pending(u)
    else
      distances ← Dijkstra(t-Spanner, u)
      for v ∈ pending(u)
        if distances(v) ≤ t · d(u, v)  pending ← pending − {(u, v)}
      smallest ← H_1/4 smallest edges ∈ pending(u)
      t-Spanner ← t-Spanner ∪ smallest
      pending ← pending − smallest

Stage 3:  t-Spanner ← t-Spanner ∪ pending
```

Figure 4: Massive edges insertion algorithm (t-Spanner 2), *pending*(u) denotes $\{e \in pending,\ \exists v,\ e = (u, v)\}$.

subsets, builds sub-t-spanners in the subsets, and then merges them.

For the initial set division we take two far away objects, p_1 and p_2, that we call *representatives*, and then generate two subsets: objects nearer to p_1 and nearer to p_2. Figure 6 shows the concept graphically. For the recursive divisions we reuse the representative as one of the two objects, and the element farthest to it as the other. The recursion finishes when we have less than 3 objects.

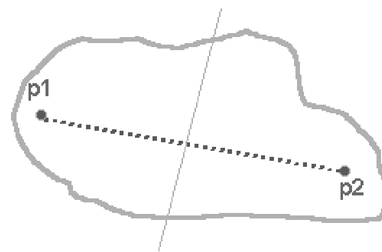


Figure 6: We select p_1 and p_2, and then divide the set.

The merge step also takes into account the spatial proximity among the objects. When we merge the sub-t-spanners, we have two node subsets V_1 and V_2, where $|V_1| \geq |V_2|$ (otherwise we swap the subsets). Then, in the sub-t-spanner represented by p_2 ($stsp_2$), we choose the object closest to p_1 (u), and insert it into the sub-t-spanner represented by p_1 ($stsp_1$) verifying that all the

distances towards V_1 are well t-estimated. Note that this is equivalent to consider that we use the incremental algorithm, where we insert u into the growing t-spanner $stsp_1$. We continue with the second closest and repeat the procedure until all the $stsp_2$ nodes are inserted into $stsp_1$. Figure 7 illustrates. Note that the edges already present in $stsp_2$ are conserved.

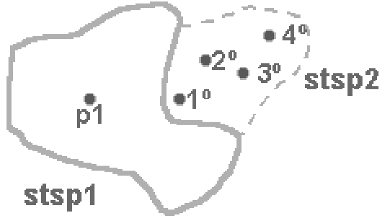


Figure 7: The merge step takes the objects according to their distances towards p_1.

This algorithm also uses an incremental Dijkstra's algorithm with limited propagation, but this time we are only interested in limiting the propagation towards $stsp_1$ nodes (because we know that towards $stsp_2$ we already satisfy the t-condition). Hence, Dijkstra's algorithm takes an array with precomputed distances initialized at $t \cdot d(u_i, u_j) + \varepsilon$ for $(u_i, u_j) \in V_2 \times V_1$, and ∞ for $(u_i, u_j) \in V_2 \times V_2$, where ε is a small positive constant. For the next iterations, Dijkstra's algorithm reuses the previously computed array.

Figure 8 depicts the recursive algorithm and the auxiliary functions used to build and merge sub-t-spanners. This algorithm takes $O(n^2)$ distance evaluations, $O(nm \log m)$ CPU time, and $O(n+m) = O(m)$ memory. The cost of dividing the sets does not affect that of the underlying incremental construction.

8 Experimental Results

We have tested our algorithms on synthetic and real-life metric spaces. The synthetic set is formed by 2,000 points in a 20-dimensional space with coordinates in the range $[-1, 1]$, with Gaussian distribution forming 256 randomly placed clusters. We consider three different standard deviations to make more crisp or more fuzzy clusters (σ = 0.1, 0.3, 0.5). Of course, we have not used the fact that the space has coordinates, but have treated the points as abstract objects in an unknown metric space.

Two real-life data sets were tested. The first is a string metric space using the edit distance (a discrete function that measures the minimum number of character insertions, deletions and replacements needed to make the strings equal). The strings form an English dictionary, where we index a subset of n = 24,000 words. The second is a space of 1,215 documents under the Cosine similarity, which is used to retrieve documents more similar to a query under the vector space model. In this model the space has one coordinate per term and documents are seen as vectors in this high dimensional space. The similarity corresponds to the cosine of the angle (inner product) among the vectors, and a suitable distance measure is the angle itself. Both spaces are of interest to Information Retrieval applications [3].

The experiments were run on an Intel Pentium IV of 2 GHz, with 512 MB of RAM and a local disk. We are interested in measuring the CPU time needed and the amount of edges generated by each algorithm. For shortness we have called t-Spanner 1 the optimized basic algorithm, t-Spanner 2 the massive edges insertion algorithm, t-Spanner 3 the incremental algorithm, and t-Spanner 4 the recursive algorithm.

Figures 9 and 10 show a comparison among the four algorithms on the Gaussian data set where we vary the stretch t and the amount of nodes, respectively. As it can be seen, all the algorithms produce t-spanners of about the same quality, although the optimized basic algorithm is consistently better, as explained. It is interesting to note that in the case of $\sigma = 0.1$ the t-spanner 2 has the worst edge performance. This is because, in its first stage, the algorithm inserts a lot of intra-cluster edges and then it tries to connect both inner and peripheral objects among the clusters. Since we need to connect just the peripheral objects, there are a lot of redundant edges that do not improve other distance estimations in the resulting t-spanner.

In the construction time, on the other hand, there are large differences. The optimized basic algorithm is impractically costly, as expected. Also, the massive edges insertion algorithm is still quite costly in comparison to the incremental and recursive algorithms. This is due to its large number of distance computations. This reinforces the idea that the t-spanner 2 is not suitable for metric spaces with highly expensive distance evaluation functions. However, we notice that, unlike all the others, this algorithm *improves* instead of degrading as the clusters become more fuzzy, becoming a competitive choice on uniformly distributed datasets. The quality of the t-spanner also varies from (by far) the worst t-spanner on crisp clusters to the second best on more fuzzy clusters. This could be due to two phenomena. The first is that there are less redundant edges among the clusters, and the second is that, on an uniform space, the t-spanner 2 inserts "better" edges since they come from MSTs (that usies the shortest possible edges).

The incremental and recursive algorithms are quite close in both measures, being by far the fastest al-

```
t-Spanner4 (Stretch t, Vertices 𝕌)

t-Spanner ← ∅ // t-spanner edge structure
(p_1, p_2) ← two distant objects
(V_1, V_2) ← 𝕌 divided according to distances towards (p_1, p_2)
stsp_1 ← makeSubtSpanner(p_1,V_1)
stsp_2 ← makeSubtSpanner(p_2,V_2)
t-Spanner ← mergeSubtSpanner(stsp_1, stsp_2)

makeSubtSpanner(representative p, Vertices V)
if |V| = 1 return t-spanner (nodes = {p}, edges = ∅)
else if |V| = 2 return t-spanner (nodes = V = {v_1, v_2}, edges = {(v_1, v_2)})
else
  p_remote ← argmax_{v∈V} {d(p, v)}
  (V, V_remote) ← V divided according to distances towards (p, p_remote)
  stsp_p ← makeSubtSpanner(p,V)
  stsp_remote ← makeSubtSpanner(p_remote,V_remote)
  return mergeSubtSpanner(stsp_p, stsp_remote)

mergeSubtSpanner (t-Spanner stsp_1, t-Spanner stsp_2)
if |nodes(stsp_1)| ≤ |nodes(stsp_2)| stsp_1 ⇔ stsp_2
nodes ← nodes(stsp_1) ∪ nodes(stsp_2)
edges ← edges(stsp_1) ∪ edges(stsp_2)
δ ← |E_{t−Spanner1}(|nodes|, d, t)| / (i · 5) // incremental H_1
p_1 ← representative(stsp_1)
for u ∈ nodes(stsp_2) in increasing order of d(u, p_1)
  // defining the propagation limit towards stsp_1
  for v ∈ nodes(stsp_1) do distances(v) ← t · d(u, v) + ε
  for v ∈ nodes(stsp_2) do distances(v) ← ∞
  while u has wrongly t-estimated edges towards stsp_1
    distances ← Dijkstra(edges, u, distances) // incremental Dijkstra
    pending_u ← {(u, v), v ∈ stsp_1/distance(v) > t · d(u, v)}
    smallest ← δ cheapest edges ∈ pending_u
    edges ← edges ∪ smallest
return t-Spanner (nodes = nodes, edges = edges)
```

Figure 8: Recursive algorithm (t-Spanner 4).

gorithms. The recursive algorithm usually produces slightly better t-spanners thanks to the more global edge analysis. Note that, for t as low as 1.5, we obtain t-spanners whose size is 5% to 15% of the full graph.

It is interesting to notice that, for crisp clusters, there is a big jump in the construction time and t-spanner size when we move from $t = 1.5$ to $t = 1.4$. The effect is much smoother for more fuzzy clusters. A possible explanation is that, for crisp clusters and large enough t, a single edge among cluster centers is enough to obtain a t-spanner. However, when t is reduced below 1.5, this becomes suddenly insufficient and we start having many edges across cluster pairs.

We show in Table 2 our least squares fittings on the data using the model $|E| = an^{1+\frac{b}{t-1}}$ and $time = an^{2+\frac{b}{t-1}}$ microseconds. This model has been chosen according to the analytical results of [1, 2]. As it can be seen, t-spanner sizes are slightly superlinear and times are slightly superquadratic. This shows that our algorithms represent in practice a large improvement over the current state of the art.

We show now some results on the metric space of strings, this time focusing on the behavior in terms of the database size n. Since these tests are more massive, we leave out the optimized basic and the massive edge insertion algorithms: They were really slow even for small subsets. This means, in particular for the massive edges insertion algorithm, that this space

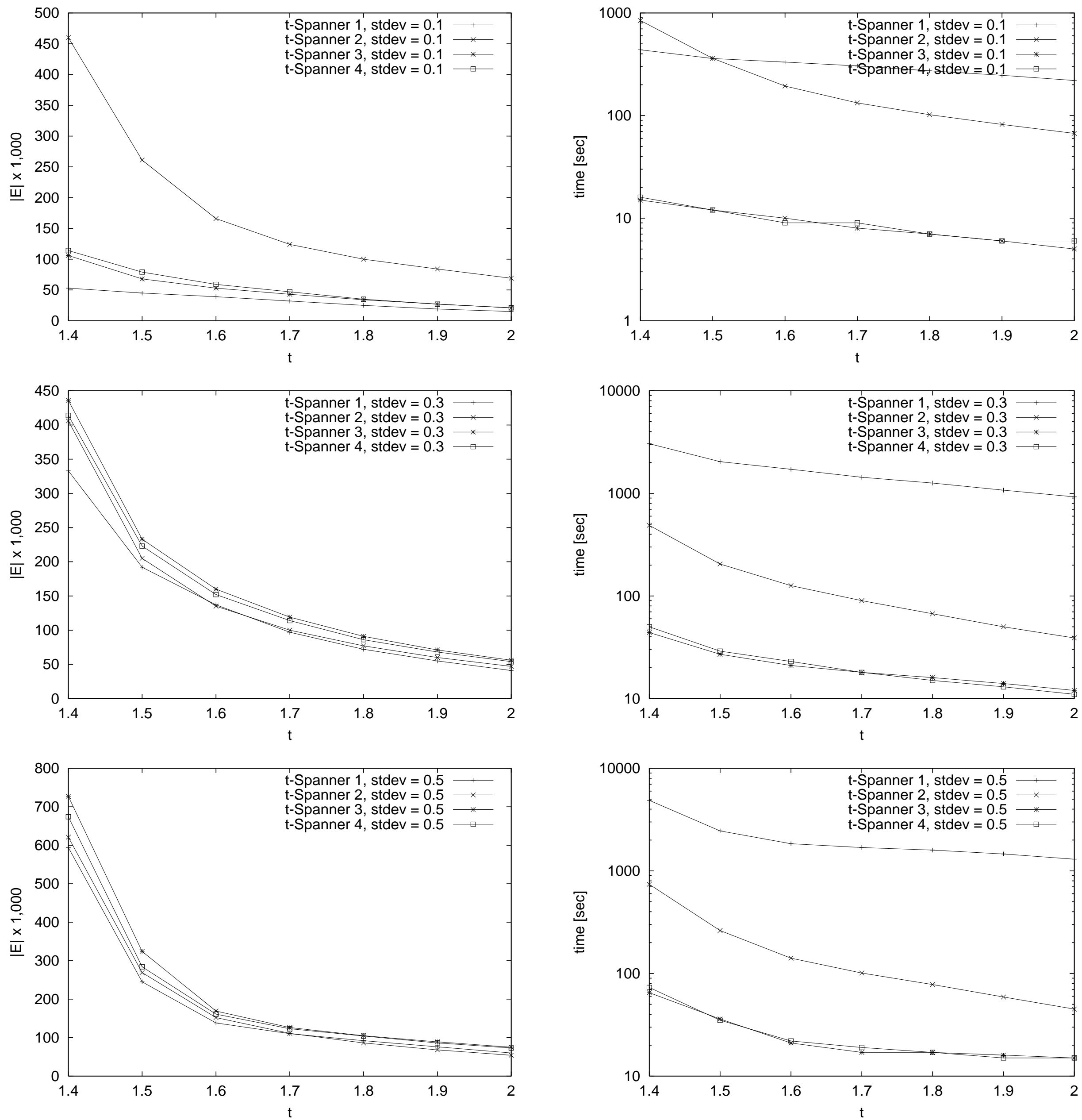


Figure 9: t-Spanner construction in the synthetic metric space of 2,000 nodes, as a function of t. On the left, edges generated (t-spanner quality). On the right, construction time. Each row corresponds to a different variance.

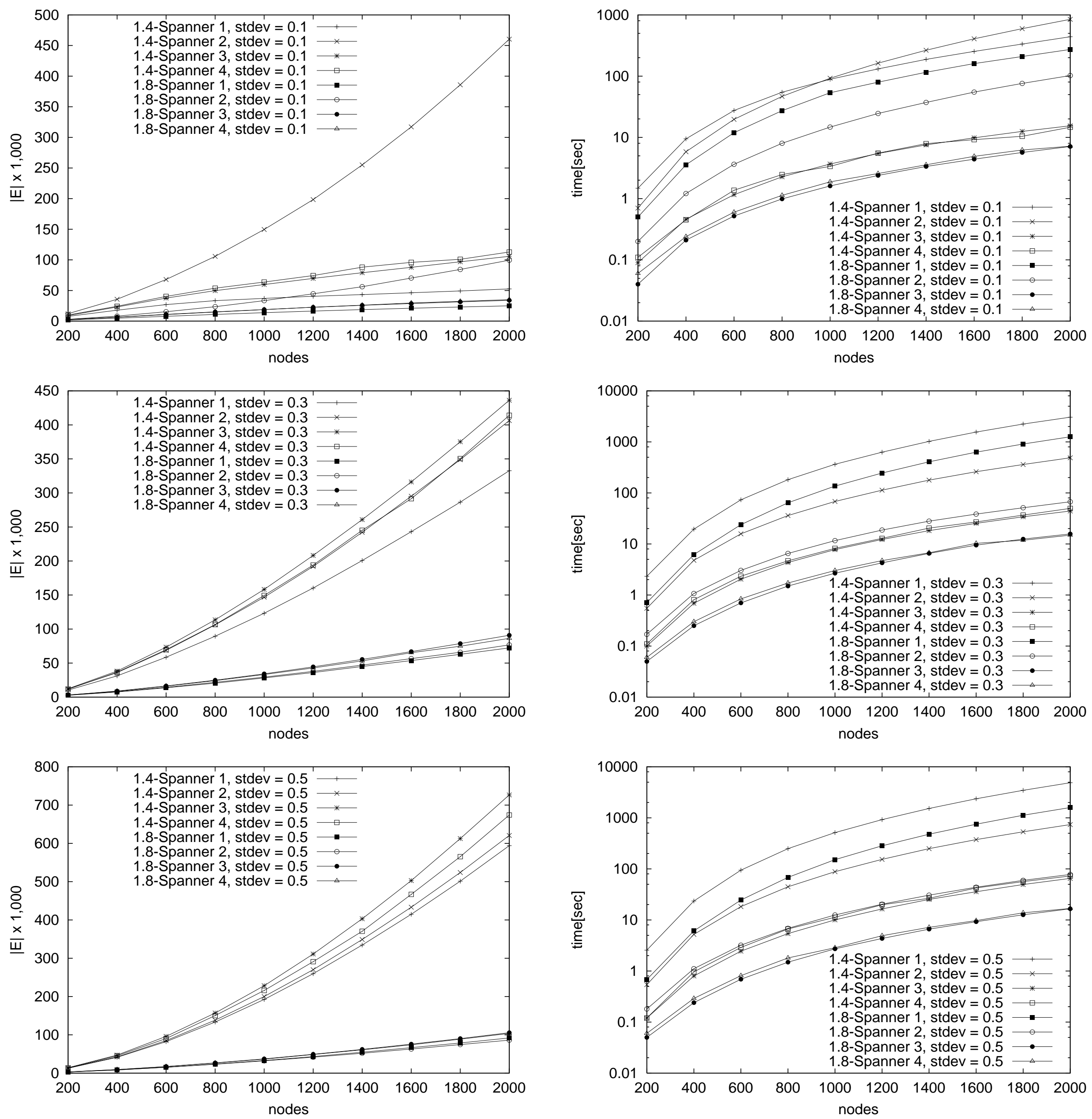


Figure 10: t-Spanner construction in the synthetic metric space of 2,000 nodes, as a function of the number of nodes. On the left, edges generated (t-spanner quality). On the right, construction time. Each row corresponds to a different variance.

	Basic optimized	Massive edge insertion	Incremental	Recursive
	Stdev 0.1			
CPU time	$17.8\ n^{2+\frac{0.09}{t-1}}$	$1.67\ n^{2+\frac{0.24}{t-1}}$	$0.670\ n^{2+\frac{0.10}{t-1}}$	$0.909\ n^{2+\frac{0.08}{t-1}}$
Edges	$5.76\ n^{1+\frac{0.10}{t-1}}$	$6.50\ n^{1+\frac{0.18}{t-1}}$	$6.17\ n^{1+\frac{0.13}{t-1}}$	$5.77\ n^{1+\frac{0.14}{t-1}}$
	Stdev 0.3			
CPU time	$25.0\ n^{2+\frac{0.16}{t-1}}$	$1.52\ n^{2+\frac{0.22}{t-1}}$	$0.771\ n^{2+\frac{0.13}{t-1}}$	$0.865\ n^{2+\frac{0.13}{t-1}}$
Edges	$5.69\ n^{1+\frac{0.18}{t-1}}$	$5.41\ n^{1+\frac{0.19}{t-1}}$	$6.52\ n^{1+\frac{0.19}{t-1}}$	$6.50\ n^{1+\frac{0.18}{t-1}}$
	Stdev 0.5			
CPU time	$21.0\ n^{2+\frac{0.19}{t-1}}$	$1.33\ n^{2+\frac{0.25}{t-1}}$	$0.587\ n^{2+\frac{0.17}{t-1}}$	$0.650\ n^{2+\frac{0.17}{t-1}}$
Edges	$4.89\ n^{1+\frac{0.21}{t-1}}$	$4.50\ n^{1+\frac{0.22}{t-1}}$	$5.20\ n^{1+\frac{0.22}{t-1}}$	$5.37\ n^{1+\frac{0.21}{t-1}}$

Table 2: Empirical complexities of our algorithms, as a function of n and t. Time is measured in microseconds.

is far from uniform. Figure 11 shows that, also for strings, the number of edges generated is slightly superlinear ($8.03\ n^{1+\frac{0.16}{t-1}}$ for the incremental algorithm and $8.45\ n^{1+\frac{0.15}{t-1}}$ for the recursive one), and the construction time is slightly superquadratic ($1.46\ n^{2+\frac{0.10}{t-1}}$ microseconds for the incremental algorithm and $1.67\ n^{1+\frac{0.09}{t-1}}$ for the recursive one). The recursive algorithm is almost always a bit better than the incremental algorithm in both aspects.

Finally, Figure 12 shows experiments on the space of documents. We have excluded the massive edges insertion algorithm, which was too slow. The reason this time is that it is the algorithm that makes, by far, more distance computations, which was clearly the dominant term in this space (comparing two document vocabularies takes several milliseconds). We can see again that, although all the algorithms produce t-spanners of about the same quality, the optimized basic algorithm is much more expensive than the other two, which are rather similar.

9 Conclusions

We have presented several algorithms for t-spanner construction when the underlying graph is the complete graph representing distances in a metric space. This is motivated by our recent research on searching metric spaces and shows that t-spanners are well suited as data structures for this problem. For this sake, we need practical construction algorithms for $1 < t \leq 2$. To the best of our knowledge, no existing technique has been shown to work well under this scenario (complete graph, metric distances, small t, practical construction time) and no practical study has been carried out on the subject. However, our algorithms are also well suited to general graphs.

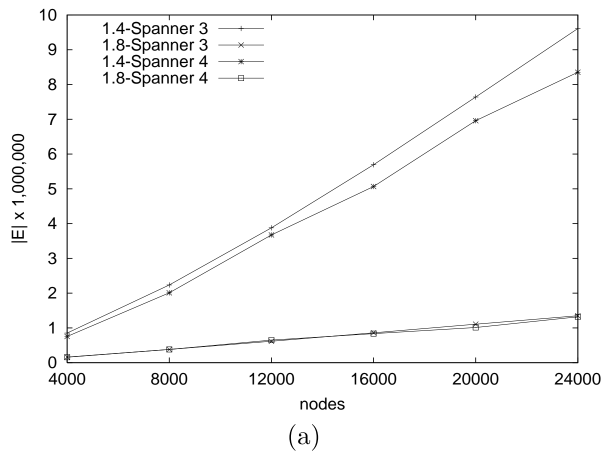
Our focus has been on practical algorithms. We have shown that it is possible to build good quality t-spanners in reasonable time. We have empirically obtained time costs of the form $C_t \cdot n^{2+\frac{0.1\ldots0.2}{t-1}}$ and number of edges of the form $C_e \cdot n^{1+\frac{0.1\ldots0.2}{t-1}}$. Note that just computing the minimum spanning tree requires $O(n^2)$ time. Moreover, just computing all the distances in a general graph requires $O(n^3)$ time. Compared to the existing algorithms, our contribution represents in practice a large improvement over the current state of the art. Note that in our case we do not provide a guarantee in the number of edges. Rather, we show that in practice we generate t-spanners with few edges with fast algorithms.

It is possible to add and remove elements from the t-spanner in reasonable time while preserving its quality. The incremental algorithm permits adding new elements. Remotion of a node can be arranged by adding a clique among its neighbors and periodically reconstructing the t-spanner with the recursive algorithm.

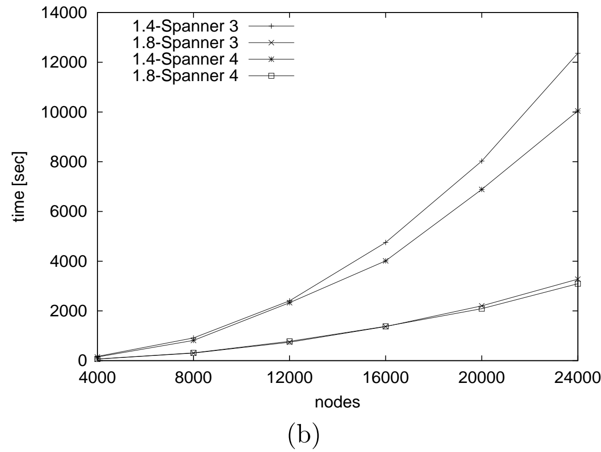
Future work involves using t-spanners where t depends on the actual distance between the nodes. Basically, we are more interested in approximating well short rather than long distances. On the other hand, we are investigating on fully dynamic t-spanners, which means that the t-spanner allows object insertions and deletions while preserving its quality. This is important when we use t-spanners in order to build an index for metric databases in real applications. Another trend is probabilistic t-spanners, where distances are well t-estimated with high probability, so that with much less edges we find most of the results.

Acknowledgement

The second author wishes to thank AT&T LA Chile for the use of their computer to run the experiments and for letting him continue his doctoral studies.

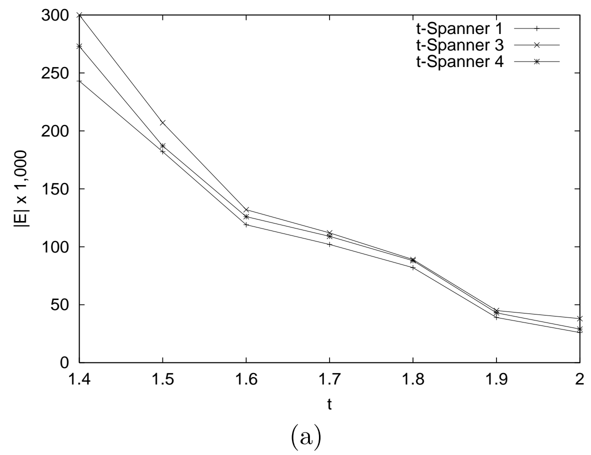


(a)

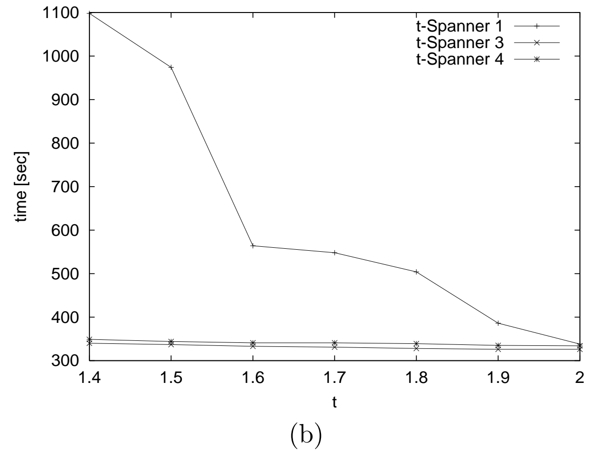


(b)

Figure 11: t-Spanner construction on the space of strings, for increasing n. (a) number of edges generated, (b) construction time.



(a)



(b)

Figure 12: t-Spanner construction on the set of documents, as a function of t. (a) number of edges generated, (b) construction time.

References

[1] I. Althöfer, G. Das, D. Dobkin, and D. Joseph. Generating sparse spanners for weighted graphs. In *Proc. 2nd Scandinavian Workshop on Algorithm Theory (SWAT'90)*, LNCS 447, pages 26–37, 1990.

[2] I. Althöfer, G. Das, D. Dobkin, D. Joseph, and J. Soares. On sparse spanners of weighted graphs. *Discrete Computational Geometry*, 9:81–100, 1993.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[4] Y. Bartal. On approximating arbitrary metrics by tree metrics. In *Proc. 30th Symposium on the Theory of Computing (STOC'98)*, pages 161–168, 1998.

[5] M. Charikar, C. Chekuri, A. Goel, S. Guha, and S. Plotkin. Approximating a finite metric by a small number of tree metrics. In *Proc. 39th Symp. on Foundations of Computer Science (FOCS'98)*, pages 379–388, 1998.

[6] E. Chávez, G. Navarro, R. Baeza-Yates, and J.L. Marroquin. Proximity searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, September 2001.

[7] E. Cohen. Fast algorithms for constructing t-spanners and paths with stretch t. *SIAM J. on Computing*, 28:210–236, 1998.

[8] D. Eppstein. Spanning trees and spanners. In *Handbook of Computational Geometry*, pages 425–461. Elsevier, 1999.

[9] J. Gudmundsson, C. Levcopoulos, and G. Narasimhan. Improved greedy algorithms for constructing sparse geometric spanners. In *Proc. 7th Scandinavian Workshop on Algorithm Theory (SWAT 2000)*, LNCS v. 1851, pages 314–327, 2000.

[10] J.M. Keil. Approximating the complete Euclidean graph. In *Proc. 1st Scandinavian Workshop in Algorithm Theory (SWAT'88)*, LNCS 318, pages 208–213, 1988.

[11] W. Liang and R. Brent. Constructing the spanners of graphs in parallel. Technical Report TR-CS-96-01, Dept. of CS and CS Lab, The Australian National University, January 1996.

[12] G. Navarro, R. Paredes, and E. Chávez. t-Spanners as a data structure for metric space searching. In *Proceedings of the 9th International Symposium on String Processing and Information Retrieval (SPIRE 2002)*, LNCS 2476, pages 298–309. Springer, 2002.

[13] D. Peleg and A. Schaffer. Graph spanners. *Journal of Graph Theory*, 13(1):99–116, 1989.

[14] D. Peleg and J. Ullman. An optimal synchronizer for the hypercube. *SIAM J. on Computing*, 18:740–747, 1989.

[15] J. Ruppert and R. Seidel. Approximating the d-dimensional complete Euclidean graph. In *3rd Canadian Conference on Computational Geometry*, pages 207–210, 1991.

[16] Mikkel Thorup and Uri Zwick. Approximate distance oracles. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 183–192. ACM Press, 2001.

[17] E. Vidal. An algorithm for finding nearest neighbors in (approximately) constant average time. *Patt. Recog. Lett.*, 4:145–157, 1986.

[18] Mark Allen Weiss. *Data Structures and Algorithm Analysis*. Addison-Wesley, 2nd edition, 1995.