Probabilistic embedding into trees: definitions and applications.
Fall 2011
Lecture 4

Instructor: Mohammad T. Hajiaghayi
Scribe: Anshul Sawant

September 21, 2011

1 Overview

Some problems which are hard on general metrics, can be trivially solved on more constrained metrics such as trees. This serves as the motivation for 'embedding' graphs into trees. In this lecture, we define what an embedding is and then study some applications of probabilistic embeddings of a graph into a distribution of trees.

2 Metric spaces and graphs

Definition 1 *A* metric space *is a set* X*, along with a metric* $d : X^2 \mapsto \mathbb{R}_{\geq 0}$*, such that:*

$$d(i,i) = 0 \ \forall i \in X$$
$$d(i,j) = d(j,i) \ \forall i,j \in X \qquad \text{[Symmetry]}$$
$$d(i,k) + d(k,j) \geq d(i,j) \ \forall i,j,k \in X \qquad \text{[Triangle inequality]}$$

A metric space is often represented as the pair (X,d). An example of metric spaces is $(\mathbb{R}^n, L_k)$, where L_k is the k-norm over $\mathbb{R}^n$ for given $n, k \in \mathbb{Z}_{\geq 1}$. We can represent a finite metric space (X, d) by a symmetric matrix S, of size $n \times n$, where $S_{i,j} = d(i,j)$ and $|X| = n$. Metric spaces can be visualized using undirected graph G, where S is distance matrix for G. Conversely, given a graph $G(V, E)$, we can represent it as a metric space (V, d), where $d(i,j)$ is length of the shortest path between $i, j \in V$.

2.1 Metric completion of a graph

Not all undirected graphs obey the metric property as distances between nodes can be assigned arbitrarily. However if we define $d_G(x, y)$ to be the length of shortest path between nodes of a graph G, then d_G satisfies the metric property. (G, d_G) is called the *metric closure* of G and this technique of converting a general graph to a metric is called *metric completion.*

3 Embeddings

Definition 2 *Given metric spaces* (X, d_X) *and* (Y, d_Y)*, an* embedding *is a mapping* $f : X \mapsto Y$*. A distance preserving mapping is called an* isometric mapping*.*

As pointed out earlier, we can use an embedding to map more general metrics to more constrained metrics. Such simplification usually has to be paid for my some 'distortion' of the original metric and hence isometric embeddings are generally not useful for this purpose.

Definition 3 *Contraction of an embedding is equal to* $\max_{x,y \in X} \left\{ \frac{d_X(x,y)}{d_Y(f(x),f(y))} \right\}$*. Expansion of an embedding is equal to* $\max_{x,y \in X} \left\{ \frac{d_Y(f(x),f(y))}{d_X(x,y)} \right\}$*.* Distortion*,* α*, of an embedding is defined as the product of its contraction and expansion.*

Distortion is invariant under scaling. Hence, we can assume an embedding to be non-contractive without affecting its distortion. In which case, distortion is equal to expansion and:

$$d_X(x, y) \leq d_Y(f(x), f(y)) \leq \alpha d_X(x, y) \tag{1}$$

4 Embedding into a distribution of trees

It would be ideal if we could embed a graph into a tree without too much distortion, but it is generally not possible. Even an embedding of C_n, a cycle of length n, into a tree has a distortion $\Omega(n)$. However, much better results can be obtained if one tries to embed a graph into a distribution of trees instead of one tree, so that expected distances are preserved.

Definition 4 *The* support *of a distribution* D*, is the set of all* x*, such that* $D(x) > 0$*.*

Definition 5 *We say a metric* (X, d) *embeds probabilistically into a distribution,* D*, of trees with distortion* α*, if and only if:*

- *Each tree,* $T(V_T, E_T)$*, in support of* D *contains points of the metric, i.e.,* $V_T \supseteq X$*. Furthermore, the distances in* T *dominate those in* d*, i.e.,*

$$d_T(x, y) \geq d(x, y) \tag{2}$$

- *Let* P_i *be probability of* T_i*. Then* $\forall x, y \in G$*, the expected distance between* x *and* y *is within a factor* α *of the distance between them in* G*.*

$$E(d_T(x,y)) = \sum_{i=1}^{k} P_i d_{T_i}(f(x), f(y)) \leq \alpha d_G(x,y) \forall x, y \in X \qquad (3)$$

For a cycle, C_n, we can find such an embedding with $\alpha = 2$. D would consist of n distinct trees that can be formed by deleting one edge from the cycle. The probability distribution would be uniform with each tree assigned a probability of $\frac{1}{n}$. The expected distance between two vertices can then easily be calculated to be $2(1 - \frac{1}{n})$.

Bartal provided an algorithm for embedding with distortion of $O(\log^2(n))$. The bound for general graphs is $\Omega(\log(n))$. This is the bound even for diamond graphs. An algorithm with this distortion was given by **Fakcharoenphol, Rao, and Talwar**. FRT algorithm requires existence of steiner nodes in the trees of embedding. The best known result, when all the trees in support of D are subgraphs of G is $\Omega(\log(n).\log(\log(n)).(\log(\log(\log(n))))^3)$ due to **Abraham, Bartal, Neiman**. It was an improvement over earlier result of **Elkin, Emek, Spielman, Teng**. We can, however, delete steiner nodes from trees by incurring a penalty of factor 8 (**Anupam Gupta**). It is easier to work with algorithm of Abraham et al. but we can usually get better bound by working with the FRT algorithm.

5 Steiner tree problem

Consider a graph, G, and a given cut (S, T) on G and a root node, r. We are required to find a tree of minimum cost that spans all nodes in $T \cup \{r\}$. We can use any subset of nodes from S. Nodes in T are called *terminal nodes* and nodes in S are called *steiner nodes*. Steiner tree problem on trees can be solved trivially. We simply remove all the nodes that don't lie on a path from a terminal node to the root. However, the problem is not trivial on general graphs. Using the result of FRT, we can easily find O(log(n)) approximation to this problem. Note that, for connectivity problems such as steiner tree problem, if the original graph is not a metric, we can always work with metric closure of the original graph.

5.1 Algorithm

Input: Metric G with root r and its probabilistic embedding into $\{P_1, .., P_k\}$ with support $\{T_1, .., T_k\}$; a cut (S,T) on G. Let α be the distortion of the embedding.
Output: A tree spanning $T \cup \{r\}$, which is $O(\alpha)$ approximation to the steiner tree problem.

1. Solve steiner tree problem on each of the trees in support of D. Let C_k be the solution for tree, T_k.

2. Let C_{min} be the minimum cost solution among all the solutions computed in the above step. If the input embedding was an embedding into subtrees (eg., if we used Abraham et al. to find the embedding), Output C_{min}. Else (eg., if we used FRT), goto step 3.

3. For any edge (u, v) in C_{min} that is not in G, we replace it by the shortest path connecting u with v in G. Due to domination property of probabilistic embeddings, cost of solution can only go down.

Proof: Let OPT be the optimum solution. Let C_i be the optimum solution on T_i. Let OPT_i be partial mapping of OPT in tree T_i. Let C_{min} be the optimum solution on the embedding and the C^G_{min} be the solution obtained in the final step of the above algorithm. Now,

$$\begin{aligned}
&\alpha \sum_{e \in OPT} d(e) \\
&\geq \sum_{e \in OPT} \sum_{i=1}^{k} P_i d_{T_i}(e) && \text{[From (3)]} \\
&= \sum_{i=1}^{k} P_i \sum_{e \in OPT} d_{T_i}(e) \\
&= \sum_{i=1}^{k} P_i OPT_i && \text{[From definition of } OPT_i\text{]} \\
&\geq \sum_{i=1}^{k} P_i C_i && \text{[} C_i \text{ is the optimal solution in } T_i\text{]} \\
&\geq C_{min} \sum_{i=1}^{k} P_i \\
&= C_{min} \\
&\geq C^G_{min} && \text{[From (2)]}
\end{aligned}$$

■

Therefore, with FRT, we get an $O(\log(n))$-approximation and with Abraham et al., we get an $\tilde{O}(\log(n))$-approximation.

A related problem is the steiner forest problem. Here instead of being given one root, we are given pairs of terminals to be connected and we have to find the minimum cost forest connecting all the terminals. For this problem too, we can easily find an $O(\log(n))$-approximation with the same approach. It should be noted that the best know approximation factors for the steiner trees and the steiner forest problems are 1.38 and $2 - \frac{1}{n}$ respectively.

However, for another related problem the above embedding leads to the best possible approximation of $O(\log^3(n))$. This problem is the group steiner problem.

scribe: Anshul Sawant
Lecture 4 **Date:** 09/21/2011

Group steiner problem: Given a graph $G(V, E)$ and $S_1, .., S_p \in V$, we are required to find a minimum cost tree that has atleast one vertex from each $S_i, i \in \{1, .., p\}$. For the tree case we have an $O(\log^2(n))$ algorithm for this problem, which matches the theoretical bound as this problem is $\Omega(\log^{2-\epsilon}(n))$ hard. A further factor of $\log(n)$ comes from the probabilistic embedding.

For directed graphs we cannot use this technique since directed graphs are not metrics. The best approximation for directed steiner on trees is $O(n^\epsilon)$ and on general metrics the problem is $O(2^{\log_{1-\epsilon}(n)})$ hard.