The Prize-Collecting Generalized Steiner Tree Problem Via A New Approach Of Primal-Dual Schema

MohammadTaghi Hajiaghayi[*] *Kamal Jain*[†]

Abstract

In this paper we study the prize-collecting version of the Generalized Steiner Tree problem. To the best of our knowledge, there is no general *combinatorial technique* in approximation algorithms developed to study the prize-collecting versions of various problems. These problems are studied on a case by case basis by Bienstock et al. [5] by applying an LP-rounding technique which is not a combinatorial approach. The main contribution of this paper is to introduce a general combinatorial approach towards solving these problems through novel primal-dual schema (without any need to solve an LP). We fuse the primal-dual schema with Farkas lemma to obtain a combinatorial 3-approximation algorithm for the Prize-Collecting Generalized Steiner Tree problem. Our work also inspires a combinatorial algorithm [12] for solving a special case of Kelly's problem [21] of pricing edges.

We also consider the k-forest problem, a generalization of k-MST and k-Steiner tree, and we show that in spite of these problems for which there are constant factor approximation algorithms, the k-forest problem is much harder to approximate. In particular, obtaining an approximation factor better than $O(n^{1/6-\varepsilon})$ for k-forest requires substantially new ideas including improving the approximation factor $O(n^{1/3-\varepsilon})$ for the notorious densest k-subgraph problem. We note that k-forest and prize-collecting version of Generalized Steiner Tree are closely related to each other, since the latter is the Lagrangian relaxation of the former.

1 Introduction

Consider a mailing company that wishes to ship packets overnight between several pairs of cities. To this end, this company can build connecting carriers between cities such that at the end by scheduling the carriers, the company is able to ship the packets overnight between pairs of connected cities. Assume the cost of connecting city i to city j is c_{ij} and the costs are symmetric. In addition, the company has the choice of leasing other companies for some pairs (i, j) of cities with cost π_{ij} so that without any worry the leased company do the shipment between cities i and j overnight. The goal is to build some carriers and lease some other companies such that the company do the shipments overnight with minimum total cost.

The above problem which has also several applications in design of telecommunications networks and wiring with different cable types which offer different amounts of capacity and carry different costs (see e.g. [2, 26]) is called the *prize-collecting generalized Steiner tree (PCGST)* problem. In this problem, given a graph $G = (V, E)$, a set of pairs $\mathcal{P} = \{(s_1, t_1), (s_1, t_1), \ldots, (s_\ell, t_\ell)\}$, a non-negative cost function $c : E \rightarrow \mathbf{Q}_+$, and finally a non-negative penalty function $\pi : \mathcal{P} \rightarrow \mathbf{Q}_+$, our goal is a minimum-cost way of buying a set of edges and paying the penalty for those pairs which are not connected via bought edges. Without loss of generality we assume that $\mathcal{P} = V \times V$, since the penalty of any pair which is not required to be connected will be zero (In this paper, by $V \times V$ we mean all unordered pairs (i, j) where $i \neq j$.) When all sinks are identical, this problem is the classic prize-collecting Steiner tree problem. When all penalties are infinity, this problem is the classic generalized Steiner tree problem. The best approximation algorithm for both of these problems is a $2 - \frac{1}{n-1}$ approximation algorithm (n is the number of vertices of the graph) due to Goemans and Williamson [13].

Another variant of the PCGST problem is the *k-forest* problem in which with the same setting mentioned for PCGST, our goal is to buy a minimum-cost set of edges of G which connect at least k pairs in $\mathcal{P}$. The problem is a generalization of the classic k-MST and k-Steiner tree (when we have a common source/sink) for which there are 2-approximation and 4-approximation algorithms, respectively (see [6] for a complete literature review of these problems).

1.1 Related Work. Other problems with the same flavor of the PCGST problem which attracted so much interest recently are the *multicommodity connected facility location (MCFL)* problem and the *multicommodity rent-or-buy (MRoB)* problem (see e.g. [2, 4, 15, 16, 20, 23]). In both of these problems, we are given an undirected graph $G = (V, E)$ with edge e possessing renting cost c_e and buying cost Mc_e for a parameter $M > 1$, and a set $V \times V = \{(s_1, t_1), (s_1, t_1), \ldots, (s_\ell, t_\ell)\}$ of pairs. In the former problem, our goal is to open a set F of facilities, assign sources and sinks to open facilities (on the vertices), and find a subgraph (V, H) of G such that for each h, if s_h is assigned to

[*]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, MA 02139, U.S.A., E-mail: hajiagha@theory.csail.mit.edu. This work was done while the author visited the Microsoft Research Theory Group.

[†]Microsoft Research. One Microsoft Way, Redmond, WA 98052, USA, E-mail: kamalj@microsoft.com.

facility i_1 and t_h is assigned to facility i_2, then there is a path in (V, H) between i_1 and i_2. The cost of a solution is the sum of the cost of buying edges in H and the cost of renting edges in the shortest path (with respect to c) of each demand to its assigned facility. In the latter problem, our goal is to find a minimum-cost way of buying some edges, which allow unlimited use after payment of a large fixed cost, and renting some edges, with cost incurred on a per-unit of capacity basis, so that a unit of demand can be sent simultaneously from each source s_i to the corresponding sink t_i. Interestingly, Kumar et al. [23] show that any α-approximation algorithm for the MCFL problem gives a 2α-approximation algorithm for the MRoB problem. The best approximation algorithm for these problems is the randomized 6.828-approximation algorithm of Becchetti et al. [4]. To the best of our knowledge, the only deterministic algorithm with extremely large constant approximation ratio [1] is a complicated primal-dual algorithm due to Kumar et al. [23] which is a combination of Goemans-Williamson [13] and Jain-Vazirani [18] primal-dual algorithms. Intuitively, the main difference between the MCFL problem and the PCGST problem, the problem considered in this paper, is that in the former problem the path between any source-sink pair (s_h, t_h) should go through rented edges, bought edges and then rented edges while in the latter problem the path either entirely should go through bought edges or entirely should go through leased (rented) edges.

Finally, it is worth mentioning that Bienstock et al. [5] present an LP-rounding technique to solve prize-collecting traveling salesman and prize-collecting Steiner tree. A similar approach can be used to obtain a 3-approximation algorithm for the PCGST problem. Using a randomized LP-rounding approach, one can further improve the approximation factor to 2.54 (see Appendix A that we present the approach for the sake of completeness). However, the LP-rounding technique uses Ellipsoid method which is known to be non-combinatorial and impractical [2]. In the rest of the paper, we mainly focus on a general combinatorial technique through primal-dual schema for the prize-collecting problems.

1.2 Our Results. We show a novel primal-dual framework which gives a 3-approximation algorithm for the prize-collecting generalized Steiner tree problem. In addition, we show that factor three is tight for our primal-dual framework. Indeed, our primal-dual approach is a general technique for prize-collecting versions of combinatorial optimization problems. The prize-collecting generalized Steiner tree problem is just an example for which we have used our general approach. It is worth mentioning that prize-collecting problems naturally appear in game theory and more specifically in the context of cost-sharing (in which cost share of an agent cannot be more than the utility derived). These problems also appear naturally if one wants to use the Lagarangian Relaxation technique of Jain-Vazirani [18] and Chudak, Roughgarden and Williamson [6] to obtain an approximation algorithm for the corresponding k-version of the problem.

Our primal-dual algorithm has some similarities with Goemans-Williamson algorithm for the prize-collecting Steiner tree (PCST) problem, however our primal-dual LPs are considerably different from theirs. In fact, the primal-dual algorithm for PCST makes a crucial use of the assumption that all commodities have the same sink vertex and does not obviously extend to the PCGST problem. This is the same hardness mentioned by Kumar et al. [23], when they want to extend primal-dual algorithms for single-sink MCFL/MRoB to general MCFL/MRoB. Indeed, if the graph G is a tree the PCST problem can be solved in polynomial time using a simple dynamic programming. However, such a simple dynamic-programming approach does not work for the PCGST problem on trees, and we reduce the problem to an instance of max flow/min cut to solve it polynomially (see Appendix D). It is worth mentioning, by having the solution of this problem on trees, one can obtain a logarithmic approximation ratio for the PCGST problem based on Bartal's machinery [3] (or its slight improvement by Fakcharoenphol et al. [9]) for probabilistically embedding general metrics into tree metrics.

The first barrier which arises when we consider the natural generalization of Goemans-Williamson primal-dual algorithm is dividing the initial penalty (which is in fact the initial potential) between vertices of each pair. We show that an incorrect dividing of the initial penalty can make the approximation factor very large. The second barrier arises in cost-sharing, the task of allocating the cost of an object to many users of the object in a "fair" manner, when two balls (connected components) merge in GW-algorithm. The last not the least barrier arises in the final removal of (unnecessary) edges in GW-algorithm. We overcome all these barriers by writing novel primal and dual LPs and then make a crucial use of Farkas Lemma. In short, we do not see any direct and general solution to these barriers which are also major hindrances in developing primal-dual schema based algorithm for many other problems (e.g., other version of the rent-or-buy problem, the Steiner tree problem based on bidirected LP, the generalized Steiner network problem). In this paper we propose a clean solution to these issues based on Farkas lemma. The problem that we solve here should be taken just as an exemplification of our approach.

Indeed some of the above barriers can be overcome if we can both increase and decrease the dual variables. To the

[1] According to the authors, the approach of the paper only seems suitable to prove an approximation factor of at least several hundreds.

[2] One can write the corresponding LPs to be flow-based rather than cut-based, and then use other LP-solver algorithms. However, during this reduction the number of variables will be cubic in terms of the number of original variables and still we cannot obtain a practical algorithm.

best of our knowledge, the only primal-dual (exact and not even approximation) algorithms which both increase and decrease the dual variables are the max-flow algorithm and the seminal Edmonds' algorithm for minimum weighted perfect matching. In a separate paper, we show that we can obtain primal and dual LPs for these problems such that in the corresponding algorithms we only increase the dual variables (however, when translated back on the original LP both increase and decrease the dual variables) and still we can solve these problems in polynomial time. Thus, our framework makes a significant step towards answering the important open question that asks whether both increasing and decreasing of dual variables can be used to obtain better primal-dual approximation algorithms for combinatorial optimization problems.

Last but not least, we show that in spite of classic k-MST and k-Steiner tree for which there are constant factor approximation algorithms, the k-forest problem is much harder to approximate. In particular, obtaining an approximation factor better than $O(n^{1/6-\varepsilon})$ for k-forest requires substantially new ideas including improving the approximation factor $O(n^{1/3-\varepsilon})$ for the notorious densest k-subgraph problem. We note that k-forest and PCGST are closely related to each other, since the latter is the Lagrangian relaxation of the former.

Our work also inspires a combinatorial algorithm [12] for solving a special case of Kelly's problem [21] of pricing edges.

2 Suitable LP Relaxation Using Farkas Lemma

The traditional LP relaxation for the PCGST problem can be written as:

$$\text{OPT} = \min \sum_{e\in E} c_e x_e + \sum_{i,j\in V} \pi_{ij} z_{ij} \qquad (2.1)$$

subject to

$$\sum_{e\in\delta(S)} x_e + z_{ij} \geq 1 \qquad \forall S \subset V, (i,j) \in V\times V, S \odot (i,j)$$

$$x_e \geq 0 \qquad \forall e \in E$$

$$z_{ij} \geq 0 \qquad \forall (i,j) \in V \times V$$

Here for a set $S \subset V$, we denote $|\{i,j\} \cap S| = 1$ by $S \odot (i,j)$.

The dual of the above LP is:

$$\max \sum_{S\subset V, S\odot(i,j)} y_{Sij} \qquad (2.2)$$

subject to

$$\sum_{S:e\in\delta(S), S\odot(i,j)} y_{Sij} \leq c_e \qquad \forall e \in E$$

$$\sum_{S:S\odot(i,j)} y_{Sij} \leq \pi_{ij} \qquad \forall (i,j) \in V\times V$$

$$y_{Sij} \geq 0 \qquad \forall S \subset V, S \odot (i,j)$$

The problem in this dual LP is that there are different dual variables for each pair of vertices. Consider a situation where two vertices i and j are replaced by a million of copies i.e., i is replaced by $i_1, i_2, \cdots$ and j is replaced by $j_1, j_2, \cdots$. Also the original penalty between i and j is replaced by its millionth between i_1 and j_1, and i_2 and j_2, and so on. By this reduction we did not logically change the original instance but the execution of a primal dual algorithm which treats each $y_{Si_1j_1}, y_{Si_2j_2}, \cdots$ independently can change significantly. Ability to raise dual variables independently to a specific goal is what really constitutes a primal-dual schema. In fact, it is not hard to create counterexamples that such an approach will not work. On the other hand if we raise $y_{Si_1j_1}, y_{Si_2j_2}, \cdots$ in some dependent manner then we will need to deal with cost-sharing, i.e., distributing the rate of increase. If we raise uniformly then a pair of vertices can bias the dual raise in its favor by creating multiple copies as shown in the above example. We do not see any direct solution to this problem, which is also a major hindrance in developing primal-dual schema based algorithm for many other problems (e.g., other version of the rent-or-buy problem, the Steiner tree problem based on bidirected LP, the generalized Steiner network problem). In this paper we propose a clean solution to this issue based on Farkas lemma. As mentioned before, the problem we solve should be taken as an exemplification of our approach.

We create a new dual variable y_S which is the sum of all y_{Sij}'s, where i and j are separated by S. We will treat y_S as an independent dual variable in our LP. We will not fix in advance any arrangement of distributing y_S among the y_{Sij}'s. Instead we will let the dual solution adapts itself to accommodate y_{Sij}'s. A more rigorous way is to eliminate the issue of cost-sharing altogether by eliminating the dual variables y_{Sij}'s. With this substitution of y_S, the dual LP 2.2becomes:

$$\max \sum_{S\subset V} y_S \qquad (2.3)$$

subject to

$$y_S \leq \sum_{(i,j):S\odot(i,j)} y_{Sij} \qquad \forall S \subset V$$

$$\sum_{S:e\in\delta(S)} y_S \leq c_e \qquad \forall e \in E$$

$$\sum_{S:S\odot(i,j)} y_{Sij} \leq \pi_{ij} \qquad \forall i,j \in V$$

$$y_{Sij} \geq 0 \; \forall S \subset V, S \odot (i,j)$$

$$y_S \geq 0 \qquad \forall S \subset V$$

LEMMA 2.1. *Dual LPs 2.2 and 2.3 are equivalent.*

Proof. Suppose we have a solution for dual LP 2.2; we can assign $y_S = \sum_{(i,j):S\odot(i,j)} y_{Sij}$ for each $S \subset V$, which is

a solution to dual LP 2.3. Suppose we have a solution for dual LP 2.3. Without loss of generality we can assume that the first set of constraints is tight; if not then we can decrease y_{Sij} without affecting the objective function (here we use that $y_S \geq 0$). Once we have $y_S = \sum_{(i,j):S\odot(i,j)} y_{Sij}$, we can replace y_S to get a solution for dual LP 2.2. □

Note that y_{Sij}'s do not appear in the objective function, so they can be classified as auxiliary variables. So the idea of the primal dual algorithm is to raise dual variables y_S's (according to some rule) and let the LP itself keep accommodating the auxiliary variables (which may increase or decrease). In other words we raise dual variables y_S's and keep testing whether feasible y_{Sij}'s exist, so in essence we do not decrease the actual dual variable but may decrease the auxiliary dual variables. Intuitively, we freeze some y_S if increasing it by infinitesimally small epsilon results in the non-existence of feasible y_{Sij}'s. Formally, we need to find the condition when feasible y_{Sij}'s do not exist. This can be done using Farkas lemma.

Suppose $\alpha : 2^V \rightarrow \mathcal{R}^+$ is a function from the powerset of V to the non-negative real numbers. For convenience we use the notation α_S to denote the value of α at S. Consider a feasible solution of the dual LP 2.3. Using the first set of constraints we get:

$$\sum_{S\subset V} \alpha_S y_S \leq \sum_{S\subset V} \alpha_S \sum_{(i,j):S\odot(i,j)} y_{Sij}$$

Changing the order of summation on the right-hand side we get:

$$\sum_{S\subset V} \alpha_S y_S \leq \sum_{(i,j)\in V\times V} \sum_{S:S\odot(i,j)} \alpha_S y_{Sij}$$

The right hand side can be increased further to yield:

$$\sum_{S\subset V} \alpha_S y_S \leq \sum_{i,j\in V} \sum_{S:S\odot(i,j)} (\max_{S:S\odot(i,j)} \alpha_S) y_{Sij}$$

This simplifies to:

$$\sum_{S\subset V} \alpha_S y_S \leq \sum_{i,j\in V} (\max_{S:S\odot(i,j)} \alpha_S) \sum_{S:S\odot(i,j)} y_{Sij}$$

Along with the third set of constraints in dual LP 2.3 this gives us:

$$\sum_{S\subset V} \alpha_S y_S \leq \sum_{i,j\in V} (\max_{S:S\odot(i,j)} \alpha_S) \pi_{ij}$$

This inequality does not have y_{Sij}. Farkas lemma tells us that if we fix the y_S's so that the feasible y_{Sij}'s do not exist then we must have a function α (dual variables corresponding to the first set of inequalities in dual LP 2.3) so that the above inequality is violated (see Schrijver books on linear and integer programming [27], Corollary 7.1.f.) Here, dual variables corresponding to the third set of inequalities are $(\max_{S:S\odot(i,j)} \alpha_S)$. Thus in essence if we enforce the above inequality for all functions α, we get the condition on those y_S's for which feasible y_{Sij}'s exist. Hence using Farkas lemma, the dual LP 2.3becomes:

$$\max \sum_{S\subset V} y_S \qquad (2.4)$$

subject to

$$\sum_{S:e\in\delta(S)} y_S \leq c_e \qquad \forall e \in E$$

$$\sum_{S\subset V} \alpha_S y_S \leq \sum_{i,j\in V} (\max_{S:S\odot(i,j)} \alpha_S)\pi_{ij} \qquad \forall \alpha$$

$$y_S \geq 0 \qquad \forall S \subset V$$

The second set of constraints is actually infinity in number, but only a finite number of them are actually necessary. This is because the set of feasible y_S's is a polytope. Let us now try to simplify dual LP 2.4.

LEMMA 2.2. *It is sufficient to consider α's having only one positive value in its range.*

Proof. Take an α. Suppose it is taking two positive values in its range. Suppose M_1 is the maximum value in its range and M_2 is the second maximum value in its range. Consider α_1 which takes value $M_1 - M_2$ whenever α takes value M_1, otherwise α_1 takes value zero. Consider $\alpha_2 = \alpha - \alpha_1$. Note that the constraint corresponding to α can be derived by adding the constraints corresponding to α_1 and α_2. Furthermore, the cardinality of the range of α_2 is decreased by 1. Rest of the proof can be completed by induction. □

COROLLARY 2.1. *It is sufficient to consider α's taking value from $\{0, 1\}$ i.e., α can be thought of denoting a family of subsets of V.*

We denote a family of subsets of V by $\mathcal{S} = \{S_1, S_2, \ldots, S_\ell\}$. For a family $\mathcal{S}$, we write $\mathcal{S} \odot (i, j)$, if there exists an $S \in \mathcal{S}$ such that $S \odot (i, j)$. The dual LP 2.4 becomes:

$$\max \sum_{S\subset V} y_S \qquad (2.5)$$

subject to

$$\sum_{S:e\in\delta(S)} y_S \leq c_e \qquad \forall e \in E$$

$$\sum_{S\in\mathcal{S}} y_S \leq \sum_{(i,j)\in V\times V, \mathcal{S}\odot(i,j)} \pi_{ij} \qquad \forall \text{ family } \mathcal{S}$$

$$y_S \geq 0 \qquad \forall S \subset V$$

Define a function $f : 2^{2^V} \rightarrow \mathcal{R}^+$. For every family, $\mathcal{S}$, $f(\mathcal{S})$ is defined as the right-hand side of the inequality corresponding to $\mathcal{S}$ in dual LP 2.5, i.e.,

$$f(\mathcal{S}) = \sum_{(i,j)\in V\times V, \mathcal{S}\odot(i,j)} \pi_{ij}.$$

LEMMA 2.3. *f is a submodular function, i.e., $f(\mathcal{S}_1) + f(\mathcal{S}_2) \geq f(\mathcal{S}_1 \cap \mathcal{S}_2) + f(\mathcal{S}_1 \cup \mathcal{S}_2)$*

Proof. We note that for $(i,j) \in V \times V$, if $\mathcal{S}_1 \odot (i,j)$ or $\mathcal{S}_2 \odot (i,j)$, then $\mathcal{S}_1 \cup \mathcal{S}_2 \odot (i,j)$, and visa versa. However, for $(i,j) \in V \times V$, if $\mathcal{S}_1 \cap \mathcal{S}_2 \odot (i,j)$, then $\mathcal{S}_1 \odot (i,j)$ and $\mathcal{S}_2 \odot (i,j)$. The proof immediately follows from these two facts. □

Note that this lemma is analogous to the submodularity of the st-cut. Indeed this lemma is also stronger than the notion of proper functions, weak supermodularity, and weak submodularity considered in the context of Steiner edge connectivity and Steiner element connectivity problems. There we usually have two inequalities and the function needs to satisfy one of them. The simplicity and the strength of this lemma is not a surprise. If we go back and observe the second set of inequalities in Dual LP 2.5,it looks like the Hall's condition for bipartite graphs. This is more formally captured as the following lemma:

LEMMA 2.4. *Construct a bipartite graph. One side of the bipartite graph is 2^V and the other side of the bipartite graph is $V \times V$. We put a directed edge of infinite capacity from S to ij if $S \odot ij$. In addition we have two more nodes s and t. We put directed edge of infinite capacity from s to all S's. For every ij we put a directed from ij to t of capacity π_{ij}. The Dual LP 2.5is maximizing the flow in this graph from s to t subject to the first set of constraint in Dual LP 2.5.*

Proof. Proof follows from the classic max-flow min-cut theorem. □

We say that an edge $e \in E$ is *tight*, if for e, the first constraint of dual LP 2.5 holds with equality. We call a family $\mathcal{S} \in \mathcal{F}$ is *tight*, if for $\mathcal{S}$, the second constraint of dual LP 2.5holds with equality.

COROLLARY 2.2. *Suppose y is a feasible solution to dual LP 2.5. Suppose the constraints corresponding to family's $\mathcal{S}_1$ and $\mathcal{S}_2$ are tight, then the constraint corresponding to family's $\mathcal{S}_1 \cup \mathcal{S}_2$ and $\mathcal{S}_1 \cap \mathcal{S}_2$ are also tight.*

Proof. By Lemma 2.3, since $f(\mathcal{S})$ is submodular, we have

$$\sum_{(i,j)\in V\times V:\ \mathcal{S}_1\odot(i,j)} \pi_{ij} + \sum_{(i,j)\in V\times V:\ \mathcal{S}_2\odot(i,j)} \pi_{ij}$$

$$\geq \sum_{(i,j)\in V\times V:\ \mathcal{S}_1\cap\mathcal{S}_2\odot(i,j)} \pi_{ij} + \sum_{(i,j)\in V\times V:\ \mathcal{S}_1\cup\mathcal{S}_2\odot(i,j)} \pi_{ij}.$$

Since y is feasible and the left hand side is modular,

$$\sum_{S\in\ \mathcal{S}_1\cap\mathcal{S}_2} y_S + \sum_{S\in\ \mathcal{S}_1\cup\mathcal{S}_2} y_S = \sum_{S\in\ \mathcal{S}_1} y_S + \sum_{S\in\ \mathcal{S}_2} y_S$$

These two facts show that both $\mathcal{S}_1 \cap \mathcal{S}_2$ and $\mathcal{S}_1 \cup \mathcal{S}_2$ are tight. □

3 The Primal-Dual Algorithm

Now, we are ready to present a primal-dual algorithm for the PCGST problem.

Algorithm A
Input: An undirected graph $G = (V, E)$, edge costs $c_e \geq 0$ for $e \in E$, and penalty $\pi_{ij} \geq 0$ for $(i,j) \in V \times V$.
Output: A forest F', and a set of pairs $\mathcal{Q}$ not connected via F'.
begin
let $F \leftarrow \emptyset$
implicitly set $y_S \leftarrow 0$ *for all* $S \subset V$
let $\mathcal{G} \leftarrow \{\{v\} : v \in V\}$
for each $(i,j) \in V \times V$
  set (i,j) *unmarked*
for each connected component $C \in \mathcal{G}$
  set C *active*
  let $y_C \leftarrow 0$
while there exists an *active* connected component $C \in \mathcal{G}$
  find minimum ε_1 s.t. if we increase y_C of each *active* $C \in \mathcal{G}$ by ε_1 we get a new tight edge $e = \{i,j\} \in E$ with $i \in C_p \in \mathcal{G}$, $j \in C_q \in \mathcal{G}$, $C_p \neq C_q$
  (see section 4)
  find minimum ε_2 s.t. if we increase y_C of each *active* $C \in \mathcal{G}$ by ε_1
  we get a new tight family $\mathcal{S}$ (see Section 4)
  let $\varepsilon = \min(\varepsilon_1, \varepsilon_2)$
  let $y_C = y_C + \varepsilon$ for all *active* $C \in \mathcal{G}$
  if $\varepsilon = \varepsilon_2$
    set all pairs $(i,j) \in V \times V$ which have run out of potential *marked* (see Section 4)
    set all $S \in \mathcal{S}$ *inactive*
  else
    let $F \leftarrow F \cup \{e\}$
    let $\mathcal{G} \leftarrow \mathcal{G} \cup \{C_p \cup C_q\} - \{C_p\} - \{C_q\}$
    let $y_{C_p \cup C_q} \leftarrow 0$
    set $C_p \cup C_q$ *active*, except when there is no $(i,j) \in V \times V$ such that $C_p \cup C_q \odot (i,j)$
let F' is derived from F by removing as many edges as possible so that every *unmarked* pair is connected in F'
let $\mathcal{Q}$ be all pairs not connected via F'
end

Since we only grow y variables for a polynomial number of sets, checking whether an edge is tight is easy. In Section 4, we show how we can find out the next tightness event for families. Note that this is especially important, since the number of families is doubly exponential.

4 Algorithm for Finding Next Event

Let y^*_S's be the current solution to Dual LP 2.5and $\mathcal{S}$ be the current family of active sets. We want to find out what is

the maximum ε we can add in all the active dual. There are two restrictions on maximum ε, first set of constraints and the second set of constraints in Dual LP 2.5.There are only linear number of first set of constraints and one can find maximum ε, say ε_1, which this set of constraints allow to add in all the active duals. So we only need to find maximum ε, say ε_2, which the second set of constraints allow to add in all the active sets. Following lemma 2.4 we can route y_S^*'s dual in the bipartite graph described in lemma 2.4. We want to route additional maximum flow, which should be equal through all the active sets. We call such a flow *equal flow*, which in spirit is quite similar to the balanced flow used in [8]. The following lemma characterize the maximum equal flow (ε_2) through the active sets.

LEMMA 4.1. *Construct the bipartite graph as done in lemma 2.4. Delete all the vertices except those which corresponds to either positive duals (according to y_S^*) or active duals. Note that we kept only polynomial number of vertices and have deleted all others. Change the capacity on edges from s to S to y_S^*, if S is not an active set; and to $y_S^* + \varepsilon_2$ if S is an active set. No other change is required.*

ε_2 is maximum equal-flow which can be routed if and only if there are at least two minimum st-cut, one separating all active sets (in fact all sets) from s and another keeping at least one active set on the s side.

Proof. Note that y_S^* flow through inactive sets and $y_S^* + \varepsilon_2$ flow through active sets can be routed, hence if these are the corresponding capacities then s separated from all other vertices is a minimum st-cut (its capacity is the same as a flow).

Suppose there is another minimum st-cut which keeps an active set on s side. Suppose there are k active sets. If we increase ε_2 by some amount, say δ, then the capacity of this cut will be strictly smaller than the capacity of the cut in which s is separated from all other vertices. Hence we will not be able to route additional $\varepsilon_2 + \delta$ equal-flow.

On the other hand suppose all the minimum st-cuts separate all the active sets from s. In other words if we insist to keep at least one active set on the s side then the capacity of the minimum cut will be strictly more than the capacity of the actual minimum cut. Let D be the difference between these capacities. If we increase ε_2 by D/k then the capacity of all the minimum cut increases by D. Hence we can route an additional $\varepsilon_2 + (D/k)$ equal-flow. □

This lemma gives us an iterative algorithm to find maximum equal-flow. We start with an upper bound on ε_2. A good upperbound is $(\sum_{ij} \pi_{ij} - \sum_S y_S^*)/k$. Set $\varepsilon_2 = (\sum_{ij} \pi_{ij} - \sum_S y_S^*)/k$. We compute the minimal minimum st-cut (i.e., s side is set theoretically minimal). Suppose this minimum cut, call it T, separates all the active sets from s then we have the condition of the lemma. So assume that T separates only k' number of active sets from s, where $k' < k$. Let the capacity of T is D less than the capacity of the minimum cut which separates s from all the sets. We decrease ε_2 by $D/(k - k')$, which makes the capacity of T equals to the capacity of the cut which separates s from all other vertices. We compute the minimal minimum cut again. If this new cut separates s from all the other vertices then again we have the condition of the lemma. If not then the claim is that the new cut separates strictly more than k' active vertices from s. Indeed, otherwise the decrease in the capacity of cuts, which separates less than or equal to k' active vertices, is at most $k'D/(k - k')$, which is also the decrease in the capacity of T. So among these cuts T is still the minimum cut. This proves the claim.

THEOREM 4.1. *ε_2 can be found by at most linear number of maxflow computations.*

The active sets which need to be freeze can also be obtained from this bipartite graph. Use ε_2 as found above and find the maximal minimum st-cut. All the active sets on the s side need to be frozen. The reason is if we add any further ε to any of the active dual on s side then S separated from all other vertices does not remain a minimum st-cut.

Finally it is worth mentioning that in the full version of this paper, we propose a more general combinatorial algorithm to find tight sets which also can be applicable for a wide range of prize-collecting problems.

5 Analysis

First we consider the time complexity of the algorithm. The main loop of the algorithm terminates when all connected components of F are inactive. Since in each iteration the sum of the number of components and the number of active components decreases, the loop terminates after at most $2n - 1$ iterations. Since the most time-consuming part of the loop is finding a tight family which can be done in polynomial time, the total running time of the algorithm is polynomial.

Now, we consider the approximation factor of the algorithm.

THEOREM 5.1. *The Algorithm A outputs a forest F' and a set of pairs $\mathcal{Q}$ which are not connected via F' such that*

$$\sum_{e \in F'} c_e + \sum_{(i,j) \in \mathcal{Q}} \pi_{ij} \leq (3 - \frac{2}{n}) \sum_{S \subset V} y_S \leq (3 - \frac{2}{n})\text{OPT}$$

The second inequality is easy since y_S's form a feasible solution for the dual. The proof of the first inequality follows immediately from the following two lemmas.

LEMMA 5.1. *The sum of penalties of marked pairs in F is at most $\sum_{S \subset V} y_S$.*

Proof. We mark a pair only if it belongs to a tight family. Using corollary 2.2 union of tight families is a tight family. The constraint corresponding to this tight family proves the lemma. □

LEMMA 5.2. $\sum_{e\in F'} c_e \leq (2-\frac{2}{n})\sum_{S\subset V} y_S$.

Proof. The idea of the proof is similar to that of Goemans-Williamson [13] and is also not an emphasizing point of this paper. The proof is moved to the appendix C. □

Finally it is worth mentioning that the ratio three between the cost of the resulting solution from Algorithm A and the sum of dual variables, i.e., $\sum_{S\subset V} y_S$ is tight. Consider the following example. Graph G contains a path $a, s_1, \ldots, s_n, b$ where all edges of this path have length $2+\varepsilon$. Also there is a node t whose distance from all vertices in the path is ∞. The penalty for pair (a, b) is ∞, the penalty for pair (s_i, t_i), $1 \leq i \leq \ell$ is 1, and the rest of the penalties are zero. One can easily observe that for this graph and these penalties, the algorithm stops when $\sum_{S\subset V} y_S = n + o(1)$. However the total cost of the optimal solution, which buys all edges of the path and pays the penalties 1 for pairs (s_i, t_i), $1 \leq i \leq \ell$, is $3n + o(1)$. Note that in this case, Algorithm A outputs the optimal solution also.

6 Hardness For k-Forest

In this section[3], we show an interesting relation between the k-forest problem and the densest k-subgraph problem. Formally, we show that if there is a polynomial time f-approximation algorithm $\mathcal{A}$ for the k-forest problem, then there is a polynomial time $2f^2$-approximation algorithm for the densest k-subgraph problem. Given a graph G and a parameter k, the densest k-subgraph problem is to find a set of k vertices with maximum number of induced edges. The densest k-subgraph problem is well-studied in the literature [11, 22]. The best known approximation factor for the densest k-subgraph problem is $O(n^{1/3-\varepsilon})$ for some small $\varepsilon > 0$ and improvement is known to be difficult [10, 22]. The connection between k-forest and the densest k-subgraph problem suggests that obtaining an approximation factor better than $O(n^{1/6-\varepsilon})$ for k-forest would require substantially new ideas.

We state the reduction in two steps. First we consider *minimum k-edge coverage (MkEC)*, which is the minimum number of vertices in a graph G whose induced subgraph has at least k edges, and mention its relation to densest k-subgraph.

THEOREM 6.1. *If there is a polynomial time f-approximation algorithm $\mathcal{A}$ for MkEC, then there is a polynomial time $2f^2$-approximation algorithm for the densest k-subgraph problem.*

[3]This section is based on a joint work with Lap Chi Lau.

Proof. Given a graph G with m edges, we would like to find a set of k vertices with maximum number of edges in the subgraph induced by this set. We use $\mathcal{A}$ to find the approximate solution for MkEC of the graph. Suppose l is the maximum number of edges in the MkEC problem for which $\mathcal{A}$ outputs a solution Y with at most k vertices. That is, there are l edges in the subgraph induced by Y, and the approximate solution returned by $\mathcal{A}$ when $l+1$ edges are required to be covered contains at least $k+1$ vertices. Let the optimal solution to the densest k-subgraph problem contain *opt* edges. We shall prove that $opt \leq 2f^2 l$ and thus Y is a solution to the densest k-subgraph problem which is within a factor of $\frac{1}{2f^2}$ to the optimal solution.

By our choice of l and the fact that $\mathcal{A}$ is an f-approximation algorithm, any $\frac{k}{f}$ vertices of G can induce at most l edges. Consider a subset X with k vertices. The total number of edges induced by all possible subsets of $\frac{k}{f}$ elements of X is at most $\binom{k}{\frac{k}{f}} l$. Notice that each edge is counted exactly $\binom{k-2}{\frac{k}{f}-2}$ times. So, the total number of edges in X is at most

$$\frac{\binom{k}{\frac{k}{f}} l}{\binom{k-2}{\frac{k}{f}-2}} = \frac{k(k-1)}{\frac{k}{f}(\frac{k}{f}-1)} l \leq f^2 l(\frac{k-1}{k-f}) < 2f^2 l$$

(The last inequality holds since we can assume without loss of generality that $k > 2f$, otherwise, any single edge is a $2f^2$-approximation). Since X is an arbitrary set with k vertices, $opt \leq 2f^2 l$ and this completes the proof. □

Next we mention the relation between MkEC and k-forest.

THEOREM 6.2. *If there is a polynomial time f-approximation algorithm $\mathcal{A}$ for k-forest, then there is a polynomial time f-approximation algorithm for MkEC.*

Proof. For a graph instance $G = (V, E)$ of MkEC, we construct an instance of the k-forest problem as follows. First, we construct a star graph S with a center c and a unit-cost edge $\{c, v\}$ for each $v \in V(G)$. For each edge $\{u, v\} \in E(G)$, we put a commodity pair (u, v) in $\mathcal{P}$. Now, it is easy to observe that any subgraph of S which connects at least k commodity pairs corresponds to a set of vertices whose induced subgraph has at least k edges and visa versa. The desired result follows immediately. □

The following theorem follows immediately from Theorems 6.1 and 6.2.

THEOREM 6.3. *If there is a polynomial time f-approximation algorithm $\mathcal{A}$ for k-forest, then there is a polynomial time $O(f^2)$-approximation algorithm for the densest k-subgraph problem.*

In particular, Theorem 6.3 shows that any algorithm with approximation factor in $o(n^{1/6-\varepsilon})$ improves the best current approximation factor for the densest k-subgraph problem. Also note that Theorem 6.2 shows that the k-forest problem even on stars is hard.

7 Discussion and Open Problems

In this paper we see that a straightforward LP sometimes can raise cost sharing issues. These issues could be tackled more easily with a different LP which altogether could be impossible to imagine. In Appendix B, we give the primal LP for dual LP 2.5, which is really the LP we used. As one can see Primal LP B.1 is very hard to imagine directly - it has doubly exponential number of variable. We also see a systematic procedure to eliminate cost sharing issues using Farkas lemma. If we go back, we will see there is not a lot we used from the Goemans-Williamson's paper on constraint forest [13]. As it seems this technique of primal-dual is a general combinatorial and practical method to solve other prize-collecting problems[4]. Indeed this is also the case with the rounding technique we presented in this paper (see Appendix A). Prize-collecting problems appear naturally in cooperative game theory. Penalty function π can be thought of the maximum price a user is ready to pay for a shared resource. So we would like to distribute a cost of shared resource(s) among its users. As shown in numerous papers [17, 19, 25, 7] primal-dual schema has an important role to play in designing cost sharing methods. This makes our primal-dual technique even more useful.

An immediate open problem is that whether we can obtain a primal-dual algorithm for the PCGST problem with a better approximation factor (ideally two). Considering the PCGST problem when we have costs on the nodes instead of edges (see e.g. [1, 14, 24]) is another interesting open area. The other important problem is to get better approximation ratios for multicommodity connected facility location and multicommodity rent-or-buy problems. The framework proposed in this paper could play an instrumental role not only in getting better understanding of the current algorithms for these problems but also in getting a better algorithm or a better analysis of a current algorithm.

8 Acknowledgement

The authors would like to thank Baruch Awerbuch for his help in the early phase of this work. The authors would also like to thank Michel Goemans, Laci Lovasz, Mohammad Mahdian, and Vahab Mirrokni for fruitful discussions. Special thanks go to Lap Chi Lau for his help in obtaining hardness results of Section 6.

[4] Algorithm to find tight sets in section 4 also should be applicable for a wide range of problems. We propose even a more general algorithm to find tight sets in the full version of this paper.

References

[1] A. AGRAWAL, P. KLEIN, AND R. RAVI, *When trees collide: an approximation algorithm for the generalized Steiner problem on networks*, SIAM J. Comput., 24 (1995), pp. 440–456.

[2] B. AWERBUCH AND Y. AZAR, *Buy-at-bulk network design*, in Proceedings of the 38th Annual Symposium on Foundations of Computer Science (FOCS '97), IEEE Computer Society, 1997, p. 542.

[3] Y. BARTAL, *On approximating arbitrary metrices by tree metrics*, in Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC'98), ACM Press, 1998, pp. 161–168.

[4] L. BECCHETTI, J. KONEMANN, S. LEONARDI, AND M. PAL, *Sharing the cost more efficiently: Improved approximation for multicommodity rent-or-buy*, in Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'05), 2005, pp. 375–384.

[5] D. BIENSTOCK, M. X. GOEMANS, D. SIMCHI-LEVI, AND D. WILLIAMSON, *A note on the prize collecting traveling salesman problem*, Math. Programming, 59 (1993), pp. 413–420.

[6] F. A. CHUDAK, T. ROUGHGARDEN, AND D. P. WILLIAMSON, *Approximate k-msts and k-steiner trees via the primal-dual method and lagrangean relaxation*, in Proceedings of the 8th International Conference on Integer Programming and Combinatorial Optimization (IPCO'01), London, UK, 2001, Springer-Verlag, pp. 60–70.

[7] N. R. DEVANUR, M. MIHAIL, AND V. V. VAZIRANI, *Strategyproof cost-sharing mechanisms for set cover and facility location games*, in Proceedings of ACM Conference on Electronic Commerce (EC'05), 2003, pp. 108–114.

[8] N. R. DEVANUR, C. H. PAPADIMITRIOU, A. SABERI, AND V. V. VAZIRANI, *Market equilibrium via a primal-dual-type algorithm*. Manuscript, 2003.

[9] J. FAKCHAROENPHOL, S. RAO, AND K. TALWAR, *A tight bound on approximating arbitrary metrics by tree metrics*, in Proceedings of the 35th ACM Symposium on Theory of Computing (STOC'03), ACM Press, 2003, pp. 448–455.

[10] U. FEIGE, *Relations between average case complexity and approximation complexity*, in Proceedings of the 34th annual ACM Symposium on Theory of Computing (STOC'02), ACM Press, 2002, pp. 534–543.

[11] U. FEIGE, G. KORTSARZ, AND D. PELEG, *The dense k-subgraph problem*, Algorithmica, 29 (2001), pp. 410–421.

[12] D. GARG, K. JAIN, K. TALWAR, AND V. V. VAZIRANI, *A primal-dual algorithm for computing fisher equilibrium in the absence of gross substitutability property*, in Proceedings of the 1st Workshop on Internet and Network Economics (WINE'05), 2005. To appear.

[13] M. X. GOEMANS AND D. P. WILLIAMSON, *A general approximation technique for constrained forest problems*, SIAM J. Comput., 24 (1995), pp. 296–317.

[14] S. GUHA, A. MOSS, J. S. NAOR, AND B. SCHIEBER, *Efficient recovery from power outage (extended abstract)*, in Proceedings of the 31st Annual ACM Symposium on Theory of Computing (STOC'99), ACM Press, 1999, pp. 574–582.

[15] A. GUPTA, A. KUMAR, M. PAL, AND T. ROUGHGARDEN, *Approximation via cost-sharing: a simple approximation al-*

gorithm for the multicommodity rent-or-buy problem, in Proceedings of the 44rd Symposium on Foundations of Computer Science (FOCS'03), IEEE Computer Society, 2003, p. 606.

[16] A. GUPTA, A. KUMAR, AND T. ROUGHGARDEN, *Simpler and better approximation algorithms for network design*, in Proceedings of the 35th ACM Symposium on Theory of Computing (STOC'03), ACM Press, 2003, pp. 365–372.

[17] K. JAIN AND V. V. VAZIRANI, *Application of approximation algorithms to cooperative games*, in Proceedings of the 33rd Annual ACM Symposium of the Theory of Computing (STOC'01), 2001, pp. 364–372.

[18] ———, *Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and Lagrangian relaxation*, J. ACM, 48 (2001), pp. 274–296.

[19] ———, *Equitable cost allocations via primal-dual-type algorithms*, in Proceedings of the 34th Annual ACM Symposium of the Theory of Computing (STOC'02), 2002, pp. 313–321.

[20] D. R. KARGER AND M. MINKOFF, *Building steiner trees with incomplete global knowledge*, in Proceedings of the 41st Annual Symposium on Foundations of Computer Science (FOCS'01), IEEE Computer Society, 2000, p. 613.

[21] F. P. KELLY, *Charging and rate control for elastic traffic*, European Transactions on Telecommunications, 8 (1997), pp. 33–37.

[22] S. KHOT, *Ruling out PTAS for graph min-bisection, densest subgraph and bipartite clique*, in Proceedings of the 44th Annual IEEE Symposium on the Foundations of Computer Science (FOCS'04), 2004, pp. 136–145.

[23] A. KUMAR, A. GUPTA, AND T. ROUGHGARDEN, *A constant-factor approximation algorithm for the multicommodity*, in Proceedings of the 43rd Symposium on Foundations of Computer Science (FOCS'02), IEEE Computer Society, 2002, p. 333.

[24] A. MOSS AND Y. RABANI, *Approximation algorithms for constrained node weighted steiner tree problems*, in Proceedings of the 33rd Annual ACM Symposium on Theory of Computing (STOC'01), ACM Press, 2001, pp. 373–382.

[25] M. PAL AND E. TARDOS, *Strategy proof mechansims via primal-dual algorithm*, in Proceedings of 44th Annual IEEE Symposium on the Foundations of Computer Science (FOCS'03), 2003, pp. 584–593.

[26] F. S. SALMAN, J. CHERIYAN, R. RAVI, AND S. SUBRAMANIAN, *Approximating the single-sink link-installation problem in network design*, SIAM J. on Optimization, 11 (2000), pp. 595–610.

[27] A. SCHRIJVER, *Theory of linear and integer programming*, John Wiley & Sons, New York, 1986.

[28] D. B. SHMOYS, *Using linear programming in the design and analysis of approximation algorithms: two illustrative problems*, in Approximation algorithms for combinatorial optimization (Approx'98), Springer, Berlin, 1998, pp. 15–32.

[29] D. B. WEST, *Introduction to Graph Theory*, Prentice Hall Inc., Upper Saddle River, NJ, 1996.

A LP Rounding

In this section, we use the LP-rounding technique to give an approximation algorithm with ratio $\frac{1}{1-e^{-1/2}} \approx 2.54$ for the PCGST problem. First, we write the LP as follows

$$\text{OPT} = \min \sum_{e\in E(G)} c_e x_e + \sum_{(s,t)\in V\times V} \pi_{st} z_{st}$$

such that

$$\sum_{e\in\delta(S)} x_e + z_{st} \geq 1$$

$$\forall (s,t) \in V\times V \;\&\; \forall S \subset V \text{ for which } S \odot (s,t)$$

$$x_e \geq 0 \quad \forall e \in E(G)$$

$$z_{st} \geq 0 \quad \forall (s,t) \in V\times V$$

Consider a parameter $0 \leq \alpha < 1$. For all pairs (s,t) with $z_{st} \geq \alpha$, we round z_{st} to one, i.e., we pay the penalties for these pairs. We call these pairs $\mathcal{Q}$. One can easily observe that if we set $x'_e = \frac{1}{1-\alpha} x_e$. Then $\sum_{e\in\delta(S)} x'_e \geq 1$ for all $(s,t) \in \mathcal{Q}$ and for all $S \subset V$ such that $S \odot (s,t)$. It means x' is a fractional solution of generalized Steiner tree for pairs in $\mathcal{Q}$. Using the 2-approximation algorithm of Goemans-Williamson we can obtain an integer solution with cost at most $2\sum_{e\in E(G)} c_e x'_e$. It means we can obtain a solution of total cost at most $2\sum_{e\in E(G)} \frac{1}{1-\alpha} c_e x_e + \sum_{(s,t)\in\mathcal{Q}} \pi_{st}$. Let $0 \leq \beta < 1$ be a parameter to be fixed later. We choose α uniformly at random within interval $[0,\beta]$. Then we have

$$E[\text{SOL}] = E[2\sum_{e\in E(G)} \frac{1}{1-\alpha} c_e x_e + \sum_{(s,t)\in\mathcal{Q}} \pi_{st}]$$

$$= E[2\sum_{e\in E(G)} \frac{1}{1-\alpha} c_e x_e] + E[\sum_{(s,t)\in V\times V \& z_{st}\geq\alpha} \pi_{st}]$$

$$= E[\frac{2}{1-\alpha}] \sum_{e\in E(G)} c_e x_e + \sum_{(s,t)\in V\times V} \pi_{st} \Pr[z_{st} \geq \alpha]$$

$$\leq (\int_0^\beta \frac{1}{\beta} \frac{2}{1-\alpha} d\alpha) \sum_{e\in E(G)} c_e x_e + \sum_{(s,t)\in V\times V} \pi_{st} (\int_0^{z_{st}} \frac{1}{\beta} d\alpha)$$

$$= \frac{2\ln(1/(1-\beta))}{\beta} \sum_{e\in E(G)} c_e x_e + \frac{1}{\beta} \sum_{(s,t)\in V\times V} \pi_{st} z_{st}$$

Now, by setting $\beta = 1 - e^{-1/2}$, $E[\text{SOL}] \leq \frac{1}{1-e^{-1/2}} (\sum_{e\in E(G)} c_e x_e + \sum_{(s,t)\in V\times V} \pi_{st} z_{st}) = \frac{1}{1-e^{-1/2}} \text{OPT}$ and thus we obtained the claim approximation ratio $\frac{1}{1-e^{-1/2}}$.

We note that here we applied the general idea of Goemans according to [28] to choose the best α for each input instead of setting it once (otherwise, we get 3-approximation). Indeed, it is not hard to show that our distribution is the best distribution for this instance of LP rounding.

B The Primal LP

Our primal LP whose dual is dual LP 2.5.is as follows (let $\mathcal{F}$ be the powerset of the powerset of V):

$$\text{OPT} = \min \sum_{e \in E} c_e x_e \qquad (B.1)$$
$$+ \sum_{\mathcal{S} \in \mathcal{F}} z_{\mathcal{S}} \Big(\sum_{(i,j) \in V \times V : \mathcal{S} \odot (i,j)} \pi_{ij} \Big)$$
$$\text{such that}$$
$$\sum_{e \in \delta(S)} x_e + \sum_{\mathcal{S} \in \mathcal{F} :\, S \in \mathcal{S}} z_{\mathcal{S}} \geq 1 \quad \forall S \subset V \quad (1)$$
$$x_e \geq 0 \quad \forall e \in E \quad (2)$$
$$z_{\mathcal{S}} \geq 0 \quad \forall \mathcal{S} \in \mathcal{F} \quad (3)$$

We note that the optimum solution of the PCGST problem is a feasible solution for the primal LP (by setting x_e one for each bought edge e and all z's zero except $z_{\mathcal{S}} = 1$ when $\mathcal{S}$ is the family corresponding to partition of connected components of bought edges (an isolated vertex is a connected component by itself).

C Proof of Lemma 5.2

Proof. The idea of the proof is similar to that of Goemans and Williamson [13]. Since all edges $e \in F'$ are tight, i.e., $\sum_{S \subset V : e \in \delta(S)} y_S = c_e$, we need to show that

$$\sum_{e \in F'} \sum_{S \subset V : e \in \delta(S)} y_S \leq (2 - \frac{2}{n}) \sum_{S \subset V} y_S,$$

or by rewriting terms,

$$\sum_{S \subset V} y_S |F' \cap \delta(S)| \leq (2 - \frac{2}{n}) \sum_{S \subset V} y_S.$$

We can prove this invariant by induction on the number of iterations of the main loop. Consider an iteration and the set of components of $\mathcal{G}$ in this iteration. Form a graph H in which active and inactive components are vertices and the edges are $e \in F' \cap \delta(A)$ for active A. We remove all isolated inactive vertices. Now let N_a be the set of active vertices and N_i be the set of inactive vertices in H. In this iteration, we increase the left-hand side of the inequality by $\varepsilon(\sum_{v \in N_a} d(v))$, while we increase the right-hand side by $\varepsilon(2 - \frac{2}{n})|N_a|$ ($d(v)$ is the degree $v \in V(H)$.) Thus we only need to show that $\sum_{v \in N_a} d(v) \leq (2 - \frac{2}{n})|N_a|$. To do this, we show that all leaves of H must be active vertices. Let v be an inactive leaf of H, adjacent to edge e, and let C_v be the inactive component corresponding to v which is deactivated at some time before. Any pair $(i,j) \in V \times V$ such that $C_v \odot (i,j)$ is marked. Furthermore, since v is a leaf, no vertex in C_v can lie on the path between vertices of an unmarked pair. Therefore, no inactive node can be a leaf. It means,

$$\sum_{v \in N_a} d(v) \leq \sum_{v \in N_a \cup N_i} d(v) - \sum_{v \in N_i} d(v)$$
$$\leq 2(|N_a \cup N_i| - 1) - 2|N_i| = (2 - \frac{2}{n})|N_a|.$$

Here the inequality holds since all inactive vertices has degree at least two, and since the number of active components is always at most n. □

D PCGST on Trees

THEOREM D.1. *The PCGST problem on trees can be solved in polynomial time.*

Proof. We reduce the problem to an instance of max flow/min cut as follows. We form a bipartite graph $G' = (X \cup Y, E')$ where there is a vertex $x_e \in X$ with weight of c_e for each edge $e \in E(G)$ (G is our original graph which is a tree), there is a vertex y_{ij} with weight π_{ij} for each $(i,j) \in V \times V$, and there is an edge from x_e to y_{ij} if and only if edge e is on the (unique) path between i and j in G. Now, it is easy to see that solving the PCGST problem on G is equivalent to find a minimum weighted vertex cover[5] in G' (for each selected vertex $x_e \in X$ we buy edge e, and for each selected vertex $y_{ij} \in Y$ we pay the penalty.) The minimum weighted vertex cover on bipartite graphs can be easily solved by the max flow/min cut theorem, see e.g. [29]. □

It is worth mentioning, by having the solution of this problem on trees, one can obtain a logarithmic approximation ratio for the PCGST problem based on Bartal's machinery [3] (or its slight improvement by Fakcharoenphol et al. [9]) for probabilistically embedding general metrics into tree metrics. Finally as mentioned in Section 7, solving the k-forest problem on trees is an important open problem.

[5] A minimum weighted vertex cover is a set of vertices with minimum weight such that for each edge, at least one of its endpoints is in the set.