Covering Minimum Spanning Trees of Random Subgraphs*

Michel X. Goemans and Jan Vondrák

Abstract

We consider the problem of finding a sparse set of edges containing the minimum spanning tree (MST) of a random subgraph of G with high probability. The two random models that we consider are subgraphs induced by a random subset of vertices, each vertex included independently with probability p, and subgraphs generated as a random subset of edges, each edge with probability p.

Let n denote the number of vertices, choose $p \in (0,1)$ possibly depending on n and let $b = 1/(1-p)$. We show that in both random models, for any weighted graph G, there is a set of edges Q of cardinality $O(n \log_b n)$ which contains the minimum spanning tree of a random subgraph of G with high probability. This result is asymptotically optimal. As a consequence, we also give a bound of $O(kn)$ on the size of the union of all minimum spanning trees of G with some k vertices (or edges) removed. More generally, we show a bound of $O(n \log_b n)$ on the size of a covering set in a matroid of rank n, which contains the minimum-weight basis of a random subset with high probability. Also, we give a randomized algorithm which calls an MST subroutine only a polylogarithmic number of times, and finds the covering set with high probability.

1 Introduction

In a variety of optimization settings, one has to repeatedly solve instances of the same problem in which only part of the input is changing. It is important in such cases to perform a precomputation that involves only the static part of the input and possibly assumptions on the dynamic part, and which allows to speed-up the repeated solution of instances. The precomputation could possibly be computationally intensive.

In telecommunication networks for example, the topology may be considered fixed but the demands of a given customer (in a network provisioning problem) may vary over time. The goal is to exploit the topology without knowing the demands. The same situation happens in performing multicast in telecommunication networks; we need to solve a minimum spanning tree or Steiner tree problem to connect a group of users, but the topology or graph does not change when connecting different groups of users. Or, in flight reservation systems, departure and arrival locations and times change for each request but schedules do not (availability and prices do change as well but on a less frequent basis). Yet another example is for delivery companies; they have to solve daily vehicle routing problems in which the road network does not change but the locations of customers to serve do.

*This work was supported by NSF contracts ITR-0121495 and CCR-0098018. Authors' address: MIT, Dept of Mathematics, Cambridge, MA 02139. {goemans,vondrak}@math.mit.edu

Examples of such *repetitive* optimization problems with both static and dynamic inputs are countless and in many cases it is unclear whether one can take advantage of the advance knowledge of the static part of the input. One situation which has been much studied (especially from a practical point of view) is the s-t shortest path problem in large-scale navigation systems or Geographic Information Systems. In that setting, it is too slow to compute the shortest path from scratch whenever a query comes in. Various preprocessing steps have been proposed, often creating a hierarchical view of the network, see for example [7]. Here, we have a modest (but nevertheless challenging) goal, to study another simple combinatorial optimization problem: the minimum spanning tree (MST), for instances repeatedly drawn either randomly or deterministically from a fixed given graph.

The MST-covering problem. Assume we are given an edge-weighted graph $G = (V, E)$ with n vertices and m edges and we would like to (repeatedly) find the minimum-weight spanning tree of either a vertex-induced subgraph $H = G[W],\ W \subseteq V$ (the *vertex case*) or a subgraph $H = (V, F),\ F \subseteq E$ (the *edge case*). In general, we need to consider the minimum spanning forest, i.e. the minimum spanning tree on each component, since the subgraph might not be connected. We denote this by $MST(W)$ or $MST(F)$.

Our primary focus is a random setting where each vertex appears in W independently with probability p (in the vertex case; we denote $W = V(p)$); secondly, a setting where each edge appears in F independently with probability p (in the edge case; we denote $F = E(p)$). The question we ask is whether there exists a *sparse set* of edges Q which contains the minimum spanning forest of the random subgraph *with high probability.* This is what we refer to as the MST-covering problem.

We also address a deterministic setting where we assume that W is obtained from V by removing a fixed number of vertices, or F is obtained from E by removing a fixed number of edges (by a *malicious adversary*). Then we seek a sparse set of edges Q containing the minimum spanning forests of *all such subgraphs.*

In the above models, if the minimum spanning tree is not unique, we ask that Q contains *some* minimum spanning tree. Alternatively, we can break ties by an arbitrary fixed ordering of edges, and require that Q contains the unique minimum spanning tree. This is a stronger requirement and in the following, we will indeed assume that the minimum spanning trees are unique (e.g. by assuming that the edge weights are distinct).

Example. Consider a complete graph G on vertices $V = \{1, 2, \ldots, n\}$ where the weight of edge $(i, j), i < j$ is $w(i, j) = 2^i$ (see Figure 1). Assume that $W \subseteq V$ is sampled uniformly, each vertex with probability 1/2. It is easy to see that $MST(W)$ is a star of edges centered at the smallest i in W and connecting i to the remaining vertices in W. The probability that $(i, j) \in MST(W)$ (for $i < j$) is $1/2^{i+1}$ since $\{i, j\}$ must be in W and no vertex smaller than i can be in W. Note that when we order the edges (i, j) lexicographically, their probabilities of appearance in MST(W) decrease exponentially, by a factor of 2 after each block of roughly n edges. An example of an MST-covering set here is

$$Q = \{(i, j) \in E : i < 3 \log_2 n\},$$

since any edge in $E \setminus Q$ appears in $MST(W)$ with probability at most $1/n^3$.

In general, we show a similar behavior. For an arbitrary weight function, if we order the edges by their non-increasing probability of appearance in the MST, these probabilities drop exponentially. As a result, we are able to take $O(n \log n)$ edges with the largest probability of appearance in $MST(W)$, and the probability that $MST(W)$ contains one of the remaining edges is polynomially small.

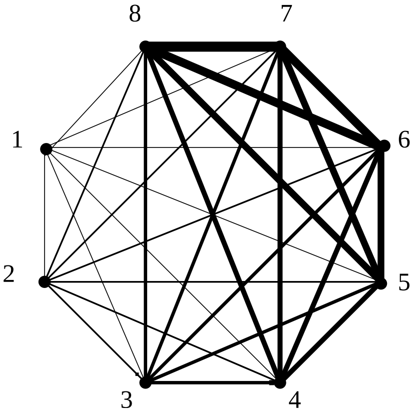

Figure 1: A complete graph K_8 with lexicographically ordered edges. The edge weights are marked by thickness.

Also, our example demonstrates that we need to include $\Omega(n \log n)$ edges in Q if $\Pr[MST(W) \setminus Q \neq \emptyset]$ should be polynomially small. More generally, this is true for any weighted graph - just consider the event that a vertex is isolated in $Q[W]$. Unless Q contains at least $\log_2 n$ edges incident with every vertex, some vertex gets isolated in $Q[W]$ with probability at least $1/n$. Then Q cannot be a good MST-covering set. We will make a more precise argument in Section 7 but this example indicates that $|Q| = O(n \log n)$ is the correct bound to aim for.

Overview of results. Consider the random setting where either vertices or edges are sampled with probability p (possibly a function of n). Let $b = 1/(1-p)$. We prove that, for any weighted graph, there exists a sparse set Q of $e(c+1)n \log_b n + O(n)$ edges which contains $MST(W)$ (or $MST(F)$) with probability at least $1 - 1/n^c$. On the other hand, for $p \geq \frac{1}{n^\gamma}$ with $\gamma < 1$, we show the existence of weighted graphs for which one needs $\Omega(n \log_b n)$ edges even to achieve a constant probability of covering the MST of a random subgraph. So $O(n \log_b n)$ is the best size of an MST-covering set that we can achieve. For low probability of failure, in particular $p = 1 - 1/n^\gamma$, we obtain a covering set of linear size $O(n)$.

We believe that our proof technique is quite interesting in its own right. We define $\sigma_p(u,v)$ to be the probability that $(u,v) \in MST(W)$, conditioned on $u, v \in W$. (Note that $1 - \sigma_p(u,v)$ represents the u-v reliability in the subgraph with all edges lighter than (u,v).) We observe that $\sigma_p(u,v)$ is the probability of a *down-monotone event* which is crucial for our analysis. We show a *boosting lemma* which states that for any down-monotone event, as p decreases, σ_p increases very rapidly. More precisely, we show that $\sigma_p = (1-p)^{f(p)}$ where $f(p)$ is a non-decreasing function (see Lemma 3). From the boosting lemma, we deduce that not many edges (u,v) can have a "large" value of $\sigma_p(u,v)$, otherwise the expected number of edges in the minimum spanning forest of a suitably chosen random subgraph would be larger than possible. This means that we can include all the edges with sufficiently large $\sigma_p(u,v)$ in our set Q, which implies the MST-covering property and Q is still not too large.

Our boosting lemma bears some similarity to a result of Bollobás and Thomason [3]; however, they use a different random model, in which a random subset of a fixed cardinality is sampled. They use the Kruskal-Katona theorem (and its simplification due to Lovász) to derive their lemma, while our lemma has an elementary probabilistic proof. We could use the Bollobás-Thomason lemma in some of our proofs as well (for instance, it would seem quite natural to apply it to prove Corollary

7, which is concerned with the minimum spanning trees after removing a fixed number of vertices), but this would produce an additional factor of $\log n$ which we are able to shave off with our boosting lemma.

In order to find Q, we could try to calculate $\sigma_p(u, v)$ for each edge; however, as $1 - \sigma_p(u, v)$ corresponds to the u-v reliability in an arbitrary graph, this is #P-hard [10]. It is even unknown how to efficiently approximate the u-v reliability. However, in our case, we only need to check if $\sigma_p(u, v)$ is (polynomially) large enough and this can be done by random sampling. This leads to a randomized algorithm for computing Q that makes a polynomial number of calls to an MST subroutine. We can reduce the number of MST calls to polylogarithmic by using the boosting lemma, and choosing the edges which appear sufficiently often. With high probability, we find a covering set of asymptotically optimal size $O(n \log_b n)$, by invoking an MST procedure $O(\log_b n \log n)$ times. If we are interested in deterministic algorithms only, we are only able to construct in polynomial time a covering set Q of cardinality $O(n^{3/2} \ln n)$ in the vertex case and $O(n\ e^{3\sqrt{\ln n}})$ in the edge case. These latter constructions are not described in this paper. The reader can refer to [11] for more information.

Going back to our original motivation, since we are able to construct a set Q of size $O(n \log_b n)$ covering almost all MSTs, we can therefore with high probability find the MST of any random subgraph by focusing on this precomputed set of $O(n \log_b n)$ edges, hence leading to an algorithm whose running time is near-linear in n instead of m. This is almost a quadratic speed-up if the original graph is dense.

In the deterministic setting where the subgraph is obtained by deleting at most $k - 1$ vertices (or edges), we show that there exists a set Q_k of cardinality at most ekn which contains the MST of *all* these subgraphs, see Corollary 7. In the edge case, we prove that the cardinality of Q_k can be actually bounded by $kn - \binom{k+1}{2}$, which is tight and the proof is by elementary induction. For the vertex case, however, we can only obtain a linear bound by probabilistic reasoning.

Our results only use the fact that the event $e \in MST(H)$ for a random subgraph H, conditioned on $e \in E(H)$, is down-monotone. This holds for any matroid; thus all our results extend to matroids, see Section 6.

Literature discussion. Not assuming that all the input data is known in advance or assuming it changes over time is a typical paradigm in the areas of optimization and algorithms. For example, in stochastic programming, part of the input is stochastic and one has to make decisions in the first stage without knowing the realization of the stochastic components; further decision are made when the complete input is revealed. Although the minimum spanning tree problem (as a prototypical combinatorial optimization problem) has been considered in a wide variety of settings with incomplete or changing data, it has not been under the particular viewpoint considered here.

In dynamic graph algorithms, one assumes that the graph is dynamically changing and one needs to update the solution of the problem after each input update. For a minimum spanning tree problem in which edges can be inserted or deleted, the best known dynamic algorithm has amortized cost $O(\log^4 n)$ per operation [6]. This is not efficient here though, since our instances are changing too drastically.

In practice, graph optimization problems are often solved on a sparse subgraph, and edges which are not included are then *priced* to see if they could potentially improve the solution found, see for example [1] for the matching problem. Our results can therefore be viewed as a theoretical basis for this practice in the case of the MST, and give precise bounds on the sparsity required.

2 The boosting lemma

We start by analyzing the event that a fixed edge appears in the minimum spanning tree of a random induced subgraph. We would like to show that the probability of this event cannot be too high for too many edges. We prove this statement by a random sampling argument. It turns out that the only property of MST that we use is the observation that for any given edge, being contained in the minimum-weight spanning forest of a random subgraph is a *down-monotone event.* The following lemma is an easy consequence of the fact that an edge is in the minimum spanning tree unless its endpoints are connected by a path containing only edges of smaller weight.

Lemma 1. *For an edge $(u, v) \in E$, let $X = V \setminus \{u, v\}$ and let $\mathcal{F}$ denote the family of vertex sets $A \subseteq X$ for which (u, v) is in the minimum spanning forest of the induced subgraph $G[A \cup \{u, v\}]$. Then $\mathcal{F}$ is a down-monotone family:*

$$A \in \mathcal{F}, B \subseteq A \Longrightarrow B \in \mathcal{F}.$$

For a random $A \subseteq X$, we say that $A \in \mathcal{F}$ is a *down-monotone event.* We prove a general inequality for down-monotone events. We call this inequality the *boosting lemma*, since it states how the probability of a down-monotone event is boosted, when we decrease the sampling probability. We first give a general version in which the sampling probability of each element could be different (asymmetric version), and then we specialize it to the case in which every element is sampled with the same probability (Lemma 3).

Lemma 2 (The Boosting Lemma, asymmetric version)**.** *Let X be a finite set and $\mathcal{F} \subseteq 2^X$ a down-monotone family of subsets of X. Let $\vec{p} \in [0, 1]^n$ and sample a random subset $X(\vec{p})$ by choosing element i independently with probability p_i. Define*

$$\sigma_p = \Pr[X(\vec{p}) \in \mathcal{F}].$$

Let $\gamma \in (0, 1)$ and similarly define $\sigma_q = \Pr[X(\vec{q}) \in \mathcal{F}]$ where element i is sampled with probability $q_i = 1 - (1 - p_i)^\gamma$. Then

$$\sigma_q \geq (\sigma_p)^\gamma.$$

Proof. We proceed by induction on $|X|$. For $X = \emptyset$ the statement is trivial ($\sigma_p = \sigma_q = 0$ or $\sigma_p = \sigma_q = 1$). Otherwise, let $a \in X$, $Y = X \setminus \{a\}$ and define

- $\mathcal{F}_0 = \{A \subseteq Y : A \in \mathcal{F}\}$
- $\mathcal{F}_1 = \{A \subseteq Y : A \cup \{a\} \in \mathcal{F}\}$

By down-monotonicity, we have $\mathcal{F}_1 \subseteq \mathcal{F}_0$. Next, we express σ_q by the law of conditional probabilities:

$$\sigma_q = \Pr[X(\vec{q}) \in \mathcal{F}] = q_a \Pr[Y(\vec{q}) \in \mathcal{F}_1] + (1 - q_a) \Pr[Y(\vec{q}) \in \mathcal{F}_0]$$

where $Y(\vec{q})$ denotes a subset of Y sampled with the respective probabilities q_i; $Y(\vec{q}) = X(\vec{q}) \setminus \{a\}$. We denote $\omega_p = \Pr[Y(\vec{p}) \in \mathcal{F}_1]$ and $\tau_p = \Pr[Y(\vec{p}) \in \mathcal{F}_0]$. By induction, we can apply the boosting lemma to the down-monotone events $\mathcal{F}_0, \mathcal{F}_1 \subseteq 2^Y$: $\omega_q \geq \omega_p^\gamma, \tau_q \geq \tau_p^\gamma$. We get

$$\sigma_q = q_a \omega_q + (1 - q_a)\tau_q \geq (1 - (1 - p_a)^\gamma)\omega_p^\gamma + (1 - p_a)^\gamma \tau_p^\gamma.$$

Note that $\omega_p \leq \tau_p$ because $\mathcal{F}_1 \subseteq \mathcal{F}_0$. It remains to prove the following:

$$(1 - (1 - p)^\gamma)\omega^\gamma + (1 - p)^\gamma \tau^\gamma \geq (p\omega + (1 - p)\tau)^\gamma \tag{1}$$

for any $p \in [0,1], \gamma \in (0,1), 0 \leq \omega \leq \tau$. Then we can conclude that

$$\sigma_q \geq (p_a \omega_p + (1-p_a)\tau_p)^\gamma = \sigma_p^\gamma$$

using the law of conditional probabilities for $\sigma_p = \Pr[X(\vec{p}) \in \mathcal{F}]$.

We verify Equation (1) by analyzing the difference of the two sides: $\phi(\tau) = (1-(1-p)^\gamma)\omega^\gamma + (1-p)^\gamma \tau^\gamma - (p\omega + (1-p)\tau)^\gamma$. We show that $\phi(t) \geq 0$ for $t \geq \omega$. For $t = \omega$, we have $\phi(t) = 0$. By differentiation,

$$\begin{aligned} \phi'(t) &= \gamma(1-p)^\gamma t^{\gamma-1} - \gamma(1-p)(p\omega + (1-p)t)^{\gamma-1} \\ &= \gamma(1-p)^\gamma t^{\gamma-1} \left(1 - \left(\frac{(1-p)t}{p\omega + (1-p)t}\right)^{1-\gamma}\right) \geq 0. \end{aligned}$$

Therefore $\phi(t) \geq 0$ for any $t \geq \omega$ which completes the proof. □

Note. The boosting lemma is tight for $\mathcal{F} = 2^A, A \subset X$ in which case $\sigma_p = \prod_{i \in X \setminus A}(1-p_i)$ and $\sigma_q = (\sigma_p)^\gamma$. This form of the lemma is the most general we have found; more restricted versions are easier to prove. For probabilities p_i, q_i satisfying $(1-p_i) = (1-q_i)^k$, $k \in \mathbb{Z}$, we give now a simple probabilistic proof using repeated sampling. Sample independently subsets $Y_j = X(\vec{q}), j = 1, 2, \ldots, k$, and set $Y = Y_1 \cup Y_2 \cup \ldots \cup Y_k$. Element i has probability $(1-q_i)^k = (1-p_i)$ that it does not appear in any Y_j, therefore Y is effectively sampled with probabilities p_i. Then we get, from the monotonicity of $\mathcal{F}$: $\sigma_p = \Pr[Y \in \mathcal{F}] \leq \Pr[\forall j; Y_j \in \mathcal{F}] = \sigma_q^k$. This is actually sufficient for the asymptotic results on covering MSTs of random subgraphs. See [5] for details.

In the remainder of this paper, we use a symmetric version of the boosting lemma where the sampling probabilities p_i are uniform.

Lemma 3 (The Boosting Lemma, symmetric version)**.** *Let X be a finite set and $\mathcal{F}$ a down-monotone family of subsets of X. For $p \in (0,1)$, define*

$$\sigma_p = Pr[X(p) \in \mathcal{F}]$$

where $X(p)$ is a random subset of X, each element sampled independently with probability p. Then

$$\sigma_p = (1-p)^{f(p)}$$

where $f(p)$ is a non-decreasing function for $p \in (0,1)$.

Proof. Consider $p, q \in (0,1)$ where $q < p$. Write $\sigma_p = (1-p)^x$ and $(1-q) = (1-p)^\gamma$ where $\gamma \in (0,1)$. Then Lemma 2, with $p_i = p$ and $q_i = q$, implies:

$$\sigma_q \geq (\sigma_p)^\gamma = (1-p)^{\gamma x} = (1-q)^x.$$

□

Connection with the Kruskal-Katona theorem. Consider another special case, where $\sigma_p = (1-p)^k$ for some $k \in \mathbb{Z}$. Denote by F_j the number of sets of size j in $\mathcal{F}$. The Kruskal-Katona theorem [2] says that $F_i \geq \binom{n-k}{i}$ implies $F_j \geq \binom{n-k}{j}$ for $i \geq j$, and this (together with $\sigma_p = (1-p)^k$) can be shown to imply that

$$\sum_{j=0}^{l} F_j p^j (1-p)^{n-j} \geq \sum_{j=0}^{l} \binom{n-k}{j} p^j (1-p)^{n-j}$$

for any $l \leq n-k$. An affine combination of these inequalities then implies a similar statement for any $q \leq p$. Therefore,

$$\sigma_q = \sum_{j=0}^{n} F_j q^j (1-q)^{n-j} \geq \sum_{j=0}^{n-k} \binom{n-k}{j} q^j (1-q)^{n-j} = (1-q)^k$$

which proves the symmetric boosting lemma in this case. It is not clear if this argument applies to $\sigma_p = (1-p)^k$ for non-integer k.

In [3], Bollobás and Thomason prove a lemma about down-monotone events which applies to random subsets of fixed size: If P_r is the probability that a random subset of size r is in $\mathcal{F}$, then for any $s \leq r$,

$$P_s^r \geq P_r^s.$$

Considering that the two random models are roughly equivalent (instead of sampling with probability p, take a random subset of size pn), this lemma has a very similar flavor to ours. However, putting the two random models side by side, the Bollobás-Thomason lemma is weaker; for example, compare $p = 1 - 1/n, q = 1/2$ and $r = n-1, s = n/2$. Our boosting lemma implies $\sigma_q \geq (\sigma_p)^{1/\log n}$. The Bollobás-Thomason lemma says only $P_s \geq \sqrt{P_r}$. For our purposes, the boosting lemma applies in a cleaner way and we gain a factor of $\log n$ compared to using the Bollobás-Thomason lemma.

3 Covering MSTs of random vertex-induced subgraphs

Now we prove the bound on the size of covering sets for the case of randomly sampled vertices. As noted before, for any edge $(u,v) \in E$, the event that $(u,v) \in MST(W)$, conditioned on $u,v \in W$, is down-monotone. Let's denote

$$\sigma_p(u,v) = Pr_W[(u,v) \in MST(W) \mid u,v \in W]$$

where $W = V(p)$ contains each vertex independently with probability p.

Lemma 4. *For a weighted graph G on n vertices, $0 < p < 1$, and any $k \geq 1/p$, let*

$$Q_k^{(p)} = \{(u,v) \in E : \sigma_p(u,v) \geq (1-p)^{k-1}\}.$$

Then

$$|Q_k^{(p)}| < ekn.$$

Proof. Sample a random subset $S = V(q)$, with probability $q = 1/k \leq p$. For every edge $(u,v) \in Q_k^{(p)}$, we have $\sigma_p(u,v) \geq (1-p)^{k-1}$, and therefore by the boosting lemma, $\sigma_q(u,v) \geq (1-q)^{k-1}$ implying

$$Pr[(u,v) \in MST(S)] = q^2 \sigma_q(u,v) \geq q^2 (1-q)^{k-1} = \frac{1}{k^2}\left(1 - \frac{1}{k}\right)^{k-1} > \frac{1}{ek^2},$$

and

$$\mathbf{E}[|MST(S)|] \geq \sum_{(u,v) \in Q_k^{(p)}} Pr[(u,v) \in MST(S)] > \frac{|Q_k^{(p)}|}{ek^2}.$$

On the other hand, the size of the minimum spanning forest on S is at most the size of S, and so

$$\mathbf{E}[|MST(S)|] \leq \mathbf{E}[|S|] = \frac{n}{k}.$$

Combining these two inequalities, we get $|Q_k^{(p)}| < ekn$. □

This lemma shows that the exponential decrease observed in the example in Section 1 always occurs. In particular, choosing $k = \frac{i}{en}$ shows that the ith largest σ_p value is less than $(1-p)^{\frac{i}{en}-1}$.

The constant factor e in the above lemma can be improved; we give here an improved version in the case of integral k.

Lemma 5. *For a weighted graph G on n vertices, $0 < p < 1$, and an integer $k \geq 1/p$, let*

$$Q_k^{(p)} = \{(u,v) \in E : \sigma_p(u,v) \geq (1-p)^{k-1}\}.$$

Then

$$|Q_k^{(p)}| < \left(1 + \frac{e}{2}\right) kn.$$

Proof. For every l, we have that $\sigma_p(u,v) \geq (1-p)^{l-1}$ for $(u,v) \in Q_l^{(p)}$ and hence by the boosting lemma, $\sigma_q(u,v) \geq (1-q)^{l-1}$ for $q = \frac{1}{k} \leq p$. By considering more carefully the argument bounding $|Q_k^{(p)}|$ in the proof of the previous lemma, we get that, for $S = V(q)$:

$$\begin{aligned} qn &\geq E[|MST(S)|] \geq \sum_{l=1}^{\infty} \sum_{(u,v) \in Q_l^{(p)} \setminus Q_{l-1}^{(p)}} Pr[(u,v) \in MST(S)] \\ &\geq \sum_{l=1}^{\infty} |Q_l^{(p)} \setminus Q_{l-1}^{(p)}| q^2 (1-q)^{l-1} = \sum_{l=1}^{\infty} |Q_l^{(p)}| q^3 (1-q)^{l-1}, \end{aligned}$$

implying

$$n \geq \sum_{l=1}^{\infty} |Q_l^{(p)}| q^2 (1-q)^{l-1}. \qquad (2)$$

Our previous argument replaced all the terms for $l < k$ by zero, which leads to the factor e. We can however improve this by using a better lower bound on $|Q_l^{(p)}|$ for $l < k$. Let us assume that G is a complete graph; this is without loss of generality as this can only increase $|Q_k^{(p)}|$. But $Q_l^{(p)}$ must be l-vertex-connected; otherwise there would be an edge $(u,v) \in E \setminus Q_l^{(p)}$ without l vertex-disjoint paths in $Q_l^{(p)}$ between u and v, which would imply $\sigma_p(u,v) \geq (1-p)^{l-1}$. Therefore, for any l, we have that $|Q_l^{(p)}| \geq ln/2$. Using this lower bound in (2), we get:

$$\begin{aligned} n &\geq \sum_{l=1}^{k-1} \frac{ln}{2} q^2 (1-q)^{l-1} + |Q_k^{(p)}| \sum_{l=k}^{\infty} q^2 (1-q)^{l-1} \\ &= \frac{n}{2} \sum_{l=1}^{k-1} \sum_{j=l}^{k-1} q^2 (1-q)^{j-1} + |Q_k^{(p)}| q (1-q)^{k-1} \\ &= \frac{n}{2} \sum_{l=1}^{k-1} q((1-q)^{l-1} - (1-q)^{k-1}) + |Q_k^{(p)}| q (1-q)^{k-1} \\ &= \frac{n}{2} \left(1 - (1-q)^{k-1}(1 + q(k-1))\right) + |Q_k^{(p)}| q (1-q)^{k-1}. \end{aligned}$$

Using the fact that $q = \frac{1}{k}$, we get:

$$|Q_k^{(p)}| \leq \frac{n}{2q(1-q)^{k-1}} + \frac{n}{2q}(1 + q(k-1)) \leq \frac{ekn}{2} + kn = \left(1 + \frac{e}{2}\right) kn.$$

□

We are now ready to prove our MST-covering result.

Theorem 6. *Let G be a weighted graph on n vertices, $0 < p < 1$, and $c > 0$. Let $b = 1/(1-p)$. Then there exists a set $Q \subseteq E$ of size*

$$|Q| \leq \left(1 + \frac{e}{2}\right)(c+1) n \log_b n + O(n)$$

such that for a random $W = V(p)$,

$$Pr_W[MST(W) \subseteq Q] > 1 - \frac{1}{n^c}.$$

Proof. Assume that $n > e^2$ (otherwise Q can be chosen to contain all edges). Order the edges in E by decreasing values of $\sigma_p(u,v)$. Partition the sequence into blocks $B_1, B_2, \ldots \subset E$ of size $\lceil (1+e/2)n \rceil$. Lemma 5 implies that for any $(u,v) \in B_{k+1}$, $k \geq 1/p$,

$$Pr[(u,v) \in MST(W)] = p^2 \; \sigma_p(u,v) < p^2 (1-p)^{k-1}.$$

Define Q to contain the first $h = \lceil (c+1) \log_b n \rceil + 2$ blocks, $Q = \bigcup_{k=1}^{h} B_k$. We have $h \geq 1/p$ (for $p \geq 1/2$ it's obvious that $h \geq 2 \geq 1/p$, and for $p < 1/2$, $h > \log_b n > \frac{\ln n}{2p} > 1/p$ for $n > e^2$). So we can apply the above bound to blocks starting from $h+1$:

$$\begin{aligned} Pr[MST(W) \setminus Q \neq \emptyset] \quad &\leq \quad \sum_{(u,v) \in E \setminus Q} Pr[(u,v) \in MST(W)] \\ \leq \sum_{k=h+1}^{\infty} \lceil (1+e/2)n \rceil p^2 (1-p)^{k-2} \quad &= \quad \lceil (1+e/2)n \rceil p (1-p)^{h-1} < \frac{\lceil (1+e/2)n \rceil p (1-p)}{n^{c+1}} < \frac{1}{n^c}. \end{aligned}$$

□

4 Covering MSTs of subgraphs of fixed size

Directly from Lemma 5, we also get the following interesting implication for the deterministic version of the problem, where at most $k-1$ vertices can be removed arbitrarily.

Corollary 7. *For any weighted graph G on n vertices, and $k \in \mathbb{Z}_+$, there exists a set $Q_k \subseteq E$ of size*

$$|Q_k| < \left(1 + \frac{e}{2}\right) kn$$

which contains $MST(W)$ for any $|W| > n-k$.

Proof. Let $Q_k = \bigcup_{|W| > n-k} MST(W)$. For $k = 1$ the lemma is trivial as Q_1 is the minimum spanning forest of G and thus $|Q_1| \leq n-1$. For $k \geq 2$, choose $p = 1/2$ and consider $Q_k^{(p)}$ as defined in Lemma 5, which states that $|Q_k^{(p)}| < (1+e/2)kn$. For any edge (u,v) which appears in $MST(W)$ for $|W| > n-k$, $\sigma_p(u,v) \geq (1-p)^{k-1}$, since the vertices in $V \setminus W$ are removed with probability at least $(1-p)^{k-1}$; therefore $Q_k \subseteq Q_k^{(p)}$. □

Observe that the set Q_k can be found in polynomial time. For every edge (u,v), its membership in Q_k can be tested by computing the vertex connectivity between vertices u, v in the subgraph G_{uv} of edges lighter than (u,v). By Menger's theorem, $(u,v) \in Q_k$ if and only if there are no k

vertex-disjoint u-v paths in G_{uv}. This, however, does not seem to imply a bound on the size of Q_k easily. The only way we can prove our bound is through probabilistic reasoning.

It is not difficult to see that $|Q_1| \leq n-1$ and $|Q_2| \leq 2n-3$. It is also possible to define edge weights so that Q_k must contain $(n-1)+(n-2)+\cdots+(n-k) = kn - \binom{k+1}{2}$ edges (see Section 7 for an example). We conjecture this to be the actual tight upper bound. Similarly, we conjecture that $kn - \binom{k+1}{2}$ is the best possible bound on $|Q_k^{(p)}|$ in Lemma 5 (and this would be achieved for the graph described in Section 7).

The same question in the edge case is easier to answer. The number of edges in all MSTs obtained after removing at most $k-1$ edges can be upper bounded by $k(n-1)$, by finding the minimum spanning tree and removing it from the graph repeatedly k times. (Which also works for multigraphs, and more generally matroids.) For simple graphs, we can prove a bound of $kn - \binom{k+1}{2}$ which is tight (see the weighted graphs constructed in Section 7).

Lemma 8. *For any (simple) weighted graph on n vertices, m edges and integer $1 \leq k \leq n$, there exists a set $Q_k \subseteq E$ of size*

$$|Q_k| \leq \sum_{i=1}^{k} (n-i) = kn - \binom{k+1}{2}$$

which contains the minimum spanning forest $MST(F)$ for any $|F| > m-k$.

Proof. Let $Q_k = \bigcup_{|F|>m-k} MST(F)$. For given k, we proceed by induction on $n \geq k$. For $n = k$, it is trivial that $|Q_k| \leq \binom{n}{2} = n^2 - \binom{n+1}{2}$. So assume $n > k$.

Consider the heaviest edge $e^* \in Q_k$. Since $e^* \in MST(F)$ for some $|F| > m-k$, there is a cut $\delta(V_1) = \{(u,v) \in E : u \in V_1, v \notin V_1\}$ such that e^* is the *lightest* edge in $\delta(V_1) \cap F$. Consequently $Q_k \cap \delta(V_1) \subseteq (E \setminus F) \cup \{e^*\}$, which means that at most k edges of Q_k are in the cut $\delta(V_1)$. Let $V_2 = V \setminus V_1$ and apply the inductive hypothesis on $G_1 = G[V_1]$ and $G_2 = G[V_2]$, and their respective MST-covering sets $Q_{k,1}, Q_{k,2}$. We use the following characterization of Q_k: $(u,v) \in Q_k \Leftrightarrow$ there are no k edge-disjoint u-v paths in the subgraph of edges lighter than (u,v) (again by Menger's theorem, for edge connectivity). Since the edge connectivity in G is at least as strong as the edge connectivity in G_1 or G_2, it follows that $Q_k[V_1] \subseteq Q_{k,1}$, $Q_k[V_2] \subseteq Q_{k,2}$ and we get

$$|Q_k| \leq |Q_{k,1}| + |Q_{k,2}| + k.$$

Let $n_1 = |V_1|, n_2 = |V_2|$; $n = n_1 + n_2 > k$. We distinguish two cases:

- If one of n_1, n_2 is at least k, assume it is n_1. By the inductive hypothesis, $|Q_{k,1}| \leq \sum_{i=1}^{k} (n_1 - i)$, and $|Q_{k,2}| \leq k(n_2-1)$ (for any n_2, smaller or larger than k), so

$$|Q_k| \leq \sum_{i=1}^{k} (n_1 - i) + k(n_2 - 1) + k = \sum_{i=1}^{k} (n-i).$$

- If both $n_1, n_2 < k$, then we estimate simply $|Q_{k,1}| \leq \binom{n_1}{2} < \frac{k(n_1-1)}{2}$, $|Q_{k,2}| \leq \binom{n_2}{2} < \frac{k(n_2-1)}{2}$. We get

$$|Q_k| < \frac{k(n_1-1)}{2} + \frac{k(n_2-1)}{2} + k = \frac{kn}{2} \leq \sum_{i=1}^{k} (n-i).$$

□

5 Algorithmic construction of covering sets

It is natural to ask whether the MST-covering sets can be found efficiently. In the deterministic case, we have shown that this is quite straightforward. However, in the probabilistic case, it is not possible to test whether $(u, v) \in Q_k^{(p)}$ directly. This would amount to calculating the *u-v-reliability* in the graph of edges lighter than (u, v), which is a #P-complete problem [10].

However, we can find a covering set Q using an efficient randomized algorithm, which takes advantage of the boosting lemma as well. It is a Monte Carlo algorithm, in the sense that it finds a correct solution with high probability, but the correctness of the solution cannot be verified easily.

The algorithm: Given $G = (V, E)$, $w : E \to \mathbb{R}$, $0 < p < 1$, $c > 0$.

- Let $b = 1/(1 - p)$ and $k = \lceil (c + 2) \log_b n \rceil + 1$.
- Repeat the following for $i = 1, \ldots, r = \lceil 32ek^2 \ln n \rceil$:
  - Sample $S_i \subseteq V$, each vertex independently with probability $q = 1/k$.
  - Find $T_i = MST(S_i)$.
- For each edge, include it in Q if it appears in at least $16 \ln n$ different T_i's.

The running time of the algorithm is determined by the number of calls to an MST procedure, which is $O(\log_b^2 n \ln n)$. Since a minimum spanning forest can be found in time $O(m\alpha(m, n))$ deterministically [4] or $O(m)$ randomized [8], for constant $b = 1/(1-p)$ we get a running time near-linear in m.

Theorem 9. *This algorithm finds with high probability a set $Q \subseteq E$ such that*

$$|Q| \leq 2e(c + 2)n \log_b n + O(n)$$

and for a random $W = V(p)$,

$$Pr_W[MST(W) \subseteq Q] > 1 - \frac{1}{n^c}.$$

Proof. Let $k = \lceil (c+2) \log_b n \rceil + 1$, $r = \lceil 32ek^2 \ln n \rceil$ and $Q_k^{(p)} = \{(u, v) \in E : \sigma_p(u, v) \geq (1 - p)^{k-1}\}$. We will argue that (1) $Q_k^{(p)} \subseteq Q$ with probability $> 1 - \frac{1}{n^2}$, (2) $Q_k^{(p)}$ is a good covering set, and (3) $|Q| \leq 2ekn + O(n)$ with probability $> 1 - \frac{1}{n^4}$.

Let $S_i = V(q)$, $q = 1/k$, and $T_i = MST(S_i)$. As in the proof of Theorem 6, $k \geq 1/p$ (for n large enough), therefore $q \leq p$ and by the boosting lemma, for any $(u, v) \in Q_k^{(p)}$,

$$Pr[(u, v) \in T_i] \geq q^2(1 - q)^{k-1} \geq \frac{1}{ek^2}.$$

Denoting by $t(u, v)$ the number of T_i's containing edge (u, v), we get $\mathbf{E}[t(u, v)] \geq r/(ek^2) \geq 32 \ln n$. By Chernoff bound (see [9, Theorem 4.2]; $Pr[X < (1 - \delta)\mu] < e^{-\mu\delta^2/2}$), with $\mu \geq 32 \ln n$, $\delta = 1/2$: $Pr[t(u, v) < 16 \ln n] < e^{-4 \ln n} = 1/n^4$, and thus $Pr[\exists(u, v) \in Q_k^{(p)}; t(u, v) < 16 \ln n] < 1/n^2$. Therefore with high probability, all edges in $Q_k^{(p)}$ are included in Q. On the other hand, $Q_k^{(p)}$ contains $MST(W)$ with high probability (with respect to a random $W = V(p)$):

$$Pr[MST(W) \setminus Q_k^{(p)} \neq \emptyset] \leq \sum_{(u,v) \in E \setminus Q_k^{(p)}} Pr[(u, v) \in MST(W)] < n^2p^2(1 - p)^{k-1} < \frac{1}{n^c}.$$

Now we estimate the size of Q. For $k \geq n/(4e)$, the condition $|Q| \leq 2ekn + O(n)$ is satisifed trivially. So assume $k < n/(4e)$. Since we are sampling $S_i = V(q)$, we have $\mathbf{E}[|S_i|] = qn$, and $\mathbf{E}\left[\sum_{i=1}^{r} |T_i|\right] \leq \mathbf{E}\left[\sum_{i=1}^{r} |S_i|\right] \leq rqn$. We can use the Chernoff bound again ([9, Theorem 4.1]; $Pr[X > (1+\delta)\mu] < e^{-\mu\delta^2/3}$), with $\mu \leq rqn$ and $\delta = 10 \ln n/(rq)$:

$$Pr\left[\sum_{i=1}^{r} |S_i| > (rq + 10 \ln n)n\right] < e^{-100n \ln^2 n/(3rq)} < e^{-n \ln n/ek} < \frac{1}{n^4}.$$

In Q, we include only edges which appear in at least $16 \ln n$ different T_i's, and $|T_i| \leq |S_i|$, so the number of such edges is, with high probability,

$$|Q| \leq \frac{\sum |S_i|}{16 \ln n} \leq \frac{(rq + 10 \ln n)n}{16 \ln n} = 2ekn + O(n).$$

□

6 Covering minimum-weight bases in matroids

Next, we consider the variant of the problem where the subgraph is generated by taking a random subset of edges $E(p)$. We approach this problem more generally, in the context of *matroids*. The matroid in this case would be the graphic matroid defined by all forests on the ground set E. In general, consider a weighted matroid $(E, \mathcal{M}, w)$, where $w : E \to \mathbb{R}$. Let m denote the size of the ground set E and n the rank of $\mathcal{M}$, i.e. the size of a largest independent set. If the weights are distinct, then any subset $F \subseteq E$ has a unique minimum-weight basis $MB(F)$, which in the case of graphs corresponds to the minimum-weight spanning forest. These bases satisfy exactly the monotonicity property that we used previously.

Lemma 10. *For an element $e \in E$, let $X = E \setminus \{e\}$ and let $\mathcal{F}$ denote the family of sets $A \subseteq X$ for which e is in the minimum-weight basis of the matroid induced by $A \cup \{e\}$. Then $\mathcal{F}$ is a down-monotone family:*

$$A \in \mathcal{F}, B \subseteq A \Longrightarrow B \in \mathcal{F}.$$

Proof. If $e \in MB(A \cup \{e\})$, it means that there is no circuit in $A \cup \{e\}$ in which e is the largest-weight element. However, then there is no such circuit in $B \cup \{e\}$ either, and therefore $e \in MB(B \cup \{e\})$. □

Thus, we can apply the same machinery to matroids. Define

$$\sigma_p(e) = Pr_F[e \in MB(F) \mid e \in F]$$

where $F = E(p)$ is a random subset of elements, sampled with probability p. We get statements analogous to the vertex case. It is interesting to notice that the bounds given in these statements depend only on the rank of the matroid, irrespective of the size of the ground set.

Lemma 11. *For a weighted matroid $(E, \mathcal{M}, w)$, of rank n, $0 < p < 1$ and $k \geq 1/p$, let*

$$Q_k^{(p)} = \{e \in E : \sigma_p(e) \geq (1-p)^{k-1}\}.$$

Then

$$|Q_k^{(p)}| < ekn.$$

Proof. The proof is similar to the proof of Lemma 4. Sample a random subset $S = E(q)$, each element with probability $q = 1/k \leq p$. For any $e \in Q_k^{(p)}$, $\sigma_p(e) \geq (1-p)^k$, therefore the boosting lemma implies that

$$Pr[e \in MB(S)] \geq q\ \sigma_q(e) \geq q(1-q)^{k-1} = \frac{1}{k}\left(1 - \frac{1}{k}\right)^{k-1} > \frac{1}{ek}.$$

Summing over all $e \in Q_k^{(p)}$, we get

$$\mathbf{E}[|MB(S)|] \geq \sum_{e \in Q_k^{(p)}} Pr[e \in MB(S)] > \frac{|Q_k^{(p)}|}{ek}.$$

On the other hand, any independent set in $\mathcal{M}$ has size at most n, therefore $\mathbf{E}[|MB(S)|] \leq n$ which implies $|Q_k^{(p)}| < ekn$. □

In the case of matroids, we don't get the equivalent to Lemma 5 as the improvement there was based on connectivity properties.

Theorem 12. *For any weighted matroid $(E, \mathcal{M}, w)$ of rank n, $0 < p < 1$, $c > 0$, and $b = 1/(1-p)$, there exists a set $Q \subseteq E$ of size*

$$|Q| \leq e(c+1)n \log_b n + O(n/p)$$

such that for a random $F = E(p)$,

$$Pr_F[MB(F) \subseteq Q] > 1 - \frac{1}{n^c}.$$

Proof. Order the elements of E by decreasing values of $\sigma_p(e)$. Partition the sequence into blocks $B_1, B_2, \ldots \subset E$ of size $\lceil en \rceil$. Lemma 11 implies that for any $e \in B_{k+1}$, $k \geq 1/p$:

$$Pr[e \in MB(F)] = p\ \sigma_p(e) < p(1-p)^{k-1}.$$

Take the first $h = \lceil (c+1) \log_b n + 2/p \rceil + 1$ blocks: $Q = \bigcup_{k=1}^{h} B_k$. Then, since $h \geq 1/p$:

$$\begin{aligned} Pr[MB(F) \setminus Q \neq \emptyset] &\leq \sum_{e \in E \setminus Q} Pr[e \in MB(F)] \leq \sum_{k=h+1}^{\infty} \lceil en \rceil p(1-p)^{k-2} \\ &= \lceil en \rceil (1-p)^{h-1} \leq \frac{\lceil en \rceil (1-p)^{2/p}}{n^{c+1}} < \frac{1}{n^c}. \end{aligned}$$

□

The forests in a graph on $n+1$ vertices form a matroid of rank n, and minimum-weight bases correspond to minimum spanning forests. Therefore this solves the edge version of the MST-covering problem as well:

Corollary 13. *For any weighted graph G on $n+1$ vertices, $0 < p < 1$, $c > 0$ and $b = 1/(1-p)$, there exists a set $Q \subseteq E(G)$ of size*

$$|Q| \leq e(c+1)n \log_b n + O(n/p)$$

such that for a random $F = E(p)$,

$$Pr_F[MST(F) \subseteq Q] > 1 - \frac{1}{n^c}.$$

Also, we have a randomized algorithm finding the covering set for any weighted matroid $(E, \mathcal{M}, w)$; the algorithm makes $O(\log_b n \ln m)$ calls to a minimum-weight basis procedure.

- Let $b = 1/(1-p)$ and $k = \lceil (c+2) \log_b n \rceil + 1$.
- Repeat the following for $i = 1, \ldots, r = \lceil 16ek \ln m \rceil$:
  - Sample $S_i \subseteq E$, each element independently with probability $q = 1/k$.
  - Find $T_i = MB(S_i)$.
- For each edge, include it in Q if it appears in at least $8 \ln m$ different T_i's.

Theorem 14. *This algorithm finds with high probability a set $Q \subseteq E$ such that*

$$|Q| \leq 2e(c+2)n \log_b n + O(n)$$

and for a random $F = E(p)$,

$$Pr_F[MB(F) \subseteq Q] > 1 - \frac{1}{n^c}.$$

Proof. Let $k = \lceil (c+2) \log_b n \rceil + 1$, $r = \lceil 16ek \ln m \rceil$ and $Q_k^{(p)} = \{e \in E : \sigma_p(e) \geq (1-p)^{k-1}\}$. We claim that (1) $Q_k^{(p)} \subseteq Q$ with high probability, (2) $Q_k^{(p)}$ is a good covering set and (3) Q is not too large. As in the proof of Theorem 6, $k \geq 1/p$ for n large enough, therefore for any $e \in Q_k^{(p)}$, for $S_i = E(q)$ and $T_i = MB(S_i)$, the boosting lemma implies

$$Pr[e \in T_i] = q\ \sigma_q(e) > \frac{1}{ek}.$$

Letting $t(e)$ denote the number of T_i's containing element e, we obtain $\mathbf{E}[t(e)] \geq r/(ek) \geq 16 \ln m$. By the Chernoff bound, $Pr[t(e) < 8 \ln m] < e^{-2 \ln m} = 1/m^2$, implying that $Pr[\exists e \in Q_k^{(p)}; t_e < 8 \ln m] < 1/m$. Therefore with high probability, all edges in $Q_k^{(p)}$ are included in Q.

On the other hand, $Q_k^{(p)}$ contains $MB(F)$ with high probability. Consider the elements in $E \setminus Q$. We order them in a sequence of decreasing values of $\sigma_p(e)$, and again divide them into blocks $B_1, B_2, \ldots$ as before. Since we have included all edges with $\sigma_p(e) \geq (1-p)^{k-1}$ in Q, in the first k blocks the values of $\sigma_p(e)$ cannot be larger than $(1-p)^{k-1}$. Then, the l-th block can have values of $\sigma_p(e)$ at most $(1-p)^{l-2}$ (by Lemma 11). Thus

$$Pr[MB(F) \setminus Q \neq \emptyset] \leq p \left(k(1-p)^{k-1} + \sum_{l=k+1}^{\infty} (1-p)^{l-2} \right) \lceil en \rceil \leq \frac{(kp+1)\lceil en \rceil}{n^{c+2}} = \frac{O(\ln n)}{n^{c+1}} < \frac{1}{n^c}.$$

Finally, we estimate the size of Q. We have $\sum_{i=1}^{r} |T_i| \leq rn$. Every element $e \in Q$ appears in $8 \ln m$ different T_i's, therefore

$$|Q| \leq \frac{\sum |T_i|}{8 \ln m} \leq 2ekn + O(n).$$

□

7 Lower bounds

For both variants of the problem, we have a closely matching lower bound on the size of Q, even if we only want to achieve a constant probability of covering the MST. We get a lower bound of $\Omega(n \log_b(n/\ln n))$ for $p > \ln n/n$ in the edge case and $\Omega(n \log_b(pn/5))$ for $p > 5/n$ in the vertex case. Both bounds reduce to $\Omega(n \log_b n)$, for a wide range of p, namely the lower bound of $\Omega(n \log_b n)$ holds for $p \geq 1/n^{\gamma}, \gamma < 1$.

The constructions for the vertex and edge variants are different; first let's describe the edge variant which is simpler.

Lemma 15. *For any $n > e^{e^2}$ and $\frac{\ln n}{n} \leq p < 1$, $b = 1/(1-p)$, there is a weighted graph G on n vertices, such that if $Pr[MST(E(p)) \subseteq Q] \geq \frac{1}{e}$ then $|Q| > ln - \binom{l+1}{2}$ where $l = \lfloor \log_b(n/\ln n) \rfloor$.*

Proof. Consider the complete graph K_n with edge weights ordered lexicographically. For $i < j$, let

$$w_{ij} = ni + j$$

(see Figure 1 in Section 1). Let $F = E(p)$ be a random subset of edges. For each edge (j,k), $j < k$, consider an event A_{jk}, which occurs when

$$(j,k) \in F \ \& \ \forall i < j; (i,k) \notin F.$$

Due to the ordering of edge weights, A_{jk} implies that $(j,k) \in MST(F)$, since it is the lightest edge in F, incident with vertex k. Also, for given k, A_{jk} can occur only for one value of j. For a set Q of given size, we estimate the probability that A_{jk} occurs for some $(j,k) \in E \setminus Q$.

Let $J_k = \{j : j < k, (j,k) \in E \setminus Q\}$. Since the events A_{jk} for different elements of J_k are disjoint,

$$Pr[\bigcup_{j \in J_k} A_{jk}] = \sum_{j \in J_k} p(1-p)^{j-1}.$$

The events $\bigcup_{j \in J_k} A_{jk}$ for different k's are mutually independent, since the sets of edges involved for different J_k's are disjoint. Therefore:

$$Pr[MST(F) \subseteq Q] \leq Pr[\bigcap_{(j,k) \in E \setminus Q} \overline{A_{jk}}] = \prod_k Pr[\bigcap_{j \in J_k} \overline{A_{jk}}]$$

$$= \prod_k (1 - \sum_{j \in J_k} p(1-p)^{j-1}) \leq \exp\left(- \sum_{(j,k) \in E \setminus Q} p(1-p)^{j-1} \right).$$

For a given size of Q, the last expression is maximized when Q contains edges (j,k) with minimum possible values of j. Assume that Q contains all the edges (j,k) for $j = 1, 2, \ldots l$. Then $|Q| = \sum_{j=1}^{l} (n-j) = ln - \binom{l+1}{2}$ and

$$\sum_{(j,k) \in E \setminus Q} p(1-p)^{j-1} = \sum_{j=l+1}^{n-1} (n-j) p (1-p)^{j-1}.$$

Let's denote this sum by $S(l)$. As can be verified by backward induction on l,

$$S(l) = \left(n - l - \frac{1}{p}\right)(1-p)^l + \frac{1}{p}(1-p)^n.$$

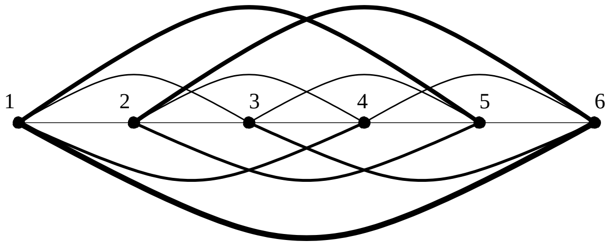

Figure 2: The lower bound example for random sampling of vertices. Edge weights are marked by thickness.

We have that for any Q of size at most $ln - \binom{l+1}{2}$, $Pr[MST(F) \subseteq Q] \leq e^{-S(l)}$.

Let's choose $l = \lfloor \log_b(n/\ln n) \rfloor$. Then, for $p \geq \frac{\ln n}{n}$,

$$S(l) \geq \left(n - \log_b\left(\frac{n}{\ln n}\right) - \frac{1}{p}\right)\frac{\ln n}{n} \geq \left(n - \frac{\ln n - \ln\ln n + 1}{p}\right)\frac{\ln n}{n} \geq \ln\ln n - 1.$$

Therefore, for any set Q of size at most $ln - \binom{l+1}{2}$, $Pr[MST(W) \subseteq Q] \leq e^{-S(l)} < 1/e$ for $n > e^{e^2}$. □

Note. For $p \geq 1/n^{\gamma}, \gamma < 1$, we have $l = (1 - o(1)) \log_b n << n$ and therefore any Q achieving at least a constant probability of MST-covering must have size $|Q| > (1 - o(1))n \log_b n$.

We now describe our lower bound in the vertex case.

Lemma 16. *For any $n > 5$, and $p \geq 5/n$, there exists a weighted graph G on n vertices, such that if $Pr[MST(V(p)) \subseteq Q] \geq \frac{1}{e}$ then $|Q| > ln - \binom{l+1}{2}$ where $l = \lfloor \log_b(np/5) \rfloor$.*

Proof. Let G be the complete graph K_n. Consider the vertices placed on a line uniformly and define edge weights by distances between pairs of vertices, breaking ties arbitrarily, for example:

$$w_{ij} = n|j - i| + i.$$

Let $W = V(p)$ be a random subset of vertices. For each edge (j, k), $j < k$, consider an event A_{jk} as a boolean function of W:

$$A_{jk}(W) \Longleftrightarrow j \in W,\ k \in W \ \&\ \forall i; j < i < k \Rightarrow i \notin W.$$

This event is equivalent to $(j, k) \in MST(W)$, since (j, k) must be in $G[W]$ and no path connecting j, k via a vertex in between can be in $G[W]$. However, we have to be more careful here, because these events are not necessarily independent, which might ruin our bound on the probability of their union. Therefore, we have to impose an additional condition which we deal with later. Assume that C is a set of edges satisfying

(*) *There is no pair of edges $(i, j), (j, k) \in C$ such that $i < j < k$.*

Then we claim that the events $\overline{A_{jk}}$ for $(j, k) \in C$ are never positively correlated. More specifically, if $(u_0, v_0), (u_1, v_1), \ldots (u_k, v_k) \in C, u_i < v_i$, then

$$Pr[\overline{A_{u_1v_1}} \cap \overline{A_{u_2v_2}} \cap \ldots \overline{A_{u_kv_k}} \mid \overline{A_{u_0v_0}}] \leq Pr[\overline{A_{u_1v_1}} \cap \overline{A_{u_2v_2}} \cap \ldots \overline{A_{u_kv_k}}]. \tag{3}$$

We prove this in the following way: For any $W \subseteq V$, define $W' = W \cup \{u_0, v_0\} \setminus \{i : u_0 < i < v_0\}$. Then $A_{u_0v_0}(W')$ is true by definition. In fact, if $W = V(p)$ is a random subset, W' is a random subset sampled in the same way as W, but conditioned on $A_{u_0v_0}$. Now consider the other edges, $(u_1,v_1),\ldots(u_k,v_k)$. Let's call an edge (u_i,v_i) "interfering" with (u_0,v_0), if its interval $[u_i,v_i]$ intersects $[u_0,v_0]$. Property $(*)$ implies that the intervals $[u_0,v_0]$, $[u_i,v_i]$ cannot share exactly one point, so either one of u_i,v_i is an internal point of $[u_0,v_0]$, or one of u_0,v_0 is an internal point of $[u_i,v_i]$. Either way, $A_{u_iv_i}(W')$ cannot be true, because then u_i and v_i ought to be in W' and all vertices inside $[u_i,v_i]$ ought not to be in W' which is impossible. Therefore, $A_{u_iv_i}(W')$ is always false for (u_i,v_i) interfering with (u_0,v_0). On the other hand, if an edge (u_i,v_i) is not interfering with (u_0,v_0), then $A_{u_iv_i}(W')$ if and only if $A_{u_iv_i}(W)$, because W' does not differ from W on the vertices outside of $[u_0,v_0]$.

Thus we have demonstrated that in any case, $\overline{A_{u_iv_i}(W)} \Rightarrow \overline{A_{u_iv_i}(W')}$, and W' corresponds to random sampling conditioned on $A_{u_0v_0}$, which implies

$$Pr[\overline{A_{u_1v_1}} \cap \overline{A_{u_2v_2}} \cap \ldots \overline{A_{u_kv_k}} \mid A_{u_0v_0}] \geq Pr[\overline{A_{u_1v_1}} \cap \overline{A_{u_2v_2}} \cap \ldots \overline{A_{u_kv_k}}].$$

This is equivalent to Eq. 3. As a consequence, we get

$$Pr[\bigcap_{e\in C} \overline{A_e}] \leq \prod_{e\in C} Pr[\overline{A_e}].$$

For a set C satisfying $(*)$, we can now estimate the probability that none of these edges appear in $MST(W)$:

$$Pr[C\cap MST(W)=\emptyset] = Pr[\bigcap_{(j,k)\in C} \overline{A_{jk}}] \leq \prod_{(j,k)\in C} Pr[\overline{A_{jk}}]$$

$$= \prod_{(j,k)\in C} (1-p^2(1-p)^{|k-j|-1}) < \exp\left(-\sum_{(j,k)\in C} p^2(1-p)^{|k-j|-1}\right).$$

Suppose that Q has size at most $\sum_{j=1}^{l}(n-j) = ln - \binom{l+1}{2}$. The optimal way to minimize $\sum_{(j,k)\in E\setminus Q} p^2(1-p)^{|k-j|-1}$ is to choose Q to contain all the edges of length at most l. Then we have

$$\sum_{(j,k)\in E\setminus Q} p^2(1-p)^{|k-j|-1} = \sum_{j=l+1}^{n-1}(n-j)p^2(1-p)^{j-1} = p\ S(l)$$

where $S(l)$ is defined in the previous proof, $S(l) = (n-l-\frac{1}{p})(1-p)^l + \frac{1}{p}(1-p)^n$. We choose $l = \lfloor \log_b(np/5)\rfloor$ and then:

$$p\ S(l) \geq (p(n-l)-1)(1-p)^l \geq (p(n-\log_b(np/5))-1)\frac{5}{np} \geq 5\left(1-\frac{\ln(np/5)+1}{np}\right) \geq 4$$

for $np \geq 5$. Thus we have $\sum_{(j,k)\in E\setminus Q} p^2(1-p)^{|k-j|-1} \geq 4$ for any Q of size at most $ln - \binom{l+1}{2}$. Now we apply the probabilistic method to choose a suitable subset of $E\setminus Q$. We sample a uniformly random subset of vertices S. Let $C = \{(j,k)\in E\setminus Q : j<k, j\in S, k\notin S\}$. For each edge in $E\setminus Q$, there is probability $1/4$ that it appears in C. Therefore

$$\mathbf{E}[\sum_{(j,k)\in C} p^2(1-p)^{|k-j|-1}] \geq \frac{1}{4}\sum_{(j,k)\in E\setminus Q} p^2(1-p)^{|k-j|-1} \geq 1$$

and there exists a set C which achieves at least this expectation. Due to the construction of C, it satisfies condition (*), and we can conclude that

$$Pr[MST(W)\subseteq Q] \leq Pr[C\cap MST(W)=\emptyset] < e^{-1}.$$

This is a contradiction, and so Q must be larger than $ln - \binom{l+1}{2}$. □

Note. This bound becomes void as p approaches $5/n$. On the other hand, for $p = c/n$, $c > 5$ fixed, we get $|Q| = \Omega(n^2)$. For $p = 1/n^\gamma, \gamma < 1$, we get $|Q| > (1 - \gamma - o(1))n \log_b n$. For p constant, we get $|Q| > (1 - o(1))n \log_b n$. This confirms the optimality of our results up to a constant factor, for a wide range of sampling probabilities p.

References

[1] D. Applegate and W. Cook: Solving large-scale matching problems, In D. Johnson and C.C. McGeoch, eds., Network Flows and Matchings, AMS 1993.

[2] B. Bollobás: Combinatorics - set systems, hypergraphs, families of vectors, and combinatorial probability, Cambridge University Press 1986.

[3] B. Bollobás and A. Thomason: Threshold functions, Combinatorica 7 (1987), 35–38.

[4] B. Chazelle: A minimum spanning tree algorithm with inverse-Ackermann type complexity, J. ACM 47 (2000), 1028–1047.

[5] M.X. Goemans and J. Vondrák: Covering the minimum spanning trees of random subgraphs, The ACM-SIAM Symposium on Discrete Algorithms 2004, 927–934.

[6] J. Holm, K. De Lichtenberg and M. Thorup: Poly-logarithmic deterministic fully-dynamic algorithms for connectivity, minimum spanning tree, 2-edge, and biconnectivity, J. ACM 48(4) (2001), 723–760.

[7] N. Jing, Y.W. Huang and E.A.Rundensteiner: Hierarchical encoded path views for path query processing: An optimal model and its performance evaluation, IEEE T. on Knowledge and Data Engineering 10(3) (1998), 409–432.

[8] D.R. Karger, P.N. Klein and R.E. Tarjan: A randomized linear-time algorithm to find minimum spanning trees, J. ACM 42(2) (1995), 321–328.

[9] R. Motwani and P. Raghavan: Randomized algorithms, Cambridge University Press 1995.

[10] L. Valiant: The complexity of enumeration and reliability problems, SIAM J. on Computing 8 (1979), 410–421.

[11] J. Vondrák: Probabilistic methods in combinatorial and stochastic optimization, Ph.D. thesis, MIT (2005).