Quantifying Variational Approximation for Log-Partition Function

Romain Cosson   COSSON@MIT.EDU
Massachusetts Institute of Technology

Devavrat Shah   DEVAVRAT@MIT.EDU
Massachusetts Institute of Technology

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

Variational methods, such as mean-field (MF) and tree-reweighted (TRW), provide computationally efficient approximations of the log-partition function for generic graphical models but their approximation ratio is generally not quantified. As the primary contribution of this work, we provide an approach to quantify their approximation ratio for any discrete pairwise graphical model with non-negative potentials through a property of the underlying graph structure G. Specifically, we argue that (a variant of) TRW produces an estimate within factor $1/\sqrt{\kappa(G)}$ where $\kappa(G) \in (0, 1]$ captures how far G is from tree structure. As a consequence, the approximation ratio is 1 for trees, $\sqrt{(d+1)/2}$ for graphs with maximum average degree d and $1 + 1/(2\beta) + o_{\beta \to \infty}(1/\beta)$ for graphs with girth at least $\beta \log N$. The quantity $\kappa(G)$ is the solution of a min-max problem associated with the spanning tree polytope of G that can be evaluated in polynomial time for any graph. We provide a near linear-time variant that achieves an approximation ratio depending on the minimal (across edges) effective resistance of the graph. We connect our results to the graph partition approximation method and thus provide a unified perspective.

Keywords: variational inference, log-partition function, spanning tree polytope, minimum effective resistance, balanced covering of graph, min-max spanning tree, local inference

1. Introduction

The Setup. We consider a collection of N discrete valued random variables on a discrete alphabet, $\mathbf{X} = (X_1, \ldots, X_N)$, whose joint distribution is modeled as a pair-wise graphical model. Let $G = (V, E)$ represent the associated graph with vertices $V = \{1, \ldots, N\}$ representing N variables and $E \subset V \times V$ representing edges. Let each variable take value in a discrete alphabet $\mathcal{X}$. For $e \in E$, let $\phi_e : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ denote the edge potential that we assume takes only non-negative values and let $\theta_e \in \mathbb{R}_+$ denote the associated weight. This leads to a joint distribution with probability mass function

$$\mathbb{P}(\mathbf{X} = \boldsymbol{x}; \boldsymbol{\theta}) \propto \exp\Big(\sum_{e \in E} \theta_e \phi_e(x_e)\Big) = \frac{1}{Z(\boldsymbol{\theta})} \exp\Big(\sum_{e \in E} \theta_e \phi_e(x_e)\Big), \qquad (1)$$

where $\boldsymbol{x} = (x_1, \ldots, x_N) \in \mathcal{X}^N$, x_e is short hand for (x_s, x_t) if $e = (s, t) \in E$, $\boldsymbol{\theta} = (\theta_e : e \in E) \in \mathbb{R}_+^{|E|}$ and the normalizing constant or partition function $Z(\boldsymbol{\theta})$ is defined as

$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x} \in \mathcal{X}^N} \exp\Big(\sum_{e \in E} \theta_e \phi_e(x_e)\Big). \qquad (2)$$

Such pairwise graphical models provide succinct description for complicated joint distributions. However, the key challenge in utilizing them (e.g. for inference) arises in estimating the partition function $Z(\boldsymbol{\theta})$. In this work, our interest is in computing logarithm of $Z(\boldsymbol{\theta})$, precisely

$$\Phi(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta}) \;=\; \log\left[\sum_{\boldsymbol{x}\in\mathcal{X}^N}\exp\left(\sum_{e\in E}\theta_e\phi_e(x_e)\right)\right]. \tag{3}$$

Computing $Z(\boldsymbol{\theta})$ is known to be computationally hard in general, i.e. #P-complete due to relation to counting discrete objects such as independent sets cf. Valiant (1979); Jerrum and Sinclair (1989). Due to reductions from discrete optimization problems to log-partition function computation, approximating $\Phi(\boldsymbol{\theta})$, even up to a multiplicative error, can be NP-hard cf. Weitz (2006); Wainwright and Jordan (2008); Dembo et al. (2013). Therefore, the goal is to develop polynomial time (in N) approximation method for $\Phi(\boldsymbol{\theta})$ with provable guarantees. Specifically, let ALG denote such an approximation method that takes problem description $\mathcal{P} = (G, \mathcal{X}, (\phi_e)_{e\in E}, \boldsymbol{\theta})$ as input and produces estimate $\widehat{\Phi}^{\mathrm{ALG}}(\boldsymbol{\theta})$ for $\Phi(\boldsymbol{\theta})$. Then, we define approximation ratio associated with ALG, $\alpha(G, \mathrm{ALG}) \geq 1$ as

$$\alpha(G, \mathrm{ALG}) = \sup_{\mathcal{P}}\max\left(\frac{\Phi(\boldsymbol{\theta})}{\widehat{\Phi}^{\mathrm{ALG}}(\boldsymbol{\theta})}, \frac{\widehat{\Phi}^{\mathrm{ALG}}(\boldsymbol{\theta})}{\Phi(\boldsymbol{\theta})}\right). \tag{4}$$

Prior Work. There is a long literature on computationally efficient approximation methods for the log-partition function with significant progress in the past two decades. We recall some relevant prior works here.

A collection of methods, classified as variational approximations, utilize the (Gibbs) variational characterization of the log-partition function when distribution (1) is viewed as a member of an exponential family, cf. Georgii (2011); Wainwright and Jordan (2008). Specifically, $\Phi(\boldsymbol{\theta})$ can be viewed as a solution of a high-dimensional constrained maximization problem. By solving the problem with additional constraints, one obtains a valid lower bound such as that given by Mean-Field methods. By utilizing the convexity of $\Phi(\cdot)$ and restricting it to tree-structured sub-graphs of G, one obtains a valid upper bound such as that given by the tree-reweighted (TRW) method. By relaxing the constraints and adapting the objective to allow for pairwise pseudo-marginals, one obtains heuristics such as Belief Propagation (BP) via Bethe approximation Yedidia et al. (2001, 2003). While BP does not provide provable upper or lower bounds in general, for graphs with large-girth such as sparse random graphs and distributions with spatial decay of correlation, it provides an excellent approximation cf. Mezard and Montanari (2009). The spatial decay of correlation property has been further exploited to obtain a deterministic Fully Polynomial Time Approximation Schemes (FPTAS) for various counting problems, i.e. computing partition functions cf. Weitz (2006); Gamarnik and Katz (2012); Bayati et al. (2007); Gamarnik and Katz (2009). The approximation error of belief propagation has been studied through connection to loop calculus as well cf. Chertkov and Chernyak (2006); Chandrasekaran et al. (2011).

In another line of works, graph partitioning based methods have been proposed to provide Polynomial Time Approximation Schemes (PTAS) for classes of graphs that satisfy certain graph partitioning properties which includes minor-excluded graphs Jung and Shah (2006) or graphs with polynomial growth Jung et al. (2009).

In summary, despite the progress, the approximation ratio $\alpha(G, \mathrm{ALG})$ for any of the known variational approximation methods ALG remains undetermined.

Summary of Contributions. As the main contribution, for a simple variant of tree-reweighted (TRW) method, denoted as TRW$'$, we quantify $\alpha(G, \text{TRW}')$ for any G. TRW$'$ is described in Section 3 and produces an estimate of $\Phi(\cdot)$ in polynomial time. Specifically, we establish

Theorem 1 *For any graph G, the approximation ratio of TRW$'$ is such that*

$$\alpha(G, \textit{TRW}') \leq 1/\sqrt{\kappa(G)} \quad \textit{where} \quad \kappa(G) = \min_{S \subset V} \frac{|S| - 1}{|E(S)|}, \tag{5}$$

with $E(S) = E \cap (S \times S)$ for any $S \subset V$.

The term $\kappa(G)$ captures the proximity of G with respect to the tree structure across all of its induced sub-graphs: for $S \subset V$, the induced subgraph $(S, E(S))$ would have at most $|S| - 1$ edges if it were cycle free, but it has $|E(S)|$ edges. Therefore, the ratio $(|S| - 1)/|E(S)|$ measures how far it is from a tree. It is equal to 1 for a connected tree and $2/|S|$ for the complete graph. The minimum over all possible $S \subset V$ of this ratio captures how far G is from a tree structure.

Using this characterization, we provide bounds on $\alpha(G, \text{TRW}')$ in terms of various simpler graph properties in Section 4.3. Specifically, we show that for any graph with maximum average degree $d \geq 1$, $\alpha(G, \text{TRW}') \leq \sqrt{(d+1)/2}$. And for graphs with girth (i.e. length of shortest cycle) $g > 3$, $\alpha(G, \text{TRW}') \leq \sqrt{\frac{1 + N^{2/(g-3)}}{2(1 - 1/g)}}$, which implies $\alpha(G, \text{TRW}') \leq \left(1 + \frac{1}{2\beta} + o_{\beta \to \infty}(\frac{1}{\beta})\right)$ if $g \geq \beta \log N$. This means that for *any G* with large ($\gg \log N$) girth, $\alpha(G, \text{TRW}') \approx 1$.

In general, we establish that $\kappa(G)$ can be evaluated in polynomial time for any graph G by solving an appropriate linear program on the spanning tree polytope. This is explained in Section 4.

The tree-reweighted variant TRW$'$ uses a linear solver over the tree polytope of G, which can be hard to implement in practice. With an eye towards near linear-time (in $|E|$) computation, a variant that instead of optimizing over the tree polytope simply considers a feasible point that corresponds to the uniform distribution over spanning trees of G. Using the near-linear time sampling of spanning tree from Schild (2018), we provide a randomized approximation method for $\Phi(\boldsymbol{\theta})$. With high probability, its approximation ratio $\alpha(G)$ is bounded above by $1/\sqrt{\min_{e \in E} r_e}$ where $r_e \geq 0$ is the effective resistance of $e \in E$ for the graph $G = (V, E)$ (see (37) for precise definition). While in general, this provides a weaker approximation guarantee, for graphs with degree bounded by d it leads to a similar guarantee of $\alpha(G) \leq \sqrt{(d+1)/2}$.

We show that the results based on graph partitioning cf. Jung and Shah (2006); Jung et al. (2009) can be recovered as a natural extension of the variant of TRW introduced in this work by allowing for general graphs with bounded tree-width beyond trees.

We take note of the fact that though results discussed in this work are primarily for the variant of TRW described in Section 3, as an immediate consequence of our results, $\alpha(G, \text{TRW}) \leq 1/\kappa(G)$, i.e. it is bounded by the square of that derived in Theorem 1. As discussed in Section 8, understanding the tightness of this characterization especially for TRW remains an important open direction.

Outline of Paper. In Section 2, we provide some preliminaries and recall the tree-reweighted (TRW) method. In Section 3, we provide a modification of TRW and characterize its approximation guarantee. In Section 4, we provide a linear optimization characterization of the approximation guarantee which leads to the proof of Theorem 1. We also discuss implications of Theorem 1 for various classes of graphs. In Section 5, we present a near linear-time variant based on sampling

3

from the uniform distribution of spanning trees over G. We derive approximation guarantees for the resulting method in terms of the effective resistance of the graph and derive its implications. In Section 6, we discuss connection with graph partitioning methods by extending the modified TRW of Section 3 to allow for bounded tree-width subgraphs beyond trees. We argue how results of Jung and Shah (2006); Jung et al. (2009) follow naturally. Section 8 discusses directions for future work.

2. Preliminaries and Background

2.1. Variational Characterization, Mean-Field Approximation and Belief Propagation

We start by recalling the variational characterization of the log-partition function $\Phi(\cdot)$. Let $\mathcal{P}(\mathcal{X}^N)$ denote the space of all probability distributions over $\mathcal{X}^N$. Then, the Gibbs variational characterization states that

$$\Phi(\boldsymbol{\theta}) = \sup_{q \in \mathcal{P}(\mathcal{X}^N)} \mathbb{E}_{\mathsf{x} \sim q} \left(\sum_e \theta_e \phi_e(\mathsf{x}_e) \right) + H(q), \tag{6}$$

where $H(q) = -\mathbb{E}_{\mathsf{x} \sim q}(\log(q(\mathsf{x})))$ is the entropy of q. While computationally (6) does not provide tractable solution for evaluating $\Phi(\cdot)$, it provides a framework to develop approximation methods that we refer to as *variational approximations*.

As mentioned earlier, the classical mean-field (MF) consists in relaxing $\mathcal{P}(\mathcal{X}^N)$ to the space of independent distributions over $\mathcal{X}^N$ denoted as $\mathcal{I}(\mathcal{X}^N) \subset \mathcal{P}(\mathcal{X}^N)$. By restricting optimization in (6) to $\mathcal{I}(\mathcal{X}^N)$, one obtains a lower bound on $\Phi(\boldsymbol{\theta})$.

It turns out that (6) can be solved efficiently for tree-structured graph. Specifically, if G is a connected tree, i.e. G is connected with no cycle, then any distribution satisfying (1) can be re-parametrized as

$$\mathbb{P}(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{u \in V} \mathbb{P}_{X_u}(x_u) \prod_{(u,v) \in E} \frac{\mathbb{P}_{X_u, X_v}(x_u, x_v)}{\mathbb{P}_{X_u}(x_u) \mathbb{P}_{X_v}(x_v)}. \tag{7}$$

In the expression above, $\mathbb{P}_{X_u}(\cdot)$ denotes the marginal distribution of $X_u, u \in V$ and $\mathbb{P}_{X_u, X_v}(\cdot, \cdot)$ denotes the pairwise marginal distribution of (X_u, X_v) for any edge $e = (u, v) \in E$. The Belief Propagation (or sum-product) algorithm can compute these marginal distributions efficiently for tree graphs using only knowledge of $\boldsymbol{\theta}$ and $\phi_e, e \in E$. It utilizes $O(|\mathcal{X}|^2 N)$ computation time. Therefore, $Z(\boldsymbol{\theta})$ and hence $\Phi(\boldsymbol{\theta})$ can be computed for tree graphs using $O(|\mathcal{X}|^2 N)$ computations.

Indeed, the re-parametrization of the form (7) was a basis for the Belief Propagation (BP) algorithm for generic graphical models and also led to the so called Bethe Approximation of (6), cf. Yedidia et al. (2001). However, it does not result in a provably upper or lower bound in general (with few exceptions).

To obtain an upper bound on $\Phi(\cdot)$, its convexity was exploited in Wainwright et al. (2005) along with the fact that (6) is solvable efficiently for tree-structured graph. This resulted into tree-reweighted (TRW) algorithm which we describe next.

2.2. Tree-Reweighted (TRW): An Upper Bound on $\Phi(\cdot)$

Recall that a spanning tree T is a subgraph of $G = (V, E)$ that contains all vertices and a subset of the edges so that T does not have a cycle. Let $\mathcal{T}(G)$ be the set of all spanning trees of G. We

shall denote a distribution on $\mathcal{T}(G)$ as $\boldsymbol{\rho} = (\rho^T)_{T \in \mathcal{T}(G)}$ where $\rho^T \geq 0$ for all $T \in \mathcal{T}(G)$ and $\sum_{T \in \mathcal{T}(G)} \rho^T = 1$. The space of all distributions on $\mathcal{T}(G)$ is denoted by $\mathcal{P}(\mathcal{T}(G))$. For simplicity, we shall drop the notation of G at times when it is clear from the context and denote it simply as $\mathcal{P}(\mathcal{T})$. A distribution $\boldsymbol{\rho} \in \mathcal{P}(\mathcal{T})$ induces for all edge $e \in E$ an edge probability ρ_e that this edge will appear in a tree selected from $\boldsymbol{\rho}$,

$$\rho_e = \mathbb{P}_{\mathsf{T} \sim \boldsymbol{\rho}}\big(e \in \mathsf{T}\big) = \sum_{T \in \mathcal{T}(G)} \rho^T \mathbb{1}(e \in T). \tag{8}$$

Note that in the above, we have abused notation using T as a spanning tree as well as the set of edges constituting it. We shall continue using this notation since the all spanning trees have the same set of vertices and only the edges differ (among subsets of E). Also note another convenient abuse of notation: given $\boldsymbol{\rho}$, ρ^T denotes probability of $T \in \mathcal{T}(G)$ while ρ_e is the probability of edge $e \in E$ being present in a random tree. Note that because all spanning trees have $N - 1$ edges, edge probabilities satisfy $\sum_{e \in E} \rho_e = N - 1$. Given $\boldsymbol{\rho} \in \mathcal{P}(\mathcal{T}(G))$, we now define $\kappa(G, \boldsymbol{\rho})$ as

$$\kappa(G, \boldsymbol{\rho}) = \min_{e \in E} \rho_e. \tag{9}$$

For any $\boldsymbol{\theta} \in \mathbb{R}_+^{|E|}$, define its support as $s(\boldsymbol{\theta}) = \{e \in E : \theta_e \neq 0\}$. Given a spanning tree $T \in \mathcal{T}(G)$, let $\theta^T \in \mathbb{R}_+^{|E|}$ be such that $s(\theta^T) \subset T$ and let $\boldsymbol{\rho} \in \mathcal{P}(\mathcal{T})$ along with $(\theta^T)_{T \in \mathcal{T}}$ be such that $\sum_{T \in \mathcal{T}} \rho^T \theta^T = \theta$. That is, $\mathbb{E}_{\mathsf{T} \sim \boldsymbol{\rho}}\big[\boldsymbol{\theta}^{\mathsf{T}}\big] = \boldsymbol{\theta}$. Therefore, we can write

$$\Phi(\boldsymbol{\theta}) = \Phi\big(\mathbb{E}_{\mathsf{T} \sim \boldsymbol{\rho}}\big[\theta^{\mathsf{T}}\big]\big). \tag{10}$$

It has been established that $\Phi : \mathbb{R}_+^{|E|} \to \mathbb{R}$ is a convex function because its Hessian is a covariance matrix and hence is positive semidefinite (see Wainwright et al. (2005)). From Jensen's inequality applied to (10) it follows that

$$\Phi(\boldsymbol{\theta}) \leq \mathbb{E}_{\mathsf{T} \sim \boldsymbol{\rho}}\big[\Phi(\theta^{\mathsf{T}})\big] = \sum_{T \in \mathcal{T}} \rho^T \Phi(\theta^T). \tag{11}$$

Since the upper bound (53) holds for any $\boldsymbol{\rho} \in \mathcal{P}(\mathcal{T})$ and $(\theta^T)_{T \in \mathcal{T}}$ such that $\sum_{T \in \mathcal{T}} \rho^T \theta^T = \boldsymbol{\theta}$ we can optimize on these two parameters to obtain the tree-reweighted upper bound

$$\Phi(\boldsymbol{\theta}) \leq \inf_{\sum_{T \in \mathcal{T}} \rho^T \theta^T = \boldsymbol{\theta}} \left(\sum_{T \in \mathcal{T}} \rho^T \Phi(\theta^T) \right) \equiv U^{\mathrm{TRW}}(\boldsymbol{\theta}). \tag{12}$$

As established in Wainwright et al. (2005), this seemingly complicated bound, $U^{\mathrm{TRW}}(\boldsymbol{\theta})$, can be computed via an iterative *tree-reweighted message-passing* algorithm through the dual of the above optimization problem. While this is a valid upper bound, how tight the upper bound is for a given graphical model is not quantified in the literature. And this is precisely the primary contribution of this work.

3. Algorithm and Approximation Guarantee

Modified Tree-Reweighted: TRW′. We describe a simple variant of TRW that enables us to bound the approximation ratio of the estimation of Φ using properties of G. We start with some useful

notations. Given $\boldsymbol{\theta} = (\theta_e)_{e \in E} \in \mathbb{R}_+^{|E|}$, $\boldsymbol{\rho} \in \mathcal{P}(\mathcal{T}(G))$ and a spanning tree $T \in \mathcal{T}(G)$ of graph G, define "projection" operations

$$\Pi^T : \mathbb{R}_+^{|E|} \to \mathbb{R}_+^{|E|} \quad \text{where} \quad \Pi^T(\boldsymbol{\theta}) = \big(\mathbb{1}(e \in T)\theta_e\big)_{e \in E}$$

$$\Pi_{\boldsymbol{\rho}}^T : \mathbb{R}_+^{|E|} \to \mathbb{R}_+^{|E|} \quad \text{where} \quad \Pi_{\boldsymbol{\rho}}^T(\boldsymbol{\theta}) = \big(\mathbb{1}(e \in T)\theta_e/\rho_e\big)_{e \in E}. \tag{13}$$

With these notations, for a given $\boldsymbol{\rho} \in \mathcal{P}(\mathcal{T}(G))$ define

$$L_{\boldsymbol{\rho}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathsf{T} \sim \boldsymbol{\rho}}(\Phi(\Pi^{\mathsf{T}}(\boldsymbol{\theta}))) = \sum_{T \in \mathcal{T}(G)} \rho^T \Phi(\Pi^T(\boldsymbol{\theta})), \tag{14}$$

$$U_{\boldsymbol{\rho}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathsf{T} \sim \boldsymbol{\rho}}(\Phi(\Pi_{\boldsymbol{\rho}}^{\mathsf{T}}(\boldsymbol{\theta}))) = \sum_{T \in \mathcal{T}(G)} \rho^T \Phi(\Pi_{\boldsymbol{\rho}}^T(\boldsymbol{\theta})). \tag{15}$$

For a given $\boldsymbol{\rho} \in \mathcal{P}(\mathcal{T}(G))$, one obtains an estimate of $\Phi(\boldsymbol{\theta})$ as follows:

$$\widehat{\Phi}_{\boldsymbol{\rho}}(\boldsymbol{\theta}) = \sqrt{L_{\boldsymbol{\rho}}(\boldsymbol{\theta})U_{\boldsymbol{\rho}}(\boldsymbol{\theta})}. \tag{16}$$

TRW$'$ outputs $\widehat{\Phi}_{\boldsymbol{\rho}^\star(G)}(\boldsymbol{\theta})$ where $\boldsymbol{\rho}^\star(G) \in \mathcal{P}(\mathcal{T}(G))$ is defined as:

$$\boldsymbol{\rho}^\star(G) \in \underset{\boldsymbol{\rho} \in \mathcal{P}(\mathcal{T}(G))}{\arg\max} \kappa(G, \boldsymbol{\rho}) \quad \text{with} \quad \kappa(G, \boldsymbol{\rho}) = \min_{e \in E} \rho_e. \tag{17}$$

Interpretation of $\boldsymbol{\rho}^*(G)$. The distribution $\boldsymbol{\rho}^*(G) \in \mathcal{P}(\mathcal{T}(G))$ can be viewed as a "balanced cover" of G by a convex combination of its spanning trees (see Appendix A for more intuition). The problem of computing $\rho^*(G)$ in polynomial time and characterizing $\kappa(G) = \kappa(G, \boldsymbol{\rho}^*(G))$ is addressed in Section 4. The lemma below quantifies the approximation ratio for TRW$'$. Its proof is in Appendix B.

Lemma 2 *Given $\boldsymbol{\theta} \in \mathbb{R}_+^{|E|}$, TRW$'$ produces $\widehat{\Phi}_{\boldsymbol{\rho}^\star}(\boldsymbol{\theta})$ with $\boldsymbol{\rho}^\star = \boldsymbol{\rho}^\star(G)$ as defined in (17). Then,*

$$\alpha(G, TRW') \leq \frac{1}{\sqrt{\kappa(G)}} \quad \text{where} \quad \kappa(G) = \max_{\boldsymbol{\rho} \in \mathcal{P}(\mathcal{T}(G))} \big(\min_{e \in E} \rho_e\big). \tag{18}$$

4. $\kappa(G)$: Efficient computation, characterization

Lemma 2 establishes the approximation guarantee for TRW$'$ as claimed in Theorem 1 with caveat that it is in terms of $\kappa(G) = \max_{\boldsymbol{\rho} \in \mathcal{P}(\mathcal{T}(G))} \big(\min_{e \in E} \rho_e\big)$ while Theorem 1 relates it to a structural property of the graph defined in (5). In this section, we shall establish this identity which will allow us to bound $\kappa(G)$ for certain classes of graphs and obtain meaningful intuitions. In the process, we will argue that $\boldsymbol{\rho}^*(G)$ can be computed in polynomial time for any graph G.

4.1. Computing $\boldsymbol{\rho}^\star(G)$ and $\kappa(G)$ efficiently

Spanning Tree Polytope. We will use a notion of spanning tree polytope for a given graph G. Recall that $\mathcal{T}(G)$ is the set of all spanning trees of G. For any tree $T \in \mathcal{T}(G)$, we shall utilize the notation of $\chi^T = [\chi_e^T] \in \{0, 1\}^E$ to represent the characteristic vector of the tree T defined such that

$$\forall e \in E : \chi_e^T = \mathbf{1}(e \in T). \tag{19}$$

Given this notation, we define the polytope of spanning trees of G, denoted $\mathsf{P}^{\text{tree}}(G)$, as the convex hull of their characteristic vectors. That is,

$$\mathsf{P}^{\text{tree}}(G) = \Big\{ \mathbf{v} \in [0,1]^E : \mathbf{v} = \sum_{T \in \mathcal{T}(G)} \rho^T \chi^T, \ \sum_{T \in \mathcal{T}(G)} \rho^T = 1, \ \rho_T \geq 0, \ \forall T \in \mathcal{T}(G) \Big\}. \quad (20)$$

The weights $(\rho^T)_{T \in \mathcal{T}(G)}$ can be viewed as probability distribution on $\mathcal{T}(G)$, i.e. an element of $\mathcal{P}(\mathcal{T}(G))$. Therefore $\mathbf{v} = \sum_{T \in \mathcal{T}(G)} \rho^T \chi^T$ corresponds to a vector representing the probabilities that edges in E will be present in in $\mathsf{T} \sim \boldsymbol{\rho} = (\rho^T)$, i.e. $\mathbf{v} = \mathbb{E}_{\mathsf{T} \sim \boldsymbol{\rho}}[\mathbb{1}(e \in \mathsf{T})]$. That is, $\mathbf{v} = (\rho_e)_{e \in E}$ as defined in (8). Therefore, we shall abuse notation and write

$$\mathsf{P}^{\text{tree}}(G) = \big\{ (\rho_e)_{e \in E} \mid (\rho^T)_{T \in \mathcal{T}(G)} \in \mathcal{P}(\mathcal{T}(G)) \big\}. \quad (21)$$

Edmonds (1971) gave the following characterization of the spanning tree polytope:

$$\mathsf{P}^{\text{tree}}(G) = \left\{ (v_e)_{e \in E} \in \mathbb{R}_+^E \ \middle| \ \begin{array}{r} \forall S \subset E : v(E(S)) \leq |S| - 1 \\ v(E) = |V| - 1 \end{array} \right\}, \quad (22)$$

where $v(E(S)) = \sum_{e \in E(S)} v_e$.

Efficient Separation Oracle. A polytope $\mathsf{P} \subset \mathbb{R}^n$, defined through a set of linear constraints, is said to have a separation oracle if there exists a polynomial time algorithm in n which for any $x \in \mathbb{R}^n$ can determine whether $x \in \mathsf{P}$ or not; and output a violated constraint if $x \notin \mathsf{P}$. Edmond's characterization of the spanning tree polytope, though it has an exponential number of constraints, admits an efficient separation oracle. Such an efficient separation oracle is defined explicitly via a min-cut reduction, see (Lau et al., 2011, Chapter 4.1).

Complexity of Linear Programming. Consider a linear program where the goal is to find the minimum of a linear objective function over a polytope defined by finitely many linear constraints. Such a linear program can be solved in polynomial time (in size of problem description) via the Ellipsoid method if the polytope admits an efficient separation oracle, see (Bertsimas and Tsitsiklis, 1997, Theorem 8.5) for example. Given that the spanning tree polytope has an efficient separation oracle, optimizing a linear objective over it can be solved efficiently. Of course, due to the structure of the trees, a greedy algorithm like that of Kruskal's may be a lot more direct for solving such a linear program. Having said that, the benefit of efficient separation oracle becomes apparent as soon as we consider additional linear constraints beyond those described in $\mathsf{P}^{\text{tree}}(G)$. Indeed, such approaches have found utility in solving other problems, like solving bounded-degree maximum-spanning-tree relaxations like in Goemans (2006).

Augmented Spanning Tree Polytope. We consider a reformulation of the max-min problem in (17). To that end consider the following augmented spanning tree polytope:

$$\mathsf{P}^{\text{tree}}_{\min}(G) = \left\{ (z, (v_e)_{e \in E}) \in \mathbb{R} \times \mathbb{R}_+^{|E|} \ \middle| \ \begin{array}{r} \forall e \in E : z \leq v_e \\ \forall S \subset E : v(E(S)) \leq |S| - 1 \\ v(E) = |V| - 1 \end{array} \right\}. \quad (23)$$

With this notation, we can re-write $\kappa(G)$ as per (17) as

$$\kappa(G) = \max_{(v_e)_{e \in E} \in \mathsf{P}^{\mathrm{tree}}} \{ \min_{e \in E} v_e \} = \max_{(z,(v_e)_{e \in E}) \in \mathsf{P}^{\mathrm{tree}}_{\min}} z. \tag{24}$$

Next, we argue that $\mathsf{P}^{\mathrm{tree}}_{\min}$ admits an efficient separation oracle as follows. The separation oracle for $\mathsf{P}^{\mathrm{tree}}_{\min}$ takes $(z, (v_e)_{e \in E})$ as input. It first checks that all $|E|$ constraints of the form $z \leq v_e$ are satisfied. If one is not satisfied, then the oracle outputs this constraint. If all constraints are satisfied, the algorithm runs the separation oracle of $\mathsf{P}^{\mathrm{tree}}$ on $(v_e)_{e \in E}$ and reproduces its output. Since $|E| \leq N^2$ and $\mathsf{P}^{\mathrm{tree}}$ has an efficient separation oracle, this leads to polynomial time separation oracle for $\mathsf{P}^{\mathrm{tree}}_{\min}$.

Efficient computation of $\boldsymbol{\rho}^\star(G)$ and $\kappa(G)$. From the linear program formulation (24) and from the efficient separation oracle as defined above, we can compute $\kappa(G)$ in polynomial time using the Ellipsoid algorithm. Note that this does not directly provides $\boldsymbol{\rho}^\star(G) \in \mathcal{P}(\mathcal{T}(G))$ since the representation in $\mathsf{P}^{\mathrm{tree}}$ corresponds to the edge probabilities $(\rho^\star(G)_e)_{e \in E}$. However, $(\rho^\star(G)_e)_{e \in E}$ is a convex combination of extreme points of $\mathsf{P}^{\mathrm{tree}}$, which correspond to the spanning trees of G. Since $\mathsf{P}^{\mathrm{tree}}$ has efficient separation oracle, we can recover a decomposition of $(\rho^\star(G)_e)_{e \in E}$ in terms of convex combination of characteristic vectors weighted by $(\rho^\star(G)^T)_{T \in \mathcal{T}(G)}$ and such that at most $|E|$ of these weights are strictly positive, see details in (Grötschel et al., 1981, Theorem 3.9).

4.2. Characterizing $\kappa(G)$

We wish to establish that

$$\kappa(G) = \max_{(v_e)_{e \in E} \in \mathsf{P}^{\mathrm{tree}}} \{ \min_{e \in E} v_e \} = \min_{S \subset V} \frac{|S| - 1}{|E(S)|}. \tag{25}$$

Upper bound: $\kappa(G) \leq \frac{|S|-1}{|E(S)|}$. The upper bound is immediately given by Edmond's characterisation of the spanning tree polytope. For any $(\rho_e)_{e \in E} \in \mathsf{P}^{\mathrm{tree}}$ and any $S \subset V$:

$$|E(S)| \big(\min_{e \in E} \rho_e \big) \leq \sum_{e \in E(S)} \rho_e = \rho(E(S)) \leq |S| - 1. \tag{26}$$

That is, for any $\boldsymbol{\rho} \in \mathcal{P}(\mathcal{T}(G))$

$$\kappa(G, \boldsymbol{\rho}) \leq \min_{S \subset V} \frac{|S| - 1}{|E(S)|}. \tag{27}$$

And hence it holds for $\boldsymbol{\rho}^\star(G)$ as well.

Lower bound: $\kappa(G) \geq \frac{|S|-1}{|E(S)|}$. To establish the lower bound, we need a few additional results. To start with, we define a dual of the optimization problem (24) to characterize $\kappa(G)$. By strong duality it follows that

$$\kappa(G) = \max_{\boldsymbol{\rho} \in \mathcal{P}(\mathcal{T})} \min_{e \in E} \sum_{T \in \mathcal{T}} \mathbb{1}(e \in T) \rho^T = \min_{\mathbf{w} \in \mathcal{P}(E)} \max_{T \in \mathcal{T}} \sum_{e \in E} \mathbb{1}(e \in T) w_e, \tag{28}$$

where $\mathcal{P}(E) = \{ \mathbf{w} = (w_e)_{e \in E} : \sum_{e \in E} w_e = 1, w_e \geq 0 \ \forall \ e \in E \}$. Table 1 provides the precise primal and dual formulation associated with $\kappa(G)$ justifying (28). More intuition and examples are given in Appendix A. We state the following Lemma characterizing an optimal solution of **Dual**, whose proof is in Appendix C.

	Primal	**Dual**
Objective	$\max z$	$\min y$
Variables / Constraints	$\begin{aligned} z &\in \mathbb{R} \\ \forall T \in \mathcal{T} : \rho_T &\in \mathbb{R}_+ \end{aligned}$	$\begin{aligned} \sum_{e \in E} w_e &= 1 \\ \forall T \in \mathcal{T} : y - \sum_{e \in T} w_e &\geq 0 \end{aligned}$
Constraints / Variables	$\begin{aligned} \sum_{T \in \mathcal{T}} \rho_T &= 1 \\ \forall e \in E : \sum_{T \ni e} \rho_T - z &\geq 0 \end{aligned}$	$\begin{aligned} y &\in \mathbb{R} \\ \forall e \in E : w_e &\in \mathbb{R}_+ \end{aligned}$

Table 1: The primal (cf. (24)) and dual formulation of $\kappa(G)$.

Lemma 3 *There exists an optimal solution of **Dual**,*

$$\mathbf{w}^\star \in \arg\min_{\mathbf{w} \in \mathcal{P}(E)} \max_{T \in \mathcal{T}} \sum_{e \in E} \mathbb{1}(e \in T) w_e,$$

such that all non-zero components of $\mathbf{w}^\star$ take the identical values.

As per Lemma 3, consider an optimal solution $\mathbf{w}^\star$ of **Dual**, that assigns constant value to a subset $F \subset E$ edges and 0 to edges $E \backslash F$. It follows that $w_e^\star = \frac{1}{|F|}$ for $e \in F$ and $w_e^\star = 0$ for $e \in E \backslash F$. Let $V(F) \subset V$ be set of all vertices corresponding to the end points of edges in F making a subgraph $(V(F), F)$ of G. Let $c(F) \geq 1$ denote the number of connected components of $(V(F), F)$. Per **Dual**, given $\mathbf{w}^\star$, $\kappa(G)$ equals the weight of the maximum weight spanning tree in G with edges assigned weights as per $\mathbf{w}^\star$. Such a maximum weight spanning tree must select as many edges as possible from F: it can select at most $|V(F)| - c(F)$ such edges and any each such edge has weight $1/|F|$ whereas the other selected edges have weight 0. Thus, the total weight of such a maximum weight spanning tree is $(|V(F)| - c(F))/|F|$. This gives us an equivalent characterization for $\kappa(G)$ as

$$\kappa(G) = \min_{F \subset E} \frac{|V(F)| - c(F)}{|F|}. \tag{29}$$

Now we state a Lemma, whose proof is in Appendix D, that concludes on Theorem 1.

Lemma 4 *For any graph G,*

$$\min_{S \subset V} \frac{|S| - 1}{|E(S)|} = \min_{F \subset E} \frac{|V(F)| - c(F)}{|F|}. \tag{30}$$

4.3. Evaluating $\kappa(G)$ For Certain Graphs

As established in Section 4.1, $\kappa(G)$ can be computed in polynomial time for any G. Here, we attempt to obtain a (lower) bound on $\kappa(G)$ in terms of simple graph properties. To that end, we obtain the following for graphs with bounded maximum average degree and for graphs with bounded girth.

Lemma 5 *Consider a graph $G = (V, E)$. If G has maximum average degree $\bar{d} = \max_{S \subset V} \frac{2|E(S)|}{|S|}$,*

$$\frac{2}{\bar{d} + 1} \leq \kappa(G). \tag{31}$$

If G has girth (length of its shortest cycle) $g > 3$,

$$\frac{2}{1 + N^{\frac{2}{g-3}}}(1 - 1/g) \leq \kappa(G). \tag{32}$$

The proof of Lemma 5 is presented in Appendix E. As an immediate consequence, for $g = \beta \log N$,

$$\kappa(G) = 1 - \frac{1}{\beta} + o_{\beta \to \infty}(\frac{1}{\beta}) \quad \text{therefore} \quad \alpha(G, \text{TRW}') \leq \frac{1}{\sqrt{\kappa(G)}} = 1 + \frac{1}{2\beta} + o_{\beta \to \infty}(\frac{1}{\beta}). \tag{33}$$

5. A Near Linear-Time Variant of TRW$'$

5.1. Algorithm

TRW$'$ requires finding $\rho^\star(G)$. As discussed in Section 4, it can be computed efficiently. However it can be cumbersome and having near-linear (in $|E|$) time variant can be more attractive in practice. With this as a motivation, we propose utilizing the uniform distribution on $\mathcal{T}(G)$, denoted as $\mathbf{u}(G) \equiv \mathcal{U}(\mathcal{T}(G))$, in place of $\rho^\star(G)$ in TRW$'$. The challenge is it has very large support and hence it is difficult to compute $L_\mathbf{u}(\boldsymbol{\theta}), U_\mathbf{u}(\boldsymbol{\theta})$. But, both of these quantities are averages, with respect to $\mathbf{u}$, of a certain functional. And it is feasible to sample spanning trees uniformly at random for any G in near-linear time. Therefore, we can draw n samples from the distribution $\mathbf{u}$ and consider the empirical distribution $\hat{\mathbf{u}}^n$ to compute estimates $L_{\hat{\mathbf{u}}^n}(\boldsymbol{\theta}), U_{\hat{\mathbf{u}}^n}(\boldsymbol{\theta})$ with few samples. This is precisely the algorithm.

To that end, consider n trees $\mathsf{T}_1, \ldots, \mathsf{T}_n$ sampled uniformly at random from $\mathcal{T}(G)$. Compute

$$\hat{u}_e^n = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(e \in \mathsf{T}_i), \quad L_{\hat{\mathbf{u}}^n}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \Phi(\Pi^{\mathsf{T}_i}(\boldsymbol{\theta})), \quad U_{\hat{\mathbf{u}}^n}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \Phi(\Pi_{\hat{\mathbf{u}}^n}^{\mathsf{T}_i}(\boldsymbol{\theta})), \tag{34}$$

where $\hat{\mathbf{u}}^n = (\hat{u}_e^n)_{e \in E}$. Given this, produce the estimate,

$$\widehat{\Phi}_{\hat{\mathbf{u}}^n}(\boldsymbol{\theta}) = \sqrt{L_{\hat{\mathbf{u}}^n}(\boldsymbol{\theta})U_{\hat{\mathbf{u}}^n}(\boldsymbol{\theta})}. \tag{35}$$

5.2. Guarantees

Given a graph G, remember that $\kappa(G, \mathbf{u}) = \min_{e \in E} u_e$ with $u_e = \mathbb{P}_{\mathsf{T} \sim \mathbf{u}}(e \in \mathsf{T})$. We obtain the following guarantee (proof can be found in Appendix F).

Lemma 6 *Given $\epsilon > 0$ and $\delta > 0$, for $n \geq O\left(\log(\frac{N}{\delta})\kappa(G, \mathbf{u})^{-2}\epsilon^{-2}\right)$ and ϵ sufficiently small, with probability at least $1 - \delta$,*

$$\max\left(\frac{\Phi(\boldsymbol{\theta})}{\widehat{\Phi}_{\hat{\mathbf{u}}^n}(\boldsymbol{\theta})}, \frac{\widehat{\Phi}_{\hat{\mathbf{u}}^n}(\boldsymbol{\theta})}{\Phi(\boldsymbol{\theta})}\right) \leq \frac{1 + \epsilon}{\sqrt{\kappa(G, \mathbf{u})}}. \tag{36}$$

5.3. Computation Cost

To sample trees uniformly at random from $\mathcal{T}(G)$, Schild (2018) recently proposed a "short-cutting" method that has $O(|E|^{1+o(1)})$ runtime. The earliest polynomial time algorithm has been known since Guenoche (1983). While we do not recall either of these here, we briefly recall algorithm

from Broder (1989) due to its elegance even though it is not the optimal (it has $O(N|E|)$ run time): (1) starting with any $u \in V$ run a random walk on G until it covers all vertices, (2) for every vertex $v \neq u$, select the edge through which v was reached for the first time during the walk, and (3) output the $N - 1$ edges (which form tree) thus selected.

Given n such samples, in order to compute $\widehat{\Phi}_{\hat{\mathbf{u}}^n}(\boldsymbol{\theta})$, we have to compute $2n$ log-partition functions for tree structured graph. As noted in Section 2, each such computation requires $O(N|\mathcal{X}|^2)$ operations. By Lemma 6, we therefore need total of $O(|E|^{1+o(1)} + N|\mathcal{X}|^2) \times O(\kappa(G, \mathbf{u})^{-2}\epsilon^{-2} \log 1/\epsilon)$ runtime for $(1 + \epsilon)/\sqrt{\kappa(G, \mathbf{u})}$ approximation with probability $1 - \epsilon$.

5.4. $\kappa(G, \mathbf{u})$ and Effective Resistance

Recall that $\kappa(G, \mathbf{u}) = \min_{e \in E} u_e$ where $u_e = \mathbb{E}_{\mathsf{T} \sim \mathbf{u}}[\mathbb{1}(e \in \mathsf{T})]$ is equal to the so called "effective resistance" associated with edge $e \in E$ for the graph G. The notion was introduced by Klein and Randić (1993) and has multiple interpretations. We present one here. For $e = (s, t) \in E$, the effective resistance u_e is equal to the amount of electric energy dissipated by the network when all edges are seen as electric wire of resistance $R_e = 1$ and a generator guarantees a total current flow ($\iota_{\mathrm{gen}} = 1$) from s to t. The distribution of the current ι across the network must minimize the dissipated energy while respecting the constraints imposed by Kirchoff's laws (also see (Lyons and Peres, 2017, Chapter 2)). Below we provide variational characterization of it.

$$\forall e = (s, t) \in E : \quad u_e = \min \left\{ \sum_{\{u,v\} \in E} \iota(u, v)^2 \middle| \begin{array}{l} \forall \{u, v\} \in E : \iota(u, v) + \iota(v, u) = 0 \\ \forall u \in V \setminus \{s, t\} : \sum_{v|(u,v) \in E} \iota(u, v) = 0 \\ \sum_{v|(s,v) \in E} \iota(s, v) = \sum_{u|(u,t) \in E} \iota(u, t) = 1 \end{array} \right\}.$$

$$(37)$$

We observe that the minimum effective resistance of the graph, $\kappa(G, \mathbf{u}) = \min_{e \in E} u_e$ can also be connected to structural properties of the graph (proofs are in G).

Lemma 7 *Given $G = (V, E)$, if G has maximum degree d,*

$$\frac{2}{d+1} \leq \kappa(G, \mathbf{u}) \tag{38}$$

, and if G has girth at least $g > 3$,

$$\frac{1}{1 + |E|/(g-1)^2} \leq \kappa(G, \mathbf{u}) \tag{39}$$

6. Beyond Trees

This far, we have restricted to approximating $\Phi(\boldsymbol{\theta})$ by decomposing $\boldsymbol{\theta} = \mathbb{E}_{\mathsf{T} \sim \boldsymbol{\rho}}[\Pi_{\boldsymbol{\rho}}^T(\boldsymbol{\theta})]$ and then using properties of Φ to produce an approximation guarantee. Such arguments would hold even if we decompose $\boldsymbol{\theta}$ using subgraphs of G beyond trees. The choice of trees was particularly useful since they allow for an efficient computation of Φ. In general, graphs with bounded tree-width lend themselves to efficient computation of Φ, cf. Chandrasekaran et al. (2011).

To that end, let $\mathcal{T}_k(G)$ denote the set of all subgraphs of G that have treewidth bounded by $k \geq 1$. Let $\mathcal{P}(\mathcal{T}_k(G))$ denote the distribution over all such subgraphs. For any $H \in \mathcal{T}_k(G)$ and $\boldsymbol{\rho} \in \mathcal{P}(\mathcal{T}_k(G))$, define $\Pi^H(\cdot)$ and $\Pi^H_{\boldsymbol{\rho}}(\cdot)$ similar to that in (13) in Section 3 and define,

$$L_{\boldsymbol{\rho}}(\boldsymbol{\theta}) = \mathbb{E}_{H \sim \boldsymbol{\rho}}(\Phi(\Pi^H(\boldsymbol{\theta}))), \quad U_{\boldsymbol{\rho}}(\boldsymbol{\theta}) = \mathbb{E}_{H \sim \boldsymbol{\rho}}(\Phi(\Pi^H_{\boldsymbol{\rho}}(\boldsymbol{\theta}))), \quad \widehat{\Phi}_{\boldsymbol{\rho}}(\boldsymbol{\theta}) = \sqrt{L_{\boldsymbol{\rho}}(\boldsymbol{\theta})U_{\boldsymbol{\rho}}(\boldsymbol{\theta})}. \quad (40)$$

Using identical arguments as in Theorem 1, it follows that $\widehat{\Phi}_{\boldsymbol{\rho}}(\boldsymbol{\theta})$ is $1/\sqrt{\kappa(G,\rho)}$-approximation. By optimizing over the choice $\rho_k^* \in \mathcal{P}(\mathcal{T}_k(G))$ one would theoretically attain

$$\kappa_k(G) = \max_{\boldsymbol{\rho} \in \mathcal{P}(\mathcal{T}_k(G))} \min_{e \in E} \rho_e. \quad (41)$$

(ϵ, k)-*partitioning.* While such generality is pleasing, the space of $\mathcal{P}(\mathcal{T}_k(G))$ seems too vast and complex to compute $\kappa_k(G)$ or the associated distribution $\rho_k^*(G) \in \arg\max_{\boldsymbol{\rho} \in \mathcal{P}(\mathcal{T}_k(G))} \min_{e \in E} \rho_e$ for $k > 1$. In Jung and Shah (2006); Jung et al. (2009) a seemingly different approach was proposed using graph partitioning. It resulted in an approximation method for a large family of graphs including minor-excluded graphs and graphs with polynomial growth. It considers the set of k-partitions of G, $\mathsf{Part}_k(G) \subset \mathcal{T}_k(G)$ defined as

$$\mathsf{Part}_k(G) = \{H = (V, \bigcup_{i=1}^{K} E(S_i)) \mid (S_i)_{1 \leq i \leq k} \text{ is a partition of } V \text{ and } \forall i : |S_i| \leq k\}. \quad (42)$$

A distribution $\boldsymbol{\rho} \in \mathcal{P}(\mathsf{Part}_k(G)) \subset \mathcal{P}(\mathcal{T}_k(G))$ is called an (ϵ, k)-partitioning of G if $\forall e \in E : 1 - \epsilon \leq \mathbb{E}_{H \sim \boldsymbol{\rho}}[\mathbb{1}(e \in H)] \leq 1$. The proof of the following result can be found in Appendix H.

Theorem 8 *If G is such that there exists an (ϵ, k)-partitioning of G, $\boldsymbol{\rho} \in \mathsf{Part}_k(G)$, then*

$$\sqrt{1 - \epsilon} \leq \frac{\Phi(\boldsymbol{\theta})}{\widehat{\Phi}_{\boldsymbol{\rho}}(\boldsymbol{\theta})} \leq \frac{1}{\sqrt{1 - \epsilon}}. \quad (43)$$

We note that $\frac{1}{\sqrt{1-\epsilon}} = 1 + \frac{1}{2}\epsilon + o(\epsilon)$ and hence it improves upon the result given in Jung and Shah (2006); Jung et al. (2009) which achieves a $1 + \epsilon$ approximation error.

7. Relation to hardness of approximation

As mentioned in earlier, the task of computing log-partition function $\Phi(\boldsymbol{\theta}) = \log(Z(\boldsymbol{\theta}))$ is computationally hard. Further finding constant factor approximation to it is hard in general. We provide details to this effect.

To that end, let us consider a graph G with $\mathcal{X} = \{-1, 1\}$ and potentials such that $\forall e = (u, v) : \phi_e(x_e) = \beta \mathbb{1}(x_u \neq x_v)$ where $\beta \in \mathbb{R}^+$. The corresponding partition function $Z(\beta)$ satisfies:

$$Z(\beta) = \sum_{\boldsymbol{x} \in \{-1,1\}^n} \exp\left(\beta \sum_{(u,v) \in E} \mathbb{1}(x_u \neq x_v)\right), \quad (44)$$

$$= \sum_{0 \leq k \leq |E|} N_k \exp(\beta k), \quad (45)$$

12

where N_k denotes the number of k-cuts of G. This gives the following bound:

$$\beta \text{MAXCUT}(G) \leq \Phi(\beta) \leq N \log(2) + \beta \text{MAXCUT}(G), \tag{46}$$

where $\text{MAXCUT}(G) = \max\{k \mid N_k \neq 0\}$ is the cardinality of the maximum cut of G. As a consequence, for β sufficiently large ($\beta = N \log(2)/\epsilon$ for some small $\epsilon > 0$) observe that any constant factor guarantee on $\widehat{\Phi}(\beta)$ will yield a similar constant factor guarantee on the size of the maximum cut of G. The task of approximating the maximum cut of a graph was shown to be NP-hard for any constant factor greater than $16/17 \approx 0.941$ by Håstad (2001) for general graphs, although this ratio was improved under structural assumptions for G (such as bounded degree) by Feige et al. (2002). Through the above connection, our result provides $1/\kappa(G)$-approximation guarantee for any G.

The example above generalizes to binary constraint satisfaction problems (CSP). For a graph $G = (V, E)$ and an alphabet $\mathcal{X}$, a binary constraint satisfaction problem $\mathcal{C}$ is defined by a collection of constraints on the edges of the graph $(C_e)_{e \in E}$, where each constraint corresponds to a subset $C_e \subset \mathcal{X}^2$ of "acceptable" values for the corresponding pair of variables. For a given constraint satisfaction problem $\mathcal{C} = (C_e)_{e \in E}$, a quantity of interest is its maximum satifiability $\text{MAXCSP}(\mathcal{C})$, which is the highest fraction of satisfied constraints, over all possible assignments for V. Notice that the maximum cut problem mentioned above is a particular case. By the same derivations, we obtain that for the choice of potential functions $\phi_e(x_e) = \beta \mathbb{1}(x_e \in C_e)$ and β sufficiently large ($\beta \geq N \log(|\mathcal{X}|)/\epsilon$) the corresponding log-partition function satisfies:

$$\Phi(\beta)/\beta - \epsilon \leq \text{MAXCSP}(\mathcal{C}) \leq \Phi(\beta)/\beta. \tag{47}$$

Maximum constraint satisfaction problems are known to be hard to approximate. A particular case of constraint satisfaction problems (where the constraints are of the form $C_e = \{(z, \pi(z)) \mid z \in \mathcal{X}\}$ with π a permutation of $\mathcal{X}$) are called "unique games" (see Khot and Vishnoi (2005) for precise definition). The unique games conjecture (Khot, 2002) states that no constant factor approximation guarantee is achievable in polynomial time for maximum constraint satisfaction problems on unique games in general. That is, one should not expect constant factor approximation for log partition function computation, even with non-negative potentials as considered in this work, for general graph. Indeed, this work provides such non-constant factor approximation, $1/\kappa(G)$, for any graph G for tree-reweighted method. The question remains, whether such is the tightest possible for tree-reweighted method in general.

8. Conclusions

We presented an approach to quantify the approximation ratio of variational methods for estimating the log-partition function of discrete pairwise graphical models. As the main contribution, we quantified the approximation error as a function of the underlying graph properties. In particular, for a variant of the tree-reweighted algorithm, for graphs with bounded degree the approximation ratio is a constant factor (function of degree) and for graphs with large ($\gg$ logarithmic) girth, the approximation ratio is close to 1. The method naturally extends beyond trees unifying prior works on graph partitioning based approach.

In this work, we restricted the analysis to non-negative valued potentials and edge parameters. If potentials are bounded, we can transform the general setting into a setting with non-negative po-

tentials. However, the approximation ratio with respect to this transformed setting may not translate to that of the original setting. This may be interesting direction for future works.

Acknowledgments

References

Noga Alon, Shlomo Hoory, and Nathan Linial. The moore bound for irregular graphs. *Graphs and Combinatorics*, 18(1):53–57, 2002.

Mohsen Bayati, David Gamarnik, Dimitriy Katz, Chandra Nair, and Prasad Tetali. Simple deterministic approximation algorithms for counting matchings. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 122–127, 2007.

Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

Andrei Z Broder. Generating random spanning trees. In *FOCS*, volume 89, pages 442–447. Citeseer, 1989.

Venkat Chandrasekaran, Misha Chertkov, David Gamarnik, Devavrat Shah, and Jinwoo Shin. Counting independent sets using the bethe approximation. *SIAM Journal on Discrete Mathematics*, 25(2):1012–1034, 2011.

Michael Chertkov and Vladimir Y Chernyak. Loop series for discrete statistical models on graphs. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(06):P06009, 2006.

Amir Dembo, Andrea Montanari, Nike Sun, et al. Factor models on locally tree-like graphs. *Annals of Probability*, 41(6):4162–4213, 2013.

Jack Edmonds. Matroids and the greedy algorithm. *Mathematical programming*, 1(1):127–136, 1971.

Uriel Feige, Marek Karpinski, and Michael Langberg. Improved approximation of max-cut on graphs of bounded degree. *Journal of Algorithms*, 43(2):201–219, 2002.

David Gamarnik and Dmitriy Katz. Sequential cavity method for computing free energy and surface pressure. *Journal of Statistical Physics*, 137(2):205–232, 2009.

David Gamarnik and Dmitriy Katz. Correlation decay and deterministic fptas for counting colorings of a graph. *Journal of Discrete Algorithms*, 12:29–47, 2012.

Hans-Otto Georgii. *Gibbs measures and phase transitions*, volume 9. Walter de Gruyter, 2011.

Michel X Goemans. Minimum bounded degree spanning trees. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 273–282. IEEE, 2006.

Martin Grötschel, László Lovász, and Alexander Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.

Alain Guenoche. Random spanning tree. *Journal of Algorithms*, 4(3):214–220, 1983.

Johan Håstad. Some optimal inapproximability results. *Journal of the ACM (JACM)*, 48(4):798–859, 2001.

Mark Jerrum and Alistair Sinclair. Approximating the permanent. *SIAM journal on computing*, 18 (6):1149–1178, 1989.

Kyomin Jung and Devavrat Shah. Local approximate inference algorithms. *arXiv preprint cs/0610111*, 2006.

Kyomin Jung, Pushmeet Kohli, and Devavrat Shah. Local rules for global map: When do they work? In *NIPS*, pages 871–879, 2009.

Subhash Khot. On the power of unique 2-prover 1-round games. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 767–775, 2002.

Subhash Khot and Nisheeth K Vishnoi. On the unique games conjecture. In *FOCS*, volume 5, page 3. Citeseer, 2005.

Douglas J Klein and Milan Randić. Resistance distance. *Journal of mathematical chemistry*, 12(1): 81–95, 1993.

Lap Chi Lau, Ramamoorthi Ravi, and Mohit Singh. *Iterative methods in combinatorial optimization*, volume 46. Cambridge University Press, 2011.

Russell Lyons and Yuval Peres. *Probability on trees and networks*, volume 42. Cambridge University Press, 2017.

Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

Aaron Schild. An almost-linear time algorithm for uniform random spanning tree generation. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 214–227, 2018.

Leslie G Valiant. The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, 8(3):410–421, 1979.

Martin J Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

Martin J Wainwright, Tommi S Jaakkola, and Alan S Willsky. A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory*, 51(7):2313–2335, 2005.

Dror Weitz. Counting independent sets up to the tree threshold. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 140–149, 2006.

Jonathan S Yedidia, William T Freeman, and Yair Weiss. Generalized belief propagation. In *Advances in neural information processing systems*, pages 689–695, 2001.

Jonathan S Yedidia, William T Freeman, and Yair Weiss. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239, 2003.

Appendix A. Balanced Covering of Graphs

In section 4, we described two linear programming problems **Primal** and **Dual** of optimum equal to the main quantity of interest $\kappa(G)$. We recall their formulation given in (28) and provide some interpretation.

$$\textbf{Primal} = \max_{\rho \in \mathcal{P}(\mathcal{T})} \min_{e \in E} \sum_{T \in \mathcal{T}} \mathbb{1}(e \in T)\rho^T = \min_{\mathbf{w} \in \mathcal{P}(E)} \max_{T \in \mathcal{T}} \sum_{e \in E} \mathbb{1}(e \in T)w_e = \textbf{Dual}.$$

The objective of **Primal** is to obtain a *balanced covering* of G with its spanning trees, that is a distribution $\rho \in \mathcal{P}(\mathcal{T}(G))$ that maximizes the probability of apparition of the least likely edge. We give example of such balanced coverings in Figure 1. Interestingly, **Dual** also has a nice formulation as a *max-min spanning tree* problem. Given a total budget of 1, it aims to assign weights to the edges of G such that an adversary removing a maximum spanning tree will receive the smallest reward. It is natural to observe that the densest sub-graphs of G are the key to both these problems. They are key to **Primal** because they are harder to cover with spanning trees. They are key to **Dual** because assigning weight to dense regions is a good way to limit the adversary's reward. This is made explicit by the representations in Figure 2 and Figure 3.
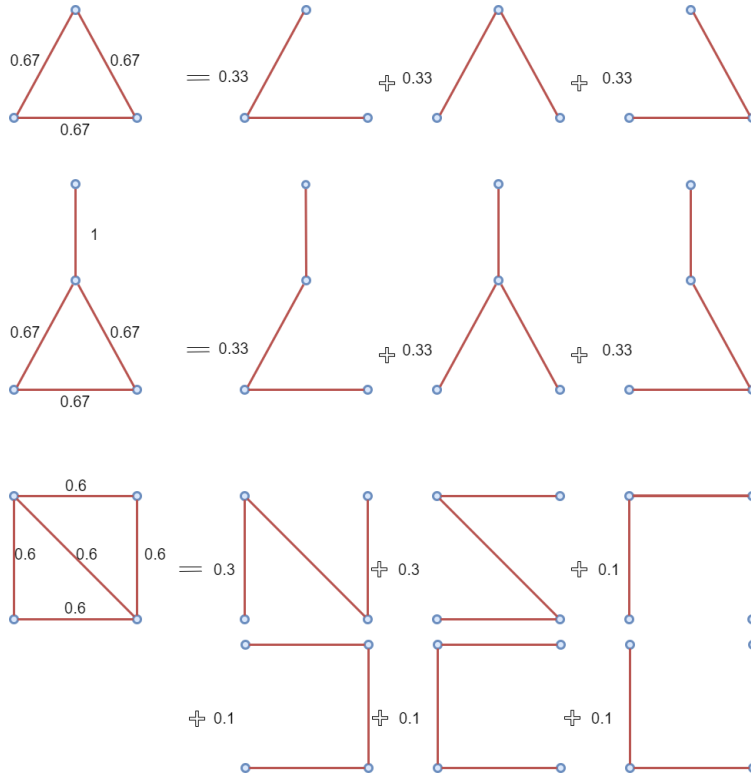

Figure 1: Three examples of balanced coverings of graphs. Notice that for the triangle, the most connected subgraph is the entire graph itself which yields $\kappa(G) = \frac{3-1}{3} = \frac{2}{3} = 0.67$ just like for the third example for which $\kappa(G) = \frac{4-1}{5} = \frac{3}{5} = 0.6$.
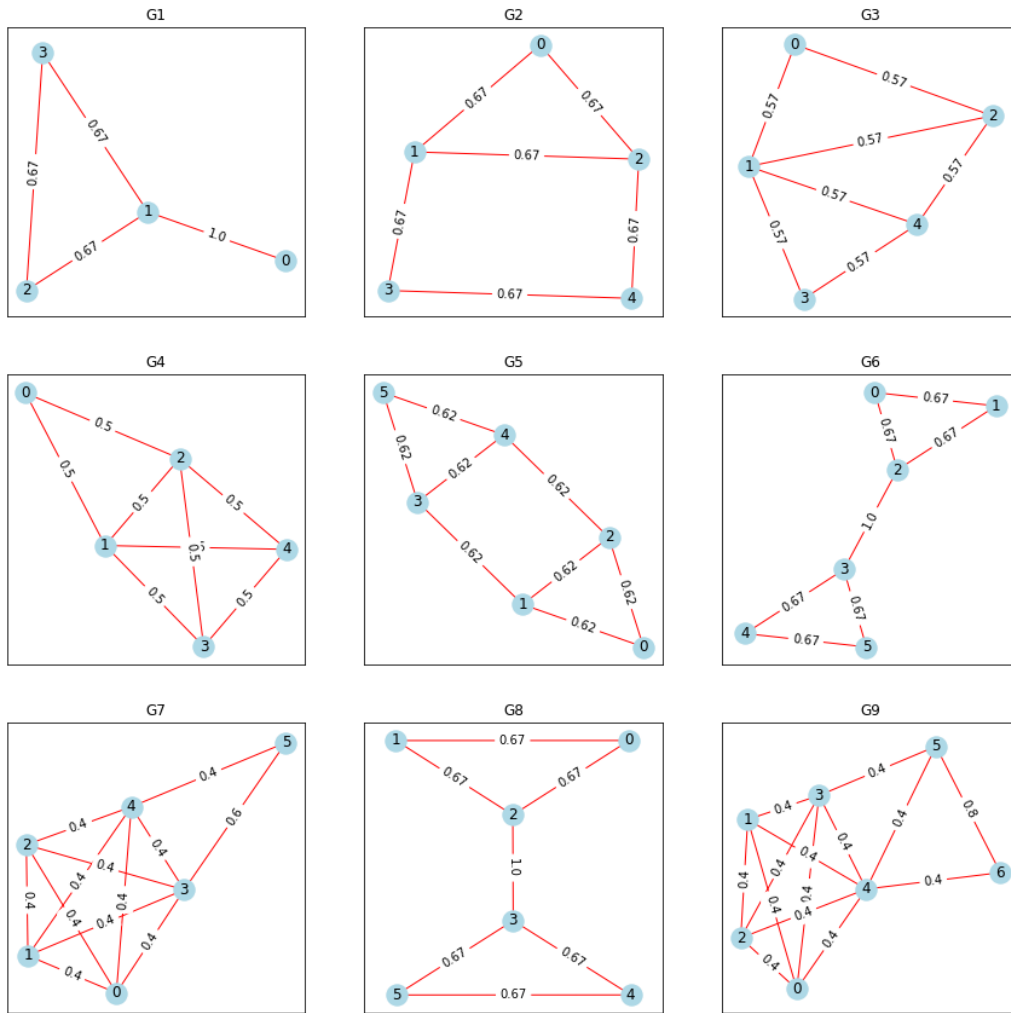
Figure 2: More examples of balanced covering of graphs, obtained with a LP solver. The number assigned to each edge corresponds to $(\rho_e)_{e \in E}$ for a balanced cover $\boldsymbol{\rho}^*(G)$. Note that for G_7, some of the edge probability of edge $(3, 5)$ could be redistributed to edge $(4, 5)$ by symmetry, but this would not improve the optimum. This is also true for G_9.

Appendix B. Proof of Lemma 2

We start by observing a few properties of function $\Phi(\cdot)$.

Property 1. Φ is non-decreasing. For $a, b \in \mathbb{R}^n$ let $a \preceq b$ denote that every component of a is less or equal to that of b, i.e. $a_i \leq b_i, \ i \in [n]$. With this, for $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}_+^{|E|}$ such that $\boldsymbol{\theta} \preceq \boldsymbol{\theta}'$, it can be easily verified that

$$\Phi(\boldsymbol{\theta}) \leq \Phi(\boldsymbol{\theta}'). \qquad \text{(monotonicity)}$$

Since $\Phi(\mathbf{0}) = N \log |\mathcal{X}|$, and $\mathbf{0} \preceq \boldsymbol{\theta} \preceq \boldsymbol{\theta}'$, we have

$$N \log(|\mathcal{X}|) \leq \Phi(\boldsymbol{\theta}) \leq \Phi(\boldsymbol{\theta}'). \qquad (48)$$

Property 2. Φ is sub-linear. For $\lambda \geq 1$ and $\boldsymbol{\theta} \in \mathbb{R}_+^{|E|}$,

$$\Phi(\lambda \boldsymbol{\theta}) \leq \lambda \Phi(\boldsymbol{\theta}). \qquad \text{(sub-linearity)}$$

The above follows from the fact that for any $\mathbf{s} = (s_i) \in \mathbb{R}_+^n$,

$$\Big(\sum_{i=1}^n s_i^\lambda\Big) \leq \Big(\sum_{i=1}^n s_i\Big)^\lambda.$$

Now consider $\boldsymbol{\rho} \in \mathcal{P}(\mathcal{T}(G))$. For any $T \in \mathcal{T}(G)$ and $\boldsymbol{\theta} \in \mathbb{R}_+^{|E|}$, by definition of Π^T, we have that $\Pi^T(\boldsymbol{\theta}) \preceq \boldsymbol{\theta}$. Therefore, using the monotonicity of the log-partition function it follows that

$$L_{\boldsymbol{\rho}}(\boldsymbol{\theta}) = \sum_{T \in \mathcal{T}(G)} \rho^T \Phi(\Pi^T(\boldsymbol{\theta})) \leq \sum_{T \in \mathcal{T}(G)} \rho^T \Phi(\boldsymbol{\theta}) \leq \Phi(\boldsymbol{\theta}). \qquad (49)$$

By definition $\boldsymbol{\theta} = \mathbb{E}_{\mathsf{T} \sim \boldsymbol{\rho}}[\Pi_{\boldsymbol{\rho}}^{\mathsf{T}}(\boldsymbol{\theta})]$, and due to convexity of Φ, it follows that

$$\Phi(\boldsymbol{\theta}) = \Phi\big(\mathbb{E}_{\mathsf{T} \sim \boldsymbol{\rho}}[\Pi_{\boldsymbol{\rho}}^{\mathsf{T}}(\boldsymbol{\theta})]\big) \leq \mathbb{E}_{\mathsf{T} \sim \boldsymbol{\rho}}[\Phi(\Pi_{\boldsymbol{\rho}}^{\mathsf{T}}(\boldsymbol{\theta}))] = U_{\boldsymbol{\rho}}(\boldsymbol{\theta}). \qquad (50)$$

By definition of $\kappa(G, \boldsymbol{\rho}) = \min_{e \in E} \rho_e$, it follows that

$$\Pi_{\boldsymbol{\rho}}^T(\boldsymbol{\theta}) \leq \frac{1}{\kappa(G, \boldsymbol{\rho})} \Pi^T(\boldsymbol{\theta}), \ \ \forall \, T \in \mathcal{T}(G). \qquad (51)$$

And, by definition $1/\kappa(G, \boldsymbol{\rho}) \geq 1$. Therefore by (monotonicity) and (sub-linearity), we have

$$\Phi(\Pi_{\boldsymbol{\rho}}^T(\boldsymbol{\theta})) \leq \Phi\big(\frac{1}{\kappa(G, \boldsymbol{\rho})} \Pi^T(\boldsymbol{\theta})\big) \leq \frac{1}{\kappa(G, \boldsymbol{\rho})} \Phi(\Pi^T(\boldsymbol{\theta})). \qquad (52)$$

Therefore,

$$U_{\boldsymbol{\rho}}(\boldsymbol{\theta}) = \sum_{T \in \mathcal{T}(G)} \rho^T \Phi(\Pi_{\boldsymbol{\rho}}^T(\boldsymbol{\theta})) \leq \frac{1}{\kappa(G, \boldsymbol{\rho})} \sum_{T \in \mathcal{T}(G)} \rho^T \Phi(\Pi^T(\boldsymbol{\theta})) = \frac{1}{\kappa(G, \boldsymbol{\rho})} L_{\boldsymbol{\rho}}(\boldsymbol{\theta}). \qquad (53)$$

As a consequence of (49), (50) and (53) we obtain that

$$\Phi(\boldsymbol{\theta}) \leq U_{\boldsymbol{\rho}}(\boldsymbol{\theta}) \leq \frac{1}{\kappa(G, \boldsymbol{\rho})} \Phi(\boldsymbol{\theta}) \quad \text{and} \quad \kappa(G, \boldsymbol{\rho}) \Phi(\boldsymbol{\theta}) \leq L_{\boldsymbol{\rho}}(\boldsymbol{\theta}) \leq \Phi(\boldsymbol{\theta}). \qquad (54)$$

From this, it follows that

$$\sqrt{\kappa(G, \boldsymbol{\rho})}\Phi(\boldsymbol{\theta}) \leq \sqrt{L_{\boldsymbol{\rho}}(\boldsymbol{\theta})U_{\boldsymbol{\rho}}(\boldsymbol{\theta})} \leq \frac{1}{\sqrt{\kappa(G, \boldsymbol{\rho})}}\Phi(\boldsymbol{\theta}). \tag{55}$$

Which can be rewritten as

$$\sqrt{\kappa(G, \boldsymbol{\rho})} \leq \frac{\widehat{\Phi}_{\boldsymbol{\rho}}(\boldsymbol{\theta})}{\Phi(\boldsymbol{\theta})} \leq \frac{1}{\sqrt{\kappa(G, \boldsymbol{\rho})}}. \tag{56}$$

By optimizing over choice of $\boldsymbol{\rho} = \boldsymbol{\rho}^{\star}(G)$, we conclude that $\alpha(G, \mathrm{TRW}') \leq \frac{1}{\kappa(G)}$.

Appendix C. Proof of Lemma 3

(See illustration in Figure 3) For $\mathbf{w} = (w_e)_{e \in E}$ denote $f(\mathbf{w})$ the number of distinct values in its support:

$$f(\mathbf{w}) = |\{w_e : e \in E, w_e \neq 0\}|. \tag{57}$$

To prove the lemma, it suffices to show that there exists an optimal solution of **Dual** such that $f(\mathbf{w}) = 1$. We will prove that if $\mathbf{w}$ is an optimal solution and $f(\mathbf{w}) > 1$ then we can build $\mathbf{w}'$ of similar objective value such that $f(\mathbf{w}') \leq f(\mathbf{w}) - 1$. By repeating this till $f(\mathbf{w}) = 1$ will conclude the proof.

Let $\mathbf{w}$ be an optimal solution with $f(\mathbf{w}) > 1$. We consider the edges $e_1, e_2, ..., e_{|E|}$ ordered by their weights, i.e.

$$w_{e_1} \geq ... \geq w_{e_{|E|}}. \tag{58}$$

In what follows, we will make sure that the ordering on the edges never changes, therefore we allow ourselves to write w_i instead of w_{e_i}. Now the objective of **Dual** achieved by such an optimal $\mathbf{w}$ corresponds to the weight of a maximum weight spanning tree. Let us utilize Kruskal's algorithm to find such an maximum weight spanning tree. Recall that Kruskal's algorithm greedily selects edges from higher to lower weight as long as they do not create a cycle with previously selected edges. We will denote $I_T = \{t_1 < ... < t_{N-1}\}$ the indices of the edges selected by the algorithm to construct tree T and let $I_{E \setminus T} = \cup_{k=1}^{N-1}\{s : t_k < s < t_{k+1}\}$ denote the indices of edges not part of T with notation $t_N = |E| + 1$. The weight of the maximum spanning tree is then $\mathbf{w}(T) = \sum_{k=1}^{N-1} w_{t_k}$. Note that $t_1 = 1$ and $t_2 = 2$ since cycle requires 3 or more edges. By definition $w_{j-1} \geq w_j$ for $2 \leq j \leq |E|$. Now if $w_{j-1} > w_j$ the we claim that $j \in I_T$. This is because for $1 \leq k \leq N - 1$ if $(w_{t_k}, ..., w_{t_{k+1}-1})$ are not equal, setting them all to their average decreases w_{t_k} strictly while preserving $\mathbf{w} \in \mathcal{P}(E)$ as well as the order on the edges and therefore contradicting the optimality of $\mathbf{w}$ for **Dual**. Therefore $\mathbf{w}$ is piece-wise constant with discontinuities only appearing for $j \in I_T$.

If $f(\mathbf{w}) = 2$ and all weights are positive, we denote $2 \leq k \leq N - 1$ such that $w_{t_k-1} > w_{t_k} > 0$ and we have:

$$w_1 = ... = w_{t_k-1} > w_{t_k} = ... = w_{|E|}. \tag{59}$$

In this case, the optimal objective value for **Dual** is equal to $(k - 1)w_1 + (N - k)w_{t_k}$. To make $\mathbf{w}$ constant on its support while preserving the order on the weights, there are two possibilities. Either transfer all weight from $(w_{t_k}, ..., w_{|E|})$ to $(w_1, ..., w_{t_k-1})$ until $(w_{t_k}, ..., w_{|E|})$ reaches

zero. The objective will then be $w_1 + \frac{|E|-t_k+1}{t_k-1} w_{t_k}$. Or transfer all weight from $(w_1, ..., w_{t_k-1})$ to $(w_{t_k}, ..., w_{|E|})$ until all weights are equal. The objective will be then $w_{t_k} + \frac{|E|-t_k+1}{t_k-1}(w_1 - w_{t_k})$. Because either $\frac{|E|-t_k+1}{t_k-1} \leq 1$ or $\frac{t_k-1}{|E|-t_k+1} \leq 1$, one of these transfers does not increase the objective and yields $f(\mathbf{w}) = 1 < 2$.

If $f(\mathbf{w}) = 2$ and some weights are 0, denote k_0 the smallest index such that $w_{t_{k_0}} = 0$. The method above still holds when replacing $|E| - t_k + 1$ by $t_{k_0} - t_k$.

Now suppose $f(\mathbf{w}) \geq 3$, making sure that the order on the weights is preserved requires extra caution. In addition to k and k_0 (if required), we denote k_1 the index of the discontinuity that follows k. We have:

$$... = w_{t_{k_1}-1} > w_{t_{k_1}} = ... = w_{t_k-1} > w_{t_k} = ... = w_{t_{k_0}-1} > w_{t_{k_0}} = ... \tag{60}$$

In the event when we want to transfer weight from $(w_{t_k}, ..., w_{t_{k_0}-1})$ to $(w_{t_{k_1}}, ..., w_{t_k-1})$, we must make sure that $(w_{t_{k_1}}, ..., w_{t_k-1})$ does not exceed $w_{t_{k_1}-1}$. If $(w_{t_{k_1}}, ..., w_{t_k-1})$ attains $w_{t_{k_1}-1}$ the transfer must stop at equality, and one should observe that we have decreased $f(\mathbf{w})$ strictly by 1 because the discontinuity at $w_{t_{k_1}}$ has disappeared and no new discontinuity was created.

In summary, we have argued that if $\mathbf{w}$ is an optimal solution and $f(\mathbf{w}) > 1$ then we can build $\mathbf{w}'$ of same objective value (optimal) and such that $f(\mathbf{w}') \leq f(\mathbf{w}) - 1$. This completes the proof of Lemma.

Appendix D. Proof of Lemma 4

We prove the equality by establishing inequalities in both direction.

Establishing $\min_{S \subset V} \frac{|S|-1}{|E(S)|} \geq \min_{F \subset E} \frac{|V(F)|-c(F)}{|F|}$: For $S \subset V$ note that $V(E(S)) \subset S$ and $c(E(S)) \geq 1$ and therefore that $\frac{|S|-1}{|E(S)|} \geq \frac{|V(E(S))|-c(E(S))}{|E(S)|}$ with $E(S) \subset E$. Thus, $\min_{S \subset V} \frac{|S|-1}{|E(S)|}$ is minimizing a larger objective function over smaller set compared to $\min_{F \subset E} \frac{|V(F)|-c(F)}{|F|}$. Therefore, the inequality follows immediately.

Establishing $\min_{S \subset V} \frac{|S|-1}{|E(S)|} \leq \min_{F \subset E} \frac{|V(F)|-c(F)}{|F|}$: Let $F^\star \subset E$ be a minimizer of $\min_{F \subset E} \frac{|V(F)|-c(F)}{|F|}$. Let $H = (V(F^\star), F^\star)$. By optimality, all connected components of H must be vertex-induced sub-graphs of G. This is because, if not then it is possible to add edges to H without changing the number of vertices or number of connected components in it, which would contradict optimality. In other words, there exists disjoint subsets $S_i, 1 \leq i \leq c(H)$ of $V(F^\star)$ with $V(F^\star) = \cup_{i=1}^{c(H)} S_i$ and $F^\star = \cup_{i=1}^{c(H)} E(S_i)$. If $c(H) = 1$, then the inequality follows immediately. If $c(H) \geq 2$, denote $H \backslash H_1$ the graph obtained by removing $H_1 = (S_1, E(S_1))$ from H. Note that $c(H \backslash H_1) = c(H) - 1$ and that $c(H_1) = 1$. By Lemma 9, $\forall a, b, c, d \in \mathbb{R}_+^4 : \min(\frac{a}{b}, \frac{c}{d}) \leq \frac{a+c}{b+d}$. Therefore,

$$\min \left(\frac{|V(H_1)| - c(H_1)}{|E(H_1)|}, \frac{|V(H \backslash H_1)| - c(H \backslash H_1)}{|E(H \backslash H_1)|} \right) \leq \frac{|V(H)| - c(H)}{|E(H)|}. \tag{61}$$

If H_1 achieves the minimum on the left hand side, then it concludes the proof. If $H \backslash H_1$ achieves the minimum simply iterate the above argument till we are left with single connected component and that would conclude the proof.
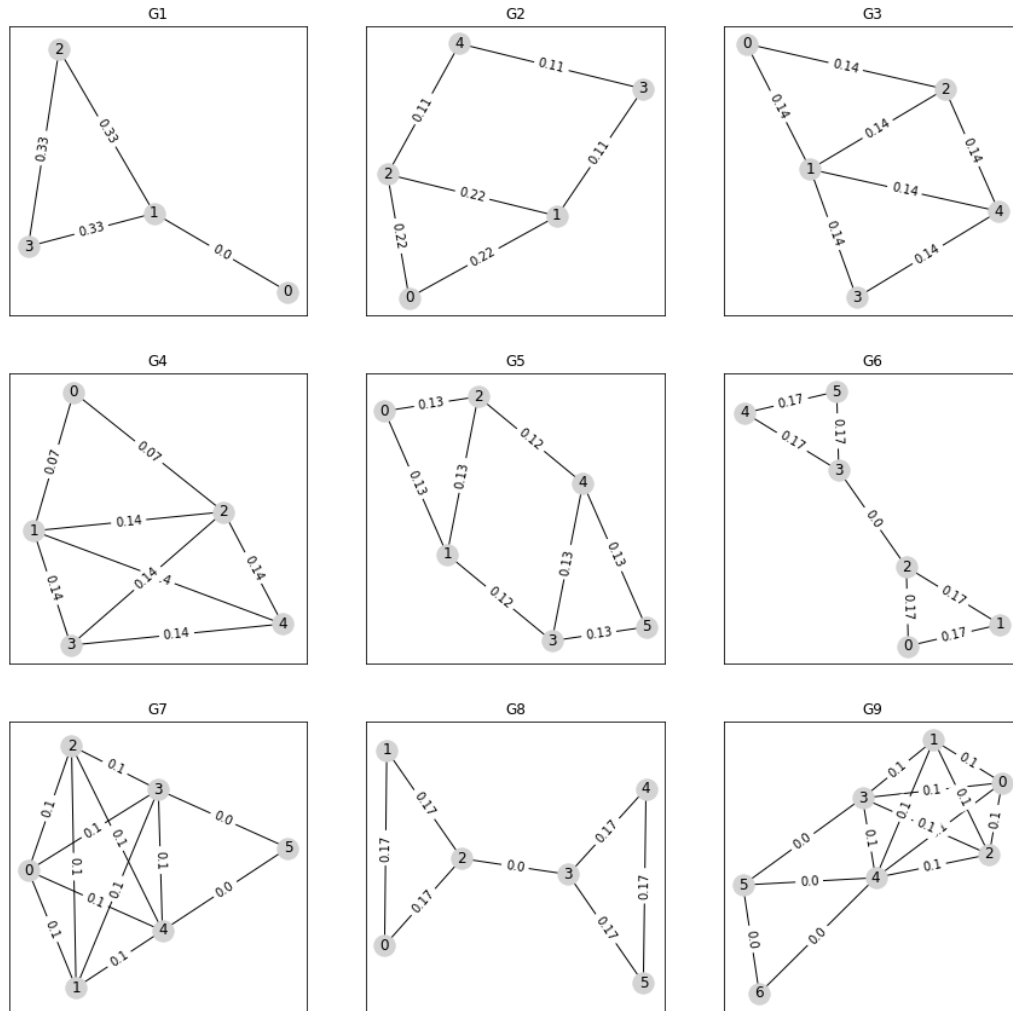
Figure 3: An example of an optimal weight assignment $(w_e)_{e \in E}$ for the problem **Dual** on nine different graphs. The solution was found by the interior point method using a linear programming solver. Note that for most graphs, the solution reached is already constant on its support. On graphs G_2 and G_4, note that evening out the weights would not increase the weight of the maximum spanning tree.

Lemma 9 *For any $a, b, c, d \in \mathbb{R}_+$,*

$$\min\left(\frac{a}{b}, \frac{c}{d}\right) \leq \frac{a+c}{b+d}. \tag{62}$$

Let $a, b, c, d \in \mathbb{R}_+$. The following sequence of statements hold leading to the proof of the claim:

$$ad \leq bc \quad \text{or} \quad bc \leq ad \tag{63}$$

$$\min(ad(b+d), cb(b+d)) \leq (a+c)bd \tag{64}$$

$$\min\left(\frac{a}{b}, \frac{c}{d}\right) \leq \frac{a+c}{b+d}. \tag{65}$$

Appendix E. Proof of Lemma 5

Case of bounded maximum average degree. Assume G has maximum average degree bounded by $\bar{d}$, where by definition

$$\bar{d} = \max_{S \subset V} \frac{2|E(S)|}{|S|}. \tag{66}$$

Therefore, for any $S \subset V$, $|E(S)| \leq \frac{\bar{d}}{2}|S|$. And there can be at most $\binom{|S|}{2}$ edges in a graph over vertices S, and hence $|E(S)| \leq \frac{|S|(|S|-1)}{2}$. Therefore, we obtain

$$\frac{|S| - 1}{|E(S)|} \geq \frac{2}{\bar{d}}\left(1 - \frac{1}{|S|}\right) = L_1(|S|), \tag{67}$$

$$\frac{|S| - 1}{|E(S)|} \geq \frac{2}{|S|} = L_2(|S|). \tag{68}$$

Therefore

$$\frac{|S| - 1}{|E(S)|} \geq \min_{x \in \mathbb{R}_+} \{\max(L_1(x), L_2(x))\}. \tag{69}$$

Note that L_1 is increasing and bounded whereas L_2 is decreasing. Therefore, $\max(L_1(x), L_2(x))$ with $x \in \mathbb{R}$ reaches its minimum for x such that $L_1(x) = L_2(x)$ which leads to minima at $x = \bar{d}+1$. Therefore, we conclude that for all $S \subset V$,

$$\frac{|S| - 1}{|E(S)|} \geq \frac{2}{\bar{d} + 1}. \tag{70}$$

Case of bounded girth. Let G with girth $g > 3$. Note that all subgraphs of G have girth at least g. The generalised Moore bound (obtained by Alon et al. (2002)) then gives $\forall S \subset V$:

$$|S| \geq 1 + d_S \sum_{i=0}^{\frac{g-3}{2}} (d_S - 1)^i \qquad \text{if g is odd,} \tag{71}$$

$$|S| \geq 2 \sum_{i=0}^{\frac{g-2}{2}} (d_S - 1)^i, \qquad \text{if g is even} \tag{72}$$

with $d_S = \frac{2|E(S)|}{|S|}$. We will only keep a weaker version of this bound that does not depend on the parity of g. Specifically, for all $S \subset V$:

$$|S| \geq \left(2\frac{|E(S)|}{|S|} - 1\right)^{\frac{g-3}{2}}. \tag{73}$$

Therefore, $|E(S)| \leq \frac{1}{2}(|S|^{\frac{2}{g-3}+1} + |S|)$ for all $S \subset V$. Subsequently, we have

$$\frac{|S|-1}{|E(S)|} \geq 2\frac{1 - \frac{1}{|S|}}{1 + |S|^{\frac{2}{g-3}}} \geq 2\frac{1 - \frac{1}{|S|}}{1 + N^{\frac{2}{g-3}}} \tag{74}$$

This bound is clearly increasing with $|S|$. Also note that if $|S| \leq g - 1$, the subgraph $(S, E(S))$ can have no cycle and therefore $\frac{|S|-1}{|E(S)|} = 1$. The worse case is therefore attained for $|S| = g$ where we have:

$$\frac{|S|-1}{|E(S)|} \geq \frac{2}{1 + N^{\frac{2}{g-3}}}\left(1 - \frac{1}{g}\right). \tag{75}$$

Appendix F. Proof of Lemma 6

We shall use Hoeffding's inequality: for any bounded random variable $a \leq X \leq b$, the deviation of its n-empirical average $\overline{X}_n$ computed from in dependant samples is such that for any $t > 0$,

$$\mathbb{P}(|\mathbb{E}(X) - \overline{X}_n| \geq t) \leq 2\exp\left(\frac{-2nt^2}{(b-a)^2}\right). \tag{76}$$

Another version of the equation when $\mathbb{E}(X) > 0$ is as follows, for any $\epsilon > 0$

$$\mathbb{P}\left(1 - \epsilon \leq \frac{\overline{X}_n}{\mathbb{E}(X)} \leq 1 + \epsilon\right) \geq 1 - 2\exp\left(\frac{-2n\epsilon^2\mathbb{E}(X)^2}{(b-a)^2}\right). \tag{77}$$

An immediate consequence is that $\hat{\mathbf{u}}^n$ is a good approximation for $\mathbf{u}$. For any $e \in E$,

$$\mathbb{P}\left(1 - \epsilon \leq \frac{\hat{u}_e^n}{u_e} \leq 1 + \epsilon\right) \geq 1 - 2\exp\left(-2n\epsilon^2 u_e^2\right), \tag{78}$$

therefore by union bound,

$$\mathbb{P}\left(\forall e \in E : 1 - \epsilon \leq \frac{\hat{u}_e^n}{u_e} \leq 1 + \epsilon\right) \geq 1 - 2|E|\exp\left(-2n\epsilon^2\kappa(G, \mathbf{u})\right). \tag{79}$$

Another consequence is that $L_{\hat{\mathbf{u}}^n}$ is a good approximation for $L_{\mathbf{u}}$. Indeed, considering the random variable $\Phi(\Pi^{\mathsf{T}}(\boldsymbol{\theta}))$ of mean $L_{\mathbf{u}}(\boldsymbol{\theta})$ and of empirical average $L_{\hat{\mathbf{u}}^n}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \Phi(\Pi^{\mathsf{T}_i}(\boldsymbol{\theta}))$ and noting that this variable is bounded as follows $0 \leq \Phi(\Pi^{\mathsf{T}}(\boldsymbol{\theta})) (\leq \Phi(\boldsymbol{\theta})) \leq \frac{1}{\kappa_{\mathbf{u}}}L_{\mathbf{u}}(\boldsymbol{\theta})$, we have

$$\mathbb{P}\left(1 - \epsilon \leq \frac{L_{\hat{\mathbf{u}}^n}(\boldsymbol{\theta})}{L_{\mathbf{u}}(\boldsymbol{\theta})} \leq 1 + \epsilon\right) \geq 1 - 2\exp\left(-2n\epsilon^2\kappa_{\mathbf{u}}^2\right). \tag{80}$$

Regarding $U_{\hat{\mathbf{u}}^n}(\boldsymbol{\theta})$, the discussion requires an additional argument because $\frac{1}{n} \sum_{i=1}^n \Phi(\Pi_{\hat{\mathbf{u}}^n}^{\mathsf{T}_i}(\boldsymbol{\theta}))$ is not a sum of independent random variables. Instead, let us focus on the close quantity, $\frac{1}{n} \sum_{i=1}^n \Phi(\Pi_{\mathbf{u}}^{\mathsf{T}_i}(\boldsymbol{\theta}))$ for which we have $0 \leq \Phi(\Pi_{\mathbf{u}}^{\mathsf{T}}(\boldsymbol{\theta})) \left(\leq \frac{1}{\kappa_{\mathbf{u}}} \Phi(\boldsymbol{\theta}) \right) \leq \frac{1}{\kappa_{\mathbf{u}}} U_{\mathbf{u}}(\boldsymbol{\theta})$ and therefore,

$$\mathbb{P}\left(1 - \epsilon \leq \frac{\frac{1}{n} \sum_{i=1}^n \Phi(\Pi_{\mathbf{u}}^{\mathsf{T}_i}(\boldsymbol{\theta}))}{U_{\mathbf{u}}(\boldsymbol{\theta})} \leq 1 + \epsilon \right) \geq 1 - 2 \exp\left(-2n\epsilon^2 \kappa_{\mathbf{u}}^2 \right). \tag{81}$$

Fortunately, if (79) is satisfied this quantity turns out to be a good approximation of $U_{\hat{\mathbf{u}}^n}(\boldsymbol{\theta})$. Indeed, assuming that $\forall e \in E : 1 - \epsilon \leq \frac{\hat{u}_e^n}{u_e} \leq 1 + \epsilon$ we have that for all $T \in \mathcal{T}(G)$,

$$(1 - \epsilon) \Pi_{\mathbf{u}}^T(\boldsymbol{\theta}) \preceq \Pi_{\hat{\mathbf{u}}^n}^T(\boldsymbol{\theta}) \preceq (1 + \epsilon) \Pi_{\mathbf{u}}^T(\boldsymbol{\theta}) \tag{82}$$

therefore by (monotonicity) and (sub-linearity),

$$(1 - \epsilon) \Phi(\Pi_{\mathbf{u}}^T(\boldsymbol{\theta})) \leq \Phi(\Pi_{\hat{\mathbf{u}}^n}^T(\boldsymbol{\theta})) \leq (1 + \epsilon) \Phi(\Pi_{\mathbf{u}}^T(\boldsymbol{\theta})), \tag{83}$$

which shows,

$$(1 - \epsilon) \leq \frac{U_{\hat{\mathbf{u}}^n}(\boldsymbol{\theta})}{\frac{1}{n} \sum_{i=1}^n \Phi(\Pi_{\mathbf{u}}^{\mathsf{T}_i}(\boldsymbol{\theta}))} \leq (1 + \epsilon). \tag{84}$$

Therefore by union bound,

$$\mathbb{P}\left((1 - \epsilon)^2 \leq \frac{U_{\hat{\mathbf{u}}^n}(\boldsymbol{\theta})}{U_{\mathbf{u}}(\boldsymbol{\theta})} \leq (1 + \epsilon)^2 \right) \geq 1 - (2|E| + 2) \exp\left(-2n\kappa_{\mathbf{u}}^2 \epsilon^2 \right). \tag{85}$$

By putting together (80) and (85), we obtain

$$\mathbb{P}\left((1 - \epsilon)^{\frac{3}{2}} \leq \frac{\widehat{\Phi}_{\hat{\mathbf{u}}^n}(\boldsymbol{\theta})}{\Phi_{\mathbf{u}}(\boldsymbol{\theta})} \leq (1 + \epsilon)^{\frac{3}{2}} \right) \geq 1 - (2|E| + 4) \exp\left(-2n\kappa_{\mathbf{u}}^2 \epsilon^2 \right), \tag{86}$$

and by arguments of Lemma 2, we can conclude that

$$\mathbb{P}\left(\sqrt{\kappa_{\mathbf{u}}}(1 - \epsilon)^{\frac{3}{2}} \leq \frac{\widehat{\Phi}_{\hat{\mathbf{u}}^n}(\boldsymbol{\theta})}{\Phi(\boldsymbol{\theta})} \leq \frac{(1 + \epsilon)^{\frac{3}{2}}}{\sqrt{\kappa_{\mathbf{u}}}} \right) \geq 1 - (2|E| + 4) \exp\left(-2n\sqrt{\kappa_{\mathbf{u}}}^{-2} \epsilon^2 \right), \tag{87}$$

This completes the proof of Lemma 6.

Appendix G. Proof of Lemma 7

Bounded degree graph G. First assume that G has maximum degree d. Consider any edge $e = (s, t) \in E$. Denote $\mathcal{N}(s), \mathcal{N}(t) \subset V$ the neighbours of s and t. Consider current $\iota : V \times V \to \mathbb{R}$ which is a solution of optimization problem corresponding to effective resistance as defined in (37). By definition, we have that the effective resistance u_e for $e \in E$ is given by

$$\begin{aligned}
u_e &= \sum_{(u,v) \in E} \iota(u, v)^2 \\
&\geq \iota(s, t)^2 + \sum_{u \in \mathcal{N}(s) \backslash \{t\}} \iota(s, u)^2 + \sum_{u \in \mathcal{N}(t) \backslash \{s\}} \iota(u, t)^2.
\end{aligned} \tag{88}$$

By constraints of the optimization problem, the sum of currents entering source s and leaving sink t is equal to 1 (whereas it is null for isolated vertices). Therefore, focusing on s, we have $\sum_{u \in \mathcal{N}(s) \setminus \{t\}} |\iota(s, u)| \geq 1 - |\iota(s, t)|$. By applying Cauchy Schwarz inequality, we have that

$$\Big(\sum_{u \in \mathcal{N}(s) \setminus \{t\}} \iota(s, u)^2 \Big) \times \Big(\sum_{u \in \mathcal{N}(s) \setminus \{t\}} 1^2 \Big) \geq (1 - |\iota(s, t)|)^2. \tag{89}$$

Recall that G has maximum vertex degree d and therefore $|\mathcal{N}(s) \setminus \{t\}| \leq d - 1$. Therefore,

$$\sum_{u \in \mathcal{N}(s) \setminus \{t\}} \iota(s, u)^2 \geq \frac{(1 - |\iota(s, t)|)^2}{d - 1}. \tag{90}$$

Because the same holds for the term $\sum_{u \in \mathcal{N}(t) \setminus (s)} \iota(u, t)^2$, we obtain from (88) that

$$u_e \geq \iota(s, t)^2 + (1 - |\iota(s, t)|)^2 \frac{2}{d - 1}. \tag{91}$$

This expression holds for all possible values of $\iota(s, t)$. We note that for any given $\lambda \in \mathbb{R}_+$,

$$\inf_{x \in \mathbb{R}} x^2 + (1 - x)^2 \lambda \geq \frac{\lambda}{1 + \lambda}. \tag{92}$$

Therefore, we conclude that for graph G with bounded degree d,

$$u_e \geq \frac{2}{d + 1}. \tag{93}$$

Graph G with girth g. We now assume that G has girth g. As before, let $e = (s, t) \in E$. Denote $G \setminus \{e\} = (V, E \setminus \{e\})$ the graph obtained by removing edge e from G. For $0 \leq k \leq g - 2$, we define

$$E_k = \{(u, v) \in E : d_{G \setminus \{e\}}(s, u) = k, d_{G \setminus \{e\}}(s, v) = k + 1\}, \tag{94}$$

where $d_{G \setminus \{e\}}(s, u)$ denotes the shortest path distance between vertices s, u in graph G excluding edge e. That is, E_k is the set of edges connecting vertices at distance k from s in $G \setminus \{e\}$ to vertices at distance $k + 1$ from s in $G \setminus \{e\}$. Since $k \leq g - 2$, all E_k are disjoint and hence current ι satisfies

$$u_e \geq \iota(s, t)^2 + \sum_{k=0}^{g-2} \sum_{(u,v) \in E_k} \iota(u, v)^2. \tag{95}$$

For $0 \leq k \leq g - 2$, note that $E_k \cup \{e\}$ defines a cut of G. Therefore by Kirchoff's law $\sum_{(u,v) \in E_k} |\iota(u, v)| \geq 1 - |\iota(s, t)|$. Using Cauchy-Schwartz inequality, we obtain:

$$\Big(\sum_{(u,v) \in E_k} \iota(u, v)^2 \Big) \times \Big(\sum_{(u,v) \in E_k} 1^2 \Big) \geq (1 - \iota(s, t))^2. \tag{96}$$

By summing-up all inequalities, we obtain

$$\Big(\sum_{k=0}^{g-2} \sum_{(u,v) \in E_k} \iota(u, v)^2 \Big) \geq (1 - |\iota(s, t)|)^2 \Big(\sum_{k=0}^{g-2} \frac{1}{|E_k|} \Big). \tag{97}$$

Note that if a sequence $(m_k) \geq 0$ respects $\sum_{k=1}^{l} m_k \leq |E|$ then, $\sum_{k=1}^{l} \frac{1}{m_k} \geq \frac{l^2}{|E|}$. Therefore, because all E_k are disjoint, $\sum_{k=0}^{g-2} \frac{1}{|E_k|} \geq \frac{(g-1)^2}{|E|}$. Inserting this in (95), we obtain

$$u_e \geq \iota(s,t)^2 + (1 - \iota(s,t))^2 \frac{(g-1)^2}{|E|}. \tag{98}$$

Using (92), we obtain

$$u_e \geq \frac{1}{1 + \frac{|E|}{(g-1)^2}}. \tag{99}$$

This completes the proof of Lemma 7.

Appendix H. Proof of Theorem 8

The proof follows by establishing that by definition if $\boldsymbol{\rho} \in \mathcal{P}(\mathsf{Part}_k(G))$ is an (ϵ, k)-partitioning of G, it implied that

$$\kappa(G, \boldsymbol{\rho}) \geq 1 - \epsilon, \tag{100}$$

Indeed, by definition of (ϵ, k) partition, we have that for any $e \in E$,

$$\rho_e = \mathbb{E}_{\mathsf{H} \sim \boldsymbol{\rho}}[\mathbb{1}(e \in \mathsf{H})] \geq 1 - \epsilon. \tag{101}$$

Subsequently, using arguments identical to that for proof of Lemma 2, it follows that $\widehat{\Phi}_{\boldsymbol{\rho}}(\boldsymbol{\theta})$ is $1/\sqrt{\kappa(G, \boldsymbol{\rho})}$ approximation. That is,

$$\sqrt{1 - \epsilon} \leq \frac{\Phi(\boldsymbol{\theta})}{\widehat{\Phi}_{\boldsymbol{\rho}}(\boldsymbol{\theta})} \leq \frac{1}{\sqrt{1 - \epsilon}}. \tag{102}$$

This completes the proof of Theorem 8.