How Many Pairwise Preferences Do We Need to Rank A Graph Consistently?

Aadirupa Saha
Indian Institute of Science, Bangalore
aadirupa@iisc.ac.in

Rakesh Shivanna
Google Inc., Mountain View
rakeshshivanna@google.com

Chiranjib Bhattacharyya
Indian Institute of Science, Bangalore
chiru@iisc.ac.in

Abstract

We consider the problem of optimal recovery of true ranking of n items from a randomly chosen subset of their pairwise preferences. It is well known that without any further assumption, one requires a sample size of $\Omega(n^2)$ for the purpose. We analyze the problem with an additional structure of relational graph $G([n], E)$ over the n items added with an assumption of *locality*: Neighboring items are similar in their rankings. Noting the preferential nature of the data, we choose to embed not the graph, but, its *strong product* to capture the pairwise node relationships. Furthermore, unlike existing literature that uses Laplacian embedding for graph based learning problems, we use a richer class of graph embeddings—*orthonormal representations*—that includes (normalized) Laplacian as its special case. Our proposed algorithm, *Pref-Rank*, predicts the underlying ranking using an SVM based approach over the chosen embedding of the product graph, and is the first to provide *statistical consistency* on two ranking losses: *Kendall's tau* and *Spearman's footrule*, with a required *sample complexity* of $O(n^2\chi(\bar{G}))^{\frac{2}{3}}$ pairs, $\chi(\bar{G})$ being the *chromatic number* of the complement graph $\bar{G}$. Clearly, our sample complexity is smaller for dense graphs, with $\chi(\bar{G})$ characterizing the degree of node connectivity, which is also intuitive due to the *locality* assumption e.g. $O(n^{\frac{4}{3}})$ for union of k-cliques, or $O(n^{\frac{5}{3}})$ for random and power law graphs etc.—a quantity much smaller than the fundamental limit of $\Omega(n^2)$ for large n. This, for the first time, relates ranking complexity to structural properties of the graph. Our theoretical guarantees are justified with extensive experimental evaluations, on different synthetic and real datasets, where our algorithm is shown to outperform the state-of-the-art methods.

1 Introduction

The problem of ranking from pairwise preferences has widespread applications in various real world scenarios e.g. web search Page et al. [1998], Kleinberg [1999], gene classification, recommender systems Theodoridis et al. [2013], image search Geng et al. [2009] and more. Its of no surprise why the problem is so well studied in various disciplines of research, be that computer science, statistics, operational research or computational biology. In particular, we study the problem of ranking (or ordering) of set of n items, given some partial information of the relative ordering of the item pairs.

It is well known from the standard results of classical sorting algorithms, for any set of n items associated to an unknown deterministic ordering, say $\boldsymbol{\sigma}_n^*$, and given the learner has access to only preferences of the item pairs, in general one requires to observe $\Omega(n \log n)$ *actively* selected pairs (where the learner can choose which pair to observe next) to obtain the true underlying ranking $\boldsymbol{\sigma}_n^*$; whereas, with *random* selection of pairs, it could be as bad as $\Omega(n^2)$.

Related Work. Over the years, numerous attempts have been made to improve the above sample complexities by imposing different structural assumptions on the set of items or the underlying ranking model. In active ranking setting, Jamieson and Nowak [2011] gives a sample complexity of $O(d \log^2 n)$, provided the true ranking is realizable in a d-dimensional embedding; Braverman and Mossel [2008] and Ailon [2012] proposed a near optimal recovery with sample complexity of $O(n \log n)$ and $O(n\text{poly}(\log n))$ respectively, under *noisy permutation* and tournament ranking model. For the non-active (random) setting, Wauthier et al. [2013] and Negahban et al. [2012] gave a sample complexity bound of $O(n \log n)$ under noisy permutation (with $O(\log n)$ repeated sampling) and BTL ranking model. Recently, Rajkumar and Agarwal [2016] showed a recovery guarantee of $O(nr \log n)$, given the preference matrix is rank r under suitable transformation.

However, existing literature on sample complexity for *graph based ranking problems* is sparse, where it goes without saying that the underlying structural representation of the data is extremely relevant in various real world applications where the edge connections model item similarities e.g. In social network, connection among friends can be modelled as a graph, or in recommender systems, movies under same the genre should lie in close neighbourhood. It is important to note that a relational graph is different from imposing item dependencies through feature representations and much more practical, since side information of exact features may not even be available to the learner as required in the later case.

Furthermore, the only few algorithmic contributions made on the problem of ranking on graphs – Page et al. [1998], He et al. [2017], Del Corso and Romani [2016], Hsu et al. [2017] have not explored their theoretical performance. Agarwal [2010, 2008] proposed an SVM-rank based algorithm, with generalization error bounds for the inductive and transductive

Table 1: Summary of sample complexities for ranking from pairwise preferences.

Reference	Assumption on the Ranking Model	Sampling Technique	Sample Complexity
Braverman and Mossel [2008]	Noisy permutation	Active	$O(n \log n)$
Jamieson and Nowak [2011]	Low d-dimensional embedding	Active	$O(d \log^2 n)$
Ailon [2012]	Deterministic tournament	Active	$O(n\text{poly}(\log n))$
Gleich and Lim [2011]	Rank-r pairwise preference with ν incoherence	Random	$O(n\nu r(\log n)^2)$
Negahban et al. [2012]	Bradley Terry Luce (BTL)	Random	$O(n \log n)$
Wauthier et al. [2013]	Noisy permutation	Random	$O(n \log n)$
Rajkumar and Agarwal [2016]	Low r-rank pairwise preference	Random	$O(nr \log n)$
Niranjan and Rajkumar [2017]	Low d-rank feature with BTL	Random	$O(d^2 \log n)$
Agarwal [2010]	Graph + Laplacian based ranking	Random	✗
Pref-Rank (This paper)	Graph + Edge similarity based ranking	Random	$O(n^2\chi(\bar{G}))^{\frac{2}{3}}$

graph ranking problems. Agarwal and Chakrabarti [2007] derived generalization guarantees for PageRank algorithm. To the best of our knowledge, we are not aware of any literature which provide *statistical consistency* guarantees to recover the true ranking and analyze the required sample complexity, which remains the primary focus of this work.

Problem Setting We precisely address the question: Given the additional knowledge of a relational graph on the set of n items, say $G([n], E)$, can we find the underlying ranking $\boldsymbol{\sigma}_n^*$ faster (i.e. with a sample complexity lesser than $\Omega(n^2)$)? Of course, in order to hope for achieving a better sample complexity, there must be a connection between the graph and the underlying ranking – question is how to model this?

A natural modelling could be to assume that similar items connected by an edge are close in terms of their rankings or similar node pairs have similar pairwise preferences. E.g. In movie recommendations, if two movies A and B belongs to thriller genre and C belongs to comedy, and it is known that A is preferred over C (i.e. the true ranking over latent topics prefers thriller over comedy), then it is likely that B would be preferred over C; and the learner might not require an explicit (B, C) labelled pair – thus one can hope to reduce the sample complexity by inferring preference information of the neighbouring similar nodes. However, how to impose such a smoothness constraint remains an open problem.

One way out could be to assume the true ranking to be a smooth function over the graph Laplacian as also assumed in Agarwal [2010]. However, why should we confine ourself to the notion of Laplacian embedding based similarity when several other graph embeddings could be explored for the purpose? In particular, we use a broader class of *orthonormal representation* of graphs for the purpose, which subsumes (normalized) Laplacian embedding as a special case, and assume the ranking to be a *smooth function* with respect to the underlying embedding (see Sec. 2.1 for details).

Our Contributions. Under the smoothness assumptions, we show a sample complexity guarantee of $O(n^2\chi(\bar{G}))^{\frac{2}{3}}$ to achieve *ranking consistency* – the result is intuitive as it indicates smaller sample complexity for densely connected graph, as one can expect to gather more information about the neighboring nodes compared to a sparse graph. Our proposed *Pref-Rank* algorithm, to the best of our knowledge, is the first attempt in proving *consistency* on large class of graph families with $\vartheta(G) = o(n)$, in terms of *Kendall's tau* and *Spearman's footrule* losses – It is developed on the novel idea of embedding nodes of the strong product graph $G \boxtimes G$, drawing inference from the preferential nature of the data and finally uses a kernelized-SVM approach to learn the underlying ranking. We summarize our contributions:

- *The choice of graph embedding*: Unlike the existing literature, which is restricted to Laplacian graph embedding Ando and Zhang [2007], we choose to embed the strong product $G \boxtimes G$ instead of G, as our ranking performance measures penalizes every pairwise misprediction; and use a general class of orthonormal representations, which subsumes (normalized) Laplacian as a special case.

- *Our proposed preference based ranking algorithm: Pref-Rank* is a kernelized-SVM based method that inputs an embedding of pairwise graph $G \boxtimes G$. The generalization error of *Pref-Rank* involves computing the transductive *rademacher complexity* of the function class associated with the underlying embedding used (see Thm. 3, Sec. 3).

- For the above, we propose to embed the nodes of $G \boxtimes G$ with 3 different orthonormal representations: (a) Kron-Lab($G \boxtimes G$) (b) PD-Lab(G) and (c) *LS*-labelling; and derive generalization error bounds for the same (Sec. 4).

- *Consistency:* We prove the existence of an optimal embedding in Kron-Lab($G \boxtimes G$) for which *Pref-Rank* is statistically consistent (Thm. 10, Sec. 5) over a large class of graphs, including power law and random graphs. To the best of our knowledge, this is the first attempt at establishing algorithmic consistency for graph ranking problems.

- *Graph Ranking Sample Complexity:* Furthermore, we show that observing $O(n^2\chi(\bar{G}))^{\frac{2}{3}}$ pairwise preferences a sufficient for *Pref-Rank* to be consistent (Thm. 12, Sec. 5.1), which implies that a *densely connected graph requires much smaller training data compared to a sparse graph* for learning the optimal ranking – as also intuitive. Our result is the first to connect the complexity of graph ranking problem to its structural properties. Our proposed bound is a significant improvement in sample complexity (for *random* selection of pairs) for dense graphs e.g. $O(n^{\frac{4}{3}})$ for union of k-cliques; and $O(n^{\frac{5}{3}})$ for random and power law graphs – a quantity much smaller than $\Omega(n^2)$.

Our experimental results demonstrate the superiority of *Pref-Rank* algorithm compared to *Graph Rank* Agarwal [2010], Rank Centrality Negahban et al. [2012] and Inductive Pairwise Ranking Niranjan and Rajkumar [2017] on various synthetic and real-world datasets; validating our theoretical claims. Table 1 summarizes our contributions.

2 Preliminaries and Problem Statement

Notations. Let $[n] := \{1, 2, \dots n\}$, for $n \in \mathbb{Z}_+$. Let x_i denote the i^{th} component of a vector $\mathbf{x} \in \mathbb{R}^n$. Let $\mathbf{1}\{\varphi\}$ denote an indicator function that takes the value 1 if the predicate φ is true and 0 otherwise. Let $\mathbf{1}_n$ denote an n-dimensional vector of all 1's. Let $S^{n-1} = \left\{\mathbf{u} \in \mathbb{R}^n \middle| \|\mathbf{u}\|_2 = 1\right\}$ denote a $(n-1)$ dimensional sphere. For any given matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$, we denote the i^{th} column by $\mathbf{M}_i$, $\forall i \in [n]$ and $\lambda_1(\mathbf{M}) \geq \dots \geq \lambda_n(\mathbf{M})$ to denote its sorted eigenvalues, $tr(\mathbf{M})$ to be its trace. Let $\mathbf{S}_n^+ \in \mathbb{R}^{n \times n}$ denote $n \times n$ square symmetric positive semi-definite matrices. $G(V, E)$ denotes a simple undirected graph, with vertex set $V = [n]$ and edge set $E \subseteq V \times V$. We denote its adjacency matrix by A_G.

Orthonormal Representation of Graphs. Lovász [1979] An orthonormal representation of $G(V, E)$, $V = [n]$ is $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_n] \in \mathbb{R}^{d \times n}$ such that $\mathbf{u}_i^\top \mathbf{u}_j = 0$ whenever $(i, j) \notin E$ and $\mathbf{u}_i \in \mathbf{S}^{d-1}$ $\forall i \in [n]$. Let $Lab(G)$ denote the set of all possible orthonormal representations of G given by $Lab(G) := \{\mathbf{U} \mid \mathbf{U} \text{ is an Orthonormal Representation}\}$. Consider the set of graph kernels $\mathcal{K}(G) := \{\mathbf{K} \in S_n^+ \mid K_{ii} = 1, \forall i \in [n];\ K_{ij} = 0, \forall (i, j) \notin E\}$. Jethava et al. [2013] showed the two sets to be equivalent i.e. for every $\mathbf{U} \in Lab(G)$, one can construct $\mathbf{K} \in \mathcal{K}(G)$ and vice-versa.

Definition 1. Lovász Number. *Lovász [1979] Orthonormal representations $Lab(G)$ of a graph G is associated with an interesting quantity – Lovász number of G, defined as*

$$\vartheta(G) := \min_{\mathbf{U} \in Lab(G)} \min_{\mathbf{c} \in S^{d-1}} \max_{i \in V} \frac{1}{(\mathbf{c}^\top \mathbf{u}_i)^2}$$

Lovász Sandwich Theorem: If $I(G)$ and $\chi(G)$ denote the independence number and chromatic number of the graph G, then $I(G) \leq \vartheta(G) \leq \chi(\bar{G})$ Lovász [1979].

Strong Product of Graphs. Given a graph $G = (V, E)$, strong product of G with itself, denoted by $G \boxtimes G$, is defined over the vertex set $V(G \boxtimes G) = V \times V$, such that two nodes $(i, j), (i', j') \in V(G \boxtimes G)$ is adjacent in $G \boxtimes G$ if and only if $i = i'$ and $(j, j') \in E$, or $(i, i') \in E$ and $j = j'$, or $(i, i') \in E$ and $(j, j') \in E$. Also it is known from the classical work of Lovász [1979] that $\vartheta(G \boxtimes G) = \vartheta^2(G)$ (see Def. 15, Appendix for details).

2.1 Problem Statement

We study the problem of graph ranking on a simple, undirected graph $G = (V, E)$, $V = [n]$. Suppose there exists a true underlying ranking $\boldsymbol{\sigma}_n^* \in \Sigma_n$ of the nodes V, where Σ_n is the set of all permutations of $[n]$, such that for any two distinct nodes $i, j \in V$, i is said to be preferred over j iff $\sigma_n^*(i) < \sigma_n^*(j)$. Clearly, without any structural assumption on how $\boldsymbol{\sigma}_n^*$ relates to the underlying graph $G(V, E)$, the knowledge of $G(V, E)$ is not very helpful in predicting $\boldsymbol{\sigma}_n^*$:

Ranking on Graphs: Locality property. A ranking $\boldsymbol{\sigma}_n^*$ is said to have *locality property* if $\exists$ at least one ranking function $\mathbf{f} \in \mathbb{R}^n$ such that $f(i) > f(j)$ iff $\sigma(i) < \sigma(j)$ and

$$|f(i) - f(j)| \leq c, \text{ whenever } (i, j) \in E, \tag{1}$$

where $c > 0$ is a small constant that quantifies the "locality smoothness" of $\mathbf{f}$. One way is to model $\mathbf{f}$ as a smooth function over the Laplacian embedding $\mathbf{L}$ Agarwal [2010] such that $\mathbf{f}^\top \mathbf{L} \mathbf{f} = \sum_{(i,j) \in E} A_G(i, j)\big(f_i - f_j\big)^2$ is small. However, we generalize this notion to a broader class of embeddings:

Locality with Orthonormal Representations: Formally, we try to solve for $\mathbf{f} \in \text{RKHS}(\mathbf{K})$[1] i.e. $\mathbf{f} = \mathbf{K}\boldsymbol{\alpha}$, for some $\boldsymbol{\alpha} \in \mathbb{R}^n$, where the *locality* here implies $\mathbf{f}$ to be a smooth function over the embedding $\mathbf{K} \in \mathcal{K}(G)$, or alternatively $\mathbf{f}^\top \mathbf{K}^\dagger \mathbf{f} \leq B$, where $\mathbf{K}^\dagger$ is the pseudo inverse of $\mathbf{K}$ and $B > 0$ is a small constant (see Appendix A for more details). Note that if G is a completely disconnected graph, $\mathcal{K}(G) = \{\mathbf{I}_n\}$ is the only choice for $\mathbf{K}$ and f_i's are independent of each other, and the problem is as hard as the classical sorting of n items. But as the density of G increases, or equivalently $\vartheta(G) \leq \chi(\bar{G}) \ll n$, then $\mathcal{K}(G)$ becomes more expressive and the problem enters into an interesting regime, as the node dependencies come to play aiding to faster learning rate. Recall that, however we only have access to G, our task is to find a suitable $\mathbf{K}$ that fits $\mathbf{f}$ on G and estimate $\boldsymbol{\sigma}_n^*$ accurately.

Problem Setup. Consider the set of all node pairs $\mathcal{P}_n = \{(i, j) \in V \times V \mid i < j\}$. Clearly $|\mathcal{P}_n| = \binom{n}{2}$. We will use $N = \binom{n}{2}$ and denote the pairwise preference label of the k^{th} pair (i_k, j_k) as $y_k \in \{\pm 1\}$, such that $y_k := \text{sign}(\sigma_n^*(i_k) - \sigma_n^*(j_k))$, $\forall k \in [N]$. The learning algorithm is given access to a set of randomly chosen node-pairs $S_m \subseteq \mathcal{P}_n$, such that $|S_m| = m \in [N]$. Without loss of generality, by renumbering the pairs we will assume the first m pairs to be labelled $S_m = \{(i_k, j_k)\}_{k=1}^m$, with the corresponding pairwise preference labels $\mathbf{y}_{S_m} = \{y_k\}_{k=1}^m$, and set of unlabelled pairs $\bar{S}_m = \mathcal{P}_n \backslash S_m = \{(i_k, j_k)\}_{k=m+1}^N$. Given G, S_m and $\mathbf{y}_{S_m}$, the goal of the learner is to predict a ranking $\hat{\boldsymbol{\sigma}}_n \in \Sigma_n$ over the nodes V, that gives an accurate estimate of the underlying true ranking $\boldsymbol{\sigma}_n^*$. We use the following ranking losses to measure performance Monjardet [1998]:

Kendall's Tau loss: $d_k(\boldsymbol{\sigma}^*, \hat{\boldsymbol{\sigma}}) = \frac{1}{N} \sum_{k=1}^N \mathbf{1}\big((\sigma^*(i_k) - \sigma^*(j_k))(\hat{\sigma}(i_k) - \hat{\sigma}(j_k)) < 0\big)$ and **Spearman's Footrule loss:** $d_s(\boldsymbol{\sigma}^*, \hat{\boldsymbol{\sigma}}) = \frac{1}{n} \sum_{i=1}^n \Big|\sigma^*(i) - \hat{\sigma}(i)\Big|$. d_k measures the average number of mispredicted pairs, whereas d_s measures the average displacement of the ranking order. By Diaconi-Graham inequality Kumar and Vassilvitskii [2010], we know for any $\boldsymbol{\sigma}, \boldsymbol{\sigma}' \in \Sigma_n$, $d_k(\boldsymbol{\sigma}, \boldsymbol{\sigma}') \leq d_s(\boldsymbol{\sigma}, \boldsymbol{\sigma}') \leq 2 d_k(\boldsymbol{\sigma}, \boldsymbol{\sigma}')$.

Now instead of predicting $\hat{\boldsymbol{\sigma}}_n \in \Sigma_n$, suppose the learner is allowed to predict a pairwise score function $\mathbf{f} : \mathcal{P}_n \mapsto \mathbb{R} \backslash \{0\}$ (note, $\mathbf{f} = [f_k]_{k=1}^N \in (\mathbb{R} \setminus \{0\})^N$ can also be realized as a vector), where f_k denotes the score for every k^{th} pair (i_k, j_k), $k \in [N]$). We measure the prediction accuracy as **pairwise** (0-1) **loss:** $\ell^{0-1}(y_k, f_k) = \mathbf{1}(f_k y_k < 0)$, or using the convex surrogate loss functions – **hinge loss:** $\ell^{\text{hinge}}(y_k, f_k) = (1 - f_k y_k)_+$ or **ramp loss:** $\ell^{\text{ramp}}(y_k, f_k) = \min\{1, (1 - f_k y_k)_+\}$, where $(a)_+ = \max(a, 0)$.

In general, given a transductive learning framework, following the notations from Ando and Zhang [2007], El-Yaniv and Pechyony [2007], for any pairwise preference loss ℓ, we denote the empirical (training) ℓ-error of $\mathbf{f}$ as $er_{S_m}^\ell(\mathbf{f}) = \frac{1}{m} \sum_{k=1}^m \ell(y_k, f_k)$, the generalization (test set) error as

[1] RKHS: Reproducing Kernel Hilbert Space

$er^{\ell}_{\bar{S}_m}(\mathbf{f}) = \frac{1}{N-m}\sum_{k=m+1}^{N}\ell(y_k, f_k)$ and the average pairwise misprediction error as $er^{\ell}_{n}(\mathbf{f}) = \frac{1}{N}\sum_{k=1}^{N}\ell(y_k, f_k)$.

2.2 Learners' Objective - Statistical Consistency for Graph Ranking from Pairwise Preferences

Let $\mathcal{G}$ be a graph family with infinite sequence of nodes $\mathcal{V} = \{v_n\}_{n=1}^{\infty}$. Let V_n denote the first n nodes of $\mathcal{V}$ and $G_n \in \mathcal{G}$ be a graph instance defined over $(V_n, E_1 \cup \ldots \cup E_n)$, where E_n is the edge information of node v_n with previously observed nodes $V_{n-1},\ n \geq 2$. Let $\boldsymbol{\sigma}^*_n \in \Sigma_n$ be the true ranking of the nodes V_n. Now given G_n and $f \in (0,1)$ a fixed number, let Π_f be a uniform distribution on the random draw of $m(f) = \lceil Nf \rceil$ pairs of nodes from N possible pairs $\mathcal{P}_n$. Let $S_{m(f)} = \{(i_k, j_k) \in \mathcal{P}_n\}_{k=1}^{m(f)}$ be an instance of the draw, with corresponding pairwise preferences $\mathbf{y}_{S_{m(f)}} = \{y_k\}_{k=1}^{m(f)}$. Given $(G_n, S_{m(f)}, \mathbf{y}_{S_{m(f)}})$, a learning algorithm $\mathcal{A}$ that returns a ranking $\hat{\sigma}_n$ on the node set V_n is said to be statistically d-rank consistent *w.r.t.* $\mathcal{G}$ if

$$Pr_{S_{m(f)}\sim\Pi_f}\left(d(\boldsymbol{\sigma}^*_n, \hat{\boldsymbol{\sigma}}_n) \geq \epsilon\right) \to 0 \quad as \quad n \to \infty,$$

for any $\epsilon > 0$ and d being the Kendall's tau (d_k) or Spearman's footrule (d_s) ranking losses. In the next section we propose *Pref-Rank* an SVM based graph ranking algorithm and prove it to be statistically d-rank consistent (Sec. 5) with 'optimal embedding' in Kron-Lab($G \boxtimes G$) (Sec. 4.1).

3 *Pref-Rank* - Preference Ranking Algorithm

Given a graph $G(V, E)$ and training set of pairwise preferences $(S_m, \mathbf{y}_{S_m})$, we design an SVM based ranking algorithm that treats each observed pair in S_m as a binary labelled training instance and outputs a pairwise score function $\mathbf{f} \in \mathbb{R}^N$, which is used to estimate the final rank $\hat{\boldsymbol{\sigma}}_n$.

Step 1. Select an embedding ($\tilde{\mathbf{U}}$): Choose a pairwise node embedding $\tilde{\mathbf{U}} = [\tilde{\mathbf{u}}_1, \cdots \tilde{\mathbf{u}}_N] \in \mathbb{R}^{d\times N}$, where any node pair $(i_k, j_k) \in \mathcal{P}_n$ is represented by $\tilde{\mathbf{u}}_k,\ \forall k \in [N]$. We discuss the suitable embedding schemes in Sec. 4.

Step 2. Predict pairwise scores ($\mathbf{f}^* \in \mathbb{R}^N$): We solve the binary classification problem given the embeddings $\tilde{\mathbf{U}}$ and pairwise node preferences $\{(\tilde{\mathbf{u}}_k, \mathbf{y}_k)\}_{k=1}^{m}$ using SVM:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{k=1}^{m}\ell^{\text{hinge}}(y_k, \mathbf{w}^\top\tilde{\mathbf{u}}_k) \tag{2}$$

where $C > 0$ is a regularization hyperparameter. Note that the dual of the above formulation is given by:

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^m_+,\ \|\boldsymbol{\alpha}\|_\infty \leq C} \sum_{k=1}^{m}\alpha_k - \frac{1}{2}\sum_{k,k'\in[m]}\alpha_k\alpha_{k'}y_k y_{k'}\tilde{\mathbf{K}}_{k,k'}$$

where $\tilde{\mathbf{K}} = \tilde{\mathbf{U}}^\top\tilde{\mathbf{U}}$ denotes the embedding kernel of the pairwise node instances. From standard results of SVM, we know that optimal solution of (2) gives $\mathbf{w}^* = \sum_{k=1}^{m} y_k\tilde{\mathbf{U}}_k\alpha_k = \tilde{\mathbf{U}}\boldsymbol{\beta}$, where $\boldsymbol{\beta} \in \mathbb{R}^N$ is such that $\beta_k = y_k\alpha_k,\ \forall k \in [m]$ and 0 otherwise. Since $y_k \in \{\pm 1\}$, $\|\boldsymbol{\alpha}\|_\infty = \|\boldsymbol{\beta}\|_\infty \leq C$. Thus for any $k \in [N]$, the score of the pair (i_k, j_k) is given by $f^*_k = \mathbf{w}^{*\top}\tilde{\mathbf{u}}_k = \sum_{l\in[m]} y_l\alpha_l\tilde{\mathbf{u}}_l^\top\tilde{\mathbf{u}}_k$ or equivalently $\mathbf{f}^* = \tilde{\mathbf{U}}^\top\mathbf{w}^* = \tilde{\mathbf{U}}^\top\tilde{\mathbf{U}}\boldsymbol{\beta} = \tilde{\mathbf{K}}\boldsymbol{\beta}$, which suggests an alternate formulation of SVM:

$$\max_{\mathbf{f}\in\mathbb{R}^N} \frac{1}{2}\mathbf{f}^\top\tilde{\mathbf{K}}^\dagger\mathbf{f} + Cm\ \hat{er}^{\ell^{hinge}}_{S_m}(\mathbf{f}) \tag{3}$$

Clearly, if $\mathbf{f}^*$ denotes the optimal solution of (3), then we have $\mathbf{f}^* \in \{\mathbf{f} \mid \mathbf{f} = \tilde{\mathbf{K}}\boldsymbol{\beta},\ \boldsymbol{\beta} \in \mathbb{R}^N,\ \|\boldsymbol{\beta}\|_\infty \leq C\}$.

Remark 1. The regularization $\mathbf{f}^\top\tilde{\mathbf{K}}^\dagger\mathbf{f}$, precisely enforces the *locality* assumption of Sec. 2.1 *(see Lem. 14, Appendix).*

Step 3. Predict $\hat{\boldsymbol{\sigma}}_n \in \Sigma_n$ from pairwise scores $\mathbf{f}^*$: Given the score vector $\mathbf{f}^* \in \mathbb{R}^N$ as computed above, predict a ranking $\hat{\boldsymbol{\sigma}}_n \in \Sigma_n$ over the nodes V of G as follows:

1. Let $c(i)$ denote the number of wins of node $i \in V$ given by $\sum_{\{k=(i_k,j_k)|i_k=i\}}\mathbf{1}\big(f^*_k > 0\big) + \sum_{\{k=(i_k,j_k)|j_k=i\}}\mathbf{1}\big(f^*_k < 0\big)$.
2. Predict the ranking of nodes by sorting *w.r.t.* $c(i)$, i.e. choose any $\hat{\boldsymbol{\sigma}}_n \in \text{argsort}(\mathbf{c})$, where $\text{argsort}(\mathbf{c}) = \big\{\boldsymbol{\sigma} \in \Sigma_n \mid \sigma(i) < \sigma(j),\ \text{if } c(i) > c(j),\ \forall i, j \in V\big\}$.

A brief outline of *Pref-Rank* is given below:

Algorithm *Pref-Rank*

Input: $G(V, E)$ and subset of preferences $(S_m, \mathbf{y}_{S_m})$
Init: Pairwise graph embedding $\tilde{\mathbf{U}} \in \mathbb{R}^{d\times N},\ d \in \mathbb{N}_+$
Compute preference scores $\mathbf{f}^* = \tilde{\mathbf{U}}^\top\mathbf{w}^*$ using (2)
Count number of wins for each node $i \in V$
$c(i) := \sum_{\{k=(i_k,j_k)|i_k=i\}}\mathbf{1}\big(f^*_k > 0\big) + \sum_{\{k=(i_k,j_k)|j_k=i\}}\mathbf{1}\big(f^*_k < 0\big)$
Return ranking of nodes $\hat{\boldsymbol{\sigma}}_n \in \text{argsort}(\mathbf{c})$

3.1 Generalization Error of *Pref-Rank*

We now derive generalization guarantees of *Pref-Rank* (Sec. 3) on its test error $er^{\ell^\rho}_{\bar{S}_m}(\mathbf{f}^*) = \frac{1}{N-m}\sum_{k=m+1}^{N}\ell^\rho(y_k, f^*_k)$, *w.r.t.* some loss function $\ell^\rho : \{\pm 1\} \times \mathbb{R} \mapsto \mathbb{R}_+$, where ℓ^ρ is assumed to be ρ-lipschitz ($\rho > 0$) with respect to its second argument i.e. $|\ell^\rho(y_k, f_k) - \ell^\rho(y_k, f'_k)| \leq \frac{1}{\rho}|f_k - f'_k|$, where $\mathbf{f}, \mathbf{f}' : \mathcal{P}_n \mapsto \mathbb{R}$ be any two pairwise score functions. We find it convenient to define the following function class complexity measure associated with orthonormal embeddings of pairwise preference strong product of graphs (as motivated in Pelckmans et al. [2007]):

Definition 2 (Transductive Rademacher Complexity). *Given a graph $G(V, E)$, let $\tilde{\mathbf{U}} \in \mathbb{R}^{d\times N}$ be any pairwise embedding of G and let $col(\tilde{\mathbf{U}})$ denote the column space spanned by $\tilde{\mathbf{U}}$. Then for any function class $\mathcal{H}_{\tilde{\mathbf{U}}} = \{\mathbf{h} \mid \mathbf{h} : col(\tilde{\mathbf{U}}) \mapsto \mathbb{R}\}$ associated with $\tilde{\mathbf{U}}$, its transductive Rademacher complexity is defined as*

$$R(\mathcal{H}_{\tilde{\mathbf{U}}}, \tilde{\mathbf{U}}, p) = \frac{1}{N}\mathbb{E}_{\boldsymbol{\gamma}}\left[\sup_{\mathbf{h}\in\mathcal{H}_{\tilde{\mathbf{U}}}}\sum_{k=1}^{N}\gamma_k\mathbf{h}(\tilde{\mathbf{u}}_k)\right],$$

where for any fixed $p \in (0, 1/2]$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)$ is a vector of i.i.d. *random variables such that $\gamma_i \sim \{+1, -1, 0\}$ with probability p, p and $1 - 2p$ respectively.*

We bound the generalization error of *Pref-Rank* in terms of the rademacher complexity. Note the result below crucially depends on the fact that any score vector $\mathbf{f}^*$ returned by *Pref-Rank*, is of the form $\mathbf{f}^* = \tilde{\mathbf{U}}^\top \mathbf{w}^*$, for some $\mathbf{w}^* \in \{\mathbf{h} \mid \mathbf{h} = \tilde{\mathbf{U}}\boldsymbol{\beta}, \boldsymbol{\beta} \in \mathbb{R}^N, \|\boldsymbol{\beta}\|_\infty \leq C\}$, where $\tilde{\mathbf{U}} \in \mathbb{R}^{d \times N}$ be the embedding used in *Pref-Rank* (refer (2), (3) for details).

Theorem 3 (**Generalization Error of** *Pref-Rank*)**.** *Given a graph $G(V, E)$, let $\tilde{\mathbf{U}} \in \mathbb{R}^{d \times N}$ be any pairwise embedding of G. For any $f \in (0, 1/2]$, let Π_f be a uniform distribution on the random draw of $m(f) = \lceil Nf \rceil$ pairs of nodes from $\mathcal{P}_n$, such that $S_{m(f)} = \{(i_k, j_k) \in \mathcal{P}_n\}_{k=1}^{m(f)} \sim \Pi_f$, with corresponding pairwise preference $\mathbf{y}_{S_{m(f)}}$. Let $\bar{S}_{m(f)} = \mathcal{P}_n \backslash S_{m(f)}$. Let $\mathcal{H}_{\tilde{\mathbf{U}}} = \{\mathbf{w} \mid \mathbf{w} = \tilde{\mathbf{U}}\boldsymbol{\beta},\ \boldsymbol{\beta} \in \mathbb{R}^N,\ \|\boldsymbol{\beta}\|_\infty \leq C,\ C > 0\}$ and $\ell^\rho : \{\pm 1\} \times \mathbb{R} \mapsto [0, B]$ be a bounded, ρ-Lipschitz loss function. For any $\delta > 0$, with probability $\geq 1 - \delta$ over $S_{m(f)} \sim \Pi_f$*

$$er^{\ell^\rho}_{\bar{S}_{m(f)}}(\mathbf{f}^*) \leq er^{\ell^\rho}_{S_{m(f)}}(\mathbf{f}^*) + \frac{R(\mathcal{H}_{\tilde{\mathbf{U}}}, \tilde{\mathbf{U}}, p)}{\rho f(1-f)} + \frac{C_1 B \sqrt{\ln\left(\frac{1}{\delta}\right)}}{(1-f)\sqrt{Nf}},$$

where $p = f(1-f)$ and $\mathbf{f}^* = \tilde{\mathbf{U}}^\top \mathbf{w}^* \in \mathbb{R}^N$ is pairwise score vector output by Pref-Rank and $C_1 > 0$ is a constant.

Remark 2. It might appear from above that a higher value of $R(\mathcal{H}_{\tilde{\mathbf{U}}}, \tilde{\mathbf{U}}, p)$ leads to increased generalization error. However, note that there is a *tradeoff* between the first and second term since a higher rademacher complexity implies a richer function class $\mathcal{H}_{\tilde{\mathbf{U}}}$, which in turn is capable of producing a better prediction estimate $\mathbf{f}^* = \tilde{\mathbf{U}}^\top \mathbf{w}$, resulting in a much lower training set error $er^{\ell^\rho}_{S_{m(f)}}[\mathbf{f}^*]$. Thus, a *higher value of $R(\mathcal{H}_{\tilde{\mathbf{U}}})$ is desired* better generalization performance.

Taking insights from Thm. 3, it follows that the performance of *Pref-Rank* crucially depends on the *rademacher complexity* $R(\mathcal{H}_{\tilde{\mathbf{U}}}, \tilde{\mathbf{U}}, p)$ of the underlying function class $\mathcal{H}_{\tilde{\mathbf{U}}}$, which boils down to the problem of finding a "good" embedding $\tilde{\mathbf{U}}$. We address this issue in the next section.

4 Choice of Embeddings

We discuss different classes of pairwise graph embeddings and their generalization guarantees. Recalling the results of Ando and Zhang [2007] (see Thm. 1), which provides a crucial characterization of the class of optimal embeddings for any graph based regularization algorithms, we choose to work with embeddings with normalized kernels, i.e. $\tilde{\mathbf{K}} = \tilde{\mathbf{U}}^\top \tilde{\mathbf{U}}$ such that $\tilde{K}_{kk} = 1, \forall k \in [N]$. The following theorem analyses the rademacher complexity of 'normalized' embeddings:

Theorem 4 (**Rademacher Complexity of Orthonormal Embeddings**)**.** *Given $G(V, E)$, let $\tilde{\mathbf{U}} \in \mathbb{R}^{d \times N}$ be any 'normalized' node-pair embedding of $G \boxtimes G$, let $\tilde{\mathbf{K}} = \tilde{\mathbf{U}}^\top \tilde{\mathbf{U}}$ be the corresponding graph-kernel, then $R(\mathcal{H}_{\tilde{\mathbf{U}}}, \tilde{\mathbf{U}}, p) \leq C\sqrt{2p\lambda_1(\tilde{\mathbf{K}})}$, where $\lambda_1(\tilde{\mathbf{K}})$ is the largest eigenvalue of $\tilde{\mathbf{K}}$.*

Note that the above result does not educate us on the choice of $\tilde{\mathbf{U}}$ – we impose more structural constraints and narrow down the search space of optimal 'normalized' graph embeddings and propose the following special classes:

4.1 Kron-Lab($G \boxtimes G$): Kronecker Product Orthogonal Embedding

Given any graph $G(V, E)$, with $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots \mathbf{u}_n] \in \mathbb{R}^{d \times n}$ being an orthogonal embedding of G, i.e. $\mathbf{U} \in \text{Lab}(G)$, its Kronecker Product Orthogonal Embedding:

$$\text{Kron-Lab}(G \boxtimes G) := \{\tilde{\mathbf{U}} \in \mathbb{R}^{d^2 \times n^2} \mid \tilde{\mathbf{U}} = \mathbf{U} \otimes \mathbf{U}, \mathbf{U} \in \mathbb{R}^{d \times n} \text{ such that } \mathbf{U} \in \text{Lab}(G)\},$$

where $\otimes$ is the kronecker (or outer) product of two matrix. The *'niceness'* of the above embedding lies in the fact that one can construct $\tilde{\mathbf{U}} \in$ Kron-Lab($G \boxtimes G$) from any orthogonal embedding of the original graph $\mathbf{U} \in \text{Lab}(G)$ – let $\mathbf{K} := \mathbf{U}^\top \mathbf{U}$ and $\tilde{\mathbf{K}} := \tilde{\mathbf{U}}^\top \tilde{\mathbf{U}}$, we see that for any two $k, k' \in [n^2]$, $\tilde{\mathbf{K}}_{kk'} = \tilde{\mathbf{u}}_k^\top \tilde{\mathbf{u}}'_k = (\mathbf{u}_{i_k} \otimes \mathbf{u}_{j_k})^\top (\mathbf{u}_{i_{k'}} \otimes \mathbf{u}_{j_{k'}}) = (\mathbf{u}_{i_k}^\top \mathbf{u}_{i_{k'}})(\mathbf{u}_{j_k}^\top \mathbf{u}_{j_{k'}}) = \mathbf{K}_{i_k i_{k'}} \mathbf{K}_{j_k j_{k'}}$, where $(i_{(\cdot)}, j_{(\cdot)}) \in [n] \times [n]$ are the node pairs corresponding to k, k'. Hence $\tilde{\mathbf{K}} = \mathbf{K} \otimes \mathbf{K}$. Note that when $k = k'$, we have $\tilde{\mathbf{K}}_{kk} = 1$, as $\mathbf{U} \in \text{Lab}(G)$, $K_{ii} = 1, \forall i \in [n]$. This ensures that the kronecker product graph kernel $\tilde{\mathbf{K}}$ satisfies the optimality criterion of 'normalized' embedding as previously discussed.

Lemma 5 (**Rademacher Complexity of Kron-Lab($G \boxtimes G$)**)**.** *Consider any $\mathbf{U} \in$ Lab(G), $\mathbf{K} = \mathbf{U}^\top \mathbf{U}$ and the corresponding $\tilde{\mathbf{U}} \in$ Kron-Lab($G \boxtimes G$). Then for any $p \in [0, 1]$ and $\mathcal{H}_{\tilde{\mathbf{U}}} = \{\mathbf{w} \mid \mathbf{w} = \tilde{\mathbf{U}}\boldsymbol{\beta},\ \boldsymbol{\beta} \in R^N,\ \|\boldsymbol{\beta}\|_\infty \leq C,\ C > 0\}$ we have, $R(\mathcal{H}_{\tilde{\mathbf{U}}}, \tilde{\mathbf{U}}, p) \leq C\lambda_1(\mathbf{K})\sqrt{2p}$.*

Above leads to the following generalization guarantee:

Theorem 6 (**Generalization Error of** *Pref-Rank* **with Kron-Lab($G \boxtimes G$)**)**.** *For the setting as in Thm. 3 and Lem. 5, for any $\tilde{\mathbf{U}} \in$ Kron-Lab($G \boxtimes G$), we have*

$$er^{\ell^\rho}_{\bar{S}}[\mathbf{f}^*] \leq er^{\ell^\rho}_{S}[\mathbf{f}^*] + \frac{C\lambda_1(\mathbf{K})\sqrt{2}}{\rho\sqrt{f(1-f)}} + \frac{C_1 B}{1-f}\sqrt{\frac{\log(\frac{1}{\delta})}{Nf}}$$

4.2 Pairwise Difference Orthogonal Embedding

Given any graph $G(V, E)$, let $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots \mathbf{u}_n] \in \mathbb{R}^{d \times n}$ be such that $\mathbf{U} \in \text{Lab}(G)$. We define the class of *Pairwise Difference Orthogonal Embedding* of G as:

$$\text{PD-Lab}(G) := \{\tilde{\mathbf{U}} \in \mathbb{R}^{d \times N} \mid \tilde{\mathbf{u}}_{ij} = \mathbf{u}_i - \mathbf{u}_j\ \forall (i, j) \in \mathcal{P}_n, \mathbf{U} \in \mathbb{R}^{d \times n} \text{ such that } \mathbf{U} \in \text{Lab}(G)\}$$

Let $\mathbb{E} = [\mathbf{e}_i - \mathbf{e}_j]_{(i,j) \in \mathcal{P}_n} \in \{0, \pm 1\}^{n \times N}$, where $\mathbf{e}_i$ denotes the i^{th} standard basis of $\mathbb{R}^n$, $\forall i \in [n]$; then it is easy to note that $\tilde{\mathbf{U}} = \mathbf{U}\mathbb{E} \in$ PD-Lab(G) and the corresponding graph kernel is given by $\tilde{\mathbf{K}} = \mathbb{E}^\top \mathbf{K} \mathbb{E}$. For PD embedding, we get:

Lemma 7 (**Rademacher Complexity of PD-Lab(G)**)**.** *Consider any $\mathbf{U} \in$ Lab(G), $\mathbf{K} = \mathbf{U}^\top \mathbf{U}$ and the corresponding $\tilde{\mathbf{U}} \in$ PD-Lab(G). Then for any $p \in [0, 1]$ and $\mathcal{H}_{\tilde{\mathbf{U}}} = \{\mathbf{w} \mid \mathbf{w} = \tilde{\mathbf{U}}\boldsymbol{\beta},\ \boldsymbol{\beta} \in R^N,\ \|\boldsymbol{\beta}\|_2 \leq tC\sqrt{N},\ C > 0\}$, we have $R(\mathcal{H}_{\tilde{\mathbf{U}}}, \tilde{\mathbf{U}}, p) \leq 2C\sqrt{pn\lambda_1(\mathbf{K})}$.*

Similarly as before, using above we can show that:

Theorem 8 (**Generalization Error of** *Pref-Rank* **with PD-Lab(G)**). *For the setting as in Thm. 3 and Lem. 7, for any $\tilde{\mathbf{U}} \in$ PD-Lab(G), we have*

$$er_S^{\ell^\rho}[\mathbf{f}^*] \leq er_S^{\ell^\rho}[\mathbf{f}^*] + \frac{2C\sqrt{n\lambda_1(\mathbf{K})}}{\rho\sqrt{f(1-f)}} + \frac{C_1 B}{1-f}\sqrt{\frac{\log(\frac{1}{\delta})}{Nf}}$$

Recall from Thm 3 that $\mathbf{f}^* = \tilde{\mathbf{U}}^\top \mathbf{w}$. Thus the *'niceness'* of PD-Lab(G) lies in the fact that it comes with the free transitivity property – for any two node pairs $k_1 := (i,j)$ and $k_2 := (j,l)$, if $\mathbf{f}^*$ scores node i higher than j i.e. $f^*_{k_1} > 0$, and node j higher than node l i.e. $f^*_{k_2} > 0$; then for any three nodes $i, j, l \in [n]$, this automatically implies $f^*_{k_3} > 0$, where $k_3 := (i,l)$ i.e. node i gets a score higher than node l.

Remark 3. Although Lem. 5 and 7 shows that both Kron-Lab($G \boxtimes G$) and PD-Lab(G) are associated to rich expressive function classes with high rademacher complexity, *the superiority of Kron-Lab($G \boxtimes G$) comes with an additional consistency guarantee*, as we will derive in Sec. 5.

4.3 *LS*-labelling based Embedding

The embedding (graph kernel) corresponding to *LS*-labelling Luz and Schrijver [2005] of graph G is given by:

$$\mathbf{K}_{LS}(G) = \frac{\mathbf{A}_G}{\tau} + \mathbf{I}_n, \text{ where } \tau \geq |\lambda_n(\mathbf{A}_G)|, \qquad (4)$$

where $\mathbf{A}_G$ is the adjacency matrix of graph G. It is known that $\mathbf{K}_{LS} \in \mathbb{R}^{n\times n}$ is symmetric and positive semi-definite, and hence defines a valid graph kernel; also $\exists \mathbf{U}_{LS} \in \text{Lab}(G)$ such that $\mathbf{U}_{LS}^\top \mathbf{U}_{LS} = \mathbf{K}_{LS}$. We denote $\mathbf{U}_{LS}$ to be the corresponding embedding matrix for *LS*-labelling. We define *LS*-labelling of the strong product of graphs as:

$$\tilde{\mathbf{K}}_{LS}(G \boxtimes G) = \mathbf{K}_{LS}(G) \otimes \mathbf{K}_{LS}(G) \qquad (5)$$

and equivalently the embedding matrix $\tilde{\mathbf{U}}_{LS}(G \boxtimes G) = \mathbf{U}_{LS}(G) \otimes \mathbf{U}_{LS}(G)$. Similar to Kron-Lab($G \boxtimes G$), we have $\tilde{\mathbf{K}}_{LS}(k,k) = 1,\ \forall k \in [n^2]$, since $\mathbf{K}_{LS}(i,i) = 1,\ \forall i \in [n]$. Following result shows that $\tilde{\mathbf{K}}_{LS}(G \boxtimes G)$ has high Rademacher complexity on random $G(n,q)$ graphs.

Lemma 9. *Let $G(n,q)$ be a Erdós-Réyni random graph, where each edge is present independently with probability $q \in [0,1]$, $q = O(1)$. Then the Rademacher complexity of function class associated with $\tilde{\mathbf{K}}_{LS}(G \boxtimes G)$ is $O(\sqrt{n})$.*

Laplacian based Embedding. This is the most popular choice of graph embedding that uses the inverse of the Laplacian matrix for the purpose. Formally, let d_i denotes the degree of vertex $i \in [n]$ in graph G, i.e. $d_i = (\mathbf{A}_G)_i^\top \mathbf{1}_n$, and $\mathbf{D}$ denote a diagonal matrix such that $D_{ii} = d_i, \forall i \in [n]$. Then the Laplacian and normalized Laplacian kernel of G is defined as follows: $\mathbf{K}_{Lap}(G) = (\mathbf{D} - \mathbf{A}_G)^\dagger$ and $\mathbf{K}_{nLap}(G) = (\mathbf{I}_n - \mathbf{D}^{-1/2}\mathbf{A}_G\mathbf{D}^{-1/2})^{\dagger\ 2}$.

Though widely used Agarwal [2010], Ando and Zhang [2007], it is not very expressive on dense graphs with high $\chi(G)$ – we observe that the Rademacher complexity of function class associated with Laplacian is an order magnitude smaller than that of *LS*-labelling. See App. C.8 for details.

[2] † denotes the pseudo inverse.

5 Consistency with Kron-Lab($G \boxtimes G$)

In this section, we show that *Pref-Rank* is provably statistically consistent while working with *kronecker product orthogonal embedding* Kron-Lab($G \boxtimes G$)(see Sec. 4.1).

Theorem 10 (**Rank-Consistency**). *For the setting as in Sec. 2.2, there exists an embedding $\tilde{\mathbf{U}}_n \in$ Kron-Lab($G_n \boxtimes G_n$) such that if $\boldsymbol{\sigma}_n \in \mathbb{R}^N$ denotes the pairwise scores returned by Pref-Rank on input $(\tilde{\mathbf{U}}_n, S_m(f), \mathbf{y}_{S_{m(f)}})$, then $\forall G_n \in \mathcal{G}$, with probability at least $\left(1 - \frac{1}{N}\right)$ over $S_{m(f)} \sim \Pi_f$*

$$d(\boldsymbol{\sigma}_n^*, \hat{\boldsymbol{\sigma}}_n) = O\left(\left(\frac{\vartheta(G_n)}{nf}\sqrt{\frac{1-f}{f}}\right)^{\frac{1}{2}} + \sqrt{\frac{\ln n}{Nf}}\right),$$

where d denotes Kendall's tau (d_k) or Spearman's footrule (d_s) ranking loss functions.

Consistency follows from the fact that for large families of graphs including random graphs Coja-Oghlan [2005] and power law graphs Jethava et al. [2013], $\vartheta(G_n) = o(n)$.

5.1 Sample Complexity for Ranking Consistency

We analyze the minimum fraction of pairwise node preferences f^* to be observed for *Pref-Rank* algorithm to be statistically ranking consistent. We refer the required sample size $m(f^*) = \lceil Nf^* \rceil$ as *ranking sample complexity*.

Lemma 11. *If $\mathcal{G}$ in Thm. 10 is such that $\vartheta(G_n) = n^c$, $0 \leq c < 1$. Then observing only $f^* = O\left(\frac{\sqrt{\vartheta(G_n)}}{n^{\frac{1}{2}-\varepsilon}}\right)^{\frac{4}{3}}$ fraction of pairwise node preferences is sufficient for Pref-Rank to be statistically rank consistent, for any $0 < \varepsilon < \frac{(1-c)}{2}$.*

Note that one could potentially choose any $\varepsilon \in (0, \frac{1-c}{2})$ for the purpose – the *tradeoff* lies in the fact that a higher ε leads to *faster convergence rate of* $d(\boldsymbol{\sigma}_n^*, \hat{\boldsymbol{\sigma}}_n) = O(\frac{1}{n^\varepsilon})$, although at the *cost of increased sample complexity*; on the contrary setting $\varepsilon \to 0$ gives a smaller sample complexity, with significantly slower convergence rate (see proof of Lem. 11 in App. for details). We further extend Lem. 11 and relate ranking sample complexity to structural properties of the graph – *coloring number* of the complement graph $\chi(\bar{G})$.

Theorem 12. *Consider a graph family $\mathcal{G}$ such that $\chi(\bar{G}_n) = o(n)$, $\forall G_n \in \mathcal{G}$. Then observing $O(n^2\chi(\bar{G}))^{\frac{2}{3}}$ pairwise preferences is sufficient for Pref-Rank to be consistent.*

Above conveys that for *dense graphs we need fewer pairwise samples compared to sparse graphs* as $\chi(\bar{G})$ reduces with increasing graph density. We discuss the sample complexities for some special graphs below where $\vartheta(G) = o(n)$.

Corollary 13 (**Ranking Consistency on Special Graphs**). *Pref-Rank algorithm achieves consistency on the following graph families, with the required sample complexities – (a) Complete graphs: $O(n^{\frac{4}{3}})$ (b) Union of k disjoint cliques: $O(n^{\frac{4}{3}}k^{\frac{2}{3}})$ (c) Complement of power-law graphs: $O(n^{\frac{5}{3}})$ (d) Complement of k-colorable graphs: $O(n^{\frac{4}{3}}k^{\frac{2}{3}})$ (e) Erdős Réyni random $G(n,q)$ graphs with $q = O(1)$: $O(n^{\frac{5}{3}})$.*

Remark 4. Thm. 10 along with Lem. 11 suggest that if the graph satisfies a crucial structural property: $\vartheta(G) = o(n)$ and given sufficient sample of $\Omega(n^2\vartheta(G))^{\frac{2}{3}}$ pairwise preferences, *Pref-Rank* yields consistency. Note that $\vartheta(G) \leq \chi(\bar{G}) \leq n$, where the last inequality is tight for completely disconnected graph – which implies one need to observe $\Omega(n^2)$ pairs for consistency, as a disconnected graph does not impose any structure on the ranking. Smaller the $\vartheta(G)$, denser the graph and we attain consistency observing a smaller number of node pairs, the best is of course when G is a clique, as $\vartheta(G) = 1$! So for sparse graphs with $\vartheta(G) = \Theta(n)$, consistency and learnability is far fetched without observing $\Omega(n^2)$ pairs.

Note that proof of Thm. 10 relies on the fact that the maximum SVM margin attained for the formulation (2) is $\vartheta(G \boxtimes G)$, which is achieved by *LS*-labelling on Erdős Réyni random graphs Shivanna and Bhattacharyya [2014]; and thus guarantee consistency, with $O(n^{\frac{5}{3}})$ sample complexity.

6 Experiments

We conducted experiments on both real world and synthetic graphs, comparing *Pref-Rank* with the following algorithms:

Algorithms. We thus used the following 5 algorithms: (a) **PR-Kron**: *Pref-Rank* with $\tilde{\mathbf{K}}_{LS}(G \boxtimes G)$ (see Eqn. (5)) (b) **PR-PD**: *Pref-Rank* with PD-Lab(G) with *LS*-labelling i.e. $\mathbf{U} = \mathbf{U}_{LS}$, (c) **GR**: Graph Rank Agarwal [2010], (d) **RC**: Rank Centrality Negahban et al. [2012] and (e) **IPR**: Inductive Pairwise Ranking, with Laplacian as feature embedding Niranjan and Rajkumar [2017].

Recall from the list of algorithms in Table 1. Except Agarwal [2010], none of the other applies directly to ranking on graphs. Moreover they work only under specific models – e.g. *noisy permutations* for Wauthier et al. [2013], Rajkumar and Agarwal [2016] requires the knowledge of the preference matrix rank r etc. We compare with **RC** (works only under BTL model) and **IPR** (requires item features), but as expected both perform poorly. For better comparison, we present plots comparing only the initial 3 methods in App. E.

Performance Measure. Note the generalization guarantee of Thm. 3 not only holds for full ranking but for any *general preference learning problem*, where the nodes of G are assigned to an underlying preference vector $\boldsymbol{\sigma}^*_n \in \mathbb{R}^n$. Similarly, the goal is to predict a pairwise score vector $\mathbf{f} \in \mathbb{R}^N$ to optimize the average pairwise mispredictions *w.r.t.* some loss function $\ell : \{\pm 1\} \times \mathbb{R} \setminus \{0\} \mapsto \mathbb{R}_+$ defined as:

$$er^{\ell}_D(\mathbf{f}) = \frac{1}{|D|} \sum_{k \in D} \ell(y_k, f_k), \tag{6}$$

where $D = \{(i_k, j_k) \in \mathcal{P}_n \mid \sigma^*_n(i_k) \neq \sigma^*_n(j_k),\ k \in [N]\} \subseteq \mathcal{P}_n$ denotes the subset of node pairs with distinct preferences and $y_k = \text{sign}(\sigma^*_n(j_k) - \sigma^*_n(i_k)),\ \forall k \in D$. In particular, *Pref-Rank* applies to *bipartite ranking* **(BR)**, where $\boldsymbol{\sigma}^*_n \in \{\pm 1\}^n$, *categorical or d-class ordinal ranking* **(OR)**, where $\boldsymbol{\sigma}^*_n \in [d]^n$, $d < n$, and the original *full ranking* **(FR)** problem as motivated in Sec. 2.1. We consider all three tasks in our experiments with **pairwise** 0-1 **loss**, i.e. $\ell(y_k, f_k) = \mathbf{1}(y_k f_k < 0)$. $er^{\ell^{0-1}}_n(\mathbf{f}^*)$ in Eqn. (6).

6.1 Synthetic Experiments

Graphs. We use 3 *types of graphs*, each with $n = 30$ nodes: (a) *Union of k-disconnected cliques* with $k = 2$ and 10, (b) *r-Regular graphs* with $r = 5$ and 15; and (c) $G(n, q)$ *Erdős Réyni random graphs* with edge probability $q = 0.2$ and 0.6.

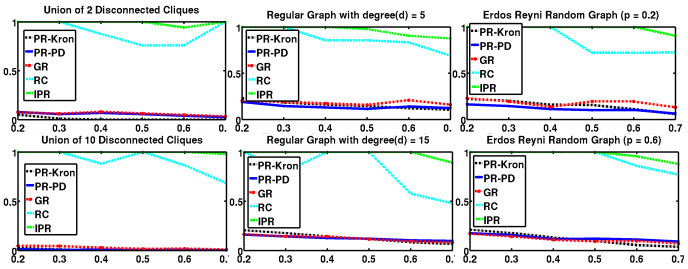

Figure 1: Synthetic Data: Average number of mispredictions ($er^{\ell^{0-1}}_D(\mathbf{f})$, Eqn. (6)) vs fraction of sampled pairs (f).

Generating $\boldsymbol{\sigma}^*_n$. For each of the above graphs, we compute $\mathbf{f}^* = \mathbf{A}_G \boldsymbol{\alpha}$, where $\boldsymbol{\alpha} \in [0, 1]^n$ is generated randomly, and set $\boldsymbol{\sigma}^*_n = \text{argsort}(\mathbf{f}^*)$ (see *Pref-Rank*, Step 3 for definition).

All the performances are averaged across 10 repeated runs. The results are reported in Fig. 6.1. In all the cases, our proposed algorithms **PR-Kron** and **PR-PD** outperforms the rest, with **GR** performing competitively well [3]. As expected, **RC** and **IPR** perform very poorly as they could not exploit the underlying graph *locality* based ranking property.

6.2 Real-World Experiments

Datasets. We use 6 standard real datasets[4] for three graph learning tasks – (a) *Heart* and *Fourclass* for **BR**, (b) *Vehicle* and *Vowel* for **OR**, and (c) *House* and *Mg* for **FR**.

Graph generation. For each dataset, we select 10 random subsets of 40 items each and construct a similarity matrix using RBF kernel, where $(i, j)^{\text{th}}$ entry is given by $\exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\mu^2}\right)$, $\mathbf{x}_i$ being the feature vector and μ the average distance. For each of the 10 subsets, we constructed a graph by thresholding the similarity matrices about the mean.

Generating $\boldsymbol{\sigma}^*_n$. For each dataset, the provided item labels are used as the score vector $\mathbf{f}^*$ and we set $\boldsymbol{\sigma}^*_n = \text{argsort}(\mathbf{f}^*)$.

For each of the task, the averaged result across 10 randomly subsets are reported in Fig. 6.2. As before, our proposed methods **PR-Kron** and **PR-PD** perform the best, followed by **GR**. Once again **RC** and **IPR** perform poorly[5]. Note that, the performance error increases from bipartite ranking **(BR)** to full ranking **(FR)**, former being a relatively simpler task. Results on more datasets are available in App. E.2 and E.3.

[3] See App. E.1 for better comparisons of only the first 3 methods.

[4] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

[5] We omit them for **BR** and **OR** for better comparisons.

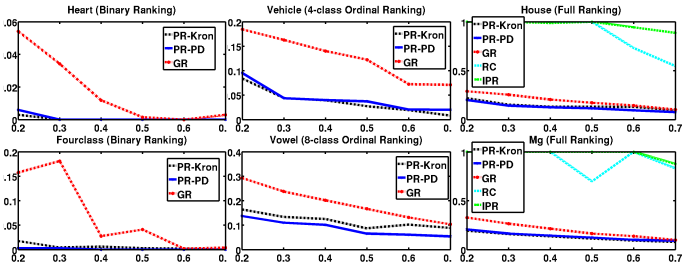

Figure 2: Real-World Data: Average number of mispredictions ($er_D^{\ell^{0\text{-}1}}(\mathbf{f})$, Eqn. (6)) vs fraction of sampled pairs (f).

7 Conclusion and Future Works

In this paper we addressed the problem of ranking nodes of a graph $G([n], E)$ given a random subsample of their pairwise preferences. Our proposed algorithm *Pref-Rank*, guarantees consistency with a required *sample complexity* of $O\big(n^2\chi(\bar{G})\big)^{\frac{2}{3}}$ – also gives novel insights by relating the ranking sample complexity with graph structural properties through chromatic number of $\bar{G}$, i.e. $\chi(\bar{G})$, for the first time. One possible future direction is to extend the setting to noisy preferences e.g. using BTL model Negahban et al. [2012], or analyse the problem with other measures of ranking losses e.g. NDCG, MAP Agarwal [2008]. Furthermore, proving consistency of *Pref-Rank* algorithm using PD-Lab(G) also remains an interesting direction to explore.

References

Alekh Agarwal and Soumen Chakrabarti. Learning random walks to rank nodes in graphs. In *Proceedings of the 24th international conference on Machine learning*, pages 9–16. ACM, 2007.

Shivani Agarwal. Transductive ranking on graphs. *Technical Report, MIT-CSAIL*, 2008.

Shivani Agarwal. Learning to rank on graphs. *Machine learning*, 81(3):333–357, 2010.

Shivani Agarwal and Partha Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10(Feb):441–474, 2009.

Nir Ailon. An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity. *Journal of Machine Learning Research*, 13(Jan): 137–164, 2012.

Rie K Ando and Tong Zhang. Learning on graph with Laplacian regularization. In *Advances in neural information processing systems*, pages 25–32, 2007.

Mark Braverman and Elchanan Mossel. Noisy sorting without resampling. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 268–276. Society for Industrial and Applied Mathematics, 2008.

Amin Coja-Oghlan. The lovász number of random graphs. *Combinatorics, Probability and Computing*, 14(04):439–465, 2005.

Gianna M Del Corso and Francesco Romani. A multi-class approach for ranking graph nodes: models and experiments with incomplete data. *Information Sciences*, 329:619–637, 2016.

Ran El-Yaniv and Dmitry Pechyony. Transductive Rademacher Complexity and its applications. In *Learning Theory*, pages 157–171. Springer, 2007.

Ran El-Yaniv and Dmitry Pechyony. Transductive Rademacher Complexity and its applications. *Journal of Artificial Intelligence Research*, 35(1):193, 2009.

Alan Frieze, Michael Krivelevich, and Cliff Smyth. On the chromatic number of random graphs with a fixed degree sequence. *Combinatorics Probability and Computing*, 16 (5):733, 2007.

Zoltán Füredi and János Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1(3):233–241, 1981.

Bo Geng, Linjun Yang, and Xian-Sheng Hua. Learning to Rank with Graph Consistency. 2009.

David F Gleich and Lek-heng Lim. Rank aggregation via nuclear norm minimization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 60–68. ACM, 2011.

Xiangnan He, Ming Gao, Min-Yen Kan, and Dingxian Wang. Birank: Towards ranking on bipartite graphs. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):57–71, 2017.

Chin-Chi Hsu, Yi-An Lai, Wen-Hao Chen, Ming-Han Feng, and Shou-De Lin. Unsupervised Ranking using Graph Structures and Node Attributes. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 771–779. ACM, 2017.

Kevin G Jamieson and Robert Nowak. Active ranking using pairwise comparisons. In *Advances in Neural Information Processing Systems*, pages 2240–2248, 2011.

Vinay Jethava, Anders Martinsson, Chiranjib Bhattacharyya, and Devdatt P Dubhashi. Lovász ϑ function, SVMs and finding dense subgraphs. *Journal of machine learning research*, 14(1):3495–3536, 2013.

Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

Ravi Kumar and Sergei Vassilvitskii. Generalized distances between rankings. In *Proceedings of the 19th international conference on World wide web*, pages 571–580. ACM, 2010.

László Lovász. On the Shannon capacity of a graph. *Information Theory, IEEE Transactions on*, 25(1):1–7, 1979.

Carlos J Luz and Alexander Schrijver. A convex quadratic characterization of the lovász theta number. *SIAM Journal on Discrete Mathematics*, 19(2):382–387, 2005.

Bernard Monjardet. On the comparison of the Spearman and Kendall metrics between linear orders. *Discrete mathematics*, 192(1-3):281–292, 1998.

Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative ranking from pair-wise comparisons. In *Advances in neural information processing systems*, pages 2474–2482, 2012.

UN Niranjan and Arun Rajkumar. Inductive Pairwise Ranking: Going Beyond the n log (n) Barrier. In *AAAI*, pages 2436–2442, 2017.

L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia, 1998. URL `citeseer.nj.nec.com/page98pagerank.html`.

Kristiaan Pelckmans, Johan AK Suykens, and BD Moor. Transductive Rademacher Complexities for Learning Over a Graph. In *MLG*, 2007.

Arun Rajkumar and Shivani Agarwal. When can we rank well from comparisons of O (n\log (n)) non-actively chosen pairs? In *Conference on Learning Theory*, pages 1376–1401, 2016.

Rakesh Shivanna and Chiranjib Bhattacharyya. Learning on graphs using Orthonormal Representation is statistically consistent. In *Advances in Neural Information Processing Systems*, pages 3635–3643, 2014.

Antonis Theodoridis, Constantine Kotropoulos, and Yannis Panagakis. Music recommendation using hypergraphs and group sparsity. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 56–60. IEEE, 2013.

Fabian Wauthier, Michael Jordan, and Nebojsa Jojic. Efficient ranking from pairwise comparisons. In *International Conference on Machine Learning*, pages 109–117, 2013.

Appendix: How Many Preference Pairs Suffice to Rank A Graph Consistently?

A Discussion of Locality Property on RKHS

By definition, any smooth function $\mathbf{f} : V \mapsto \mathbb{R}$ over a graph $G(V, E)$ implies $\mathbf{f}$ to vary slowly on the neighbouring nodes of the graph G; i.e., if $(i, j) \in E$ then $f_i \approx f_j,\ \forall i, j \in V$. The standard way of defining this is by considering $\mathbf{f}^\top \mathbf{L}\mathbf{f} = \sum_{(i,j)\in E}(f_i - f_j)^2$ to be small, say $\mathbf{f}^\top \mathbf{L}\mathbf{f} \leq B$, for some constant $B > 0$. Clearly a small value of B implies $|f_i - f_j|$ to be small for any two neighboring nodes, i.e. $(i, j) \in E$.

We first analyze the RKHS view of the above notion of smooth reward functions. Consider the SVD of the graph Laplacian $\mathbf{L} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$, where $\mathbf{Q} = [\mathbf{q}_1\ \mathbf{q}_2\ \dots\ \mathbf{q}_n] \in \mathbb{R}^{n\times n}$, $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, \dots, \lambda_n)$ and suppose the singular values $\lambda_i = 0,\ \forall i > d$, for some $d \in [N]$. Now consider the linear space of real-valued vectors–

$$\mathcal{H}(G) = \{\mathbf{g} \in \mathbb{R}^n \mid \mathbf{g}^\top \mathbf{q}_i = 0\ \ \forall i > d\}$$

Note since $\mathbf{L} \in \mathbf{S}_n^+$ is positive semi-definite, the function $\|\cdot\|_\mathbf{L} : \mathcal{H}(G) \mapsto \mathbb{R}$, such that $\|\mathbf{g}\|_\mathbf{L} = \mathbf{g}^\top \mathbf{L}\mathbf{g}$ defines a valid norm on $\mathcal{H}(G)$. In fact, one can show that $\mathcal{H}(G)$ along with the inner product $\langle\cdot,\cdot\rangle_\mathbf{L} : \mathcal{H}(G) \times \mathcal{H}(G) \mapsto \mathbb{R}$, such that $\langle \mathbf{g}_1, \mathbf{g}_2\rangle_\mathbf{L} = \mathbf{g}_1^\top \mathbf{L}\mathbf{g}_2,\ \forall \mathbf{g}_1, \mathbf{g}_2 \in \mathcal{H}(G)$, defines a valid RKHS with respect to the reproducing kernel $\mathbf{K} = \mathbf{L}^\dagger$. This can be easily verified from the fact that $\forall \mathbf{g} \in \mathcal{H}(G), \mathbf{L}^\dagger\mathbf{L}\mathbf{g} = \mathbf{g}$, and hence $\langle \mathbf{g}, \mathbf{K}_i\rangle_\mathbf{L} = \mathbf{g}^\top \mathbf{L}\mathbf{K}\mathbf{e}_i = (\mathbf{L}\mathbf{K}\mathbf{g})^\top \mathbf{e}_i = (\mathbf{L}^\dagger\mathbf{L}\mathbf{g})^\top \mathbf{e}_i = g_i,\ \forall i \in [N]$.

Thus the smoothness assumption on the reward function $\mathbf{f}$, can alternatively be interpreted as $\mathbf{f}$ being small in terms of the RKHS norm $\|\cdot\|_\mathbf{L}$. The above interpretation gives us the insight of extending the notion of "smoothness" with respect to a general RKHS norm associated to some kernel matrix $\mathbf{K} \in \mathbf{S}_n^+$. More specifically, we choose the kernel matrix $\mathbf{K}$ from the set of orthonormal kernels $\mathcal{K}(G)$ and consider $\mathbf{f}$ to be smooth in the corresponding RKHS norm. Note here the Hilbert space of functions $\mathcal{H}(\mathbf{K})$ is given by

$$\mathcal{H}(\mathbf{K}) = \{\mathbf{g} \in \mathbb{R}^n \mid \mathbf{g}^\top \mathbf{q}_i = 0\ \ \forall i > d\}, \tag{7}$$

where same as before, the SVD of $\mathbf{K} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$, $\mathbf{Q} = [\mathbf{q}_1, \dots, \mathbf{q}_n] \in \mathbb{R}^{n\times n}$ being the orthogonal eigenvector matrix of $\mathbf{K}$, $\mathbf{\Lambda} = diag(\lambda_1, \dots \lambda_n)$ be the diagonal matrix containing singular values of $\mathbf{K}$. Clearly $\lambda_i = 0,\ \forall i > d$ implies $r(\mathbf{K}) = d$. Also we define the corresponding inner product $\langle\cdot,\cdot\rangle_\mathbf{K} : \mathcal{H}(\mathbf{K}) \times \mathcal{H}(\mathbf{K}) \mapsto \mathbb{R}$, as $\langle \mathbf{g}_1, \mathbf{g}_2\rangle_\mathbf{K} = \mathbf{g}_1^\top \mathbf{K}^\dagger \mathbf{g}_2,\ \forall \mathbf{g}_1, \mathbf{g}_2 \in \mathcal{H}(\mathbf{K})$. Then similarly as above, we can show that $\mathcal{H}(\mathbf{K})$ along with $\langle\cdot,\cdot\rangle_\mathbf{K}$ defines a valid RKHS with respect to the reproducing kernel $\mathbf{K}$, as $\forall \mathbf{g} \in \mathcal{H}(\mathbf{K})$, $\langle \mathbf{g}, \mathbf{K}_i\rangle_\mathbf{K} = \mathbf{g}^\top \mathbf{K}^\dagger \mathbf{K}\mathbf{e}_i = g_i,\ \forall i \in [n]$.

The RKHS norm $\|\mathbf{g}\|_\mathbf{K} = \mathbf{g}^\top \mathbf{K}^\dagger \mathbf{g}$ defines a measure of the smoothness of $\mathbf{g}$, with respect to the kernel function $\mathbf{K}$. One way to see this is that $\forall \mathbf{g} \in \mathcal{H}(\mathbf{K})$, $\|g_i - g_j\| = \|\langle \mathbf{g}, (\mathbf{K}(i,\cdot) - \mathbf{K}(j,\cdot))\rangle\| \leq \|\mathbf{g}\|_\mathbf{K}\|\mathbf{K}(i,\cdot) - \mathbf{K}(j,\cdot)\|_\mathbf{K} = \|\mathbf{g}\|_\mathbf{K}|(K_{ii} + K_{jj} - 2K_{ij})|$, where the inequality follows from the Cauchy-Schwarz inequality of RKHS($\mathbf{K}$). Note since $\mathbf{K} \in \mathcal{K}(G)$, $K_{ii} = 1,\ \forall i \in [N]$, we have $|(K_{ii} + K_{jj} - 2K_{ij})| \leq 4\ \forall i, j \in [N]$. In particular, for two neighboring nodes i and j such that $(i, j) \in E$, it is expected that $K(i, j) \approx 1$ (i.e. $\mathbf{u}_i \approx \mathbf{u}_j$), in which case the quantity $|(K_{ii} + K_{jj} - 2K_{ij})| \approx 0$. Thus to impose a smoothness constraint on $\mathbf{g}$, it is sufficient to upper bound $\|\mathbf{g}\|_\mathbf{K} \leq B$, for some fixed $B \in \mathbb{R}, \forall \mathbf{g} \in \mathcal{H}(\mathbf{K})$.

We thus justify our assumption of $\|\mathbf{f}\|_\mathbf{K} \leq B$ which implies the ranking function (vector) $\mathbf{f}$ to be a smooth functions over the underlying graph G, with respect to embedding $\mathbf{K}$. Interestingly, $\mathcal{H}(\mathbf{K})$ incorporates $\mathcal{H}(G)$ as its special case with $\mathbf{K} = \mathbf{L}^\dagger$. Thus our space of ranking functions rightfully generalizes the Laplacian based rankings, as studied by Agarwal and Niyogi [2009], Agarwal [2010]. From the definition of $\mathcal{H}(\mathbf{K})$ in (7), it follows that the unknown ranking function $\mathbf{f} \in \mathcal{H}(\mathbf{K})$, lies in the column space of $\mathbf{K}$, i.e. $\mathbf{f} = \mathbf{K}\boldsymbol{\alpha}$, for some $\tilde{\boldsymbol{\alpha}} \in \mathbb{R}^N$. Also recall $\forall \mathbf{K} \in \mathcal{K}(G)$, there exists an $\mathbf{U} \in Lab(G)$, such that $\mathbf{K} = \mathbf{U}^\top\mathbf{U}$, $\mathbf{U} \in \mathbb{R}^{N\times N}$. Thus we have $\mathbf{f} = \mathbf{K}\boldsymbol{\alpha} = \mathbf{U}^\top\mathbf{U}\boldsymbol{\alpha} = \mathbf{U}^\top\tilde{\boldsymbol{\alpha}}$, where $\tilde{\boldsymbol{\alpha}} = \mathbf{U}\boldsymbol{\alpha}$.

Lemma 14. *If $\mathbf{f} \in$ RKHS($\mathbf{K}$), $\mathbf{f}^\top\mathbf{K}\mathbf{f} \leq B$, and we define $\tilde{\mathbf{K}} = \mathbf{K} \otimes \mathbf{K}$, $\tilde{\mathbf{f}} = \mathbf{f} \otimes \mathbf{f}$, then $\tilde{\mathbf{f}} \in$ RKHS($\tilde{\mathbf{K}}$), $\tilde{\mathbf{f}}^\top\tilde{\mathbf{K}}\tilde{\mathbf{f}} \leq B^2$.*

Proof. The proof follows from the straightforward properties of tensor products. We describe it below from completeness: Since $\mathbf{f} \in$ RKHS($\mathbf{K}$), we have $\mathbf{f} = \mathbf{K}\boldsymbol{\alpha}$ for some $\boldsymbol{\alpha} \in \mathbb{R}^n$. Now

$$\begin{aligned}\tilde{\mathbf{f}} &= \mathbf{f} \otimes \mathbf{f} = (\mathbf{K}\boldsymbol{\alpha}) \otimes (\mathbf{K}\boldsymbol{\alpha}) \\ &= (\mathbf{K} \otimes \mathbf{K}) * (\boldsymbol{\alpha} \otimes \boldsymbol{\alpha}) = \tilde{\mathbf{K}}(\boldsymbol{\alpha} \otimes \boldsymbol{\alpha}),\end{aligned}$$

and hence $\tilde{\mathbf{f}} \in$ RKHS($\tilde{\mathbf{K}}$), where the second last inequality follows due to the the properties of tensor product. Further more, using the same property, we have

$$\begin{aligned}\tilde{\mathbf{f}}^\top\mathbf{K}\tilde{\mathbf{f}} &= (\mathbf{f} \otimes \mathbf{f})^\top(\mathbf{K} \otimes \mathbf{K})(\mathbf{f} \otimes \mathbf{f}) \\ &= (\mathbf{f} \otimes \mathbf{f})^\top(\mathbf{K}\mathbf{f} \otimes \mathbf{K}\mathbf{f}) \\ &= (\mathbf{f}^\top\mathbf{K}\mathbf{f}) * (\mathbf{f}^\top\mathbf{K}\mathbf{f}) \leq B^2\end{aligned}$$

□

Definition 15. Strong Product of Graphs. *Given a graph $G = (V, E)$, strong product of G with itself, denoted by $G \boxtimes G$, is defined over the vertex set $V(G \boxtimes G) = V \times V$, such that two nodes $(i, j), (i', j') \in V(G \boxtimes G)$ is adjacent in $G \boxtimes G$ if and only if $i = i'$ and $(j, j') \in E$, or $(i, i') \in E$ and $j = j'$, or $(i, i') \in E$ and $(j, j') \in E$. Note that for every node $k \in V(G \boxtimes G)$, there exists a corresponding node pair $(i_k, j_k) \in V \times V$ in the original graph G.*

Let $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_n] \in \mathbb{R}^{d\times n}$ and $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_n] \in \mathbb{R}^{d'\times n}$ be any two orthonormal representations of G. We denote $\mathbf{u} \otimes \mathbf{v} = [u_1v_1 \dots u_1v_n\ u_2v_1 \dots u_nv_n]^\top \in \mathbb{R}^{dd'}$ to be the kronecker (or outer) product of the two vectors $\mathbf{u} \in \mathbb{R}^d, \mathbf{v} \in \mathbb{R}^{d'}$. Let $\mathbf{w}_k = \mathbf{u}_{i_k} \otimes \mathbf{v}_{j_k} \in \mathbb{R}^{dd'}$, for every node $k \in V(G \boxtimes G)$. It is easy to see that any such embedding $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots \mathbf{w}_{n^2}] \in \mathbb{R}^{dd'\times n^2}$ defines a valid orthonormal representation of $G \boxtimes G$. Using above, it can also be shown that $\vartheta(G \boxtimes G) = \vartheta^2(G)$ Lovász [1979].

B Appendix for Section 3.1

B.1 Proof of Theorem 3

Theorem 3 (**Generalization Error of** *Pref-Rank*)**.** *Given a graph $G(V, E)$, let $\tilde{\mathbf{U}} \in \mathbb{R}^{d\times N}$ be any pairwise embedding of G. For any $f \in (0, 1/2]$, let Π_f be a uniform distribution on the random draw of $m(f) = \lceil Nf \rceil$ pairs of nodes from $\mathcal{P}_n$, such that $S_{m(f)} = \{(i_k, j_k) \in \mathcal{P}_n\}_{k=1}^{m(f)} \sim \Pi_f$, with corresponding pairwise preference $\mathbf{y}_{S_{m(f)}}$. Let $\bar{S}_{m(f)} = \mathcal{P}_n \backslash S_{m(f)}$. Let $\mathcal{H}_{\tilde{\mathbf{U}}} = \{\mathbf{w} \mid \mathbf{w} = \tilde{\mathbf{U}}\boldsymbol{\beta},\ \boldsymbol{\beta} \in \mathbb{R}^N,\ \|\boldsymbol{\beta}\|_\infty \leq C,\ C > 0\}$ and $\ell^\rho : \{\pm 1\} \times \mathbb{R} \mapsto [0, B]$ be a bounded, ρ-Lipschitz loss function. For any $\delta > 0$, with probability $\geq 1 - \delta$ over $S_{m(f)} \sim \Pi_f$*

$$er^{\ell^\rho}_{\bar{S}_{m(f)}}(\mathbf{f}^*) \leq er^{\ell^\rho}_{S_{m(f)}}(\mathbf{f}^*) + \frac{R(\mathcal{H}_{\tilde{\mathbf{U}}}, \tilde{\mathbf{U}}, p)}{\rho f(1-f)} + \frac{C_1 B\sqrt{\ln\left(\frac{1}{\delta}\right)}}{(1-f)\sqrt{Nf}},$$

where $p = f(1-f)$ and $\mathbf{f}^* = \tilde{\mathbf{U}}^\top \mathbf{w}^* \in \mathbb{R}^N$ is pairwise score vector output by Pref-Rank and $C_1 > 0$ is a constant.

Proof. To proof the above result, let us first recall the error bound for learning classification models in transductive setting from El-Yaniv and Pechyony [2009].

Consider the problem of transductive binary classification over a fixed set $S_{m+u} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m+u}$ of $m + u$ points, where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the instances with their labels y_i. The learner is provided with the unlabeled (full) instance set $X_{m+u} = \{\mathbf{x}_i\}_{i=1}^{m+u}$. A set consisting of m points is selected from X_{m+u} uniformly at random among all subsets of size m. These m points together with their labels are given to the learner as a training set. Renumbering the points, suppose the unlabeled training set points are denoted by $X_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ and the labeled training set by $S_m = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$. The goal is to predict the labels of the unlabeled test points, $X_u = \{x_{m+1}, \ldots, x_{m+u}\} = X_{m+u} \setminus X_m$, given $S_m \cup X_u$.

Consider any learning algorithm generates soft classification vectors $\mathbf{h} = (h_1, \ldots, h_{m+u}) \in \mathbb{R}^{m+u}$ (or equivalently $\mathbf{h}$ can also be seen as function such that $h : \mathbf{X}_{m+u} \mapsto \mathbb{R}$). $h_i (= h(\mathbf{x}(i))) \in \mathbb{R}$ denotes the soft label for the example $\mathbf{x}_i$ given by the hypothesis $\mathbf{h}$. For actual (binary) classification of $\mathbf{x}_i$, the algorithm outputs $sgn(h_i)$. The soft classification accuracy is measured with respect to the some loss function $\ell : \{\pm 1\} \times \mathbb{R} \mapsto [0, B]$. Thus $\ell(y_i, h_i)$ denotes the loss for the i^{th} instance $\mathbf{x}_i$. We denote by ℓ^{0-1}, the 0-1 loss vector, i.e. $\ell^{0-1}(y_i, h_i) = \mathbf{1}(y_i \neq sgn(h_i))$.

Theorem 16 (Transductive test error bound (Thm. 2) El-Yaniv and Pechyony [2009])**.** *Let $\mathcal{H}_{out} \subseteq \mathbb{R}_{m+u}$ denotes the set of all possible soft classification vectors generated by the learning algorithm, upon operating on all possible training/test set partitions, the loss function ℓ^ρ is ρ-lipschitz. Then for $c = \sqrt{\frac{32\ln(4e)}{3}} < 5.05$, $Q = \left(\frac{1}{m} + \frac{1}{u}\right)$, and $S = \frac{m+u}{(m+u-1/2)(1-\frac{1}{2\max(m,u)})}$, and a fixed ρ, with probability of at least $(1-\delta)$ over the choice of the training set from X_{m+u}, for all $h \in \mathcal{H}_{out}$*

$$\frac{1}{u}\sum_{i=m+1}^{m+u} \ell^\rho(y_i, h_i) \leq \frac{1}{m}\sum_{i=1}^{m} \ell^\rho(y_i, h_i) + \frac{R(H_{out})}{\rho} + cBQ\sqrt{\min(m,u)} + B\sqrt{\frac{SQ}{2}\ln\frac{1}{\delta}}, \quad (8)$$

where $R_{m+u}(\mathcal{H}_{out}) = Q\mathbb{E}_{\boldsymbol{\gamma}}\left[\sup_{\mathbf{h}\in\mathcal{H}_{out}} \boldsymbol{\gamma}^\top \mathbf{h}\right]$ is the pairwise Rademacher complexity of the function class $\mathcal{H}_{out}$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{m+u})$ be a vector of i.i.d. random variables such that $\gamma_i \in \{\pm 1, 0\},\ i \in [m+u]$, with probability p, p and $1-2p$ respectively, with $p = \frac{mu}{(m+u)^2}$.

It is now straightforward to see that, for our current problem of interest training and test set sizes are respectively $m = Nf$ and $u = N(1-f)$. This immediately gives that $Q = \frac{1}{Nf(1-f)}$, $\min(m, u) = Nf$ and $p = f(1-f)$. The true labels of the pairwise classification problem are given by $y_k = sgn(\sigma^*(i_k) - \sigma^*(j_k)),\ \forall k \in [N]$ and the function class $\mathcal{H}_{out} = \mathcal{H}_{\tilde{\mathbf{U}}}$. Thus $R(\mathcal{H}_{\tilde{\mathbf{U}}}, \tilde{\mathbf{U}}, p) = f(1-f)R(\mathcal{H}_{out})$. Also note that for large n and $f < \frac{1}{2}$, $S = \frac{N}{(N-1/2)\left(1-\frac{1}{N}\right)} \approx 1$. Thus (8) reduces to

$$\begin{aligned} er^{\ell^\rho}_{\bar{S}}[\mathbf{f}] &= \frac{1}{u}\sum_{i=m+1}^{m+u} \ell^\rho(y_i, h_i) \\ &\leq \frac{1}{m}\sum_{i=1}^{m} \ell^\rho(y_i, h_i) + \frac{R(\mathcal{H}_{\tilde{\mathbf{U}}}, \tilde{\mathbf{U}}, p)}{\rho f(1-f)} \\ &\quad + \frac{C_1 B}{(1-f)\sqrt{Nf}}\left(1 + \sqrt{\ln\left(\frac{1}{\delta}\right)}\right), \end{aligned}$$

for $C_1 > 0$ being the appropriate constant. Thus the claim follows. □

C Appendix for Section 4

C.1 Characterization: Choice of Optimal Embedding

In this section, we discuss different classes of pairwise preference graph embeddings and the corresponding generalization guarantees. We start by recalling Thm. 1 of Ando and Zhang [2007], which provides a crucial characterization for the class of optimal embeddings:

Suppose $\mathbf{f}^*$ denotes the score function returned by the following optimization problem

$$\mathbf{f}^* = \operatorname*{argmax}_{\mathbf{f}\in\mathbb{R}^N} C'\mathbf{f}^\top \tilde{\mathbf{K}}^{-1}\mathbf{f} + \hat{er}^{\ell^\rho}_{S_m}(\mathbf{f}),$$

(note that for *Pref-Rank* (Eqn. 3), $C' = \frac{1}{2Cm}$ and $\ell^\rho = \ell^{hinge}$), then drawing a straightforward inference, we get

Corollary 17. *Suppose $\mathbf{f}^*$ denotes the optimal solution of* (3)*. Then, over the random draw of $S_m \subseteq \mathcal{P}_n$, the expected generalization error w.r.t. any ρ-Lipschitz loss function ℓ^ρ is*

given by

$$\mathbb{E}_{S_m}[er^{\ell^{0-1}}_{\bar{S}_m}(\mathbf{f}^*)] = \frac{1}{N-m}\mathbb{E}_{S_m}\left[\sum_{k=m+1}^{N}\ell^{0-1}(y_k,\mathbf{f}^*_k)\right]$$

$$\leq \inf_{\mathbf{f}\in\mathbb{R}^N}\frac{1}{c_1}\left[er^{\ell^\rho}_{S_m\cup\bar{S}_m}(\mathbf{f}) + C'\mathbf{f}^\top\tilde{\mathbf{K}}^{-1}\mathbf{f}\right] + c_2\left(\frac{tr_p(\tilde{\mathbf{K}})}{\rho m C'}\right)^p$$

where $tr_p(\tilde{\mathbf{K}}) = \left(\frac{1}{N}\sum_{k=1}^{N}\tilde{K}^p_{kk}\right)^{\frac{1}{p}}$, $er^{\ell^\rho}_{S_m\cup\bar{S}_m}[\mathbf{f}] = \frac{1}{N}\sum_{k=1}^{N}\ell^\rho(y_k,f_k)$ *and* $p,c_1,c_2 > 0$ *are fixed constants dependent on* ℓ^ρ.

Now following a similar chain of arguments as in Ando and Zhang [2007], Cor. 17 implies that *a normalized graph kernel* $\tilde{\mathbf{K}} = \tilde{\mathbf{U}}^\top\tilde{\mathbf{U}}$ *such that* $\tilde{K}_{kk} = 1, \forall k \in [N]$ *leads to improved generalization performance*, since it ensures $tr(\tilde{\mathbf{K}})_p$ to be constant. Furthermore, the following theorem shows that the class of 'normalized' graphs embeddings have high rademacher complexity.

C.2 Proof of Theorem 4

Theorem 4 (Rademacher Complexity of Orthonormal Embeddings). *Given* $G(V,E)$*, let* $\tilde{\mathbf{U}} \in \mathbb{R}^{d\times N}$ *be any 'normalized' node-pair embedding of* $G\boxtimes G$*, let* $\tilde{\mathbf{K}} = \tilde{\mathbf{U}}^\top\tilde{\mathbf{U}}$ *be the corresponding graph-kernel, then* $R(\mathcal{H}_{\tilde{\mathbf{U}}},\tilde{\mathbf{U}},p) \leq C\sqrt{2p\lambda_1(\tilde{\mathbf{K}})}$*, where* $\lambda_1(\tilde{\mathbf{K}})$ *is the largest eigenvalue of* $\tilde{\mathbf{K}}$.

Proof. Note that for any fixed realization of $\boldsymbol{\gamma} = [\gamma_1,\dots\gamma_N]$,

$$\sup_{\mathbf{h}\in\mathcal{H}_{\tilde{\mathbf{U}}}}\sum_{k=1}^{N}\gamma_k(\mathbf{h}^\top\tilde{\mathbf{u}}_k) = \sup_{\mathbf{h}\in\mathcal{H}_{\tilde{\mathbf{U}}}}\mathbf{h}^\top\left(\tilde{\mathbf{U}}\boldsymbol{\gamma}\right)$$

$$= \sup_{\boldsymbol{\beta}\in\mathbb{R}^N:\|\boldsymbol{\beta}\|_\infty\leq C}\boldsymbol{\beta}^\top\tilde{\mathbf{U}}^\top\left(\tilde{\mathbf{U}}\boldsymbol{\gamma}\right)$$

$$\leq \sup_{\boldsymbol{\beta}\in\mathbb{R}^N:\|\boldsymbol{\beta}\|_\infty\leq C}\|\tilde{\mathbf{U}}\boldsymbol{\beta}\|_2\|\tilde{\mathbf{K}}\boldsymbol{\gamma}\|_2 \ \left(\text{Cauchy-Schwarz Ineq.}\right)$$

$$\leq \sqrt{\lambda_1(\tilde{\mathbf{K}})}\sup_{\boldsymbol{\beta}\in\mathbb{R}^N:\|\boldsymbol{\beta}\|_\infty\leq C}\|\boldsymbol{\beta}\|_2\|\tilde{\mathbf{U}}\boldsymbol{\gamma}\|_2$$

$$\leq C\sqrt{N\lambda_1(\tilde{\mathbf{K}})}\|\tilde{\mathbf{U}}\boldsymbol{\gamma}\|_2$$

Using above we further get:

$$R(\mathcal{H}_{\tilde{\mathbf{U}}},\tilde{\mathbf{U}},p) \leq \frac{1}{N}\mathbb{E}_{\boldsymbol{\gamma}}\left[C\sqrt{N\lambda_1(\tilde{\mathbf{K}})}\|U\boldsymbol{\gamma}\|_2\right]$$

$$= \frac{C\sqrt{\lambda_1(\tilde{\mathbf{K}})}}{\sqrt{N}}\mathbb{E}_{\boldsymbol{\gamma}}\left[\sqrt{\boldsymbol{\gamma}^\top\tilde{\mathbf{K}}\boldsymbol{\gamma}}\right]$$

$$\leq \frac{C\sqrt{\lambda_1(\tilde{\mathbf{K}})}}{\sqrt{N}}\sqrt{\mathbb{E}_{\boldsymbol{\gamma}}\left[\boldsymbol{\gamma}^\top\tilde{\mathbf{K}}\boldsymbol{\gamma}\right]} \ \left(\text{Jensen's Inequality}\right)$$

$$= \frac{C\sqrt{\lambda_1(\tilde{\mathbf{K}})}}{\sqrt{N}}\sqrt{2p(tr(\tilde{\mathbf{K}}))} = \frac{C}{\sqrt{N}}\sqrt{2p\lambda_1(\tilde{\mathbf{K}})(tr(\tilde{\mathbf{K}}))},$$

where the second last equality follows from the fact that $\mathbb{E}_{\boldsymbol{\gamma}}\left[\boldsymbol{\gamma}^\top\tilde{\mathbf{K}}\boldsymbol{\gamma}\right] = 2p\sum_{k=1}^{N}\tilde{K}_{kk} = 2p(tr(\tilde{\mathbf{K}}))$, as $\boldsymbol{\gamma} = (\gamma_1,\dots,\gamma_N)$ be a vector of *i.i.d.* random variables such that $\gamma_i \in \{+1,-1,0\}$, $i \in [N]$, with probability p, p and $1-2p$ respectively and $tr(\tilde{\mathbf{K}}) = \sum_{k=1}^{N}\tilde{K}_{kk}$. The proof now follows from the fact that $tr(\tilde{\mathbf{K}}) = N$, since $\tilde{K}_{kk} = 1$, $\forall k \in [N]$.

□

C.3 Proof of Lemma 5

Lemma 5 (Rademacher Complexity of Kron-Lab($G\boxtimes G$)). *Consider any* $\mathbf{U} \in$ *Lab(G),* $\mathbf{K} = \mathbf{U}^\top\mathbf{U}$ *and the corresponding* $\tilde{\mathbf{U}} \in$ *Kron-Lab($G\boxtimes G$). Then for any* $p \in [0,1]$ *and* $\mathcal{H}_{\tilde{\mathbf{U}}} = \{\mathbf{w} \mid \mathbf{w} = \tilde{\mathbf{U}}\boldsymbol{\beta},\ \boldsymbol{\beta}\in R^N,\ \|\boldsymbol{\beta}\|_\infty \leq C, C>0\}$ *we have,* $R(\mathcal{H}_{\tilde{\mathbf{U}}},\tilde{\mathbf{U}},p) \leq C\lambda_1(\mathbf{K})\sqrt{2p}$.

Proof. To show this, we first proof the following lemmas.

Lemma 18. *Let* $\tilde{\mathbf{U}}_P = [\tilde{\mathbf{u}}_{ij}]_{(i,j)\in\mathcal{P}_n} \in \mathbb{R}^{d^2\times N}$ *be the embedding matrix only for the node-pairs in* $\mathcal{P}_n$. $\tilde{\mathbf{K}}_P = \tilde{\mathbf{U}}_P^\top\tilde{\mathbf{U}}_P$, $\tilde{\mathbf{K}} = \tilde{\mathbf{U}}^\top\tilde{\mathbf{U}}$. *Then* $\lambda_1(\tilde{\mathbf{K}}_P) \leq \lambda_1(\tilde{\mathbf{K}})$.

Proof. We have that $\lambda_1(\tilde{\mathbf{K}}_P) = \sup_{\mathbf{x}\in\mathbb{R}^N}\frac{\mathbf{x}^\top\tilde{\mathbf{K}}_P\mathbf{x}}{\|\mathbf{x}\|_2^2}$. Let $\mathbf{x}_1 = \operatorname{argsup}_{\mathbf{x}\in\mathbb{R}^N}\frac{\mathbf{x}^\top\tilde{\mathbf{K}}_P\mathbf{x}}{\|\mathbf{x}\|_2^2}$.

Note that $\tilde{\mathbf{U}} = [\tilde{\mathbf{U}}_1,\tilde{\mathbf{U}}_2,\dots,\tilde{\mathbf{U}}_{n^2}] \in \mathbb{R}^{d^2\times n^2}$, and $\tilde{\mathbf{U}}_P = [\tilde{\mathbf{U}}_{P_1},\tilde{\mathbf{U}}_{P_2},\dots,\tilde{\mathbf{U}}_{P_N}] \in \mathbb{R}^{d^2\times N}$, where $\tilde{\mathbf{u}}_k = \mathbf{u}_{i_k}\circ\mathbf{u}_{j_k}$, $\forall(i_k,j_k) \in [n]\times[n]$ and $(\tilde{\mathbf{u}}_P)_k = \mathbf{u}_{i_k}\circ\mathbf{u}_{j_k}$, $\forall(i_k,j_k)\in\mathcal{P}_n$.

Let us define $k'(i,j) = n(i-1)+j$, $\forall(i,j)\in[n]\times[n]$ and $k(i,j) = \sum_{l=1}^{i-1}(n-l)+(j-i)$, $\forall(i,j)\in\mathcal{P}_n$.

Clearly $\tilde{\mathbf{K}}(k'(i,j),k'(u,v)) = \tilde{\mathbf{K}}_P(k(i,j),k(u,v))$, $\forall(i,j),(u,v)\in\mathcal{P}_n$ such that Now let us consider $\tilde{\mathbf{x}}_1 \in \mathbb{R}^{n^2}$ such that

$$\tilde{\mathbf{x}}_1(k'(i,j)) = \begin{cases}\mathbf{x}_1(k(i,j)),\ \forall(i,j)\in\mathcal{P}_n,\\ 0,\ \text{otherwise}\end{cases}$$

Note that this implies $\lambda_1(\tilde{\mathbf{K}}) = \sup_{\tilde{\mathbf{x}}\in\mathbb{R}^{n^2}}\frac{\tilde{\mathbf{x}}^\top\tilde{\mathbf{K}}\tilde{\mathbf{x}}}{\|\tilde{\mathbf{x}}\|_2^2} \geq \frac{\tilde{\mathbf{x}}_1^\top\tilde{\mathbf{K}}\tilde{\mathbf{x}}_1}{\|\tilde{\mathbf{x}}_1\|_2^2} = \frac{\mathbf{x}_1^\top\tilde{\mathbf{K}}_P\mathbf{x}_1}{\|\mathbf{x}_1\|_2^2} = \lambda_1(\tilde{\mathbf{K}}_P)$, proving the claim. □

Lemma 19. *Let* $\tilde{\mathbf{K}} = \tilde{\mathbf{U}}^\top\tilde{\mathbf{U}}$, $\mathbf{K} = \mathbf{U}^\top\mathbf{U}$, *for any* $\tilde{\mathbf{U}} \in$ *SP-Lab(G), and the corresponding* $\mathbf{U} \in$ *Lab(G). Then* $\lambda_1(\tilde{\mathbf{K}}) = (\lambda_1(\mathbf{K}))^2$.

Proof. Note that $\lambda_1(\mathbf{K}) = \sup_{\mathbf{x}\in\mathbb{R}^n}\frac{\mathbf{x}^\top\mathbf{K}\mathbf{x}}{\|\mathbf{x}\|_2^2}$. Let $\mathbf{x}_1 = \operatorname{argsup}_{\mathbf{x}\in\mathbb{R}^n}\frac{\mathbf{x}^\top\mathbf{K}\mathbf{x}}{\|\mathbf{x}\|_2^2}$.

The crucial observation is that

$$\tilde{\mathbf{K}} = \mathbf{K}\circ\mathbf{K} = \begin{bmatrix}K_{11}\mathbf{K} & \cdots & K_{1n}\mathbf{K}\\ \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot\\ K_{n1}\mathbf{K} & \cdots & K_{nn}\mathbf{K}\end{bmatrix}.$$

Let us define $\tilde{\mathbf{x}}_1 = \mathbf{x}_1 \circ \mathbf{x}_1 \in \mathbb{R}^{n^2}$. Note that $\|\tilde{\mathbf{x}}\|_2 = \|\mathbf{x}\|_2^2$. Then $\lambda_1(\tilde{\mathbf{K}}) = \sup_{\tilde{\mathbf{x}} \in \mathbb{R}^{n^2}} \frac{\tilde{\mathbf{x}}^\top \tilde{\mathbf{K}} \tilde{\mathbf{x}}}{\|\tilde{\mathbf{x}}\|_2^2} = \frac{\tilde{\mathbf{x}}_1^\top \tilde{\mathbf{K}} \tilde{\mathbf{x}}_1}{\|\tilde{\mathbf{x}}_1\|_2^2} = \frac{\left(\mathbf{x}_1^\top \mathbf{K} \mathbf{x}_1\right)^2}{\|\mathbf{x}_1\|_2^4} = (\lambda_1(\mathbf{K}))^2$. □

Thus applying Lem. 18 and 19, we get, $\lambda_1(\tilde{\mathbf{K}}_P) \leq \lambda_1(\tilde{\mathbf{K}}) = (\lambda_1(\mathbf{K}))^2$. The proof of Lem. 5 now follows from Thm. 4.

□

C.4 Proof of Theorem 6

Theorem 6 (Generalization Error of *Pref-Rank* with Kron-Lab($G \boxtimes G$)). *For the setting as in Thm. 3 and Lem. 5, for any $\tilde{\mathbf{U}} \in$ Kron-Lab($G \boxtimes G$), we have*

$$er_{\tilde{S}}^{\ell^\rho}[\mathbf{f}^*] \leq er_S^{\ell^\rho}[\mathbf{f}^*] + \frac{C\lambda_1(\mathbf{K})\sqrt{2}}{\rho\sqrt{f(1-f)}} + \frac{C_1 B}{1-f}\sqrt{\frac{\log(\frac{1}{\delta})}{Nf}}$$

The proof follows by applying Lem. 5 to Thm. 3 for $p = f(1-f)$.

C.5 Proof of Lemma 7

Lemma 7 (Rademacher Complexity of PD-Lab(G)). *Consider any $\mathbf{U} \in$ Lab(G), $\mathbf{K} = \mathbf{U}^\top\mathbf{U}$ and the corresponding $\tilde{\mathbf{U}} \in$ PD-Lab(G). Then for any $p \in [0,1]$ and $\mathcal{H}_{\tilde{\mathbf{U}}} = \{\mathbf{w} \mid \mathbf{w} = \tilde{\mathbf{U}}\boldsymbol{\beta},\ \boldsymbol{\beta} \in R^N,\ \|\boldsymbol{\beta}\|_2 \leq tC\sqrt{N},\ C > 0\}$, we have $R(\mathcal{H}_{\tilde{\mathbf{U}}}, \tilde{\mathbf{U}}, p) \leq 2C\sqrt{pn\lambda_1(\mathbf{K})}$.*

Proof. Let $\mathbb{E} = [\mathbf{e}_i - \mathbf{e}_j]_{(i,j)\in\mathcal{P}_n} \in \{0, \pm 1\}^{n \times N}$, where $\mathbf{e}_i$ denotes the i^{th} standard basis of $\mathbb{R}^n$, $\forall i \in [n]$. We start by proving the following lemma:

Lemma 20. *If $\mathbf{U} \in$ Lab(G), $\mathbf{K} = \mathbf{U}^\top\mathbf{U}$, $\tilde{\mathbf{U}} = \mathbf{U}\mathbb{E} \in$ PD-Lab(G) and $\tilde{\mathbf{K}} = \tilde{\mathbf{U}}^\top\tilde{\mathbf{U}}$, then $\lambda_1(\tilde{\mathbf{K}}) = 2n\lambda_1(\mathbf{K})$.*

Proof. By definition of $\lambda_1(\tilde{\mathbf{K}})$, we know that

$$\begin{aligned}
\lambda_1(\tilde{\mathbf{K}}) &= \sup_{\mathbf{x}\in\mathbb{R}^n} \frac{\mathbf{x}^\top\tilde{\mathbf{K}}\mathbf{x}}{\|\mathbf{x}\|_2^2} \\
&= \sup_{\mathbf{x}\in\mathbb{R}^N} \frac{\mathbf{x}^\top\mathbb{E}^\top\mathbf{K}\mathbb{E}\mathbf{x}}{\|\mathbf{x}\|_2^2} \\
&= \sup_{\mathbf{x}\in\mathbb{R}^N} \frac{(\mathbb{E}\mathbf{x})^\top\mathbf{K}(\mathbb{E}\mathbf{x})}{\|\mathbf{x}\|_2^2} \\
&= \sup_{\mathbf{x}\in\mathbb{R}^N} \frac{\lambda_{\mathbf{K}}\|\mathbb{E}\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} \\
&\leq 2n\lambda_{\mathbf{K}},
\end{aligned}$$

where the last inequality follows from the fact that, for any $\mathbf{x} \in \mathbb{R}^N$, $\|\mathbb{E}\mathbf{x}\|_2^2 \leq 2n\|\mathbf{x}\|_2^2$. □

Now further applying Thm. 4 for $\tilde{\mathbf{U}} \in$ PD-Lab(G), we get $R(\mathcal{H}_{\tilde{\mathbf{U}}}, \tilde{\mathbf{U}}, p) \leq C\sqrt{2p\lambda_1(\tilde{\mathbf{K}})} \leq C\sqrt{2p\lambda_1(\tilde{\mathbf{K}})}$, since $tr(\tilde{\mathbf{K}}) \leq N$ and the result now follows from Lem. 20. □

C.6 Proof of Theorem 8

Theorem 8 (Generalization Error of *Pref-Rank* with PD-Lab(G)). *For the setting as in Thm. 3 and Lem. 7, for any $\tilde{\mathbf{U}} \in$ PD-Lab(G), we have*

$$er_{\tilde{S}}^{\ell^\rho}[\mathbf{f}^*] \leq er_S^{\ell^\rho}[\mathbf{f}^*] + \frac{2C\sqrt{n\lambda_1(\mathbf{K})}}{\rho\sqrt{f(1-f)}} + \frac{C_1 B}{1-f}\sqrt{\frac{\log(\frac{1}{\delta})}{Nf}}$$

Proof. The proof follows by applying Lem. 7 to Thm. 3 for $p = f(1-f)$. □

C.7 Proof of Lemma 9

Lemma 9. *Let $G(n,q)$ be a Erdós-Réyni random graph, where each edge is present independently with probability $q \in [0,1],\ q = O(1)$. Then the Rademacher complexity of function class associated with $\tilde{\mathbf{K}}_{LS}(G \boxtimes G)$ is $O(\sqrt{n})$.*

Proof. For $G(n,q)$ graphs, Füredi and Komlós [1981] showed that with high probability $1 - e^{-\sqrt{n}}$, $\lambda_1(\mathbf{A}_G) = nq(1+o(1))$ and $|\lambda_n(\mathbf{A}_G)| \leq 2\sqrt{nq(1-q)}$. As $q = O(1)$, note that $\lambda_1(\mathbf{A}_G) = \Theta(n)$ and $\lambda_n(\mathbf{A}_G) = \Theta(\sqrt{n})$. Thus, choosing $\tau = \Theta(\sqrt{n})$ makes $\mathbf{K}_{LS}(G)$ a positive semi-definite matrix, and clearly $\lambda_1(\mathbf{K}_{LS}(G)) = \Theta(\sqrt{n})$. Moreover since $\tilde{\mathbf{K}}_{LS} = \mathbf{K}_{LS} \otimes \mathbf{K}_{LS}$, we have $\lambda_1(\tilde{\mathbf{K}}_{LS}) = \left(\lambda_1(\mathbf{K}_{LS})\right)^2$, as follows from Lem. 19). The claim now follows from Thm. 4 and Lem. 18.

□

C.8 Embedding with graph Laplacian.

The popular choice of graph kernel uses the inverse of the Laplacian matrix. Formally, let d_i denotes the degree of vertex $i \in [n]$ in graph G, $d_i = (\mathbf{A}_G)_i^\top \mathbf{1}_n$, and $\mathbf{D}$ denote a diagonal matrix such that $D_{ii} = d_i, \forall i \in [n]$. Then the Laplacian and normalized Laplacian kernel of G is defined as follows:[6]

$$\begin{aligned}
&\mathbf{K}_{Lap}(G) = (\mathbf{D} - \mathbf{A}_G)^\dagger \text{ and} \\
&\mathbf{K}_{nLap}(G) = (\mathbf{I}_n - \mathbf{D}^{-1/2}\mathbf{A}_G\mathbf{D}^{-1/2})^\dagger.
\end{aligned}$$

Simlar to *LS*-labelling, one could define the embedding of $G \boxtimes G$ using Kron-Lab($G \boxtimes G$) or PD-Lab(G) with $\mathbf{K}_{Lap}$ and $\mathbf{K}_{nLap}$. However, we observe that the Rademahcer complexity of function associated with Laplacian is an order magnitude smaller than that of *LS*-labelling for graphs with high connectivity – we summarize our findings in Table 2. Experimental results in Section 6 illustrate our observation.

Graph	Laplacian	*LS*-labelling
Complete graph K_n	$\Theta(1)$	$\Theta(n)$
Random Graphs $G(n, 1/2)$	$\Theta(1)$	$\Theta(\sqrt{n})$
Complete Bipartite	$\Theta(1)$	$\Theta(1)$
Star S_n	$\Theta(1)$	$\Theta(1)$

Table 2: Rademacher complexity measure of Laplacian and *LS*-labelling graph embeddings (assuming $C, p = O(1)$).

[6] † denotes the pseudo inverse.

D Appendix for Section 5

D.1 Proof of Theorem 10

Theorem 10 (Rank-Consistency). *For the setting as in Sec. 2.2, there exists an embedding $\tilde{\mathbf{U}}_n \in$ Kron-Lab($G_n \boxtimes G_n$) such that if $\boldsymbol{\sigma}_n \in \mathbb{R}^N$ denotes the pairwise scores returned by Pref-Rank on input $(\tilde{\mathbf{U}}_n, S_m(f), \mathbf{y}_{S_{m(f)}})$, then $\forall G_n \in \mathcal{G}$, with probability at least $\left(1 - \frac{1}{N}\right)$ over $S_{m(f)} \sim \Pi_f$*

$$d(\boldsymbol{\sigma}_n^*, \hat{\boldsymbol{\sigma}}_n) = O\left(\left(\frac{\vartheta(G_n)}{nf}\sqrt{\frac{1-f}{f}}\right)^{\frac{1}{2}} + \sqrt{\frac{\ln n}{Nf}}\right),$$

where d denotes Kendall's tau (d_k) or Spearman's footrule (d_s) ranking loss functions.

Proof. We first bound the total number of pairwise mispredictions of f, given by $Ner_n^{\ell^{0-1}}[\mathbf{f}]$. Note that $Ner_n^{\ell^{0-1}}[\mathbf{f}] \leq Ner_n^{\ell^{ramp}}[\mathbf{f}]$. (see Sec. 2.1 for definitions of $\ell_n^{0-1}[\mathbf{f}]$ and $\ell_n^{ramp}[\mathbf{f}]$).

Now applying Cor. 6 for ramp loss ℓ^{ramp} with $\delta = \frac{1}{n}$, we get that with probability atleast $(1 - \frac{1}{N})$,

$$\begin{aligned}
Ner_n^{\ell^{ramp}}[\mathbf{f}] &= \sum_{k=1}^{N} \ell^{ramp}(y_k, f_k) \\
&= \left(\sum_{k=Nf+1}^{N} \ell^{ramp}(y_k, f_k) + \sum_{k=1}^{Nf} \ell^{ramp}(y_k, f_k)\right) \\
&= \left((N - Nf)er_{\bar{S}}^{l^{ramp}}[\mathbf{f}] + Nfer_S^{l^{ramp}}[\mathbf{f}]\right) \\
&\leq \Bigg((N(1-f) + Nf)er_S^{l^{ramp}}[\mathbf{f}] \\
&\qquad + \frac{N\sqrt{2(1-f)}C\lambda_1(\mathbf{K})}{\rho\sqrt{f}} + \frac{C_1 NB\sqrt{\ln N}}{\sqrt{Nf}}\Bigg) \\
&\qquad\qquad (\text{ from Thm. 6}) \\
&= N\left(er_S^{\ell^{ramp}}[\mathbf{f}] + \frac{C\lambda_1(\mathbf{K})\sqrt{2(1-f)}}{\sqrt{f}} + \frac{C_1\sqrt{2\ln n}}{\sqrt{Nf}}\right), \\
&\leq N\left(er_S^{\ell^{hinge}}[\mathbf{f}] + \frac{C\lambda_1(\mathbf{K})\sqrt{2(1-f)}}{\sqrt{f}} + \frac{C_1\sqrt{2\ln n}}{\sqrt{Nf}}\right)
\end{aligned}$$

where the second last inequality is because for ramp loss, ρ and B both are 1. The last equality follows from the fact that hinge loss is an upper bound of the ramp loss.

Let us define $\tilde{\mathbf{U}}_P = [\tilde{\mathbf{u}}_{ij}]_{(i,j)\in\mathcal{P}_n} \in \mathbb{R}^{d^2 \times N}$ to be the embedding matrix only for the node-pairs in $\mathcal{P}_n$. $\tilde{\mathbf{K}}_P = \tilde{\mathbf{U}}_P^\top \tilde{\mathbf{U}}_P$. Also let us define

$$\text{PSP-Lab}(G \boxtimes G) = \{\tilde{\mathbf{U}}_P \in \mathbb{R}^{d^2 \times N} \mid \mathbf{U} \in \text{Lab}(G)\}.$$

The key of the proof lies in the following derivation that maps $\vartheta(G)$ to the training set error $er_S^{\ell^{hinge}}[\mathbf{f}]$. Specifically, note that:

$$\begin{aligned}
2C(Nf)er_S^{\ell^{hinge}}[\mathbf{f}] &= 2C\sum_{i=1}^{Nf} \ell^{hinge}(y_k, f_k) \\
&\leq \min_{\tilde{\mathbf{U}}\in\text{PSP-Lab}(G\boxtimes G),\|\mathbf{c}\|_2=1} \max_{k=1}^{N} \frac{1}{(c^\top \tilde{\mathbf{U}}_k)^2} \\
&\leq \min_{\tilde{\mathbf{U}}\in\text{Kron-Lab}(G\boxtimes G),\|\mathbf{c}\|_2=1} \max_{k=1}^{n^2} \frac{1}{(c^\top \tilde{\mathbf{U}}_k)^2} \\
&= \min_{\tilde{\mathbf{U}}\in\text{Lab}(G\boxtimes G),\|\mathbf{c}\|_2=1} \max_{k=1}^{n^2} \frac{1}{(c^\top \tilde{\mathbf{U}}_k)^2} \\
&= \vartheta(G \boxtimes G) \\
&= (\vartheta(G))^2, \qquad (9)
\end{aligned}$$

where the first inequality follows from a similar derivation as given in Thm. 5.2 of Shivanna and Bhattacharyya [2014] which relates optimum SVM objective to *Lóvasz-ϑ*. The second inequality is obvious as PSP-Lab($G \boxtimes G$) $\subset$ Kron-Lab($G \boxtimes G$). Thus we get that $er_S^{\ell^{hinge}}[f] = \frac{(\vartheta(G))^2}{2CNf}$. Combining everything we now have:

$$\begin{aligned}
er_n^{\ell^{0-1}}[\mathbf{f}] &\leq er_n^{\ell^{ramp}}[f] = \frac{1}{N}\sum_{k=1}^{N} \ell^{ramp}(y_k, f_k) \\
&\leq \left(\frac{\vartheta(G_n)^2}{2CNf} + \frac{C\lambda_1(\mathbf{K})\sqrt{2(1-f)}}{\sqrt{f}} + \frac{C_1\sqrt{2\ln n}}{\sqrt{Nf}}\right) \\
&\leq \left(\frac{\vartheta(G_n)^2}{2CNf} + \frac{Cn\sqrt{2(1-f)}}{\vartheta(G)\sqrt{f}} + \frac{C_1\sqrt{2\ln n}}{\sqrt{Nf}}\right)
\end{aligned}$$

Where the last inequality follows from $\lambda_1(\mathbf{K}) \leq \vartheta(\bar{G}_n)$, $\vartheta(G_n)\vartheta(\bar{G}_n) = n$ Lovász [1979]. Further optimizing over C we get that at $C^* = \left(\frac{\vartheta(G_n)^3}{Nn\sqrt{8f(1-f)}}\right)^{\frac{1}{2}}$, using which we get

$$\begin{aligned}
er_n^{\ell^{0-1}}[\mathbf{f}] &= \frac{1}{N}\sum_{k=1}^{N} \ell^{0-1}(y_k, h_k) \\
&\leq \left(2\left(\frac{\vartheta(G_n)}{(n-1)f}\sqrt{\frac{2(1-f)}{f}}\right)^{\frac{1}{2}} + \frac{C_1\sqrt{2\ln n}}{\sqrt{Nf}}\right)
\end{aligned}$$

Above proves the first half of the result. The second result immediately follows from above with the additional observation that $d_k(\sigma^*, \hat{\sigma}) \leq 3er_n^{\ell^{0-1}}[\mathbf{f}]$ and and the fact that

$$d_s(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \leq 2d_k(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \qquad (10)$$

where d being the Kendall's tau (d_k) or Spearman's footrule (d_s) ranking loss, which concludes the proof. □

Proof of Lemma 11

Lemma 11. *If $\mathcal{G}$ in Thm. 10 is such that $\vartheta(G_n) = n^c$, $0 \leq c < 1$. Then observing only $f^* = O\left(\frac{\sqrt{\vartheta(G_n)}}{n^{\frac{1}{2}-\varepsilon}}\right)^{\frac{4}{3}}$ fraction of pairwise node preferences is sufficient for Pref-Rank to be statistically rank consistent, for any $0 < \varepsilon < \frac{(1-c)}{2}$.*

Proof. From Thm. 10 we have that there exists a constant positive $C_0 > 0$ and an positive integer $n_0 \in \mathbb{N}$ such that $\forall n \geq n_0$

$$d(\boldsymbol{\sigma}_n^*, \hat{\boldsymbol{\sigma}}_n) \leq C_0 \left(\left(\frac{\vartheta(G_n)}{nf} \sqrt{\frac{(1-f)}{f}} \right)^{\frac{1}{2}} + \frac{C_1 \sqrt{2 \ln n}}{\sqrt{Nf}} \right)$$
$$\leq C_0 \left(\frac{1}{f^{\frac{3}{4}}} \sqrt{\left(\frac{\vartheta(G_n)}{n} \right)} + \frac{C_1 \sqrt{2 \ln n}}{\sqrt{Nf}} \right) \quad (11)$$

Now that if $\vartheta(G_n) = o(n^c)$ for some $c \in [0, 1)$ (recall $\vartheta(G_n) \in [1, n]$) and if we choose $\varepsilon \leq \frac{(1-c)}{2}$ this makes $f^* = \left(\frac{\sqrt{\vartheta(G_n)}}{n^{\frac{1}{2}-\varepsilon}} \right)^{\frac{4}{3}}$ to be a valid assignment as that ensures $f^* \in [0, 1]$. Furthermore, (11) suggests that observing only f^* fraction of nodes would suffice to achieve ranking consistency since that implies $d(\boldsymbol{\sigma}_n^*, \hat{\boldsymbol{\sigma}}_n) = O(\frac{1}{n^\varepsilon}) \to 0$, as $n \to \infty$. □

Proof of Theorem 12

Theorem 12. *Consider a graph family $\mathcal{G}$ such that $\chi(\bar{G}_n) = o(n)$, $\forall G_n \in \mathcal{G}$. Then observing $O(n^2 \chi(\bar{G}))^{\frac{2}{3}}$ pairwise preferences is sufficient for Pref-Rank to be consistent.*

Proof. From Lemma 11, using f^* fraction of nodes immediately leads to the sample complexity:

$$Nf^* \leq \frac{n^2}{2} \left(\frac{\vartheta(G_n)}{n^{1-2\varepsilon}} \right)^{\frac{2}{3}} = \frac{1}{2} (n^{2+2\varepsilon} \vartheta(G_n))^{\frac{2}{3}}$$

The result now follows from Lem. 11 and Lovász sandwich theorem: $\vartheta(G) \leq \chi(\bar{G})$ for any graph G Lovász [1979]. □

Proof of Corollary 13

Corollary 13 (Ranking Consistency on Special Graphs). *Pref-Rank algorithm achieves consistency on the following graph families, with the required sample complexities* – (a) Complete graphs*:* $O(n^{\frac{4}{3}})$ (b) Union of k disjoint cliques*:* $O(n^{\frac{4}{3}} k^{\frac{2}{3}})$ (c) Complement of power-law graphs*:* $O(n^{\frac{5}{3}})$ (d) Complement of k-colorable graphs*:* $O(n^{\frac{4}{3}} k^{\frac{2}{3}})$ (e) Erdős Réyni random $G(n, q)$ graphs *with* $q = O(1)$*:* $O(n^{\frac{5}{3}})$.

Proof. The result follows from the proof of Theorem 12 upon by substituting the values of $\vartheta(G)$ or $\chi(\bar{G})$ (note $\vartheta(G) \leq \chi(\bar{G})$) in the corresponding graphs as given below:

1. *Complete graphs*: $\chi(\bar{G}) = 1$
2. *Union of k disjoint cliques*: $\chi(\bar{G}) = k$
3. *Complement of Power-law graphs*: $\vartheta(\bar{G}) = \Theta(\sqrt{n})$ Shivanna and Bhattacharyya [2014], Frieze et al. [2007]
4. *Random graphs*: $\vartheta(G) = \Theta(\sqrt{n})$, with high probability Coja-Oghlan [2005].
5. *Complement of k-colorable graphs*: $\chi(\bar{G}) = k$.

□

E Additional Experiments

E.1 Additional Results: Experiments of Synthetic Datasets

Plots comparing only PR-Kron, PR-PD, and GR

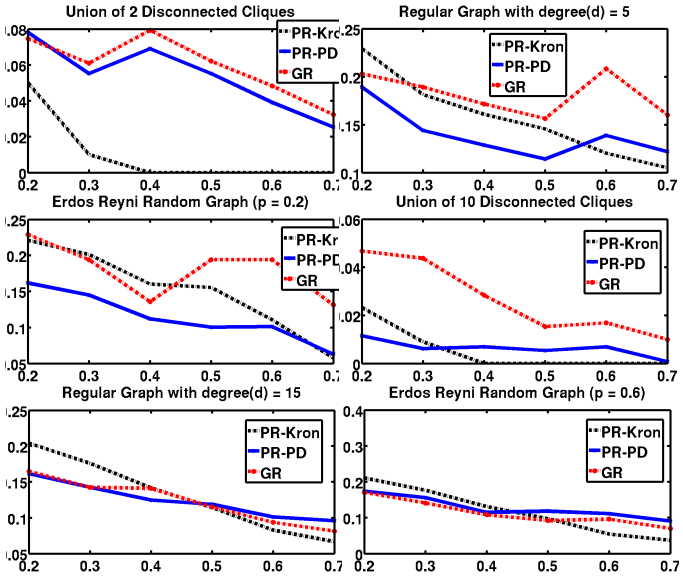

Figure 3: Synthetic Data: Average number of misprediction ($er_D^{\ell^{0\text{-}1}}(\mathbf{f})$, Eqn. 6) vs fraction of sampled pairs(f)

More Synthetic Experiments

We consider a $G(n, p, q)$ random graph with $n = 100$ nodes, $p = 0.6, q = 0.1$, where nodes [1-50] and [51-100] are densely clustered, and nodes within the same cluster are connected with edge probability p and that of two different clusters are connected with probability q. We also consider the nodes within same cluster to be closer in terms of their preference scores. More specifically, for the task of full ranking, we randomly assign a permutation to the 100 nodes such that all nodes in cluster 1 are ranked above all nodes in cluster 2 (below 50 and all nodes (51-100) are ranked above 50). Similarly for ordinal ranking we randomly assign a rating from $1 - 10$ to each graph node such that all nodes in cluster 1 are rated higher than that of cluster 2. Finally for Bipartite ranking, we randomly assign a (0,1) binary label to each node such that nodes in cluster one are 80% more likely to score higher than that of cluster 2. For each of the three tasks, we repeat the experiment for 10 times and compare the averaged performances of **PR-Kron** with **GR**. Table 3 shows that on an average *Pref-Rank* with Kron-Lab($G \boxtimes G$) performs better than *Graph Rank* for all three tasks.

Task	PR-Kron (in %)	GR (in %)
BR	**07.5**	08.2
OR(10)	**12.3**	17.6
FR	**11.8**	18.6

Table 3: Synthetic data: Average number of mispredictions.

E.2 Additonal Results: Experiments of Real Datasets

Plots comparing only PR-Kron, PR-PD, and GR

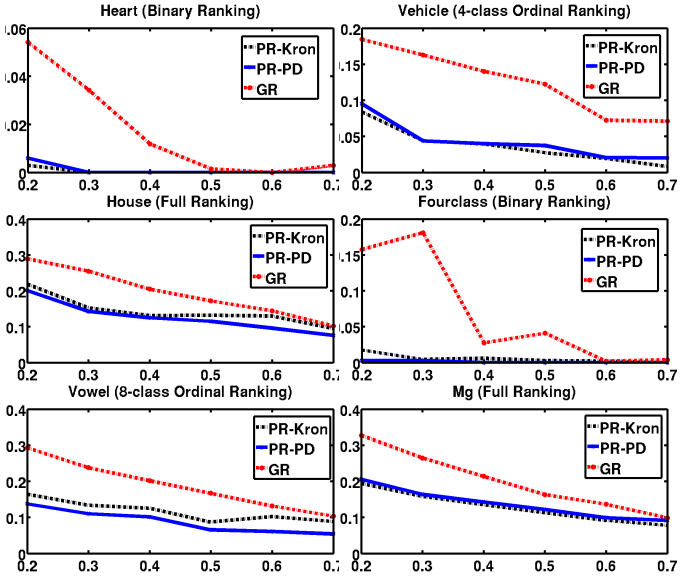

Figure 4: Real Data: Average number of misprediction ($er_D^{\ell^{0\text{-}1}}(\mathbf{f})$, Eqn. 6) vs fraction of sampled pairs(f)

E.3 More Experiments on Real Datasets

Datasets. a. *Ionosphere* and *Diabetes* for **BR** b. *Bodyfat* for **FR**.

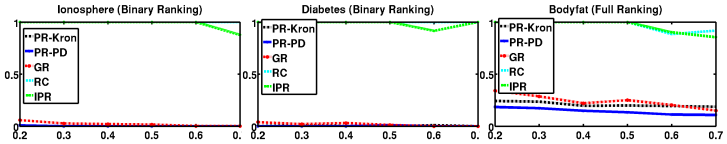

Figure 5: Real Data: Average number of misprediction ($er_D^{\ell^{0\text{-}1}}(\mathbf{f})$, Eqn. 6) vs fraction of sampled pairs(f)

Plots comparing only PR-Kron, PR-PD, and GR

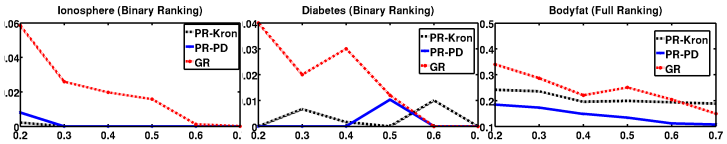

Figure 6: Real Data: Average number of misprediction ($er_D^{\ell^{0\text{-}1}}(\mathbf{f})$, Eqn. 6) vs fraction of sampled pairs(f)