Appl. Comput. Harmon. Anal. 41 (2016) 540–560

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Analysis of crowdsourced sampling strategies for HodgeRank with sparse random graphs

Braxton Osting [a,*], Jiechao Xiong [c], Qianqian Xu [b], Yuan Yao [c,*]

[a] *Department of Mathematics, University of Utah, Salt Lake City, UT 84112, USA*
[b] *Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093 & BICMR, Peking University, Beijing 100871, China*
[c] *School of Mathematical Sciences, BICMR-LMAM-LMEQF-LMP, Peking University, Beijing 100871, China*

ARTICLE INFO

Article history:
Received 28 February 2015
Received in revised form 9 March 2016
Accepted 16 March 2016
Available online 25 March 2016
Communicated by Spec. Issue Guest Editor

Keywords:
Crowdsourcing
Paired comparison
Algebraic connectivity
Erdös–Rényi random graph

ABSTRACT

Crowdsourcing platforms are now extensively used for conducting subjective pairwise comparison studies. In this setting, a pairwise comparison dataset is typically gathered via random sampling, either *with* or *without* replacement. In this paper, we use tools from random graph theory to analyze these two random sampling methods for the HodgeRank estimator. Using the Fiedler value of the graph as a measurement for estimator stability (informativeness), we provide a new estimate of the Fiedler value for these two random graph models. In the asymptotic limit as the number of vertices tends to infinity, we prove the validity of the estimate. Based on our findings, for a small number of items to be compared, we recommend a two-stage sampling strategy where a greedy sampling method is used initially and random sampling *without* replacement is used in the second stage. When a large number of items is to be compared, we recommend random sampling with replacement as this is computationally inexpensive and trivially parallelizable. Experiments on synthetic and real-world datasets support our analysis.

1. Introduction

With the advent of ubiquitous internet access and the growth of crowdsourcing platforms (e.g., MTurk, InnoCentive, CrowdFlower, CrowdRank, and AllOurIdeas), the crowdsourcing strategy is now employed by a variety of communities. Crowdsourcing enables researchers to conduct social experiments on a heterogenous set of participants and at a lower economic cost than conventional laboratory studies. For example, researchers can harness internet users to conduct user studies on their personal computers. Among various approaches to conduct subjective tests, pairwise comparisons are expected to yield more reliable results.

* Corresponding authors.
E-mail addresses: osting@math.utah.edu (B. Osting), xiongjiechao@pku.edu.cn (J. Xiong), xuqianqian@iie.ac.cn (Q. Xu), yuany@math.pku.edu.cn (Y. Yao).

http://dx.doi.org/10.1016/j.acha.2016.03.007
1063-5203/

However, in crowdsourced studies, the individuals performing the ratings are diverse compared to more controlled settings, which is difficult to control for using traditional experimental designs; researchers have recently proposed several randomized methods to conduct user studies [1–3], which accommodate incomplete and imbalanced data.

HodgeRank, as an application of combinatorial Hodge theory to the preference or rank aggregation problem from pairwise comparison data, possibly being incomplete and imbalanced, was first introduced in [4], and inspired a series of studies in statistical ranking [5–8]. Hodge theory has also found applications in game theory [9] and computer vision [10,11], in addition to traditional applications in fluid mechanics [12] etc. HodgeRank formulates the ranking problem in terms of the discrete Hodge decomposition of the pairwise data and shows that it can be decomposed into three orthogonal components: a gradient flow representing a global rating (optimal in the L_2-norm sense), a triangular curl flow representing local inconsistency, and a harmonic flow representing global inconsistency. Such a perspective generalizes various linear statistical models to provide a universal geometric description of the structure of paired comparison data, which is possibly incomplete and imbalanced due to crowdsourcing.

The two most popular random sampling schemes in crowdsourcing experiments are random sampling *with* replacement and random sampling *without* replacement. In random sampling *with* replacement, one selects a comparison pair randomly from the whole dataset regardless if the pair has been selected before; whence it is memory free. In random sampling *without* replacement, each comparison pair in the dataset has an equal chance of being selected; once selected it cannot be chosen again until all possible pairs have been chosen. The simplest model of random sampling *without* replacement in paired comparisons is the Erdös–Rényi random graph, which is a stochastic process that starts with n vertices and no edges, and at each step adds one new edge uniformly [13]. As one needs to avoid previous edges, such a sampling scheme is not memory-free and may lead to weak dependence in some estimates.

Recently, [2,3] develops the application of HodgeRank with random graph designs in subjective Quality of Experience (QoE) evaluation and shows that random graphs could play an important role in guiding random sampling designs for crowdsourcing experiments. In particular, exploiting topology evolution of clique complexes induced from Erdös–Rényi random graphs [14], [3] shows that at least $O(n \log n)$ distinct random edges are necessary to ensure the inference of a global ranking and $O(n^{3/2})$ distinct random edges are sufficient to remove the global inconsistency.

On the other hand, there are *active sampling schemes* which are designed to maximize the information in the collected dataset, potentially reducing the amount of data collected. Recently, [7,8] exploits a greedy sampling method to maximize the Fisher information in HodgeRank, which is equivalent to maximizing the smallest nonzero eigenvalue of the unnormalized graph Laplacian (a.k.a. *Fiedler value* or algebraic connectivity). Although the computational cost of such greedy sampling is prohibitive for large graphs, it effectively boosts the algebraic connectivity compared to Erdös–Rényi random graphs.

However, active sampling for data acquisition is not always feasible. For example, when data is collected from the Internet crowd or purchasing preferences, data collection is in general passive and independent. An important benefit of random sampling over active methods is that data collection can be trivially parallelized: comparisons can be collected from independent or weakly dependent processes, each selected from a pre-assigned block of object pairs. From this viewpoint, the simplicity of random sampling allows flexibility and applicability to diverse situations, such as online or distributed ranking, often desirable for crowdsourcing scenarios.

Therefore, our interest in this paper is to investigate the characteristics of these three sampling methods (i.e., random sampling with/without replacement and greedy sampling) for HodgeRank and identify an attractive sampling strategy that is particularly suitable for crowdsourcing experiments. The natural questions we are trying to address are: (i) which sampling scheme is the best, e.g., contains the most information for HodgeRank? and (ii) how do random and greedy sampling schemes compare in practice?

We approach these problems with a combination of theory and experiment in this paper. Performance of these sampling schemes is evaluated via the stability of HodgeRank, as measured by the Fiedler value. The Erdös–Rényi random graph model is associated with random sampling without replacement. For this model, an estimate of the Fiedler value was recently given in [15]. The proof of this estimate hinges on an estimate of [16], which can be shown to imply that, at first order, the Fiedler value is the minimal degree of the graph. The minimal degree of the graph can then be estimated from the binomial distribution. To analyze the random graph model associated with random sampling with replacement, we generalize the result given in [16] to multigraphs. A simple Normal approximation is then used to estimate the Fiedler value. As the graphs become increasingly dense, we prove that both random sampling methods asymptotically have the same Fiedler value. Our analysis implies:

i) For a finite graph which is sparse, random sampling *with* and *without* replacement have similar performance; for a dense finite graph, random sampling *without* replacement is superior to random sampling *with* replacement, and approaches the performance of greedy sampling.
ii) For very large graphs, the three considered sampling schemes exhibit similar performance.

In particular, the asymptotic behavior of the two random sampling schemes is rigorously proved in Theorem 1 and their discrepancy for small sample sizes is supported by heuristic estimates (see Sections 3.4 and 3.5). These analytic conclusions and the performance of the greedy sampling strategy are supported by both simulated examples and real-world datasets (see Section 4). Based on our findings, for a relatively small number of items to be compared, we recommend a two-stage sampling strategy where a greedy sampling method is used initially and random sampling without replacement is used in the second stage. When a large number of items is to be compared, we recommend random sampling with replacement as this is computationally inexpensive and trivially parallelizable.

Outline. Section 2 contains a review of related work. Then we establish some theoretical results for random sampling methods in Section 3. Proofs will be collected in Appendix A. The results of detailed experiments on crowdsourced data are reported in Section 4. We conclude in Section 5 with some remarks and a discussion of future work.

2. Related work

2.1. Crowdsourcing

The term “crowdsourcing” is a portmanteau of “crowd” and “outsourcing”. It is distinguished from outsourcing in that the work comes from an undefined public rather than being commissioned from a specific, named group. The benefits of crowdsourcing include time-efficiency and low monetary costs. Among various crowdsourcing platforms, Amazon’s Mechanical Turk (MTurk) is probably the most popular and provides a marketplace for a variety of tasks; anyone seeking help from the Internet crowd can post their task requests to the website. Another platform, Innocentive, enables organizations to engage diverse innovation communities such as employees, partners, or customers to rapidly generate novel ideas and innovative solutions to challenging research and development problems. CrowdFlower’s expertise is in harnessing the Internet crowd to provide a wide range of enterprise solutions, taking complicated projects and dividing them into smaller, simpler tasks, which are then completed by individual contributors. CrowdRank is an innovative platform that draws on the over 3 million community votes already cast to bring the crowdsourcing revolution to rankings via a novel pairwise ranking methodology that avoids the tedium of asking community members to rank every item in a category. In addition, Allourideas provides a free and open-source website that

allows groups all over the world to create and use pairwise wiki surveys. Respondents can either participate in a pairwise wiki survey or add new items that are then presented to future respondents.

With the help of these platforms, requesters post tasks (e.g. image annotation [17,18], document relevance [19], document evaluation [20], music emotion recognition [21], affection mining in computer games [22], and quality of experience evaluation [23,3]) and users are compensated in the form of micro-payments for completing these posted tasks. Several studies have been conducted to evaluate the quality of completed tasks obtained from crowdsourcing approaches. For example, researchers have investigated the reliability of non-experts and found that a single expert in the majority of cases is more reliable than a non-expert. However, using an aggregate of several, cheap non-expert judgements could approximate the performance of expensive expertise [24,25]. From this point of view, conducting subjective tests in a crowdsourcing context is a reasonable strategy.

2.2. Pairwise ranking aggregation

The problem of ranking or rating with paired comparison data has been widely studied in a variety of fields including decision science [26], machine learning [27], social choice [28], and statistics [29]. Various methods have been studied for this problem, which, among others, includes maximumlikelihood under a Bradley–Terry model, rank centrality (PageRank/MC3) [30,31], HodgeRank [4], and a pairwise variant of the Borda count [32,4]. If we consider the setting where pairwise comparisons are drawn I.I.D. from some fixed but unknown probability distribution, under a "time-reversibility" condition, the rank centrality (PageRank) and HodgeRank algorithms both converge to an optimal ranking [33]. However, PageRank is only able to aggregate the pairwise comparisons into a global ranking over the items. HodgeRank not only provides a means to determine a global ranking from paired comparison data under various statistical models (e.g., Uniform, Thurstone–Mosteller, Bradley–Terry, and Angular Transform), but also measures the inconsistency of the global ranking obtained. In particular, it takes a graph theoretic view, which maps paired comparison data to edge flows on a graph, possibly imbalanced (where different pairs may receive different number of comparisons) and incomplete (where every participant may only provide partial comparisons), and then applies combinatorial Hodge Theory to achieve an orthogonal decomposition of such edge flows into three components: a gradient flow representing the global rating (optimal in the L_2-norm sense), a triangular curl flow representing local inconsistency, and a harmonic flow representing global inconsistency. In this paper, we will analyze two random sampling methods based on the HodgeRank estimate.

2.3. Active sampling

The fundamental notion of active sampling has a long history in machine learning. To our knowledge, the first to discuss it explicitly were [34] and [35]. Subsequently, the term active learning was coined [36] and has been shown to benefit a number of multimedia applications such as object categorization [37], image retrieval [38,39], video classification [40], dataset annotation [41], and interactive co-segmentation [42], maximizing the knowledge gain while valuing the user effort [43].

Recently, several authors have studied the active sampling problems for ranking and rating, with the goal of reducing the amount of data that must be collected. For example, [44] considers the case when the true scoring function reflects the Euclidean distance of object covariates from a global reference point. If objects are embedded in $\mathbb{R}^d$ or the scoring function is linear in such a space, the active sampling complexity can be reduced to $O(d \log n)$, as demonstrated through a comparison of beer [45]. Moreover, [46] discusses the application of a polynomial time approximate solution (PTAS) for the NP-hard minimum feedback arc-set (MFAST) problem, in active ranking with sample complexity $O(n \cdot \mathrm{poly}(\log n, 1/\varepsilon))$ to achieve ε-optimum. The works mentioned above can be treated as "learning to rank" which requires a vector representation of the items to be ranked, thus can not be directly applied to crowdsourced ranking. In the crowdsourcing

scenario, the explicit feature representation of items is unavailable and the goal becomes to learn a single ranking function from the ranked items using a smaller number of samples selected actively [30]. In [47], a Bayesian framework is proposed to actively select pairwise comparison queries based on Bradley–Terry models. Furthermore, [48] addresses the problem of budget allocation in crowd labeling using the Bayesian Markov decision process and characterizing the optimal policy using the dynamic programming. Most recently, [7,8] approaches active sampling from a statistical perspective of maximizing the Fisher information, which they show to be equivalent to maximizing the Fiedler value of the graph (smallest nonzero eigenvalue of the graph Laplacian which arises in HodgeRank), subject to an integer weight constraint. In this paper, we shall focus on analyzing the Fiedler value of graphs generated based on random sampling schemes.

3. Analysis of sampling methods

Statistical preference aggregation or ranking/rating from pairwise comparison data is a classical problem, which can be traced back to the 18th century with discussions on voting and social choice. This subject area has recently undergone rapid growth in various applications due to the availability of the Internet and development of crowdsourcing techniques. In these scenarios, typically we are given pairwise comparison data on a graph $G = (V, E)$, $Y^\alpha \colon E \to \mathbb{R}$ such that $Y^\alpha_{ij} = -Y^\alpha_{ji}$ where α is the index for multiple comparisons, $Y^\alpha_{ij} > 0$ if it prefers i to j and $Y^\alpha_{ij} \leq 0$ otherwise. In the dichotomous choice, Y^α_{ij} can be taken as $\{\pm 1\}$, while multiple choices are also widely used (e.g., k-point Likert scale, $k = 3, 4, 5$).

The general purpose of preference aggregation is to look for a global score $x \colon V \to \mathbb{R}$ such that

$$\min_{\substack{x \in \mathbb{R}^{|V|} \\ x \perp 1}} \sum_{i,j,\alpha} \omega^\alpha_{ij} \mathcal{L}(x_i - x_j, Y^\alpha_{ij}), \tag{1}$$

where $\mathcal{L}(x, y) \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a loss function, ω^α_{ij} denotes the confidence weight of this comparison which is set to be 1 in this paper but other choices are also possible, and x_i (x_j) represents the global ranking score of item i (j, respectively). For a connected graph G, restricting to the subspace $\{x \in \mathbb{R}^{|V|} \colon x \perp 1\}$ guarantees a unique solution in (1). For example, $\mathcal{L}(x, y) = (\operatorname{sign}(x) - y)^2$ leads to the minimum feedback arc-set (MFAST) problem which is NP-hard, where [46] proposes an active sampling scheme whose complexity is $O(n \cdot \operatorname{poly}(\log n, 1/\varepsilon))$ to achieve ε-optimum. In HodgeRank, one benefits from the use of square loss $\mathcal{L}(x, y) = (x - y)^2$ which leads to fast algorithms to find optimal global ranking x, as well as an orthogonal decomposition of the least square residue into local and global inconsistencies [4].

To see this, let $\hat{Y}_{ij} = (\sum_\alpha \omega^\alpha_{ij} Y^\alpha_{ij})/(\sum_\alpha \omega^\alpha_{ij})$ ($\omega_{ij} = \sum_\alpha \omega^\alpha_{ij}$) be the mean pairwise comparison scores on (i, j), which can be extended to a family of generalized statistical linear models. To characterize the solution and residual of (1), we first define a 3-clique complex $X_G = (V, E, T)$ where T collects all triangular complete subgraphs in G:

$$T = \left\{ \{i, j, k\} \in \binom{V}{3} : \{i, j\}, \{j, k\}, \{k, i\} \in E \right\}.$$

Then every $\hat{Y}$ admits an orthogonal decomposition:

$$\hat{Y} = \hat{Y}^g + \hat{Y}^h + \hat{Y}^c, \tag{2}$$

where the gradient flow $\hat{Y}^g$ satisfies

$$\hat{Y}^g_{ij} = x_i - x_j, \quad \text{for some } x \in \mathbb{R}^n, \tag{3}$$

the harmonic flow $\hat{Y}^h$ satisfies

$$\hat{Y}^h_{ij} + \hat{Y}^h_{jk} + \hat{Y}^h_{ki} = 0, \text{ for each } \{i,j,k\} \in T, \tag{4}$$

$$\sum_{j:(i,j)\in E} \omega_{ij}\hat{Y}^h_{ij} = 0, \text{ for each } i \in V, \tag{5}$$

and the curl flow $\hat{Y}^c$ satisfies (5) but not (4).

The residuals $\hat{Y}^c$ and $\hat{Y}^h$ indicate whether inconsistencies in the ranking data arises locally or globally. Local inconsistency can be fully characterized by triangular cycles (e.g. $i \succ j \succ k \succ i$), while global inconsistency generically involves longer cycles of inconsistency in V (e.g. $i \succ j \succ k \succ \cdots \succ i$), which may arise due to data incompleteness and cause the fixed tournament issue. Random sampling to avoid global inconsistency, generally requires at least $O(n^{3/2})$ random samples without replacement [2,3].

The global rating score x in (3) can be obtained by solving the normal equation [4],

$$Lx = b. \tag{6}$$

Here, $L = D - A$ is the unnormalized graph Laplacian, where $A(i,j) = \omega_{ij}$ if $(i,j) \in E$, $A(i,j) = 0$ otherwise, and D is a diagonal matrix with $D(i,i) = \sum_{j:(i,j)\in E} \omega_{ij}$, as well as $b = \operatorname{div}(\hat{Y})$ is the divergence flow defined by $b_i = \sum_{j:(i,j)\in E} \omega_{ij}\hat{Y}_{ij}$. There is an extensive literature in linear algebra on solving the symmetric Laplacian equation. However, all methods are subject to the intrinsic stability of HodgeRank, characterized in the following subsection.

3.1. Stability of HodgeRank

The following classical result (see, e.g., [49]) gives a measure of the sensitivity of the global ranking score x against perturbations on L and b. Given the parameterized system

$$(L + \epsilon F)x(\epsilon) = b + \epsilon f, \quad x(0) = x$$

where $F \in R^{n\times n}$ and $f \in R^n$, then

$$\frac{\|x(\epsilon) - x\|}{\|x\|} \leq |\epsilon|\|L^{-1}\| \left(\frac{\|f\|}{\|x\|} + \|F\| \right) + O(\epsilon^2).$$

Here and throughout this paper, the matrix norm is the spectral norm and the vector norm is the Euclidean norm. In crowdsourcing, the matrix L is determined by the sampled pairs, and can be regard as fixed when given the pairwise data. However, $b = \operatorname{div}(\hat{Y})$ is random because of noise possibly induced by crowdsourcing. So there is no perturbation on L, i.e., $F = 0$, and we obtain

$$\frac{\|x(\epsilon) - x\|}{\|f\|} \leq \|L^{-1}\| \, |\epsilon| + O(\epsilon^2). \tag{7}$$

If the graph representing the pairwise comparison data is connected, the graph Laplacian, L, has a one-dimensional kernel spanned by the constant vector. In this case, the solution to (6) is understood in the minimal norm least-squares sense, i.e. $\hat{x} = L^{\dagger}b$ where $L^{\dagger}$ is the Moore–Penrose pseudoinverse of L. Hence (7) implies that the sensitivity of the estimator is controlled by $\|L^{\dagger}\| = [\lambda_2(L)]^{-1}$, the reciprocal of the second smallest eigenvalue of the graph Laplacian. $\lambda_2(L)$ is also referred to as the *Fiedler value* or *algebraic connectivity* of the graph. It follows that collecting pairwise comparison data so that $\lambda_2(L)$ is large provides an estimator which is insensitive to noise in the pairwise comparison data, $\hat{Y}$.

Remark. [7,8] show that for a fixed variance statistical error model, the Fisher information matrix of the HodgeRank estimator (6) is proportional to the graph Laplacian, L. Thus, finding a graph with large

algebraic connectivity, $\lambda_2(L)$, can be equivalently viewed in the context of optimal experimental design as maximizing the "E-criterion" of the Fisher information.

3.2. Random sampling schemes

In what follows, we study two random sampling schemes:

1. $G_0(n,m)$: *Uniform sampling with replacement*. Each edge is sampled from the uniform distribution on $\binom{n}{2}$ edges, with replacement. This is a weighted graph and the sum of weights is m.
2. $G(n,m)$: *Uniform sampling without replacement*. Each edge is sampled from the uniform distribution on the available edges without replacement. For $m \leq \binom{n}{2}$, this is an instance of the Erdös–Rényi random graph model $G(n,p)$ with $p = m/\binom{n}{2}$.

Motivated by the estimate given in (7), we will characterize the behavior of the Fiedler value of the graph Laplacians associated with these sampling methods. It is well-known that the Erdos–Renyi random graph $G(n,p)$ is connected with high probability if the sampling rate is at least $p = (1+\epsilon)\log n/n$ [13]. Therefore we use the parameter $p_0 := 2m/((n-1)\log n) \geq 1$ (where $m = n(n-1)p/2 \approx n^2 p/2$), the degree above the connectivity threshold, to compare the efficiency in boosting Fiedler values for different sampling methods.

As a comparison for random sampling schemes, we consider a *greedy sampling* method of sampling pairwise comparisons to maximize the algebraic connectivity of the graph [50,7,8]. The problem of finding a set of m edges on n vertices with maximal algebraic connectivity is an NP-hard problem. The following greedy heuristic, based on the Fiedler vector, ψ, can be used. The Fiedler vector is the eigenfunction of the graph Laplacian corresponding to the Fiedler value. We shall denote the graph with m edges on n vertices by $G_\star(n,m)$. The graph is built iteratively, at each iteration, the Fiedler vector is computed and the edge (i,j) which maximizes $(\psi_i - \psi_j)^2$ is added to the graph. The iterates are repeated until a graph of the desired sized is obtained.

3.3. Fiedler value and minimal degree

The key to evaluating the Fiedler value of random graphs is via the graph minimal degree, $d_{\min}$. This is due to the definition of graph Laplacian,

$$L = D - A,$$

whose diagonal $D(i,i) \simeq O(m/n)$ dominates as $\max_{\|v\|=1} v^T A v \simeq O(\sqrt{m/n})$. The following Lemma makes this observation precise, which is used by [15] in the study of Erdös–Rényi random graphs.

Lemma 1. *Consider the random graph $G_0(n,m)$ (or $G(n,m)$) and let λ_2 be the Fiedler value of the graph. Suppose there exists a $p_0 > 1$ so that $2m \geq p_0 n \log n$ and $C, c_1 > 0$ so that*

$$|d_{\min} - c_1 \frac{2m}{n}| \leq C\sqrt{\frac{2m}{n}}$$

with probability at least $1 - O(e^{-\Omega(\sqrt{\frac{2m}{n}})})$. Then there exists a $\tilde{C} > 0$ so that

$$|\lambda_2 - c_1 \frac{2m}{n}| \leq \tilde{C}\sqrt{\frac{2m}{n}}.$$

Lemma 1 implies that the difference between λ_2 and $d_{\min}$ (i.e., minimal degree) is small, so the Fiedler value for both random graphs can be approximated by their minimal degrees.

The proof for $G(n,m)$ follows from [15], which establishes the result for the Erdös–Rényi random model $G(n,p)$. The proof for $G_0(n,m)$, needs the following lemma.

Lemma 2. *Let A denote the adjacency matrix of a random graph from $G_0(n,m)$ and $S = \{v \perp 1 \colon \|v\| = 1\}$. There exists a constant $c > 0$, such that if $m > n\log n/2$, the estimate*

$$\max_{v\in S} v^T A v \;\le\; c\sqrt{2m/n}$$

holds with probability at least $1 - O(1/n)$.

With the aid of this lemma, one can estimate the Fiedler value by the minimal degree of $G_0(n,m)$. In fact,

$$\begin{aligned}\lambda_2(L) &= \min_{v\in S}\langle v, Lv\rangle \\ &= \min_{v\in S}\langle v, Dv\rangle - \langle v, Av\rangle \\ &\ge d_{\min} - \max_{v\in S}\langle v, Av\rangle.\end{aligned}$$

Also Cheeger's inequality tells that $\lambda_2(L) \le \frac{n}{n-1} d_{\min}$. These bounds show the validity of Lemma 1. The proof of Lemma 2 is given in Appendix A.

3.4. A heuristic estimate of the minimal degree

In this section, we estimate the minimal degree. First, consider the Erdös–Rényi random graph model $G(n,p)$ with $p = m/\binom{n}{2}$. Then $d_i \sim B(n,p)$, so $\frac{d_i - np}{\sqrt{np(1-p)}} \sim N(0,1)$. The degrees are *weakly dependent*. If the degrees were *independent*, the following concentration inequality for Gaussian random variables,

$$\mathbf{Prob}(\max_{1\le i\le n} |X_i| > t) \le n\exp\left(-\frac{t^2}{2}\right), \quad X \sim N(0, I_n),$$

would imply that the minimal value of n copies of $N(0,1)$ is about $-\sqrt{2\log n}$. In this case,

$$d_{\min} \approx np - \sqrt{2\log(n) np(1-p)},$$

implying that

$$\frac{d_{\min}}{np} \approx 1 - \sqrt{\frac{2\log n}{np}}\sqrt{1-p}.$$

A similar approximation can be employed for $G_0(n,m)$. Here, $d_i \sim B(m, 2/n)$, so $\frac{d_i - np}{\sqrt{np(1-2/n)}} \sim N(0,1)$. Again, the d_i are only weakly dependent, so

$$d_{\min} \approx np - \sqrt{2\log(n) np(1-2/n)},$$

which implies that

$$\frac{d_{\min}}{np} \approx 1 - \sqrt{\frac{2\log n}{np}}\sqrt{1-2/n}.$$

Collecting these results and using $\frac{d_{\min}}{np} \simeq \frac{\lambda_2}{np}$, we have the following estimates.

Key estimates.

$$G_0(n,m): \quad \frac{\lambda_2}{np} \approx a_1(p_0,n) := 1 - \sqrt{\frac{2}{p_0}}\sqrt{1-\frac{2}{n}} \tag{8}$$

$$G(n,m): \quad \frac{\lambda_2}{np} \approx a_2(p_0,n) := 1 - \sqrt{\frac{2}{p_0}}\sqrt{1-p} \tag{9}$$

where $p = \frac{p_0 \log n}{n}$.

Remark. As $n \to \infty$, both (8) and (9) become $\frac{\lambda_2}{np} \approx 1 - \sqrt{\frac{2}{p_0}}$. But for finite n and dense p, $G(n,m)$ may have larger Fiedler value than $G_0(n,m)$.

The above reasoning (falsely) assumes independence of d_i, which is only valid as $n \to \infty$. In the following section, we make this precise with an asymptotic estimate of the Fiedler value in the two random sampling schemes.

3.5. Asymptotic analysis of the Fiedler value

In the last section, we gave a heuristic estimator of the Fiedler value. The following theorem gives an asymptotic estimate of the Fiedler value as $n \to \infty$.

Theorem 1. *Consider a random graph $G_0(n,m)$ (or $G(n,m)$) on n vertices corresponding to uniform sampling with (without) replacement and $m = p_0 n \log(n)/2$. Let λ_2 be the Fiedler value of the graph. Then*

$$\frac{\lambda_2}{2m/n} \simeq a(p_0) + O(\frac{1}{\sqrt{2m/n}}), \tag{10}$$

with high probability, where $a(p_0) \in (0,1)$ denotes the solution to

$$p_0 - 1 = a p_0 (1 - \log a).$$

The proof of Theorem 1 is given in Appendix A.

Remark. For $p_0 \gg 1$, $a(p_0) = 1 - \sqrt{2/p_0} + O(1/p_0)$ [15]. Thus, Theorem 1 implies that the two sampling methods have the same asymptotic algebraic connectivity, $\frac{\lambda_2}{2m/n} \simeq 1 - \sqrt{2/p_0}$ as $n \to \infty$ and $p_0 \gg 1$. Note from (8) and (9), that $\lim_{n\to\infty} a_1(p_0,n) = \lim_{n\to\infty} a_2(p_0,n) = 1 - \sqrt{2/p_0}$.

The main difference between $G(n,p)$ and $G_0(n,m)$ is the weak dependence pattern; the dependence of d_i and d_j only occurs on edge (i,j) which only appears at most once for $G(n,p)$, but all of the m edges can be (i,j) for $G_0(n,m)$. However, we still have d_i and d_j are almost independent when n is sufficiently large, so the heuristic estimator using I.I.D. Normal distribution as an approximation is not unreasonable.

Theorem 1 is supported by Figs. 1 and 2, where the Fiedler value, minimal degree, and various estimates, (8), (9), and (10), are plotted for varying edge sparsity, p_0. For $G_0(n,m)$, we observe that $a(p_0)$ fits $d_{\min}$ and λ_2 pretty well for all p_0. For $G(n,m)$, we observe that $a(p_0)$ fits $d_{\min}$ and λ_2 well when p_0 is small, but when p_0 is large, the estimate give in (9) is more reliable. In all cases, the Fiedler value for the graph, $G_\star$, generated by greedy sampling, is larger than that for randomly sampled graphs.

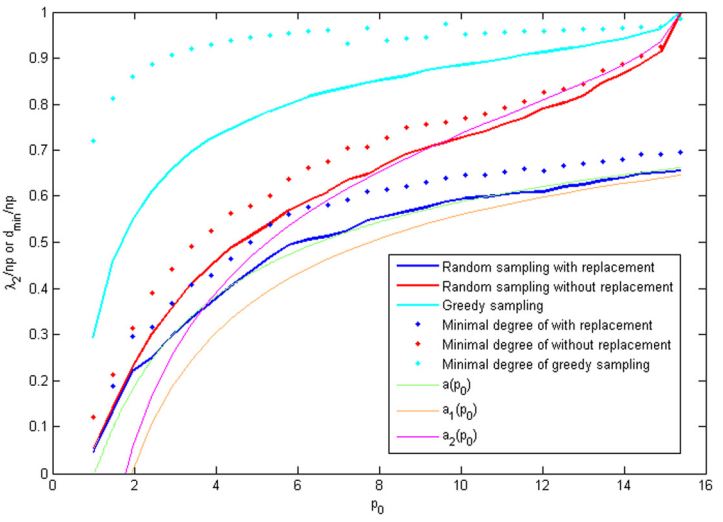


Fig. 1. A comparison of the Fiedler value, minimal degree, and estimates (8), (9), and (10) for graphs generated via random sampling with/without replacement and greedy sampling for $n = 64$.

4. Experiments

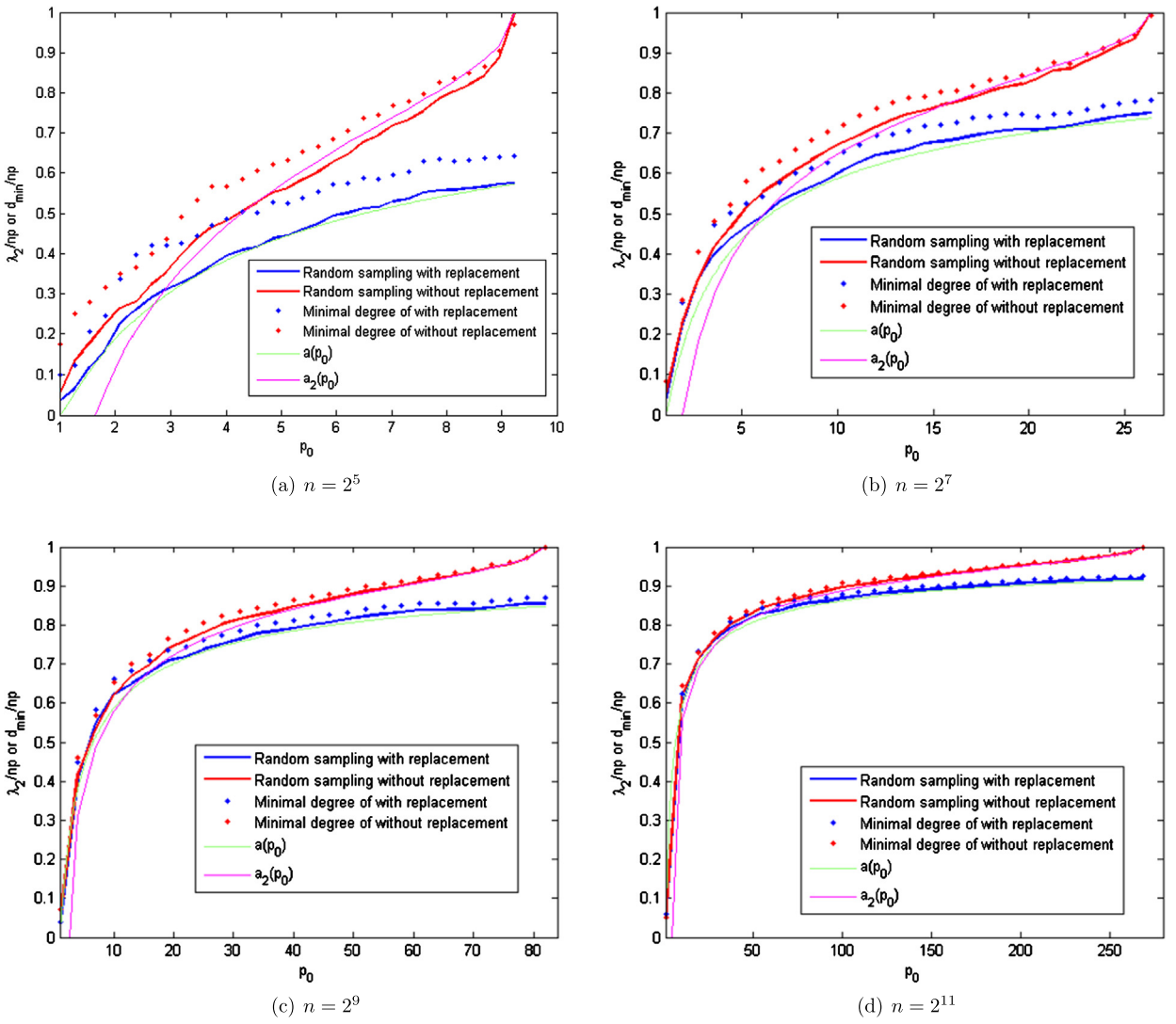
In this section, we study three examples with both simulated and real-world data to illustrate the validity of the analysis above and applications of the proposed sampling schemes. The first example is with simulated data while the latter two consider real-world data from QoE evaluation. The code for the numerical experiment and the real-world datasets can be downloaded from https://code.google.com/p/active-random-joint-sampling/.

4.1. Simulated data

This subsection uses simulated data to illustrate the performance differences among the three sampling schemes. We randomly create a global ranking score as the ground-truth, uniformly distributed on $[0, 1]$ for $|V| = n$ candidates. In this way, we obtain a complete graph with $\binom{n}{2}$ edges consistent with the true preference direction. We sample pairs from this complete graph using random sampling with/without replacement and the greedy sampling scheme. The experiments are repeated 1000 times and ensemble statistics for the HodgeRank estimator (6) are recorded. As we know the ground-truth score, the metric used here is the L_2-distance between the HodgeRank estimate and ground-true score, $\|\hat{x} - x^*\|$.

Fig. 3 (a) shows the mean L_2-distance and standard deviation associated with the three sampling schemes for $n = 16$ (chosen to be consistent with the two real-world datasets considered later). The x-axes of the graphs are the number of edges, as measured by $p_0 = \frac{pn}{\log n}$, taken to be greater than one so that the graph is connected with high probability. From these experimental results, we observe that the performance of random sampling without replacement is better than random sampling with replacement in all cases with smaller L_2-distance and smaller standard deviation. As p_0 grows, the performance of the two random sampling schemes diverge. When the graph is sparse, the greedy sampling scheme shows better performance than random sampling with/without replacement. However, when the graphs become dense, random sampling without replacement performs qualitatively similar to greedy sampling.

To simulate real-world data contaminated by outliers, each binary comparison is independently flipped with a probability, referred to as *outlier percentage* (OP). For $n = 16$, with OP = 10% and 30%, we plot the number of sampled pairs against the L_2-distance and standard deviation between the ground-truth and HodgeRank estimate in Fig. 3(b,c). As in the non-contaminated case, the greedy sampling strategy



(a) $n = 2^5$ (b) $n = 2^7$

(c) $n = 2^9$ (d) $n = 2^{11}$

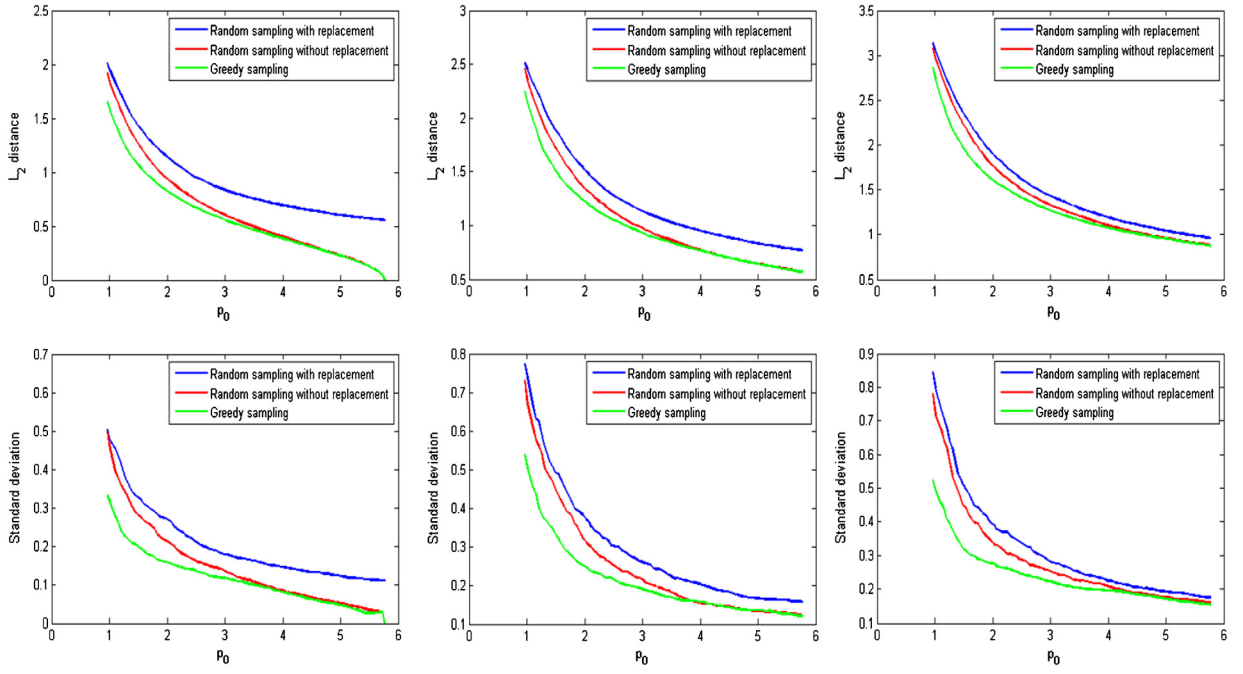
Fig. 2. Algebraic connectivity and minimal degree: Random sampling with replacement vs. Random sampling without replacement for $n = 2^5$, 2^7, 2^9, and 2^{11}. The gaps among these sampling schemes vanish as $n \to \infty$.

outperforms the random sampling strategy. As OP increases, the performance gap among the three sampling schemes decreases.

4.2. Real-world data

The second example gives a comparison of the three sampling methods on a video quality assessment dataset [2]. It contains 38 400 paired comparisons of the LIVE dataset [51] from 209 random observers. An attractive property of this dataset is that the paired comparison data is complete and balanced. As LIVE includes 10 different reference videos and 15 distorted versions of each reference (obtained using four different distortion processes — MPEG-2 compression, H.264 compression, lossy transmission of H.264 compressed bitstreams through simulated IP networks, and lossy transmission of H.264 compressed bitstreams through simulated wireless networks), for a total of 160 videos, the complete comparisons of this video database requires $10 \times \binom{16}{2} = 1200$ comparisons. Therefore, 38 400 comparisons correspond to 32 complete rounds.

As there is no ground-truth scores available, results obtained from all the paired comparisons are treated as the ground-truth. To ensure the statistical stability, for each of the 10 reference videos, we sample using



(a) $OP = 0\%$ (b) $OP = 10\%$ (c) $OP = 30\%$

Fig. 3. The L_2-distance and standard deviation between ground-truth and HodgeRank estimate for random sampling with/without replacement and greedy sampling for $n = 16$.

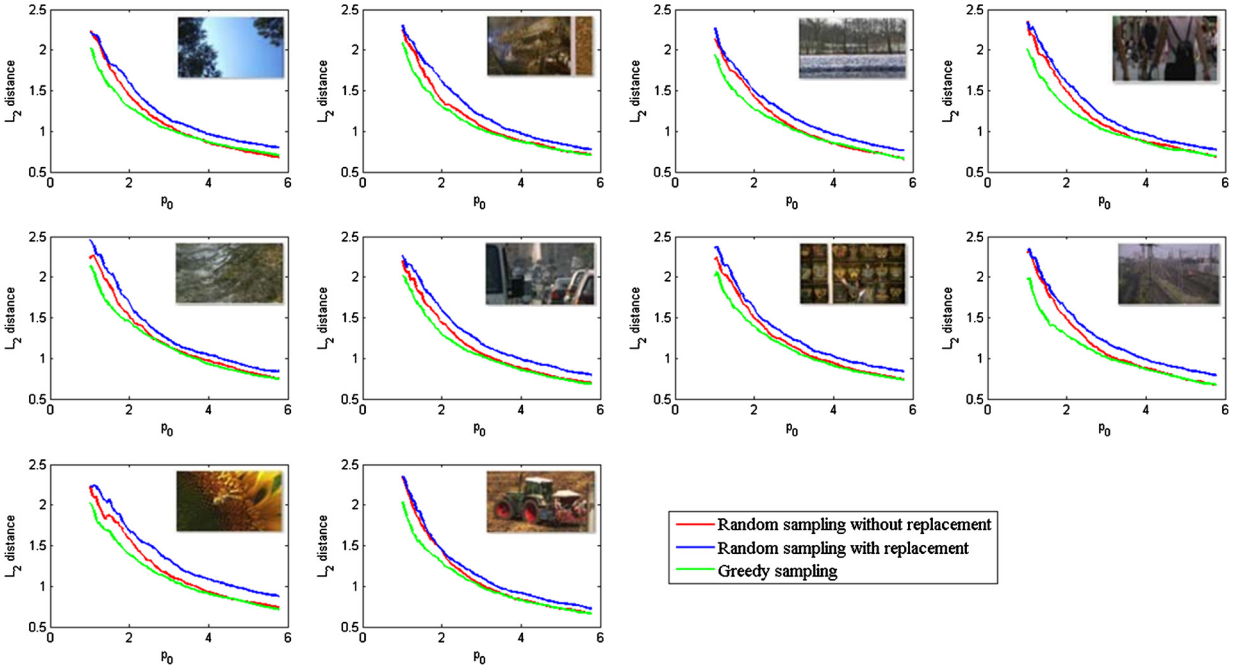


Fig. 4. Random sampling with/without replacement vs. greedy sampling for 10 reference videos in LIVE database [51].

each of the three methods 100 times. Fig. 4 shows the experimental results of the 10 reference videos in LIVE database [51]. It is interesting to obtain similar observations on all of these large scale data collections. Consistent with the simulated data, when the graph is sparse, greedy sampling performs better than both

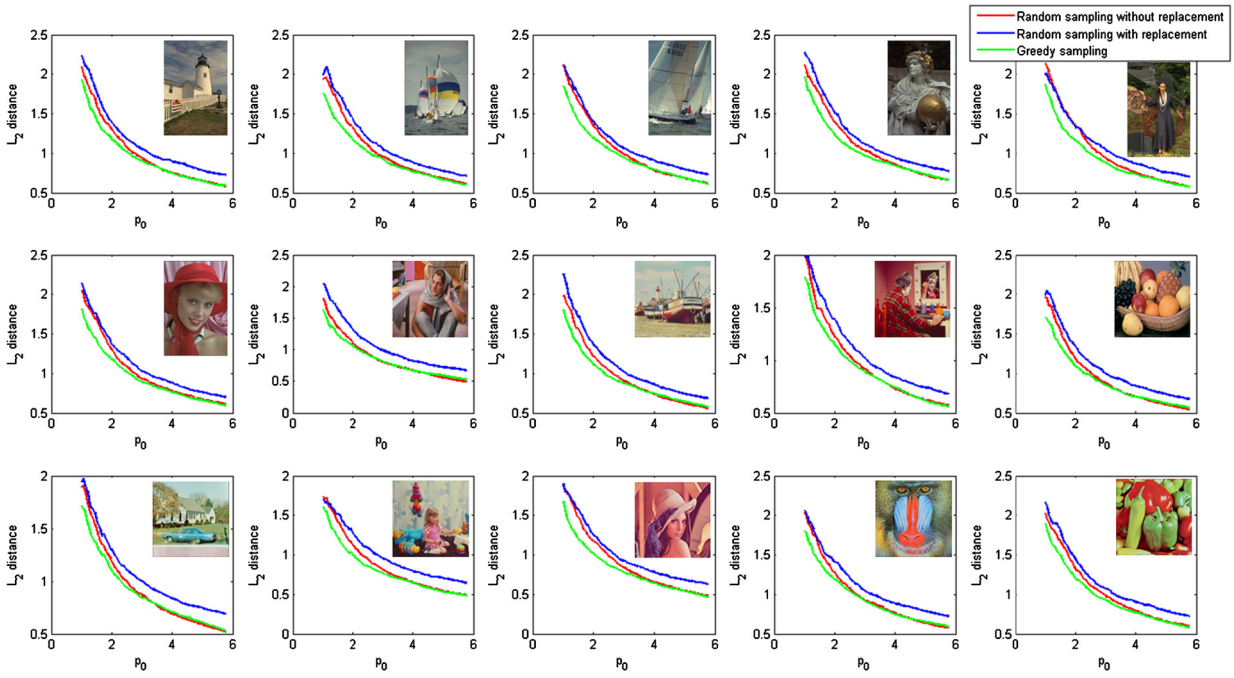


Fig. 5. Random sampling with/without replacement vs. greedy sampling for 15 reference images in LIVE [51] and IVC [52] databases.

random sampling schemes; as the number of samples increases, random sampling without replacement exhibits similar performance in the prediction of global ranking scores.

The third example shows the sampling results on an imbalanced dataset for image quality assessment, which contains 15 reference images and 15 distorted versions of each reference, for a total of 240 images which come from two publicly available datasets, LIVE [51] and IVC [52]. The distorted images in LIVE dataset [51] are obtained using five different distortion processes — JPEG2000, JPEG, White Noise, Gaussian Blur, and Fast Fading Rayleigh, while the distorted images in IVC dataset [52] are derived from four distortion types — JPEG2000, JPEG, LAR Coding, and Blurring. In total, 328 observers, each of whom performs a varied number of comparisons via the Internet, provide 43 266 paired comparisons. Since the number of paired comparisons in the is dataset is relatively large, all 15 paired comparison graphs are complete, though possibly imbalanced. This makes it possible for us to obtain comparable results of these three sampling schemes. As in the second example, quality scores obtained from all the 43 266 paired comparisons are treated as the ground-truth. Fig. 5 shows mean L^2-distance of 100 times on LIVE [51] and IVC [52] databases, and it is easy to find that all these reference images agree well with the theoretical and simulated results we have provided.

4.3. Discussion

In terms of the stability of HodgeRank, random sampling *without* replacement exhibits a performance curve between the greedy sampling scheme, proposed by [7,8], and random sampling *with* replacement. When the sampling rate is sparse, greedy sampling dominates; when the sample size is increased, random sampling *without* replacement is indistinguishable from greedy sampling, both of which dominate the random sampling *with* replacement (the simplest I.I.D. sampling).

Therefore, in practical situations, we suggest first to adopt greedy sampling in the initial stage which leads to a graph with large Fiedler value, then use random sampling *without* replacement to approximate the results of greedy sampling. Such a transition point should depend on the graph vertex set size, n, for example $p_0 := \frac{2m}{(n-1)\log n} \approx \frac{n}{2\log n}$ which suggests $p_0 \approx 3$ for $n = 16$ in our simulated and real-world

examples. After all, this random sampling scheme is simpler and more flexible than greedy sampling and does not significantly reduce the accuracy of the HodgeRank estimate.

5. Conclusion

This paper analyzed two simple random sampling schemes for the HodgeRank estimate, including random sampling *with* replacement and random sampling *without* replacement. We showed that for a finite graph when it is sparse, random sampling *without* replacement approaches its performance lower bound as random sampling *with* replacement; when it is dense, random sampling *without* replacement approaches its performance upper bound as greedy sampling. For large graphs, such performance gaps are vanishing in that all three sampling schemes exhibit similar performance.

Because random sampling relies only on a random subset of pairwise comparisons, data collection can be trivially parallelized. This simple structure makes it easy to adapt to new situations, such as online or distributed ranking. Based on these observations, we suggest in applications first adopt greedy sampling method in the initial stage and random sampling *without* replacement in the second stage. For very large graphs, random sampling *with* replacement may become the best choice, after all, it is the simplest I.I.D. sampling and when n goes to infinity, the gaps among these sampling schemes vanish. The sampling schemes enable us to derive reliable global ratings in an efficient manner, whence provide us a helpful tool for those who exploit crowdsourceable paired comparison data for subjective studies.

Acknowledgments

The research of Qianqian Xu was supported in part by National Natural Science Foundation of China: 61402019, and China Postdoctoral Science Foundation: 2015T80025. The research of Jiechao Xiong and Yuan Yao was supported in part by National Basic Research Program of China: 2015CB85600 and 2012CB825501, National Natural Science Foundation of China: 61370004 and 11421110001 (A3 project), as well as some awards from Baidu and Microsoft Research Asia.

Appendix A. Proof of Theorem 1

The following basic inequality is used extensively throughout the proofs, which can be found in [53].

Lemma 3 *(Chernoff–Hoeffding theorem). Assume that $X_i \in [0,1], i = 1, \dots, n$ are independent and $EX_i = \mu$. For $\epsilon > 0$, the following inequalities hold*

$$P(\bar{X}_n \leq \mu - \epsilon) \leq e^{-nKL(\mu-\epsilon\|\mu)},$$
$$P(\bar{X}_n \geq \mu + \epsilon) \leq e^{-nKL(\mu+\epsilon\|\mu)},$$

where $KL(p \parallel q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$, is Kullback–Leibler divergence, and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean.

Corollary 1.

$$P(\bar{X}_n \leq k\mu) \leq e^{-n\mu(k\log(k)-k+1)}, k < 1,$$
$$P(\bar{X}_n \geq k\mu) \leq e^{-n\mu(k\log(k)-k+1)}, k > 1.$$

Proof. $KL(k\mu \parallel \mu) = k\mu\log(k) + (1-k\mu)\log(\frac{1-k\mu}{1-\mu})$. Defining $f(\mu) := (1-k\mu)\log(\frac{1-k\mu}{1-\mu}) + (k-1)\mu$, we compute

$$f(0) = 0, \quad f'(0) = 0, \quad \text{and} \quad f''(\mu) = \frac{(1-k)^2}{(1-k\mu)(1-\mu)^2} > 0.$$

For all $\mu \in (0, \min(1, 1/k))$, we have that

$$f(\mu) > 0 \iff e^{-n(1-k\mu)\log(\frac{1-k\mu}{1-\mu})} < e^{n\mu(k-1)}.$$

The result then follows from Lemma 3. □

Throughout this section, $\mathbb{E}[\cdot]$ is used for expectation of random variables. Next, we prove Lemma 2.

Proof of Lemma 2. Our proof essentially follows [16] for the Erdös–Rényi random graph $G(n,p)$. For $G_0(n,m)$, consider $A = \sum_{k=1}^m A_k$ where A_k is the adjacency matrix of I.I.D. edge samples. Hence

$$v^T A v = \sum_{k=1}^m v^T A_k v = \sum_{k=1}^m 2 v_{i_k} v_{j_k}$$

is the sum of I.I.D. variables. Let $d := 2m/n$ denote the expected degree of each graph vertex.

The proof strategy is as follows. To reach a bound for $\max_{v\in S} v^T A v$, one needs a discrete cover T of the set S. We turn to an upper bound $\max_{u,v\in T} u^T A v \geq c(1-\epsilon)^2\sqrt{2m/n}$. However, any cover, T, has size $e^{O(n)}$ and therefore directly using Bernstein's inequality and the union bound doesn't work. Following [16], we divide the set $\{(u_i, v_j)\colon u, v \in T\}$ into two parts: (1) light couples with $|u_i v_j| \leq \sqrt{d}/n$, which can be bounded using Bernstein's inequality and (2) heavy couples with $|u_i v_j| > \sqrt{d}/n$ but satisfying bounded degree and discrepancy properties. These two parts make up of the variation in $u^T A v$ which will lead to the bound in Lemma 2.

Following [16], the first step is to reduce the set of vectors into a finite, yet exponentially large space. Let $S = \{v \perp 1 : \|v\| \leq 1\}$ and for fixed some $0 < \epsilon < 1$, define a grid which approximates S:

$$T = \left\{ x \in \left(\frac{\epsilon}{\sqrt{n}}\,\mathbb{Z}\right)^n : \sum_i x_i = 0, \|x\| \leq 1 \right\}.$$

Claim. (See [16].) The number of vectors in T is bounded by $e^{c_\epsilon n}$ for some c_ϵ which depends on ϵ. If for every $u, v \in T$ $u^T A v \leq c$, then for every $x \in S$, $x^T A x \leq c/(1-\epsilon)^2$.

It remains to show that

Claim. $\exists c$, almost surely, $\forall u, v \in T, u^T A v \leq c\sqrt{2m/n}$.

To prove this claim, we divide the set $\{(u_i, v_j)\colon u, v \in T\}$ into two parts:

(1) light couples with $|u_i v_j| \leq \sqrt{d}/n$ and
(2) heavy couples with $|u_i v_j| > \sqrt{d}/n$,

Let

$$L_k = u_{i_k} v_{j_k} 1_{\{|u_{i_k} v_{j_k}| \leq \frac{\sqrt{d}}{n}\}} + u_{j_k} v_{i_k} 1_{\{|u_{j_k} v_{i_k}| \leq \frac{\sqrt{d}}{n}\}}$$

$$H_k = u_{i_k} v_{j_k} 1_{\{|u_{i_k} v_{j_k}| > \frac{\sqrt{d}}{n}\}} + u_{j_k} v_{i_k} 1_{\{|u_{j_k} v_{i_k}| > \frac{\sqrt{d}}{n}\}}.$$

Then

$$u^T A v = \sum_{k=1}^{m} L_k + \sum_{k=1}^{m} H_k.$$

Bound on the contribution of light couples. It's easy to compute

$$\begin{aligned}
\mathbb{E}L_k + \mathbb{E}H_k &= \frac{2}{n(n-1)} \sum_{i \neq j} u_i v_j = -\frac{2}{n(n-1)} \langle u, v \rangle \\
|\mathbb{E}H_k| = \frac{2}{n(n-1)} |\sum_{i \neq j} u_i v_j 1_{\{|u_i v_j| \geq \frac{\sqrt{d}}{n}\}}| &\leq \frac{2}{n(n-1)} \sum_{i,j:|u_i v_j| \geq \frac{\sqrt{d}}{n}} |u_i^2 v_j^2 / \frac{\sqrt{d}}{n}| \\
&\leq \frac{2}{(n-1)\sqrt{d}} \sum_i u_i^2 \sum_j v_j^2 \\
&\leq \frac{2}{(n-1)\sqrt{d}}.
\end{aligned}$$

So $|\mathbb{E}\sum_{k=1}^{m} L_k| \leq m|\mathbb{E}H_k| + m|\mathbb{E}L_k + \mathbb{E}H_k| = O(\sqrt{d})$, as $d = 2m/n$. We also have that

$$\begin{aligned}
\mathrm{Var}(L_k) \leq \mathbb{E}(L_k)^2 &\leq 2\mathbb{E}\left((u_{i_k} v_{j_k} 1_{\{|u_{i_k} v_{j_k}| \leq \frac{\sqrt{d}}{n}\}})^2 + (u_{j_k} v_{i_k} 1_{\{|u_{j_k} v_{i_k}| \leq \frac{\sqrt{d}}{n}\}})^2 \right) \\
&\leq \frac{4}{n(n-1)} \sum_i u_i^2 \Big(\sum_{j \neq i} v_j^2 \Big) \\
&\leq \frac{4}{n(n-1)}.
\end{aligned}$$

From definition, $|L_k| \leq 2\frac{\sqrt{d}}{n}$, so $|L_k - \mathbb{E}L_k| \leq |L_k| + |\mathbb{E}L_k| \leq 4\frac{\sqrt{d}}{n} \triangleq M$. Then Bernstein's inequality gives

$$\begin{aligned}
P\left(\sum_{k=1}^{m} L_k - \mathbb{E}\sum_{k=1}^{m} L_k > c\sqrt{d} \right) &\leq \exp\left(-\frac{c^2 d}{2m\mathrm{Var}(L_k) + 2Mc\sqrt{d}/3} \right) \\
&\leq \exp\left(-\frac{c^2 d}{\frac{8m}{n(n-1)} + 8cd/3n} \right) \\
&\leq \exp(-O(cn)).
\end{aligned}$$

So taking a union bound over $u, v \in T$, the contribution of light couples is bounded by $c\sqrt{d}$ with probability at least $1 - e^{-O(n)}$.

Bound on the contribution of heavy couples. As shown in [16], if the random graph satisfies the bounded degree and discrepancy properties, then the contribution of heavy couples is bounded by $O(\sqrt{d})$. We next define these two properties and prove that they hold.

Bounded degree property. We say the *bounded degree property* holds if every vertex has a degree bounded by $c_1 d$ (for some $c_1 > 1$). Using the fact $d_i \sim B(m, \mu),\ \mu = 2/n$, together with Lemma 3 and $m > n \log n/2$, we have

$$P(d_i \geq 6m/n) \leq e^{-m\mu(3\log(3)-2)} \leq e^{-\frac{4m}{n}} \leq n^{-2}.$$

So taking a union bound over i, we get with probability at least $1 - 1/n,\ \forall i,\ d_i \leq 3d$.

Discrepancy property. Let $A, B \subseteq [n]$ be disjoint and $e(A,B)$ be a random variable which denotes the number of edges between A and B. Then $e(A,B) \sim B(m, \frac{|A|\cdot|B|}{C_n^2})$. So, $\mu(A,B) = p|A|\cdot|B|$, with $p = \frac{m}{C_n^2} = \frac{d}{n-1}$, is the expected value of $e(A,B)$. Let $\lambda(A,B) = e(A,B)/\mu(A,B)$. We say that the *dispcrepancy property* holds if there exists a constant c such that for all $A, B \subseteq [n]$ with $|B| \geq |A|$ one of the following holds:

1. $\lambda(A,B) \leq 4$,
2. $e(A,B)\cdot \log \lambda(A,B) \leq c\cdot |B|\cdot \log \frac{n}{|B|}$.

We will show that the discrepancy property holds with probability of at least $1-2/n$. Write $a = |A|, b = |B|$, and suppose $b \geq a$. We assume the bounded degree property holds with $c_1 = 3$ here.

Case 1: $b > 3n/4$. Then $\mu(A,B) = ab\cdot \frac{d}{n-1} \geq \frac{3ad}{4}$. While $e(A,B) \leq a\cdot 3d$ as each vertex in A has degree bounded by $3d$, so $\lambda(A,B) \leq 4$.

Case 2: $b \leq 3n/4$. Using the fact $e(A,B) \sim B(m,q)$, with $q = \frac{ab}{C_n^2}$ and Lemma 3, we have for $k \geq 2$,

$$P(e(A,B) \geq k\mu(A,B)) \leq e^{-mq(k\log(k)+(1-k))} \leq e^{-\mu(A,B)k\log(k)/4}.$$

Then the union bound over all A, B with size a, b for $e(A,B) \geq k\mu(A,B)$ is

$$C_n^a C_n^b e^{-\mu(A,B)k\log(k)/4} \leq e^{-\mu(A,B)k\log(k)/4} \left(\frac{ne}{a}\right)^a \left(\frac{ne}{b}\right)^b. \tag{A.1}$$

We want the right hand side of (A.1) to be smaller than $1/n^3$, so it is enough to let

$$\mu(A,B)k\log(k)/4 \geq a\left(1+\log\frac{n}{a}\right) + b\left(1+\log\frac{n}{b}\right) + 3\log n. \tag{A.2}$$

Next, we are to give an upper bound for the right hand side of (A.2). Using the fact $x\log(n/x)$ is increasing in $(0, n/e)$ and decreasing in $(n/e, 3n/4)$, we have for $b \leq n/e$,

$$a\left(1+\log\frac{n}{a}\right) + b\left(1+\log\frac{n}{b}\right) + 3\log n \leq 4b\log\frac{n}{b} + 3\log n \leq 7b\log\frac{n}{b},$$

and for $3n/4 \geq b > n/e$,

$$\begin{aligned} a\left(1+\log\frac{n}{a}\right) + b\left(1+\log\frac{n}{b}\right) + 3\log n &\leq 2\cdot 3n/4 + 2\cdot n/e + 3\log n \\ &\leq 11\cdot 3/4\cdot \log(4/3)n \\ &\leq 11b\log\frac{n}{b}. \end{aligned}$$

Therefore to make (A.2) valid, it suffices to assume $k\log k > \frac{44}{\mu(A,B)} b\log\frac{n}{b}$. Let $k_0 \geq 2$ be the minimal number that satisfies this inequality. Using the union bound over all the possible a, b, we get the following conclusion. With probability of at least $1 - 1/n$, for every choice of A, B ($b \leq 3n/4$) the following holds,

$$e(A,B) \leq k_0\mu(A,B).$$

If $k_0 = 2$ then we are done, otherwise $k_0\log(k_0)\mu(A,B) = O(1)b\log\frac{n}{b}$, so

$$e(A,B)\cdot\log\lambda(A,B) \leq e(A,B)\log(k_0) \leq k_0\log(k_0)\mu(A,B) = O(1)|B|\log\frac{n}{|B|},$$

as desired to satisfy the second condition for the discrepancy property. □

Proof of Theorem 1. The proof for $G(n,m)$ follows from [15], which establishes the result for the Erdös–Rényi random model $G(n,p)$. Here we follow the same idea for $G_0(n,m)$, using an exponential concentration inequality (Lemma 3) to derive lower and upper bounds on $d_{\min}$. The lower bound is directly from a union bound of independent argument. The upper bound deals with weak dependence using the Chebyshev–Markov inequality.

Using Lemma 1, we only need to study the asymptotic limit of $\frac{d_{\min}}{2m/n}$. As $d_i \sim B(m,\mu), \mu = 2/n$, using Lemma 3 we have

$$P(d_i \leq 2am/n) \leq e^{-m\mu(a\log(a)+(1-a))} = e^{\frac{2m}{n}\mathcal{H}(a)}. \tag{A.3}$$

where $\mathcal{H}(a) = a - a\log(a) - 1$.

In the other direction, suppose $i_0 = 2am/n$ is an integer, we have

$$\begin{aligned} P(d_i \leq 2am/n) &\geq C_m^{i_0}(2/n)^{i_0}(1-2/n)^{m-i_0} \\ &\geq \frac{(m-i_0)^{i_0}}{e\sqrt{i_0}(i_0/e)^{i_0}}(2/(n-2))^{i_0}(1-2/n)^m \\ &= \frac{1}{e\sqrt{i_0}}e^{i_0(1+\log((n/a-2)/(n-2)))+m\log(1-2/n)} \\ &\geq \frac{1}{e\sqrt{i_0}}e^{i_0(1-\log(a))-2m/n-4m/n^2} \\ &> \frac{1}{e^3\sqrt{i_0}}e^{i_0(1-\log(a)-1/a)} \\ &= \frac{1}{e^3\sqrt{a\frac{2m}{n}}}e^{\frac{2m}{n}\mathcal{H}(a)}. \end{aligned} \tag{A.4}$$

The inequality in (A.4) follows from $\log(1-x) \geq -x-x^2$ for all $x \in [0,2/3]$ and the assumption that $n \geq 3$. The last inequality is due to $m < n^2/2$.

If $2am/n$ is not an integer, we can still have

$$P(d_i \leq 2am/n) \geq \frac{c}{\sqrt{2m/n}}e^{\frac{2m}{n}\mathcal{H}(a)}. \tag{A.5}$$

Equations (A.3) and (A.5) give an estimate for $P(d_i \leq 2am/n)$.

Now let $a^{\pm} = a(p_0) \pm 1/\sqrt{2m/n}$, Taylor's theorem gives

$$\mathcal{H}(a^{\pm}) = \mathcal{H}(a(p_0)) \pm \frac{\mathcal{H}'(a_{\pm})}{\sqrt{2m/n}},$$

where $a_+ \in (a(p_0), a(p_0) + 1/\sqrt{2m/n})$, $a_- \in (a(p_0) - 1/\sqrt{2m/n}, a(p_0))$, and $\mathcal{H}'(a) = -\log(a)$ is the derivative of $\mathcal{H}$. Note $p_0\mathcal{H}(a(p_0)) = -1$, so

$$P\left(d_{\min} \leq a(p_0)\frac{2m}{n} - \sqrt{\frac{2m}{n}}\right) \leq ne^{(2m/n)\mathcal{H}(a^-)} = e^{-\sqrt{2m/n}\mathcal{H}'(a_-)} = O(e^{-\Omega(\sqrt{2m/n})}).$$

Therefore, with probability at least $1 - O(e^{-\Omega(\sqrt{2m/n})})$,

$$d_{\min} \geq a(p_0)\frac{2m}{n} - \sqrt{\frac{2m}{n}}. \tag{A.6}$$

Now we prove the reverse direction. Let $f_n = P(d_{\min} \leq a^+ \frac{2m}{n})$, $X_i = \mathbf{1}_{\{d_i \leq a^+ \frac{2m}{n}\}}$ and $N_0 = \sum_{i=1}^n X_i$. So $\mu_0 = \mathbb{E}N_0 = nf_n$. Chebyshev's inequality implies that

$$P(|N_0 - \mu_0| > nf_n/2) \leq \frac{4\mathrm{Var}(N_0)}{n^2 f_n^2}, \tag{A.7}$$

$$\mathrm{Var}(N_0) = \sum_{i=1}^n \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j) = nf_n(1-f_n) + n(n-1)\mathrm{Cov}(X_1, X_2).$$

Next, we are going to claim

$$\mathrm{Cov}(X_1, X_2) \leq O(1)pf_n^2,$$

i.e. $P(X_1 = 1, X_2 = 1) \leq (1 + O(1)p)f_n^2$.

It is enough to prove $\forall k \leq a^+ \frac{2m}{n}$

$$P(d_1 \leq a^+ \frac{2m}{n} | d_2 = k) \leq (1 + O(1)p)f_n.$$

Note the conditional distribution of d_1 given $d_2 = k$ is

$$\begin{aligned}
&B(k, 1/(n-1)) + B(m-k, 2/(n-1)),\\
&P(d_1 \leq a^+ \frac{2m}{n} | d_2 = k)\\
&= \sum_{i=0}^{k} C_k^i (\frac{1}{n-1})^i (\frac{n-2}{n-1})^{k-i} \sum_{j=0}^{a^+ \frac{2m}{n} - i} C_{m-k}^j (\frac{2}{n-1})^j (\frac{n-3}{n-1})^{m-k-j}\\
&= \sum_{s=0}^{a^+ \frac{2m}{p} n} \sum_{i=0}^{k \wedge s} C_k^i (\frac{1}{n-1})^i (\frac{n-2}{n-1})^{k-i} C_{m-k}^{s-i} (\frac{2}{n-1})^{s-i} (\frac{n-3}{n-1})^{m-k-s+i}\\
&= \sum_{s=0}^{a^+ \frac{2m}{n}} \sum_{i=0}^{k \wedge s} C_k^i C_{m-k}^{s-i} (\frac{1}{2})^i (\frac{n-2}{n-3})^{k-i} (\frac{2}{n-1})^s (\frac{n-3}{n-1})^{m-s}\\
&\leq \sum_{s=0}^{a^+ \frac{2m}{n}} \sum_{i=0}^{k \wedge s} C_k^i C_{m-k}^{s-i} (\frac{n-2}{n-3})^k (\frac{n}{n-1})^s (\frac{2}{n})^s (\frac{n-2}{n})^{m-s}\\
&\leq \sum_{s=0}^{a^+ \frac{2m}{n}} C_m^s (\frac{n}{n-1})^{s+k} (\frac{2}{n})^s (\frac{n-2}{n})^{m-s}\\
&\leq (\frac{n}{n-1})^{2a^+ \frac{2m}{n}} \sum_{s=0}^{a^+ \frac{2m}{n}} C_m^s (\frac{2}{n})^s (\frac{n-2}{n})^{m-s}\\
&= (1 + O(1)p)f_n.
\end{aligned}$$

Hence, we get $\mathrm{Var}(N_0) \leq nf_n + O(1)n^2 f_n^2 p$, and (A.7) gives

$$P(|N_0 - \mu_0| > nf_n/2) \leq \frac{4}{nf_n} + 4O(1)p.$$

Note that $nf_n \geq \frac{c}{\sqrt{2m/n}} e^{\sqrt{2m/n}\mathcal{H}'(a_+)} \to \infty$, so with probability at least $1 - O(e^{-\Omega(\sqrt{2m/n})})$, the graph has at least $nf_n/2 \to \infty$ vertices satisfying

$$d_i \leq a(p_0)\frac{2m}{n} + \sqrt{\frac{2m}{n}}.$$

Clearly $d_{\min}$ also satisfies this statement. Combining this result with (A.6), we have that with probability at least $1 - O(e^{-\Omega(\sqrt{2m/n})})$,

$$|d_{\min} - a(p_0)\frac{2m}{n}| \leq \sqrt{\frac{2m}{n}},$$

as desired. □

References

[1] A. Eichhorn, P. Ni, R. Eg, Randomised pair comparison: an economic and robust method for audiovisual quality assessment, in: The Twentieth International Workshop on Network and Operating Systems Support for Digital Audio and Video, 2010, pp. 63–68.
[2] Q. Xu, T. Jiang, Y. Yao, Q. Huang, B. Yan, W. Lin, Random partial paired comparison for subjective video quality assessment via HodgeRank, in: ACM Multimedia, 2011, pp. 393–402.
[3] Q. Xu, Q. Huang, T. Jiang, B. Yan, W. Lin, Y. Yao, HodgeRank on random graphs for subjective video quality assessment, IEEE Trans. Multimedia 14 (3) (2012) 844–857.
[4] X. Jiang, L.-H. Lim, Y. Yao, Y. Ye, Statistical ranking and combinatorial Hodge theory, Math. Program. 127 (1) (2011) 203–244.
[5] A.N. Hirani, K. Kalyanaraman, S. Watts, Least squares ranking on graphs, arXiv:1011.1716v4, 2011.
[6] B. Osting, J. Darbon, S. Osher, Statistical ranking using the L_1-norm on graphs, AIMS J. Inverse Probl. Imaging 7 (3) (2013) 907–926.
[7] B. Osting, C. Brune, S. Osher, Enhanced statistical rankings via targeted data collection, in: International Conference on Machine Learning, 2013, pp. 489–497.
[8] B. Osting, C. Brune, S. Osher, Optimal data collection for informative rankings expose well-connected graphs, J. Mach. Learn. Res. 15 (2014) 2981–3012.
[9] O. Candogan, I. Menache, A. Ozdaglar, P.A. Parrilo, Flows and decompositions of games: harmonic and potential games, Math. Oper. Res. 36 (3) (2011) 474–503.
[10] J. Yuan, G. Steidl, C. Schnorr, Convex Hodge decomposition and regularization of image flows, J. Math. Imaging Vision 33 (2) (2009) 169–177.
[11] W. Ma, J.M. Morel, S. Osher, A. Chien, An L_1-based variational model for retinex theory and its application to medical images, in: International Conference on Computer Vision Pattern Recognition, 2011, pp. 153–160.
[12] A.J. Chorin, J.E. Marsden, A mathematical introduction to fluid mechanics, in: Texts in Applied Mathematics, Springer, 1993.
[13] P. Erdös, A. Rényi, On random graphs i, Publ. Math. Debrecen 6 (1959) 290–297.
[14] M. Kahle, Topology of random clique complexes, Discrete Math. 309 (2009) 1658–1671.
[15] T. Kolokolnikov, B. Osting, J. von Brecht, Algebraic connectivity of Erdös–Rényi graphs near the connectivity threshold, UCLA CAM Report 16-14, ftp://ftp.math.ucla.edu/pub/camreport/cam16-14.pdf, 2015.
[16] U. Feige, E. Ofek, Spectral techniques applied to sparse random graphs, Random Structures Algorithms 27 (2) (2005) 251–275.
[17] A. Sorokin, D. Forsyth, Utility data annotation with Amazon Mechanical Turk, in: Computer Vision and Pattern Recognition Workshops, 2008, pp. 1–8.
[18] S. Nowak, S. Rüger, How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation, in: International Conference on Multimedia Information Retrieval, 2010, pp. 557–566.
[19] O. Alonso, D.E. Rose, B. Stewart, Crowdsourcing for relevance evaluation, ACM SIGIR Forum 42 (2) (2008) 9–15.
[20] A. Kittur, E. Chi, B. Suh, Crowdsourcing user studies with Mechanical Turk, in: SIGCHI Conference on Human Factors in Computing Systems, 2008, pp. 453–456.
[21] M. Soleymani, M. Caro, E. Schmidt, C. Sha, Y. Yang, 1000 songs for emotional analysis of music, in: ACM International Workshop on Crowdsourcing for Multimedia, 2013, pp. 1–6.
[22] G. Tavares, A. Mourao, J. Magalhaes, Crowdsourcing for affective-interaction in computer games, in: ACM International Workshop on Crowdsourcing for Multimedia, 2013, pp. 7–12.
[23] K. Chen, C. Wu, Y. Chang, C. Lei, A crowdsourceable QoE evaluation framework for multimedia content, in: ACM Multimedia, 2009, pp. 491–500.
[24] R. Snow, B. O'Connor, D. Jurafsky, A. Ng, Cheap and fast|but is it good?: evaluating non-expert annotations for natural language tasks, in: Conference on Empirical Methods on Natural Language Processing, 2008.
[25] P.Y. Hsueh, P. Melville, V. Sindhwani, Data quality from crowdsourcing: a study of annotation selection criteria, in: Workshop on Active Learning for Natural Language Processing, 2009.

[26] T. Saaty, A scaling method for priorities in hierarchical structures, J. Math. Psych. 15 (3) (1977) 234–281.
[27] R. Herbrich, T. Graepel, K. Obermayer, Large Margin Rank Boundaries for Ordinal Regression, MIT Press, 2000.
[28] K. Arrow, A difficulty in the concept of social welfare, J. Polit. Econ. 58 (4) (1950) 328–346.
[29] M. Kendall, B. Smith, On the method of paired comparisons, Biometrika 31 (3–4) (1940) 324–345.
[30] S. Negahban, S. Oh, D. Shah, Iterative ranking from pair-wise comparisons, in: Advances in Neural Information Processing Systems, 2012, pp. 2483–2491.
[31] C. Dwork, R. Kumar, M. Naor, D. Sivakumar, Rank aggregation methods for the web, in: International Conference on World Wide Web, 2001, pp. 613–622.
[32] J.C. de Borda, Mémoire sur les élections au scrutin, in: Histoire de l'Academie Royale des Sciences, 1781.
[33] A. Rajkumar, S. Agarwal, A statistical convergence perspective of algorithms for rank aggregation from pairwise data, in: International Conference on Machine Learning, 2014, pp. 118–126.
[34] H. Simon, G. Lea, Problem solving and rule induction: a unified view, Knowl. Cogn. 5 (1974) 105–127.
[35] P.H. Winston, Learning structural descriptions from examples, Technical report AI-TR-231, MIT, Cambridge, 1970.
[36] D. Cohn, L. Atlas, R. Ladner, Improved generalization with active learning, Mach. Learn. 15 (1994) 201–221.
[37] A. Kapoor, K. Grauman, R. Urtasun, T. Darrell, Active learning with Gaussian processes for object categorization, in: International Conference on Computer Vision, 2007, pp. 1–8.
[38] P.H. Gosselin, M. Cord, Active learning methods for interactive image retrieval, IEEE Trans. Image Process. 17 (7) (2008) 1200–1211.
[39] X.S. Zhou, T.S. Huang, Relevance feedback in image retrieval: a comprehensive review, Multimedia Syst. 8 (6) (2003) 536–544.
[40] R. Yan, L. Yang, A. Hauptmann, Automatically labeling video data using multi-class active learning, in: International Conference on Computer Vision, 2003, pp. 516–523.
[41] B. Collins, J. Deng, K. Li, F. Li, Towards scalable dataset construction: an active learning approach, in: European Conference on Computer Vision, 2008, pp. 86–98.
[42] D. Batra, A. Kowdle, D. Parikh, J. Luo, T. Chen, Icoseg: interactive co-segmentation with intelligent scribble guidance, in: International Conference on Computer Vision and Pattern Recogintion, 2010.
[43] S. Vijayanarasimhan, P. Jain, K. Grauman, Far-sighted active learning on a budget for image and video recognition, in: International Conference on Computer Vision and Pattern Recogintion, 2010, pp. 3035–3042.
[44] K.G. Jamieson, R.D. Nowak, Active ranking using pairwise comparisons, in: Advances in Neural Information Processing Systems, 2011, pp. 2240–2248.
[45] K.G. Jamieson, R.D. Nowak, Active ranking in practice: general ranking functions with sample complexity bounds, in: Advances in Neural Information Processing Systems Workshop, 2011.
[46] N. Ailon, An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity, J. Mach. Learn. Res. 13 (2012) 137–164.
[47] X. Chen, P.N. Bennett, K. Collins-Thompson, E. Horvitz, Pairwise ranking aggregation in a crowdsourced setting, in: ACM International Conference on Web Search and Data Mining, 2013, pp. 193–202.
[48] X. Chen, Q. Lin, D. Zhou, Statistical decision making for optimal budget allocation in crowd labeling, J. Mach. Learn. Res. 16 (1) (2015) 1–46.
[49] G.H. Golub, C.F.V. Loan, Matrix Computations, 3rd edition, The John Hopkins University Press, 1996.
[50] A. Ghosh, S. Boyd, Growing well-connected graphs, in: IEEE Conference on Decision and Control, 2006, pp. 6605–6611.
[51] LIVE image & video quality assessment database, http://live.ece.utexas.edu/research/quality/, 2008.
[52] Subjective quality assessment irccyn/ivc database, http://www2.irccyn.ec-nantes.fr/ivcdb/, 2005.
[53] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Amer. Statist. Assoc. 58 (301) (1963) 13–30.