Online HodgeRank on Random Graphs for Crowdsourceable QoE Evaluation

Qianqian Xu, Jiechao Xiong, Qingming Huang*, *Senior Member, IEEE,* and Yuan Yao*

Abstract

HodgeRank on random graphs is proposed recently as an effective framework for multimedia quality assessment problem based on paired comparison methods. With a random design on graphs, it is particularly suitable for large scale crowdsourcing experiments on the Internet. However, there still lacks a systematic study about online schemes to deal with the rising streaming and massive data in crowdsourceable scenarios. To fill in this gap, we propose in this paper an online ranking/rating scheme based on stochastic approximation of HodgeRank on random graphs for Quality of Experience (QoE) evaluation, where assessors and rating pairs enter the system in a sequential or streaming way. The scheme is shown in both theory and experiments to be efficient in obtaining global ranking by exhibiting the same asymptotic performance as batch HodgeRank under a general edge-independent sampling process. Moreover, the proposed framework enables us to monitor topological changement and triangular inconsistency in real time. Among a wide spectrum of choices, two particular types of random graphs are studied in detail, *i.e.*, Erdös-Rényi random graph and preferential attachment random graph. The former is the simplest I.I.D. (independent and identically distributed) sampling and the latter may achieve more efficient performance in ranking the top-k items due to its Rich-get-Richer property. We demonstrate the effectiveness of the proposed framework on LIVE and IVC databases.

*Corresponding author.

Q. Xu is with BICMR, Peking University & University of Chinese Academy of Sciences, Beijing 100871, China, (e-mail: qqxu@jdl.ac.cn).

J. Xiong and Y. Yao are with BICMR-LMAM-LMEQF-LMP, School of Mathematical Sciences, Peking University, Beijing 100871, China, (e-mail: xiongjiechao@pku.edu.cn; yuany@math.pku.edu.cn).

Q. Huang is with the University of Chinese Academy of Sciences & Institute of Computing Technology of Chinese Academy of Sciences, Beijing 100190, China, (e-mail: qmhuang@jdl.ac.cn).

EDICS: 3-QAUE Quality Assessment and User Experience.

Index Terms

Quality of Experience; Crowdsourcing; Paired Comparison; Online Algorithms; Robbins-Monro Procedure; Stochastic Approximation; Hodge Theory; Random Graphs; Persistent Homology

I. INTRODUCTION

The Quality of Experience (QoE) issue [1], [2], which aims at the assessment of a user's subjective expectation, feeling, perception, and satisfaction with respect to multimedia content, has drawn increasing attention from multimedia researchers during recent years. As the ultimate goal is to provide a satisfying end-user experience, there is a strong demand to investigate a technique that can measure the quality of multimedia content efficiently, reliably, and is easy to implement in reality. Since QoE results from the psychological fulfillment of the user's expectations on the utility and enjoyment of the multimedia content given the user's personality and current state, traditional subjective user studies are conducted in a laboratory environment with a tight control on influential variables. While many and possibly even diverging views on the quality of the multimedia content can be taken into account – entailing a good understanding of the QoE and its sensitivity – lab-studies can be time-consuming and costly, since the tests have to be conducted by a large number of users for statistically relevant results. Crowdsourcing arises to be a promising alternative approach. With crowdsourcing, subjective user studies can be efficiently conducted at low costs with adequate user numbers to get statistically significant QoE ranking scores. In addition, the desktop-PC based setting of crowdsourcing provides a highly realistic setting for assessing the rapidly growing online multimedia data such as Flickr and Youtube which is nearly impossible in traditional lab-studies of QoE [3], [4]. However, additional challenges emerge due to the remote test settings, among which we focus here on experimental designs with distributive sampling, reliability of users or ratings, divergent expectations of users, and the treatment of big and streaming data.

Paired comparison method is gaining rising attention in QoE recently [5], [6], since compared with the Mean Opinion Score (MOS) [7] rating scheme, it is an easier and less demanding task for raters, yielding more reliable rating data in crowdsourcing tests. In a typical MOS test, individuals are asked to give a rating from Bad to Excellent (e.g. Bad-1, Poor-2, Fair-3, Good-4, and Excellent-5) to grade the quality of a stimulus, which however may suffer from various problems such as ambiguity and even divergence in defining the scales. On the other hand, in paired comparison method raters are asked to compare two stimuli simultaneously and vote which one has the better quality, which provides more accurate results against personal scale variations. However, paired comparison method leaves a heavier burden on participants with a larger number $\binom{n}{2}$ of comparisons. Here, n represents the number of items to be assessed. While there has been a large volume of statistical literature on deterministic incomplete block design [8], these designs are not suitable for crowdsourcing on the Internet where the raters are distributive over Internet with varied backgrounds and it is hard to control with traditional designs.

To address such a challenge, we recently propose a new framework [9], [10], called HodgeRank on Random Graphs (HRRG), which exploits randomized paired comparison method [6] based on random graph theory where small subsets of all possible pairs are randomly chosen for each assessor to view. In [9], [10], we systematically answer the following two fundamental questions arising from randomization: (1) how to deal with the imbalanced and incomplete data distributed on random graphs; (2) how many samples are needed to achieve certain approximation of the complete design. Our framework exploits a recent development on a Hodge-theoretic approach to statistical ranking [11], which decomposes paired comparison data as edge flows on graphs orthogonally into three components: a gradient flow which provides a global ranking score, a triangular curl flow, and as well as a harmonic flow, both of which characterize the local and global inconsistency in the data, respectively. Such a decomposition enables us to get global ranking and investigate the reliability of pairwise ratings simultaneously. Random graphs play a central role in guiding random sampling designs for crowdsourcing experiments. For example, Erdös-Rényi random graphs select pairs of stimuli uniformly from all possible candidates, while random k-regular graphs keep a balanced sampling where each stimulus receives the same number of comparisons against others, which is important for numerical stability of global ranking [10]. Equipped with recent developments in random graph theory, $O(n \log n)$ 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. Experiments show that such a random design provides good approximations of global ratings derived from complete experimental designs.

Despite the successful developments above for subjective multimedia assessment, it remains open to explore *online algorithms* to deal with streaming data in crowdsourcing experiments on the Internet. Although most of current QoE datasets are of medium sizes which are suitable for both laboratory and crowdsourcing studies, we are witnessing a rapid growth of online multimedia data such as Flickr and Youtube with big and streaming data [3]. Such data calls for online algorithms as a sequential decision process via incremental data updates to improve its prediction accuracy which is scalable for large scale data analysis. Even though the image/video quality itself is constant in time, in subjective QoE evaluation, preferences may vary over raters and comparisons contingent on different salient features of stimuli in attention, noise from environment, and levels of attention, etc. Thus it is a fundamental question in online algorithmic design how to aggregate preferences of multiple sequential assessors into a global ranking, reflecting the statistical consensus on multimedia quality over population.

In this paper, we fill in this gap by presenting an online algorithm based on the classic Robbins-Monro procedure [12] or stochastic approximation of batch HodgeRank [10]. While there has been an extensive study of online rating algorithms in literature [13]–[15], we choose online HodgeRank mainly due to that it systematically addresses the global rating and the inconsistency of paired comparison data simultaneously, particularly suitable for crowdsourcing QoE studies.

Online algorithms could offer significant computational advantages over batch algorithms, when dealing with streaming or large-scale data. In this framework, every item (e.g. image) in comparison is regarded as a graph node and an assessor collects some samples of node pairs or edges, independently and with a fixed distribution which may vary over edges. Such an edge-independent process [16] include two important online random graph models investigated in this paper: (1) Erdös-Rényi random graph which models the simplest I.I.D. sampling scheme; (2) preferential attachment random graph which models the Rich-get-Richer scenario. We will show that our proposed online algorithm converges to the batch HodgeRank algorithm at a minimax optimal rate for all edge-independent sampling processes, and preferential attachment model is particularly useful when one expects higher accuracy and faster convergence on top-ranked items, while tolerates lower accuracy and slower convergence on bottom-ranked ones. Furthermore, we note that online algorithms can be applied to more general settings with Multiplicative-attribute random graphs, dependent sampling such as Markov sampling, and tracking time-varying environment.

We demonstrate the effectiveness and generality of the proposed framework on LIVE [17] and IVC [18] databases, which include 15 different reference images and 15 distorted versions of each reference, being widely studied in laboratory settings and here online crowdsourcing settings. Totally 186 observers have carried out the crowdsourcing tests via Internet, providing us 23,097 paired comparisons. Experimental results show that the proposed online algorithm can save the computational time-cost in magnitudes, while provides nearly the same error rates as the batch HodgeRank where all the samples in hand are processed once. Thus online HodgeRank is promising and has potentially wide applications for large scale crowdsourceable QoE evaluation.

Our contributions in this work are three-fold:

1. We propose a novel framework of online ranking/rating on random graphs for exploratory quality assessment. The framework provides the possibility of making assessment procedure significantly faster without deteriorating the accuracy, while maintaining the freedom of assessors.

2. The online rating algorithm is based on Robbins-Monro procedure or stochastic gradient descent for HodgeRank on random graphs. For every edge independent sampling process, the online rating reaches minimax optimal convergence rates hence asymptotically as efficient as a batch algorithm. Moreover, online tracking of ranking inconsistency is possible via triangular curl and persistent homology in this framework.

3. To conduct paired comparisons, two random design schemes are proposed based on Erdös-Rényi random graph and preferential attachment random graph. For Erdös-Rényi random graph, it further confirms the theoretical analysis by showing that the proposed online rating algorithm could achieve similar convergences to batch algorithms. For preferential attachment random graph, it may lead to better performance for top-*k* ranking items in HodgeRank than Erdös-Rényi random graph due to its Rich-get-Richer property.

This paper is an extension of our conference paper [19], which only studies HodgeRank with Erdös-Rényi random graph. The following distinctions are made in this paper: A new random graph called preferential attachment random graph is systematically studied in this version. The reason to choose preferential attachment random graph is that it could provide a more efficient ranking process for the top-*k* items which is important in various applications such as coding strategy and parameters selection in image/video coding community. Both random graph models favor an online fashion to generate samples. Erdös-Rényi random graph has all edges sampled independently and with an identical distribution, thus is the simplest example of I.I.D sampling, while the original preferential attachment random graph has an dependent sampling process. In this paper we adopt an edge-independent implementation of preferential attachment random graph suggested in [16], which concentrates on top-ranked items and is a special case of multiplicative-attribute random graphs [20].

The remainder of this paper is organized as follows. Section II contains a review of related work. Then we describe the proposed framework in Section III, and establish the theory of online HodgeRank based on batch HodgeRank. The detailed experiments are demonstrated in Section IV, including simulated examples and real-world data. Section V presents the conclusive remarks along with discussion for future work.

II. Related Work

A. Crowdsourcing QoE

Existing methods of QoE evaluation can be divided into two categories: subjective assessment and objective assessment. *Objective* assessment builds objective quality measurement models (refer to survey

[21] and references therein) to predict perceived quality automatically and intelligently, which may or may not reflect human's perceptual experience. On the other hand, *subjective* assessment can provide the ground-truth and verification for objective models, which is however labor-intensive and time-consuming.

In subjective viewing tests, stimuli are shown to a group of viewers, and then their opinions are recorded and averaged to evaluate the quality. Among various approaches of conducting subjective test, paired comparison is expected to yield more reliable results; however, this is an expensive and time-consuming process. To tackle the cost problem, with the growth of crowdsourcing platforms, such as Amazon Mechanical Turk (MTurk) [22], more and more researchers tend to seek help from the Internet crowd to conduct user studies on QoE evaluation [5], [9], [10], [19]. In [5], a crowdsourceable framework based on paired comparison is first proposed for QoE evaluation. However, one major shortcoming of this work lies in that it makes a strong assumption that all paired comparison data collected are complete which is impossible in practice. To address this issue, the work in [6], [9], [10], [19] all suggest a randomized paired comparison method to reduce the number of comparisons. However, crowdsourceable data are collected in a distributive and streaming way from a large population over Internet participants. Therefore, it is necessary to develop an online rating method to deal with this kind of data, which will be the main concern of this paper.

B. HodgeRank and Online Algorithms

HodgeRank, as an application of combinatorial Hodge theory to preference or rank aggregation problem, is firstly introduced in [11], which inspires a series of studies in statistical ranking [23]–[26] and game theory [27]. Most recently, we developed the application of HodgeRank with random graph designs in subjective QoE evaluation [9], [10], together with online algorithms for sequential data [19] and outlier detection [28]. Other applications of Hodge theory includes fluid mechanics [29] and computer vision [30], [31], etc.

Online learning is a well established subfield of machine learning concerned with estimation problems with limited access to the entire data. It is a sequential decision process $(f_t)_{t\in N}$ in the hypothesis space, where each f_t is decided by the current observation $z_t = (x_t, y_t)$ and f_{t-1} which only depends on previous examples, i.e. $f_t = T_t(f_{t-1}, z_t)$. As a contrast, batch learning refers to a decision utilizing the whole set of examples available at time t [32], [33]. The most famous examples of online learning algorithms can be traced back to Perceptrons [34] in classification, Adaline [35] in regression, and Kalman-Bucy filters as recursive least square methods [36], [37], etc. The online ranking scheme in this paper is based on stochastic gradient decent method in the setting of HodgeRank on random graphs, also called the

Robbins-Monro procedure [12] for mean normal equations in least square ranking, and can reach optimal convergence rates as batch learning algorithms [38].

The most famous online algorithm for ranking is probably the Elo rating system, developed by Arpad Elo [13] and adopted by United States Chess Federation in 1960. Such a rating system is later generalized by Glickman [14], [39] using Bayesian inference, which is further extended in TrueSkill [15] to multiple-team players and implemented in online games by Microsoft Co. Ltd. The scheme proposed in this paper is equivalent to the Elo rating when restricted to uniform models in both cases. However adapted from the Hodge decomposition, our scheme unifies in the same framework various statistical general linear models [10], such as Thurstone-Mosteller, Bradley-Terry, and Angular Transform etc., with a decomposition of paired comparison data into both gradient flow of global ranking and (local and global) cyclic flows as measurements of inconsistency. The latter component is ignored in the ranking algorithms above but crucial in preference aggregation as highlighted by the Arrow's impossibility theorem in economics.

Recently, there arose various studies on active sampling in ranking. Most of these are concerned with sample complexity. For example, Ailon [40] 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. Moreover, if the ranking function is decided by a Euclidean distance function from a reference point in $\mathbb{R}^d$ or a linear function in such a space, [41] shows the active sampling complexity can be reduced to $O(d \log n)$, which are successfully applied in beer taste [42] etc. Most recently, [25] approaches active sampling from a statistical perspective as Fisher information maximization, which is equivalent to maximize the smallest nonzero eigenvalue of graph Laplacian in HodgeRank with integer weights. However these works are different to preferential attachment random graph models in this paper, which is an active sampling scheme for the purpose of pursuing top-k ranking effectively.

C. Random Graphs

Random graph is a graph generated by some random process [16], [43]. It starts with a set of n vertices and adds edges between them at random. With such models we aim at crowdsourcing experimental designs where assessors may select image/video pairs at random. Different random graph models produce different probability distributions on graphs. The most commonly studied one is the Erdös-Rényi random graph [44] which is a stochastic process that starts with n vertices and no edges, and at each step adds one new edge uniformly. This kind of random graph can be viewed as a random sampling process of image/video pairs or edges independently and identically distributed (I.I.D.), and thus is well suited to our online crowdsourcing test system. In [9], [10], a random design principle based on Erdös-Rényi random graph theory is investigated to conduct crowdsourcing tests. Experimental results show that for large Erdös-Rényi random graph $G(n, q)$ with n nodes and every edge sampled with probability q, it is necessary to have $q \gg n^{-1} \log n$ such that the graph is connected and global ranking is thus possible; to avoid global inconsistency from Hodge Decomposition, it suffices to have larger sampling rates at $q \gg n^{-1/2}$. Moreover, [10] further investigates the sampling based on random k-regular graph, which may obtain a more balanced sampling and hence better performance than Erdös-Rényi random graph for small k, as well as a good approximation to Erdös-Rényi random graph for large k.

There are some other kinds of random models, such as preferential attachment random graph [45], small world random graph [46] and geometric random graph [47], etc., which may also play important roles under certain circumstances. However, in this paper, according to practical application requirements in QoE evaluation, we particularly focus on two types of them, Erdös-Rényi and preferential attachment random graph, leaving other models for future studies.

III. ONLINE HODGERANK FOR QOE

In this section, we propose a new online design to conduct paired comparison for subjective QoE evaluation and two random design principles are exploited to meet the crowdsourcing scenario, including Erdös-Rényi and preferential attachment random graphs. Specifically, we first describe HodgeRank on general graphs, and then explain how to develop the online rating algorithms based on stochastic approximation or Robbins-Monro procedure. Second, an upper bound for convergence of such online rating algorithms is given to justify the settings where the minimax optimal convergence rate is met. Finally, we discuss how to online track triangular curls and topological changement.

A. Batch HodgeRank on Graphs

HodgeRank [11] is a general framework to decompose paired comparison data on graphs, possibly imbalanced (where different pairs may receive different number of comparisons) and incomplete (where every participant may only give partial comparisons), into three orthogonal components:

$$aggregate\ paired\ ranking =$$
$$global\ ranking \bigoplus local\ inconsistency \bigoplus global\ inconsistency$$

To be precise, consider paired ranking data on a graph $G = (V, E)$, $Y_\alpha : E \to \mathbb{R}$ such that $Y_{ij}^\alpha = -Y_{ji}^\alpha$ where α is the participant index. Without loss of generality, one assumes that $Y_{ij}^\alpha > 0$ if α prefers i to

j and $Y_{ij}^{\alpha} \leq 0$ otherwise, with the magnitude representing the degree of preference. In a dichotomous choice, Y_{ij}^{α} can be taken as $\{\pm 1\}$.

In subjective multimedia assessment, it is natural to assume

$$Y_{ij}^{\alpha} = s_i^* - s_j^* + \varepsilon_{ij}^{\alpha}, \tag{1}$$

where $s^* : V \to \mathbb{R}$ is some true scaling score on V and $\varepsilon_{ij}^{\alpha}$ are independent noise of mean zero and fixed variance.

Under such assumptions, Gauss-Markov theorem tells us that the unbiased estimator of global ranking score $s : V \to \mathbb{R}$, up to a translation degree of freedom for connected graph G, is given by the following least square problem,

$$\min_{s \in \mathbb{R}^{|V|}} \sum_{i,j,\alpha} \omega_{ij}^{\alpha} (s_i - s_j - Y_{ij}^{\alpha})^2, \tag{2}$$

where ω_{ij}^{α} denotes the number of paired comparisons on $\{i, j\}$ made by rater α and s_i, s_j represent the global ranking score of item *i* and *j*, respectively.

It can be rewritten as the following weighted least square form

$$\min_{s \in \mathbb{R}^{|V|}} \sum_{i,j} \omega_{ij} (s_i - s_j - \hat{Y}_{ij})^2, \tag{3}$$

where $\hat{Y}_{ij} = (\sum_{\alpha} \omega_{ij}^{\alpha} Y_{ij}^{\alpha}) / (\sum_{\alpha} \omega_{ij}^{\alpha})$ and $\omega_{ij} = \sum_{\alpha} \omega_{ij}^{\alpha}$. Written in this form allows an extension of the linear model (1) to the following general linear model family when only binary comparisons are available.

In general linear models [8], one assumes that the probability of pairwise preference is fully decided by a linear ranking/rating function in the following way

$$\pi_{ij} = \mathbf{Prob}\{i \text{ is preferred over } j\} = \Phi(\beta_i^* - \beta_j^*), \quad \beta^* \in \mathbb{R}^V \tag{4}$$

where $\Phi : \mathbb{R} \to [0, 1]$ can be chosen as any symmetric cumulated distributed function. In an inverse direction, if an empirical preference probability $\hat{\pi}_{ij}$ is observed in experiments, one can map $\hat{\pi}$ to a skew-symmetric paired comparison data by the inverse of Φ,

$$\hat{Y}_{ij} = \Phi^{-1}(\hat{\pi}_{ij}), \tag{5}$$

where $\hat{Y}_{ij} = -\hat{Y}_{ji}$. Different choices of Φ lead to different general linear models. In [10] we compare four well known models in subjective multimedia QoE assessment: Uniform model (equivalent to the linear model (1)), Thurstone-Mosteller model, Bradley-Terry model, and Angular-Transform model, where the simplest uniform model (the linear model (1)) is nearly the best (slightly worse than the Angular-Transform model) in the setting of that paper. Therefore in the sequel we only focus on the uniform model while the principle can be applied to general linear models.

To characterize the solution and residue of (3), we first define the triangle set of G as all the 3-cliques in G:

$$T = \left\{ \{i,j,k\} \in \begin{pmatrix} V \\ 3 \end{pmatrix} | \{i,j\}, \{j,k\}, \{k,i\} \in E \right\}. \tag{6}$$

Then every $\hat{Y}$ admits an orthogonal decomposition adapted to G

$$\hat{Y} = \hat{Y}^g + \hat{Y}^h + \hat{Y}^c, \tag{7}$$

where

$$\hat{Y}^g_{ij} = \hat{s}_i - \hat{s}_j, \quad \text{for some } \hat{s} \in \mathrm{R}^V, \tag{8}$$

$$\hat{Y}^h_{ij} + \hat{Y}^h_{jk} + \hat{Y}^h_{ki} = 0, \text{ for each } \{i,j,k\} \in T, \tag{9}$$

$$\sum_{j \sim i} \omega_{ij} \hat{Y}^h_{ij} = 0, \text{ for each } i \in V, \tag{10}$$

and the residue $\hat{Y}^c$ actually satisfies (10) but not (9). Here we make some remarks about the Hodge decomposition above.

- $\hat{Y}^g$ satisfies (8) as the discrete gradient of a global ranking score $\hat{s}$, where $\hat{s}$ is given by a solution of the weighted least square problem (3).
- $\hat{Y}^h$ satisfies two conditions, the curl-free condition (9) and the divergence-free condition (10), which is called the harmonic flow and accounts for the global inconsistency of paired comparison data $\hat{Y}$. Global inconsistency generally involves loops consisting all nodes in comparisons (e.g. $i \succ j \succ k \succ$ ... $\succ i$), indicating the fixed tournament issue – arbitrary order can be achieved by manipulating the tournament schedule. Harmonic flow or global inconsistency will vanish if the underlying triangular clique complex is loop-free, i.e. the first Betti number is zero.
- $\hat{Y}^c$ with a non-vanishing curl which fails (9), hence often called the curl flow, accounts for the locally triangular inconsistency of data $\hat{Y}$. Such local inconsistency can be fully characterized by triangular cycles, such as $i \succ j \succ k \succ i$.

Two residues, $\hat{Y}^h$ and $\hat{Y}^c$, as inconsistency measurements associated with the global ranking obtained, show the validity of the ranking and can be further studied in terms of its geometric scale, namely whether inconsistency in the ranking data arises locally or globally. This provides us a quantitative tool to explore reliability of ratings given incomplete data. More details can be found in [9]–[11].

For a connected graph $G = (V, E)$, there is a translation degree of freedom for global ranking score in Eq. (1) and the minimizers of (2). To remove such a degree of freedom, we select the global ranking score estimator as the minimal norm least square solution of (2) which satisfies the following normal equation

$$\Delta_0 \hat{s} = \delta_0^* \hat{Y}, \tag{11}$$

where $\delta_0 : \mathrm{R}^V \to \mathrm{R}^E$ is a finite difference operator (matrix) on G defined by $\delta_0((i,j),i) = -1$, $\delta_0((i,j),j) = 1$, and otherwise zero, $\delta_0^* = \delta_0^T W$ ($W = \mathrm{diag}(\omega_{ij})$), $\Delta_0 = \delta_0^* \cdot \delta_0$ is the unnormalized graph Laplacian defined by $(\Delta_0)_{ii} = \sum_{j \sim i} \omega_{ij}$ and $(\Delta_0)_{ij} = -\omega_{ij}$. In fact, with the Moore-Penrose (pseudo) inverse of graph Laplacian $\Delta_0^\dagger$, we have $\hat{s} = (\Delta_0)^\dagger \delta_0^* \hat{Y}$.

An interesting variation of this l_2-norm scheme (3) is an analogous l_1-projection onto the space of gradient flows,

$$\min_{s \in \mathrm{R}^{|V|}} \sum_{i,j} \omega_{ij} |s_i - s_j - \hat{Y}_{ij}|. \tag{12}$$

This optimization problem is applied to the case that the noise is sparse but can be large, often regarded as outliers. It is more robust to outliers when compared with the l_2-norm, and thus can be regarded as a kind of robust ranking. For more details, readers may refer to [11], [24].

As the input of this HodgeRank framework is a paired comparison multigraph (the whole set of paired comparison data in one batch) provided by participants, we may call this type of work as batch HodgeRank. For details of the theoretical development, readers may refer to [11]. The work in [9], [10] adopt such batch HodgeRank to obtain quality scores of videos. However, for crowdsourcing test on the Internet, participants and pairs enter the system one by one in a dynamic and random way. Therefore, batch HodgeRank is not an efficient tool for crowdsourcing. To meet this challenge, we propose an online HodgeRank as Robins-Monro procedure or stochastic approximation of (11).

B. Online HodgeRank Algorithms

The online rating algorithm considered in this paper is constructed from Robbins-Monro procedure [12] to solve linear operator equation $\bar{A}x = \bar{b}$,

$$x_{t+1} = x_t - \gamma_t (A_t x_t - b_t), \tag{13}$$

where A_t and b_t are matrix- and vector-valued random variables whose expectations satisfy $E(A_t) = \bar{A}$ and $E(b_t) = \bar{b}$, respectively. Such a scheme has been widely exploited in online learning, e.g., [38], [48].

Now consider the normal equation (11) for the least square problem (2), $\Delta_0 s = \delta_0^* \hat{Y}$. In this case, at time t when a new rating $Y_t(i_t, j_t) = -Y_t(j_t, i_t)$ entered with pair (i_t, j_t), we have

- A_t is a $|V| \times |V|$ matrix defined by $A_t(i_t, i_t) = A_t(j_t, j_t) = -A_t(i_t, j_t) = -A_t(j_t, i_t) = 1$ and otherwise zero;
- b_t is a $|V|$-dimensional vector defined by $b_t(i_t) = -b_t(j_t) = Y_t(i_t, j_t)$ and otherwise zero.

Let $s_t = x_t$. With the realization above, Eq. (13) leads to

$$\begin{aligned} s_{t+1}(i_t) &= s_t(i_t) - \gamma_t[s_t(i_t) - s_t(j_t) - Y_t(i_t, j_t)] \\ s_{t+1}(j_t) &= s_t(j_t) + \gamma_t[s_t(i_t) - s_t(j_t) - Y_t(i_t, j_t)] \end{aligned} \tag{14}$$

where the initial choice is $s_0 = 0$ or any vector such that $\sum_i s_0(i) = 0$, and the step size γ_t is a nonnegative sequence whose choice is often taken in the following form

$$\gamma_t = \frac{a}{(t + t_0)^\theta}, \quad \theta \in [0, 1]. \tag{15}$$

The choice of step size will be discussed in more detail in the next subsection with a convergence analysis which shows minimax rates with independent and identically distributed sampling. Algorithm 1 below shows the procedure of this online rating method. Note that updates here only occur locally on the nodes associated with edge $\{i_{t+1}, j_{t+1}\}$, which is suitable for asynchronized parallel implementation.

For the sake of comparison, we also present a stochastic subgradient method for online rating with l_1-norm in (12), which is given by:

$$\begin{aligned} s_{t+1}(i_t) &= s_t(i_t) - \gamma_t \ \mathrm{sign}(s_t(i_t) - s_t(j_t) - Y_t(i_t, j_t)) \\ s_{t+1}(j_t) &= s_t(j_t) + \gamma_t \ \mathrm{sign}(s_t(i_t) - s_t(j_t) - Y_t(i_t, j_t)) \end{aligned}$$

with similar choices on initial score and steps. Compared to online HodgeRank algorithm in Algorithm 1, for l_1-based online algorithm it suffices to consider sign functions, $g_{ij} = \mathrm{sign}(s_t(i_t) - s_t(j_t) - Y_t(i_t, j_t))$.

C. Optimal Convergence Rates of Online HodgeRank

There has been extensive work on convergence analysis of subgradient methods, e.g. [49]. Typical convergence results require the conditions that step sizes $\sum_t \gamma_t^2 < \infty$ while $\sum_t \gamma_t = \infty$, and boundedness of subgradients, which are in particular $s(i) - s(j) - Y(i, j)$ and $\mathrm{sign}(s(i) - s(j) - Y(i, j))$ here. When general convex loss functions are assumed, the analysis is typically formulated as regret bounds [50].

Algorithm 1: Online HodgeRank Procedure.

```
Initialization:
$s_0 = 0$ or any vector such that $\sum_i s_0(i) = 0$;    // Initialize the quality scores of each items.
With a new rating $Y_t(i_t, j_t)$;    // A new paired comparison $(i_t, j_t)$ occurs at time t.
Compute $g_{ij} = s_t(i_t) - s_t(j_t) - Y_t(i_t, j_t)$;
Then
$s_{t+1}(i_t) = s_t(i_t) - \gamma_t * g_{ij}$;
$s_{t+1}(j_t) = s_t(j_t) + \gamma_t * g_{ij}$.    // Quality scores at time t+1.
```

In particular, when the square loss is adopted, one may achieve the following probabilistic upper bound which in fact reaches the minimax optimal rates for parametric regression, up to a logarithmic factor.

In the following, assume that $Y_t(i_t, j_t)$ is an independent and identically distributed (I.I.D.) sequence and the resulting random graph is edge-independent. Recall that a random graph G is called *edge independent* (or *independent*, for short) [16] if there is an edge-weighted function $p : E(K_n) \to [0, 1]$, satisfying

$$p(G = (V, E)) = \prod_{e \in E} p(e) \prod_{e \notin E} (1 - p(e)).$$

Here K_n denotes the complete graph of n vertices. Edge-independent processes allow different occurrence probabilities for different edges, but they have to be independent and static. Many dependent sampling process of random graphs can be well approximated by edge-independent random graphs. The following two examples are particularly important edge-independent random graphs used in this paper.

- Erdös-Rényi random graph [16]: an edge $(i, j) \in E$ is independently drawn at a fixed probability $p_{ij} = p$ uniformly. This is the simplest example of I.I.D. sampling.
- Multiplicative-Attribute random graph [20]: let each vertex $i \in E$ be associated with an attribute parameter $\theta_i \in \mathbb{R}^d$ (e.g. degree or centrality of the vertex), and each edge $(i, j) \in E$ is drawn independently according to a probability $p_{ij} = f(\theta_i, \theta_j)$ with attribute (e.g. degree) parameters θ_i and θ_j affiliated to vertex i and j, respectively. A preferential attachment random graph can be well approximated by such models with θ_i being the expected degree of vertex i [16]. In the experimental section, we will study a sort of preferential attachment random graphs with θ_i as the expected preference of vertex i (image or video).

The convergence analysis can be applied to all these examples. Define a random matrix

$$\Pi_k^t = \begin{cases} (I - \gamma_t A_t) \ldots (I - \gamma_k A_k), & k \leq t; \\ I, & k > t. \end{cases} \tag{16}$$

If we replace A_i by $\bar{A}$, we obtain a deterministic positive definite matrix, say $\bar{\Pi}_k^t$.

The following lemma leads to a martingale decomposition for error $x_t - x^*$, given in [38], [48], which is crucial to lead to the error bounds.

Lemma. For all $t \in \mathbb{N}$,

$$x_t = \Pi_1^{t-1} x_0 + \sum_{k=1}^{t-1} \gamma_k \Pi_{k+1}^{t-1} b_t \tag{17}$$

and

$$x_t - x^* = \bar{\Pi}_1^{t-1}(x_0 - x^*) - \sum_{k=1}^{t-1} \xi_k, \tag{18}$$

where

$$\xi_k = \begin{cases} \gamma_k \bar{\Pi}_{k+1}^{t-1}((A_k - \bar{A})x_k - (b_k - \bar{b})), & 1 \leq k < t; \\ 0, & k \geq t. \end{cases}$$

is a martingale difference sequence such that $E[\xi_t : \mathcal{F}_{t-1}] = 0$ for a filtration $\mathcal{F}_{t-1}$ up to time $t-1$.

The first part in error, $\bar{\Pi}_1^{t-1}(x_0 - x^*)$, is called the *initial error* and the martingale difference tail, $\sum \xi_k$, is called the *sample error*. Initial error can be bounded deterministically, while the sample error can be bounded via a Pinelis-Bernstein probabilistic inequality. Combining these bounds will lead to the following theorem, whose derivation can be found in [26].

Theorem III-C. *Let $G = (V, E, P)$ be the edge independent random graph model such as each edge $\{i, j\} \in E$ is drawn independently with probability $p_{ij} \in [0, 1]$, and $0 = \lambda_0 < \lambda_1 \leq \ldots \leq \lambda_{n-1}$ be eigenvalues of the graph Laplacian $\Delta_0 = E(A_t)$. Assume that $A = 2 \vee \lambda_{n-1}$ and $|Y_t(i,j)| \leq B$. Then there exists a choice of step size $\gamma_t = a/(t + t_0)$ (e.g. $a = 1/\lambda_1$ and $t_0 \geq B/\lambda_1$) such that the following holds for all $t \in \mathbb{N}$ with probability at least $1 - \delta$ ($\delta > 0$),*

$$\|s_t - s^*\|_2 \leq \frac{7\sqrt{A}B|E|}{\lambda_1^{3/2}} t^{-1/2} \log(t + t_0) \cdot \log \frac{2}{\delta}$$

where s_t is defined by (14).

The theorem says that the online rating algorithm converges to the underlying true score s^* under the edge independent sampling process above. The convergence rate is minimax optimal at $O(t^{-1/2})$, as good as batch HodgeRank and Kalman-Bucy filters as recursive least squares. The choice of step size $\gamma_t \sim t^{-1}$ is crucial, with large enough t_0. Although the choice of a and t_0 does not affect the asymptotic

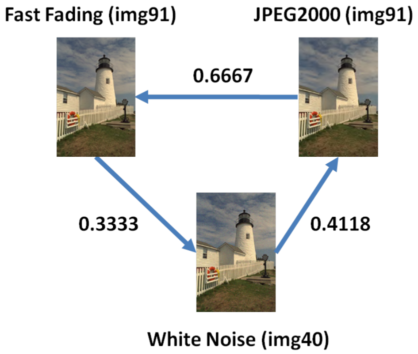


Fig. 1. Large curl due to multicriteria in paired comparisons among users. The image is undistinguishable due to its small size, so image IDs in LIVE database are printed here.

rate in theory, in practice they influence the speed of convergence when t is small. We shall see this in experimental section.

D. Online Tracking of Triangular Curls

Hodge decomposition (7) has a component $\hat{Y}^c$ which satisfies $\hat{Y}^c_{ij} + \hat{Y}^c_{jk} + \hat{Y}^c_{ki} \neq 0$ for each triangle $(i, j, k) \in T$. This encodes the information about triangular or local inconsistency. For a graph $G = (V, E)$ whose 3-clique complex $\chi_G = (V, E, T)$ does not contain a "loop" (i.e. the first Betti number $\beta_1 = 0$), global inconsistency vanishes and such triangular inconsistency explains all sorts of inconsistency. It happens when Erdös-Rényi random graphs are sufficiently dense [9], [10]. Therefore it is desired to track triangular curls:

$$curl_{ijk} = \hat{Y}^c_{ij} + \hat{Y}^c_{jk} + \hat{Y}^c_{ki} = \hat{Y}_{ij} + \hat{Y}_{jk} + \hat{Y}_{ki},$$

which is nothing but triangular trace of $\hat{Y}$ [11]. Curl is easy for online and parallel realizations. In [9], another relative curl is introduced as extensions of combinatorial intransitive triangles,

$$rel\text{-}curl_{ijk} = \frac{|\hat{Y}_{ij} + \hat{Y}_{jk} + \hat{Y}_{ki}|}{|\hat{Y}_{ij}| + |\hat{Y}_{jk}| + |\hat{Y}_{ki}|} \in [0, 1].$$

Relative curl on a triangle $(i, j, k) \in T$ is one if and only if (i, j, k) is intransitive.

The existence of large curls or intransitive triangles may be either due to noise or suggesting the existence of multicriteria in paired comparisons. If the latter case happens on a triangule (i, j, k), on each edge say $(i, j) \in E$, it will have a $\hat{Y}_{ij}$ consistently away from zero, and incur a large curl. In Figure 1, we exhibit one example of such intransitive triangle existing in the data we collected so far, which indicates a stable cyclic preference on a natural scene picture in LIVE dataset such that JPEG2000 (img91) is better than Fast Fading (img91), Fast Fading (img91) is better than White Noise (img40), and White Noise (img40) is better than JPEG2000 (img91). This is due to the fact when different pairs of images are presented to raters, different salient features are adopted by raters implicitly. Triangular curls due to noise will vanish when the sample size goes to infinity while curls due to multicriteria will persist with the increase of sample complexity. Therefore, online tracking of curls will be useful to identify such a kind of inconsistency.

Algorithm 2 outlined below shows how to track the triangular curl in an online way.

Algorithm 2: Online Tracking of Curls.

```
With a new rating $Y_{ij}^{(t)}$;
$n_{ij}^{(t+1)} = n_{ij}^{(t)} + 1$;          // $n_{ij}^{(t)}$ is the number of paired comparisons up to time $t$.
$\hat{Y}_{ij}^{(t+1)} = (1 - 1/n_{ij}^{(t+1)})\hat{Y}_{ij}^{(t)} + Y_{ij}^{(t)}/n_{ij}^{(t+1)}$;      // $\hat{Y}_{ij}^{(t)}$ follows the same definition in Section III-A.
for each k s.t. (i,j,k) is a triangle do
    $curl_{ijk}^{(t+1)} = \hat{Y}_{ij}^{(t+1)} + \hat{Y}_{jk}^{(t+1)} + \hat{Y}_{ki}^{(t+1)}$;
    $rel\text{-}curl_{ijk}^{(t+1)} = \frac{|curl_{ijk}^{(t+1)}|}{|\hat{Y}_{ij}^{(t+1)}|+|\hat{Y}_{jk}^{(t+1)}|+|\hat{Y}_{ki}^{(t+1)}|}$.
end
```

E. Online Tracking of Topology Evolution via Persistent Homology

The work in [9] shows that when the resultant graph provided by assessors is connected, we can derive global scores for all the items in comparison from batch HodgeRank. Besides, when its clique complex is loop-free, there is no global inconsistency, and as thus tracking local inconsistency (triangular curls) presented above will be enough. Motivated by these two observations, [9] adopts persistent homology [51]–[54] to check if a given graph instance satisfies the two conditions.

In fact, persistent homology is an online algorithm to check topology evolution when nodes, edges and triangles enter in a sequential way. Here we just discuss in brief the application of persistent homology to monitor the number of connected components (β_0) and loops (β_1) in our online settings. In random graph designs for image comparisons, we can assume that the images (nodes) are created at the same time, after that pairs of images (edges) are presented to assessors independently one by one. A triangle $\{i, j, k\}$ is created immediately when all the three associated edges appear. In practice with sampling of multigraph data, one may consider certain thresholds on edges and triangles for their presence, which can

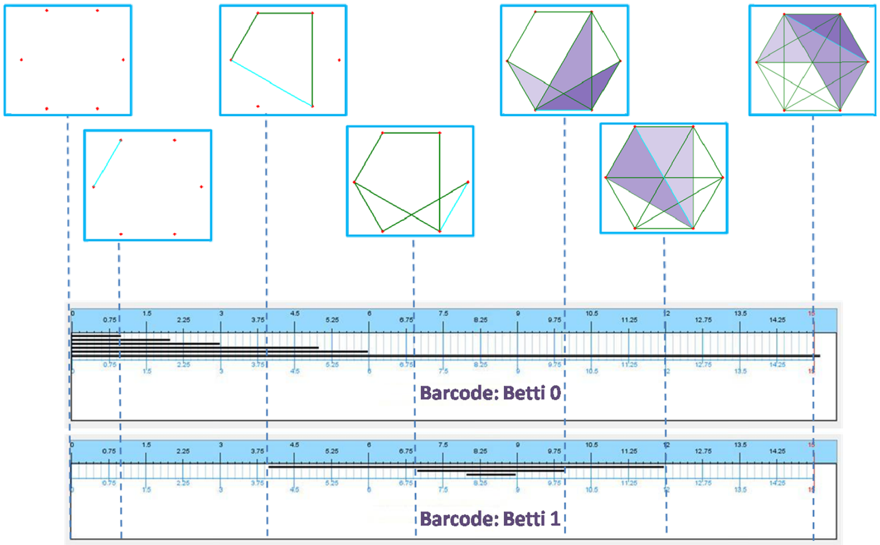


Fig. 2. An example of persistence Barcodes of Betti numbers .

be dealt with in a similar way. With such a streaming data, persistent homology may return the number of β_0 and β_1 at each time when a new node/edge/triangle is born.

Figure 2 illustrates an example of this birth process and its associated Betti numbers (β_0 and β_1) that are computed and plotted by JPlex [55]. At the first frame (say $t = 0$), 6 nodes are collected, which corresponds to $\beta_0 = 6$ at $t = 0$ in Barcode: Betti 0. On the second frame ($t = 1$), an edge connecting a pair of nodes is created which drops the number of connected components from 6 to 5, i.e. $\beta_0 = 5$ at $t = 1$ in Barcode: Betti 0. The same procedure follows and particularly at the fifth frame $t = 4$, it creates a loop and there are 3 connected components in the graph, which can be read from $\beta_0 = 3$ at $t = 4$ and $\beta_1 = 1$ at $t = 4$, respectively. Note that after the thirteenth frame $t = 12$, there is only one connected component $\beta_0 = 1$ left and no loop exists $\beta_1 = 0$ as indicated by the Barcodes.

F. Online Preferential Attachment Sampling

Here we give a brief introduction of preferential attachment sampling in this paper, which is slightly different to traditional preferential attachment random graphs. In traditional preferential attachment random graphs, an edge (i_t, j_t) is added with probability $p_{ij} \sim d_i d_j$, i.e. in proportion to the existing degree of the corresponding vertices. However, this is not an edge-independent process as d_{i_t} changes along the sampling process. To avoid this issue, in this paper we adopt the strategy used in [16], where an edge (i_t, j_t) is drawn independently at probability $p_{ij} = f(\hat{s}_i, \hat{s}_j) \sim \hat{s}_i\hat{s}_j$ for some pre-estimated global rating score $\hat{s}$. This is natural in our QoE evaluation scenario, every node has some intrinsic quality, thus can differentiate its attractiveness from other nodes with different quality. By fixing the estimation $\hat{s}$ in online sampling, we have an edge-independent process which is easy to analyze, where the number of comparisons a vertex received is in proportion to its expected preference estimated by $\hat{s}$. Therefore the higher the preference is, the more comparisons it will receive, which will lead to a faster convergence on top-ranked items confirmed by our experiments.

Algorithm 3 describes this online preferential attachment sampling. We note that in practice, one can slowly update global score estimation $\hat{s}$ after every τ samples, which may improve the performance on top-ranked estimation. Moreover, one can choose more general $p_{ij} = f(\hat{s}_i, \hat{s}_j)$, e.g. $p_{ij} = \exp(\alpha\hat{s}_i\hat{s}_j)$ where $\alpha = 0$ corresponding to Erdös-Rényi random graphs and a large $\alpha > 0$ adjusts the rates of sampling on top-ranked items. For simplicity, we adopt the scheme in Algorithm 3 which suffices to illustrate the speed-up of top-ranked convergence in the setting of this paper.

Algorithm 3: Online Preferential Attachment Sampling.

```
Begin with the initial graph $G_0$ with an estimation $\hat{s}$.   // Usually, it is taken to be an Erdös-Rényi random
    graph with connectivity ($p \gg n^{-1}\log n$) and/or loop-free requirement ($p \gg n^{-1/2}$).
For $t > 0$, at time $t$, the graph $G_t$ is formed by modifying $G_{t-1}$ as follows:
Add a new stimulus pair $\{i, j\}$ by independently choosing stimulus $i$ and stimulus $j$ with probability
    proportional to their quality scores, i.e., $p_{ij} \sim \hat{s}_i\hat{s}_j$.
```

IV. EXPERIMENTS

In this section, we systematically evaluate the performance of the proposed online HodgeRank algorithm on random graphs. Two classes of random graphs are systematically studied here: Erdös-Rényi random graph as the simplest crowdsourcing sampling scheme and the preferential attachment random graph in Algorithm 3 in favor of top-ranked candidates. Both schemes find their importance in applications. All the experiments are conducted with both simulated data and real-world datasets. Finally, we show how to online track the curls and topological evolution with persistent homology.

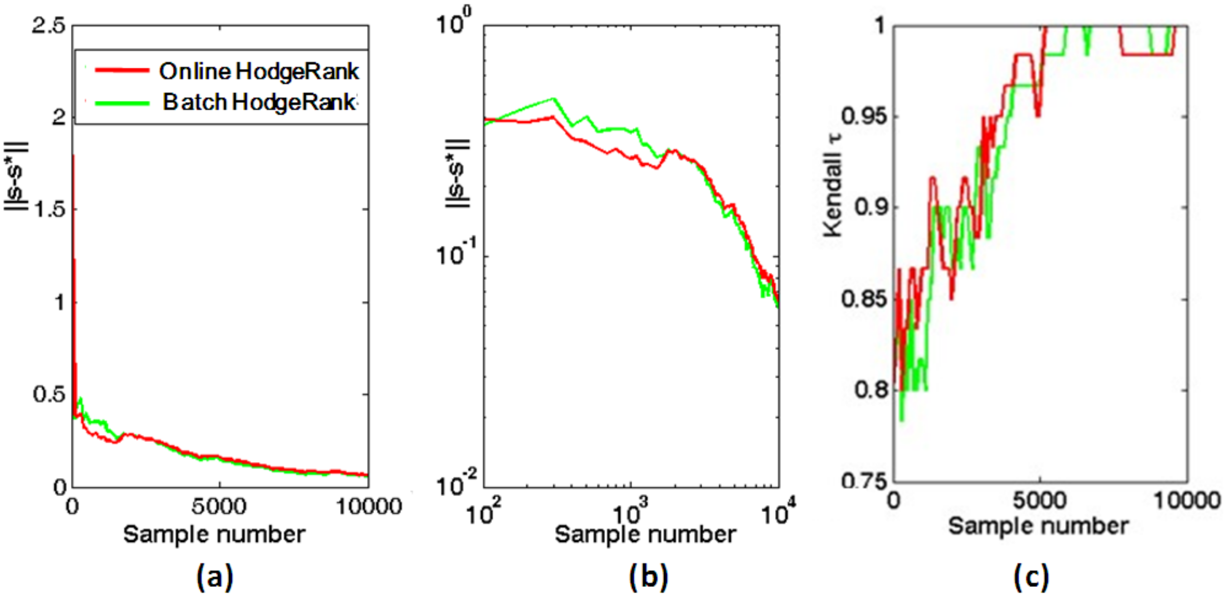


Fig. 3. Comparisons of online HodgeRank and batch HodgeRank for uniform model. (a) l_2 distance; (b) Double logarithmic coordinate axis for l_2 distance; (c) Kendall's τ.

A. Simulated Data

This subsection exploits simulation data to show that online HodgeRank may achieve optimal convergence rates in global ranking estimation, as good as batch HodgeRank, and the preferential attachment random sampling can be more efficient than Erdös-Rényi random sampling on top-k ranking.

Exp-I: Erdös-Rényi random sampling

This experiment will exhibit that for uniform models with Erdös-Rényi random sampling, online HodgeRank algorithm (14) achieves the optimal convergence rates predicted by Theorem III-C, nearly as good as the batch HodgeRank. First, we randomly create a global ranking score as the ground truth, uniformly distributed on $[0, 1]$ for $n = 16$ candidates V which is consistent with the other real-world datasets in this paper. Then paired comparison data are generated by the uniform model: $p(Y_{ij} = 1) = 1 - p(Y_{ij} = -1) = \frac{(s_i - s_j + 1)}{2}$, with each edge $(i, j) \in E$ is independent drawn from the complete graph. This procedure repeats such that we obtain a sequence of paired comparison samples. We make a note on the choice of step size parameter a and t_0 in (15): $a = 7.5 \sim 1/\lambda_1$ here to meet the condition in Theorem III-C; t_0 is less sensitive for the convergence rates which is however important for initial errors and set to be 1000 here for simplicity.

We adopt two metrics to compare the convergence performance of batch Hodge and online HodgeRank. As we know the ground-truth score here, the first metric is the l_2-distance between estimators and the true score, $\|\hat{s} - s^*\|$. Another coarse-grained metrics, called Kendall's τ [56], are also used for the comparison of the induced global ranking orders. Given two global ranking scores x_i and y_i on V, define $X_{ij} = \text{sign}(x_i - x_j)$ and $Y_{ij} = \text{sign}(y_i - y_j)$. Then Kendall's τ coefficient is defined as:

$$\tau(x, y) = \frac{\sum_{\{i,j\} \in E} X_{ij} Y_{ij}}{\sqrt{\sum X_{ij}^2 \cdot \sum {Y_{ij}}^2}}, \tag{19}$$

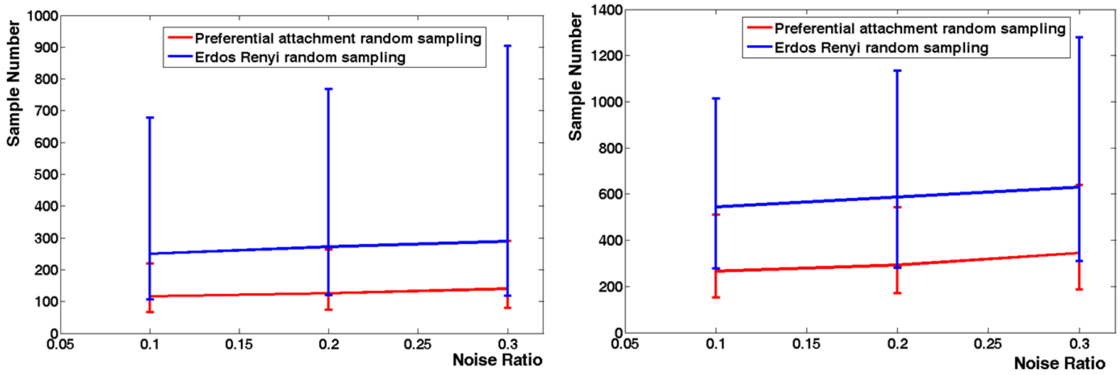
which measures the percentage of concordance ($X_{ij}Y_{ij} > 0$) minus the percentage of mismatch ($X_{ij}Y_{ij} < 0$) between two rankings.

The results are shown in Figure 3, where online HodgeRank is able to maintain competitive performances with the batch HodgeRank, in terms of both metrics above. In particular, in (b) the long term slope of the two curves in double logarithmic plot is $-1/2$, which implies both online and batch algorithms reach a convergence rate at $O(t^{-1/2})$, which is minimax optimal and thus confirms Theorem III-C. In summary, online HodgeRank can achieve a nearly optimal convergence to the true score but with much less computational cost than batch HodgeRank.

Exp-II: Preferential attachment random sampling

This experiment will exhibit some advantages of preferential attachment random sampling compared with Erdös-Rényi random sampling on the convergence speeds for top-ranked items. Similar to the experiment above we also choose $|V| = n = 16$. Besides, in order to simulate the real-world data contaminated by noise, a random subset of E is reversed in preference direction. In this way, we simulate a paired comparison graph, possibly incomplete and imbalanced, with different levels of noise to be specified below. Let the total number of paired comparisons occurred on this graph be SN (Sample Number), and the number of noisy pairs be NPN (Noisy Pairs Number). Then we define the Noise Ratio NR = NPN/SN. In the following experiment, we will show a comparison of the two sampling schemes under different level of noise ratios. Specifically, for each NR level, we add a certain number of pairs under the guidance of preferential attachment and Erdös-Rényi sampling respectively, followed by online HodgeRank until its returned top-k ranked candidates are consistent with the ground-truth. Such a random stopping time is recorded with 1000 repetitions to ensure the statistical stability. Figure 4 shows the mean stopping time together with the variances of the two sampling schemes, for $k = 3$ and 5, respectively. From these experimental results, we make the following comments.

First, it is shown that for the purpose of correct top-k ranking, preferential attachment sampling needs smaller number of samples than Erdös-Rényi sampling in all cases. This is because, unlike Erdös-Rényi



(a) Top 3

(b) Top 5

Fig. 4. Comparisons of the number of samples required by two sampling schemes when focusing on top-k items, versus noisy ratios, in Exp-II. For each NR level, the experiments are repeated 1000 times and the median number of sample pairs with [0.25, 0.75] confidence interval are plotted in the figure.

sampling in which all pairs have equal probability to be sampled, the preferential attachment sampling provides a more efficient ranking process which automatically emphasizes top-ranked candidates and truncates pairs for less important bottom-ranked ones.

Second, we can notice that with the increase of k, the performance gap between these two schemes increases, and with the increase of NR, the number of samples required also increases with a relatively slow speed.

Third, it should be noted that Erdös-Rényi sampling always shows a larger fluctuation than preferential attachment sampling, which further confirms the advantage of preferential attachment random graph.

B. *Real-world Datasets*

Two publicly available datasets, LIVE [17] and IVC [18], are used in this work. The LIVE dataset contains 29 reference images and 779 distorted images. The distorted images are obtained using five different distortion processes–JPEG2000, JPEG, White Noise, Gaussian Blur, and Fast Fading Rayleigh. Considering the resolution limit of most test computers, we only choose 6 different reference images (480×720) and 15 distorted versions of each reference, for a total of 96 images. The second dataset, IVC, which is also a broadly adopted dataset in the community of QoE evaluation, includes 10 reference images and 185 distorted images derived from four distortion types–JPEG2000, JPEG, LAR Coding, and Blurring. Following the collection strategy in LIVE, we further select 9 different reference images (512×512) and 15 distorted images of each reference. Eventually, we obtain a medium-sized image set

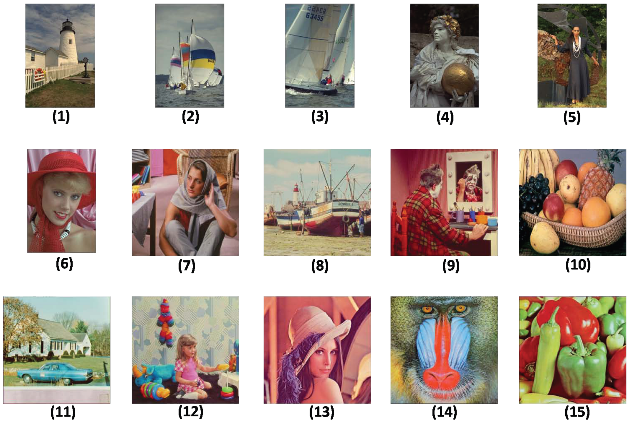


Fig. 5. Images in LIVE and IVC databases (The first six are from LIVE and the remaining ones are from IVC).

that contains a total of 240 images from 15 references, as illustrated in Figure 5. Note that we do not use the subjective scores in LIVE and IVC, but only borrow the image sources they provide. Different from them, we propose to assess image quality with paired comparison method. There are two aspects about the size of dataset: (1) number of distortion types; (2) number of reference images. The first is the number of nodes in our paired comparison graphs, which is n = 16 here. Even on such a scale, it is almost impossible for a single person to perform all $\binom{n}{2}$ paired comparisons. So it suffices to illustrate the performance of online algorithm against batch algorithm. The second does not affect the computational complexity of algorithms, thus a random choice 15 from LIVE and IVC database is to show performance consistency over these examples.

Exp-III: Erdös-Rényi guided data collection & experimental results

We now present our experiment design guided by Erdös-Rényi random graph for collecting the set of online paired data. Different from traditional complete design in paired comparison, a session in our test can have an arbitrary duration (down to a single pair) and participants are free to decide when to quit. In other words, the number of pairs ($\#pairs$) shown to participants can be adjusted according to their time constraint and preference. That is, when a participant's time is adequate, $\#pairs$ can be a bigger value. But if one is under the pressure of time or prefers not to spend more time with the experiment, $\#pairs$ will be smaller.

Before starting the experiment, each participant is briefed about the goal of the experiment and given a short training session to familiarize himself/herself with the testing procedure. In the testing process,

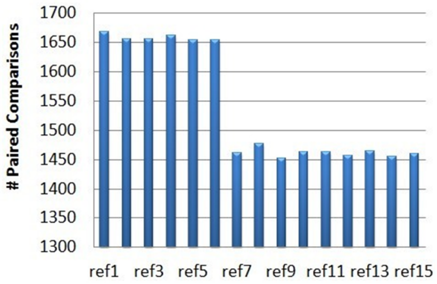


Fig. 6. Number of paired comparisons each reference received in LIVE and IVC databases.

images are displayed side by side at their native resolutions to prevent any distortions due to scaling operations performed by software or hardware. Besides, to make it impossible for participants to cheat our system by inputting "smart" answers, the order of each pair and the order within each pair are totally random for each participant. Each assessor is allowed to take as much time as needed to enter their choice. However, the assessors could not change their choice once entered or view the image again. Once the choice is entered, the next image pair is displayed.

Moreover, we hope to avoid the situation with successive pairs of test images from the same reference, to avoid contextual and memory effects in their judgments of quality. For this purpose, after the playlist for one participant is constructed, our program would go over the entire playlist to determine if adjacent pairs correspond to the same reference. If such a case is detected, one of the pairs would be swapped with another randomly chosen pair in the playlist which does not suffer from the same problem.

Finally, 186 observers of different cultural level (students, tutors, and researchers), each of whom performs a varied number of comparisons via Internet, provide 23,097 paired comparisons in total. The minimum and maximum numbers of pairs that each subject evaluated is 1 and 1552, respectively. The number of responses each reference image receives is different, as illustrated in Figure 6. Our collecting task is still on-going now for further larger-scale studies.

In the following, we will show the comparison experimental results, which involves evaluating the online algorithm (14) on various datasets against the performance of batch HodgeRank.

As there are no ground-truth scores in real-world data, one can not adopt Kendall's τ with the ground-truth as is in simulated data to evaluate the performance of the online method here. In this subsection, the metric that we used in the evaluation of the performance of various algorithms is the Mismatch Ratio (MR) ϵ_t, *i.e.*, at time t, the percentage of mismatch pairs of a global rating s_t made on all previous

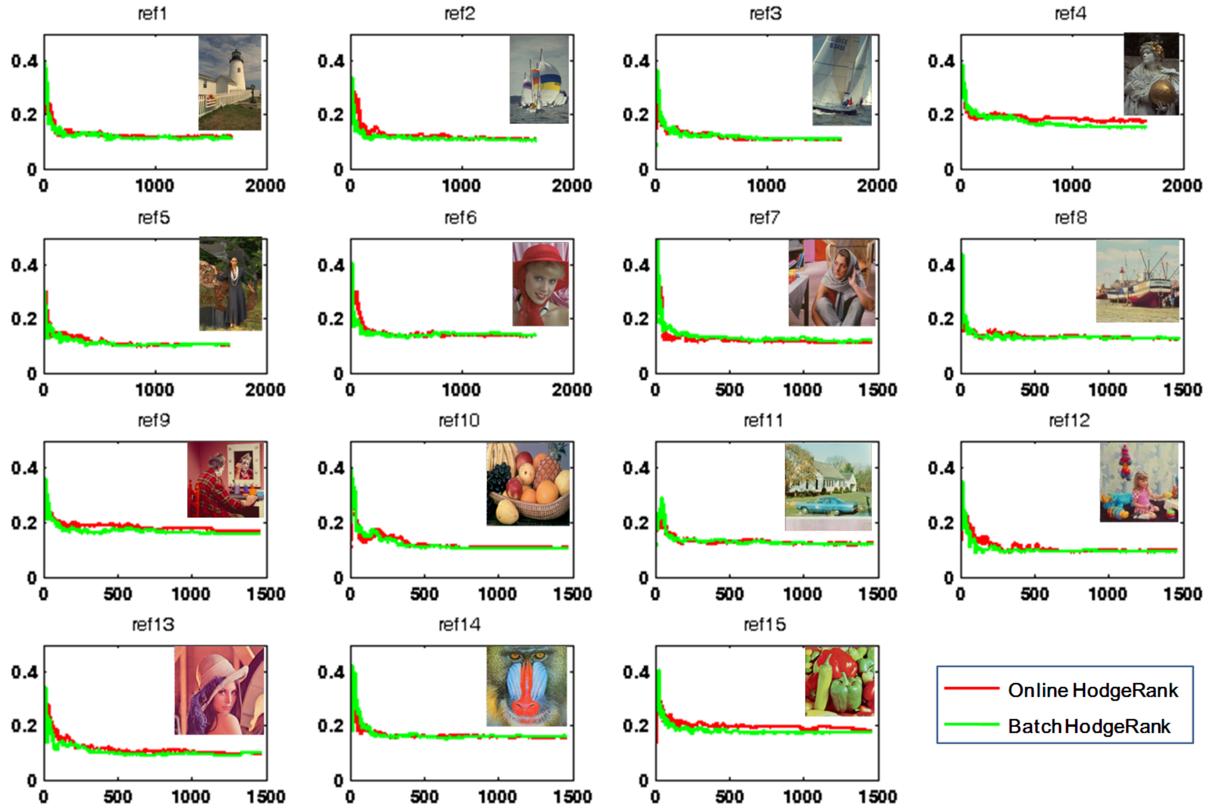


Fig. 7. Experimental results of online HodgeRank vs. batch HodgeRank. MR (y-axis) versus the number of samples (x-axis) on 15 reference images.

examples. For $Y_\tau(i,j) \in \{\pm 1\}$,

$$\epsilon_t := \frac{1}{2t}\sum_{\tau=1}^{t} |\mathrm{sign}(s_t(i_\tau) - s_t(j_\tau)) - Y_\tau(i_\tau, j_\tau)|.$$

Figure 7 shows the performance comparisons of online HodgeRank against batch HodgeRank with $t_0 = 1000$ for 15 reference datasets. It is interesting to see that on all of these large scale data collections, the online HodgeRank is able to maintain competitive performances with the batch case. Besides, Table I shows the computation complexity achieved by online HodgeRank and batch HodgeRank. It is easy to see that on our dataset, online HodgeRank can achieve up to nearly 370 times faster than batch HodgeRank, with similar prediction errors.

Exp-IV: Preferential attachment guided experimental results

In Exp-II, we have shown that preferential attachment random sampling could perform better than

TABLE I

COMPUTATION COMPLEXITY (S) COMPARISON OF ONLINE AND BATCH HODGERANK.

	ref1	ref2	ref3	ref4	ref5	ref6	ref7	ref8	ref9	ref10	ref11	ref12	ref13	ref14	ref15	mean
Online	0.166	0.158	0.159	0.162	0.164	0.163	0.125	0.128	0.131	0.125	0.130	0.128	0.133	0.135	0.133	0.143
Batch	59.28	60.78	58.25	58.65	60.09	58.22	53.15	49.58	47.45	47.81	47.84	48.01	50.29	47.40	47.43	52.95

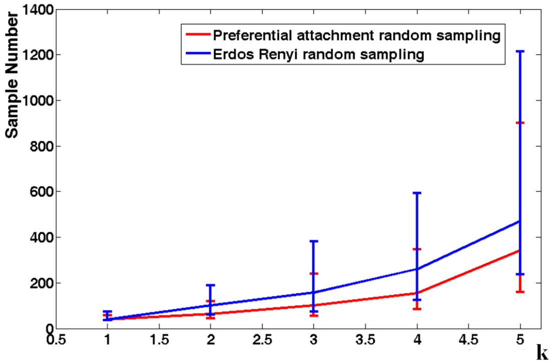


Fig. 8. Comparisons of the number of samples required by two sampling schemes when focusing on top-k items, in Exp-IV. For each k, the median number of samples required on reference 1 with [0.25, 0.75] quantile are plotted in the figure.

Erdös-Rényi random graph in simulated data when we focus on ranking the top-k items. In this subsection, we will continue to show such an improvement on real-world data collected in Exp-III. As there are no ground-truth ranking in real-world data, results obtained from all the paired comparisons collected in Exp-III are treated as the ground-truth in our experiment. Preferential attachment sampling is implemented by resampling with replacement and we run the process 1000 times to ensure the statistical stability.

For top-k ranking, Figure 8 shows the number of samples required against k ranging from 1 to 5 on a randomly selected reference image (ref (1) in Figure 5), where similar observations can be obtained from other reference images. Similar to the simulation data, it can be seen that preferential attachment sampling is more efficient than Erdös-Rényi sampling by significantly reducing the sample complexity over this range of small k. As k increases, such a benefit is increasing, but one should expect some trade-off between k and efficiency. Moreover, we notice that Erdös-Rényi sampling exhibits more significant fluctuations than preferential attachment model. By concentrating on top ranked items, preferential attachment sampling reduces the variance and thus increases the stability effectively.

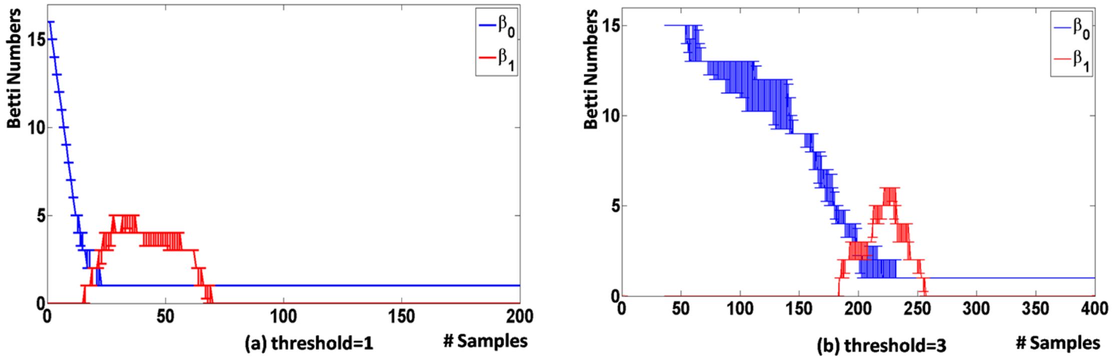


Fig. 9. Number of samples versus number of online Betti numbers. For each sample number level, the median number of Betti numbers over 15 references with [0.25, 0.75] quantile are plotted in the figure.

C. *Online Tracking of Topology and Curls*

In our online settings, due to the multiple comparisons between a pair of images, a natural question is raised that how many samples are needed to satisfy the connected & loop-free conditions? As each reference is similar in sampling scheme, we compute the online mean Betti numbers over 15 references, as illustrated in Figure 9 (a). As we can see, after about 70 samples on this multigraph, with high probability the resultant graph is connected & loop-free. In other words, it is easy to meet these two requirements and thus can avoid the possible issue of harmonic inconsistency in global ranking.

In addition, we can further set a threshold for each edge which can be treated as a confidence level. That is to say, only edges on which the number of paired comparisons are larger than this threshold will be added in our resultant graph. The bigger the threshold is set, the more robust the topological structure of the graph is. Figure 9 (b) shows the online tracking of the first two Betti numbers by persistent homology when threshold is set to be 3. One can see more examples (250) are needed to reach the connected and loop-free condition.

Triangular curls and relative curls defined in the last section are helpful to identify possible inconsistency or the existence of multicriteria adopted by raters in different paired comparisons. By online tracking of relative curls in Figure 10, we find the intransitive triangle shown in Figure 1 that JPEG2000 (img91) is better than Fast Fading (img91), Fast Fading (img91) is better than White Noise (img40), and White Noise (img40) is better than JPEG2000 (img91). The phenomenon suggests that one should explore the hidden multicriteria behind the paired comparisons among these images which will be left for future studies.

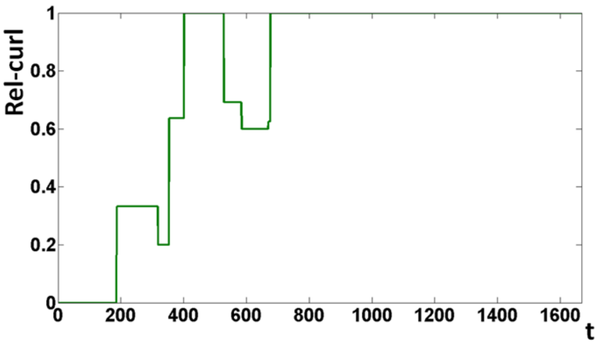


Fig. 10. Online tracking of relative curl on triangle (JPEG2000 (img91), Fast Fading (img91), White Noise (img40)). One can see the intransitive triangle constantly appears over time which suggests possible different criteria adopted by users in paired comparisons made among them.

V. CONCLUSIONS

In this paper, online algorithms are proposed for crowdsourcing QoE evaluation where the data are collected in a streaming way. The algorithms are based on Robbins-Monro procedure or stochastic approximation to solve a HodgeRank problem on random graphs. In particular, we study two random sampling schemes inspired by Erdös-Rényi and preferential attachment random graph theory, followed by online HodgeRank to analyze the streaming data collected from Internet crowd. Erdös-Rényi random graph is the simplest random sampling design bearing the I.I.D. property, while a choice of preferential attachment random sampling focuses on top-ranked items. The distinction makes preferential attachment random graph more efficient for top-k ranking problems on large-scale data collections.

Experiments with the images available in LIVE and IVC databases are conducted, including 15 different reference images and 15 distorted versions of each reference in total. It is shown that in our applications, the proposed online HodgeRank can achieve as nearly good performance as batch HodgeRank, in both theory and experiments. Furthermore, we investigate the online tracking of triangular curls and topology evolution of the paired comparison complex. In particular, we show that online tracking of triangular curls provides us important information about inconsistency, which may suggest the existence of multicriteria in rater's judgement of different object pairs.

Our studies show that online HodgeRank provides us an efficient approach to study large scale crowdsourcing QoE evaluation on the Internet. It enables us to derive global rating as well as monitor the inconsistency occurring in the data in the real time.

With the rapid growth of technologies on rich user interface, in future, we plan to assess user experience in interactive applications with an active learning setting. Besides, for other kinds of dependent sampling schemes, such as Markov sampling, etc., dynamics and convergence of online algorithms will also be our future directions.

REFERENCES

[1] R. Schatz, T. Hoßfeld, L. Janowski, and S. Egger, "From Packets to People: Quality of Experience as New Measurement Challenge," in *Data Traffic Monitoring and Analysis: From measurement, classification and anomaly detection to Quality of experience*, M. M. Ernst Biersack, Christian Callegari, Ed. Springer's Computer Communications and Networks series, 2012.

[2] C.-C. Wu, K.-T. Chen, Y.-C. Chang, and C.-L. Lei, "Crowdsourcing multimedia qoe evaluation: A trusted framework," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1121–1137, 2013.

[3] T. Hoßfeld, R. Schatz, M. Seufert, M. Hirth, T. Zinner, and P. Tran-Gia, "Quantification of youtube qoe via crowdsourcing," in *IEEE International Symposium on Multimedia (ISM) 2011)*, 12 2011.

[4] Y.-G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yang, "Understanding and predicting interestingness of videos." AAAI, 2013.

[5] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, "A crowdsourceable QoE evaluation framework for multimedia content." ACM Multimedia, 2009, pp. 491–500.

[6] A. Eichhorn, P. Ni, and R. Eg, "Randomised pair comparison: an economic and robust method for audiovisual quality assessment." ACM Workshop on Network and Operating Systems Support for Digital Audio and Video, 2010, pp. 63–68.

[7] *ITU-R Recommendation P.800. Methods for subjective determination of transmission quality*, 1996.

[8] H. David, *The method of paired comparisons*, ser. 2nd Ed., Griffin's Statistical Monographs and Courses, 41. Oxford University Press, New York, NY, 1988.

[9] Q. Xu, T. Jiang, Y. Yao, Q. Huang, B. Yan, and W. Lin, "Random partial paired comparison for subjective video quality assessment via HodgeRank." ACM Multimedia, 2011, pp. 393–402.

[10] Q. Xu, Q. Huang, T. Jiang, B. Yan, W. Lin, and Y. Yao, "HodgeRank on random graphs for subjective video quality assessment," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 844–857, 2012.

[11] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye., "Statistical ranking and combinatorial Hodge theory," *Mathematical Programming*, vol. 127, no. 1, pp. 203–244, 2011.

[12] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[13] A. E. Elo, *The rating of chess players: Past and present*. Arco Publishing, 1978.

[14] M. Glickman, "Paired comparison models with time-varying parameters," PhD Thesis, Department of Statistics, Harvard University, Tech. Rep., 1993.

[15] R. Herbrich, T. Minka, and T. Graepel, "Trueskill: A bayesian skill rating system," in *Advances in Neural Information Processing Systems*, 2006, pp. 569–576.

[16] F. Chung and L. Lu, *Complex Graphs and Networks*. CBMS Regional Conference Series in Mathematics, American Mathematical Society, 2006.

[17] "LIVE image and video quality assessment database." `http://live.ece.utexas.edu/research/quality/`, 2008.

[18] "Subjective quality assessment irccyn/ivc database." `http://www2.irccyn.ec-nantes.fr/ivcdb/`, 2005.

[19] Q. Xu, Q. Huang, and Y. Yao, "Online crowdsourcing subjective image quality assessment." ACM Multimedia, 2012, pp. 359–368.

[20] M. Kim and J. Leskovec, "Multiplicative attribute graph model of real-world networks," *Internet Mathematics*, vol. 8, no. 1-2, pp. 113–160, 2012.

[21] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297–312, 2011.

[22] A. Kittur, E. Chi, and B. Suh, "Crowdsourcing user studies with Mechanical Turk." SIGCHI conference on Human factors in computing systems, 2008, pp. 453–456.

[23] A. N. Hirani, K. Kalyanaraman, and S. Watts, "Least squares ranking on graphs," *arXiv:1011.1716*, 2011.

[24] B. Osting, J. Darbon, and S. Osher, "Statistical ranking using the l_1-norm on graphs," *AIMS J. Inverse Problems and Imaging*, 2013, preprint.

[25] B. Osting, C. Brune, and S. Osher, "Enhanced statistical rankings via targeted data collection," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 489–497.

[26] L.-H. Lim and Y. Yao, "Hodge decomposition on graphs and online ranking," *preprint*, 2013.

[27] O. Candogan, I. Menache, A. Ozdaglar, and P. A. Parrilo, "Flows and decompositions of games: Harmonic and potential games," *Mathematics of Operations Research*, vol. 36, no. 3, pp. 474–503, 2011.

[28] Q. Xu, J. Xiong, Q. Huang, and Y. Yao, "Robust evaluation for quality of experience in crowdsourcing." ACM Multimedia, 2013, preprint.

[29] A. Chorin and J. Marsden, *A Mathematical Introduction to Fluid Mechanics*, ser. Texts in Applied Mathematics. Springer, 1993.

[30] J. Yuan, C. Schnörr, and G. Steidl, "Convex hodge decomposition and regularization of image flows." *Journal of Mathematical Imaging and Vision*, vol. 33, no. 2, pp. 169–177, 2009.

[31] W. Ma, J.-M. Morel, S. Osher, and A. Chien, "An l_1-based variational model for retinex theory and its application to medical images," in *IEEE Proc. Computer Vision Pattern Recognition (CVPR)*, 2011, pp. 153–160.

[32] V. N. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.

[33] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American Mathematical Society*, vol. 29, no. 1, pp. 1–49, 2002.

[34] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[35] B. Widrow and M. Hoff, "Adaptive switching circuits," *IRE WESCON Convention Record*, no. 4, pp. 96–104, 1960.

[36] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[37] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *Journal of Basic Engineering*, vol. 83, no. 3, pp. 95–108, 1961.

[38] Y. Yao, "On complexity issue of online learning algorithms," *IEEE Transactions on Information Theory*, vol. 56, no. 12, pp. 6470–6481, 2010.

[39] M. E. Glickman, “Parameter estimation in large dynamic paired comparison experiments,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 48, no. 3, pp. 377–394, 1999.

[40] N. Ailon, “An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity,” *Journal of Machine Learning Research*, vol. 13, pp. 137–164, 2012.

[41] K. G. Jamieson and R. D. Nowak, “Active ranking using pairwise comparisons,” *Annual Conference on Neural Information Processing Systems*, 2011.

[42] K. Jamieson and R. Nowak, “Active ranking in practice: General ranking functions with sample complexity bounds.”

[43] B. Bollobas, *Random Graphs*. Cambridge University Press, 2001.

[44] P. Erdos and A. Renyi, “On random graphs i,” *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.

[45] A.-L. Barabasi and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[46] D. Watts and S. Strogatz, “Collective dynamics of ‘small-world’ networks,” *Nature*, no. 393, pp. 440–442, 1998.

[47] M. Penrose, *Random Geometric Graphs (Oxford Studies in Probability)*. Oxford University Press, 2003.

[48] S. Smale and Y. Yao, “Online learning algorithms,” *Foundation of Computational Mathematics*, vol. 6, no. 2, pp. 145–170, 2006.

[49] N. Shor, *Minimization Methods for Non-Differenliable Functions*. Springer-Verlag, 1985.

[50] A. Rakhlin, *Statistical Learning Theory and Sequential Prediction*. Lecture Notes in University of Pennsyvania, 2012.

[51] H. Edelsbrunner, D. Letscher, and A. Zomorodian, “Topological persistence and simplification,” *Discrete and Computational Geometry*, vol. 28, no. 4, pp. 511–533, 2002.

[52] A. Zomorodian and G. Carlsson, “Computing persistent homology,” *Discrete and Computational Geometry*, vol. 33, no. 2, pp. 249–274, 2005.

[53] G. Carlsson, “Topology and data,” *Bulletin of the American Mathematical Society*, vol. 46, no. 2, pp. 255–308, 2009.

[54] H. Edelsbrunner and J. Harer, “Computational topology : an introduction,” 2010.

[55] H. Sexton and M. Johansson, “JPlex: a java software package for computing the persistent homology of filtered simplicial complexes,” `http://comptop.stanford.edu/programs/jplex/`, 2009.

[56] Kendall, Maurice, and J. Gibbons, *Rank Correlation Methods*. Oxford University Press, 1990.