Random Partial Paired Comparison for Subjective Video Quality Assessment via HodgeRank [*]

Qianqian Xu
Graduate University, Chinese Academy of Sciences, Beijing 100049, China
qqxu@jdl.ac.cn

Tingting Jiang[*]
National Engineering Lab. for Video Technology, Key Lab of Machine Perception (MOE), School of EECS, Peking University, Beijing 100871, China
ttjiang@pku.edu.cn

Yuan Yao[*]
School of Mathematical Sciences, Key Lab. of Machine Perception (MoE), Peking University, Beijing 100871, China
yuany@math.pku.edu.cn

Qingming Huang[*]
Graduate University, Chinese Academy of Sciences, Beijing 100049, China
qmhuang@jdl.ac.cn

Bowei Yan
School of Mathematical Sciences, Peking University, Beijing 100871, China
yanbowei@gmail.com

Weisi Lin
School of Computer Engineering, Nanyang Technological University, Nanyang 639798, Singapore
wslin@ntu.edu.sg

ABSTRACT

Subjective visual quality evaluation provides the ground-truth and source of inspiration in building objective visual quality metrics. Paired comparison is expected to yield more reliable results; however, this is an expensive and time-consuming process. In this paper, we propose a novel framework of **HodgeRank on Random Graphs** (HRRG) to achieve efficient and reliable subjective Video Quality Assessment (VQA). To address the challenge of a potentially large number of combinations of videos to be assessed, the proposed methodology does not require the participants to perform the complete comparison of all the paired videos. Instead, participants only need to perform a random sample of all possible paired comparisons, which saves a great amount of time and labor. In contrast to the traditional deterministic incomplete block designs, our random design is not only suitable for traditional laboratory and focus-group studies, but also fit for crowdsourcing experiments on Internet where the raters are distributive over Internet and it is hard to control with precise experimental designs.

Our contribution in this work is three-fold: 1) a HRRG framework is proposed to quantify the quality of video; 2) a new random design principle is investigated to conduct paired comparison based on Erdös-Rényi random graph theory; 3) Hodge decomposition is introduced to derive, from incomplete and imbalanced data, quality scores of videos and inconsistency of participants' judgments. We demonstrate the effectiveness of the proposed framework on LIVE Database. Equipped with random graph theory and HodgeRank, our scheme has the following advantages over the traditional ones: 1) data collection is simple and easy to handle, and thus is more suitable for crowdsourcing on Internet; 2) workload on participants is lower and more flexible; 3) the rating procedure is efficient, labor-saving, and more importantly, without jeopardizing the accuracy of the results.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Evaluation/methodology*; C.4 [**Performance of Systems**]: Design studies; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human factors*

General Terms

Performance, Experimentation, Human Factors

Keywords

Subjective Video Quality Assessment, HodgeRank, Random Graphs, Persistence Homology

1. INTRODUCTION

With the rapid development and wide applications of digital media devices, the number of videos available is growing at an explosive rate. The Video Quality Assessment (VQA) issue, has drawn increasing attention from researchers during recent years, and now plays an important role in a broad range of applications, e.g. video enhancement, reconstruction, compression, communication, displaying, registration, printing, watermarking, etc.

The existing methods of VQA can be divided into two categories: subjective assessment and objective assessment. In subjective viewing tests, video sequences are shown to a group of viewers and then their opinions are recorded

---

[*]Area chair: Wei-tsang Ooi.
[*]Corresponding author.

and averaged to evaluate the quality of each video sequence. This testing process is labor-intensive and time-consuming. Therefore, there has been an increasing demand to build intelligent, objective quality measurement models [33, 24, 27, 34, 15] to predict perceived video quality automatically. Subjective experiments are often used to provide the ground-truth and verification for objective models. In typical Mean Opinion Score (MOS) test [1], individuals are asked to give a rating from Bad to Excellent (Bad-1, Poor-2, Fair-3, Good-4, and Excellent-5) to grade the quality of a video. However, such a test may suffer the following problems [8]:

1. Unable to concretely define the concept of scale;

2. Dissimilar interpretations of the scale among users;

3. Difficult to verify whether a participant gives false ratings either intentionally or carelessly.

Therefore, to address the problems above, recent investigations turn to an alternative approach with paired comparison [8]. In a paired comparison test, a participant is simply asked to compare two videos simultaneously, and vote which one has the better quality based on his/her perception. Therefore individual decision process in paired comparison is simpler than in the typical MOS test, as the five-scale rating is reduced to a dichotomous choice.

However, paired comparison approach leaves a heavier burden on participants with a larger number of comparisons. For example, we are given 15 distorted versions of 1 reference video, i.e. 16 videos in total to be compared; by adopting the MOS, it only needs to perform 15 judgments. However, it requires $\binom{16}{2} = 120$ comparisons if adopting the complete design in existing paired comparison method [8]. When the number of videos to be judged is large, it may be practically impossible, or at least unacceptable from the viewpiont of the participants. If the testing time for a single participant lasts too long [2], participants may become lack of patience and thus may input random decisions carelessly or intentionally. Therefore, how to make paired comparison method efficient and applicable in reality has become an urgent issue in the VQA community.

To address this issue, we propose in this paper a novel methodology which is not only suitable for traditional laboratory and focus-group studies, but also fit for crowdsourcing experiments, i.e. *HodgeRank on Random Graphs*.

Our rationale is that, since the number of videos to be judged can be large, we do not ask a single participant to perform the complete comparison of all the video pairs. Instead, every participant only needs to commit to a fraction of all possible comparisons. Hence it raises a question: how to choose the pairs that will be viewed by participants? There has been a large literature in statistics on deterministic incomplete block design [9]. However, these designs are not suitable for crowdsourcing on Internet where the raters are distributive over Internet with varied backgrounds and it is hard to control with traditional experimental designs. To meet this challenge, we propose a random design based on Erdös-Rényi random graph theory [12], the simplest scheme among other choices, where video pairs are independently presented to a participant for rating. Equipped with a recent new development of Hodge theoretical approach to statistical ranking [18], we can infer a reliable global ranking from such data.

In HodgeRank, it shows that every edge flow representing paired ranking can be resolved into two orthogonal components, a gradient flow that represents the l_2-optimal global ranking and a divergence-free flow (cyclic) that measures the validity of the global ranking obtained—if this is large, it indicates that the data does not have a good global ranking. This divergence-free flow can be further decomposed orthogonally into a curl flow (locally cyclic) and a harmonic flow (locally acyclic but globally cyclic); these provide information on whether inconsistency in the ranking data arises locally or globally. In applications, one should avoid large global inconsistency which indicates some serious conflicts of interests in ranking data. Through Erdös-Rényi random graphs, we can efficiently control this kind of inconsistency.

We demonstrate the effectiveness and generality of the proposed framework on LIVE Database [3]. Experimental results show that the proposed framework is promising and has potentially wide applications in subjective VQA.

Our contribution in this work is three-fold:

1. We propose a novel framework of HodgeRank with random graphs to quantify the quality of video. The advantages of our framework over the traditional ones are: 1) data collection is simple, easy to handle, and thus is more suitable for crowdsourcing on Internet; 2) workload on participants is lower and more flexible; 3) the rating procedure is efficient, labor-saving, and more importantly, without jeopardizing the accuracy of the results. As we shall see later, under such models the $O(n^2)$ complete paired comparisons can be reduced to $O(n^{3/2})$, with a lower bound of $O(n \log n)$.

2. A new random design principle is investigated to conduct paired comparison based on Erdös-Rényi random graph theory;

3. Hodge decomposition on graphs is introduced to derive, from incomplete (where every participant may only give partial comparisons) and imbalanced (where different video pairs may receive different number of comparisons) data, quality scores of videos and the inconsistency of participants' judgments.

The remainder of this paper is organized as follows. Section 2 contains a review of related works. Then Section 3 establishes the statistical ranking models based on the Hodge decomposition theory, as well as the principles for random sampling grounded in random graph theory. The detailed experiments are demonstrated in Section 4. Section 5 presents the conclusive remarks along with discussion for future work.

2. RELATED WORK

2.1 Paired Comparison

Paired comparison generally refers to any process of comparing entities in pairs by raters to judge which entity in each pair is preferred. The method of paired comparison has been widely studied in social choice, psychology, statistics, and computer science [32, 21, 9, 28, 5]. It has also drawn increasing attention from the machine learning community as it may be adapted to classification problems [14, 13, 16].

Among various approaches to analyzing paired comparison results, a recent framework based on combinatorial Hodge Theory [18] will be adopted in this paper, which is particularly suitable for the analysis of incomplete and imbalanced data distributed on a graph. Our work exploits such a HodgeRank approach with random graph models in the setting of subjective video quality assessment.

2.2 Crowdsourcing

With the advent of ubiquitous Internet access, it becomes more and more popular to ask an Internet crowd to conduct experiments on their personal computers [17, 6]. This new distributed model advocates mass collaboration and the wisdom of the commons. The main difference between crowdsourcing and ordinary outsourcing is that a task is carried out by an unspecific Internet crowd rather than a specific group of people. For example, researchers can seek help from the Internet crowd to conduct user studies on image annotation [30, 26], document relevance [4], and document evaluation [22].

Most recently, a crowdsourceable framework [8] has been proposed for VQA which adopts a paired comparison approach. However, one major shortcoming of [8] lies in that it makes a strong assumption that all paired data collected are complete which is impossible for large number of videos. In this paper, we present a new principle to deal with possibly incomplete and imbalanced data distributed on random graphs, which reduces the sampling complexity and is easy to implement in crowsourcing VQA on internet.

2.3 Inconsistency Checking

After collecting the paired data from the participants, there is a need to assess the consistency of judgment as not every participant is trustworthy. They may input random decisions carelessly or intentionally. Like traditional social choice theory with complete and balanced data, the method in [8] proposes Transitivity Satisfaction Rate (TSR) to measure the consistency of a participant's judgments, which checks all the transitive triangles such that A> B >C >A. The TSR is defined as the number of judgment triplets (e.g., the three preference relations among A, B, and C) satisfying transitivity divided by the total number of triplets where transitivity may apply; thus, the value of the TSR is always between 0 and 1. If a participant's judgments are consistent throughout all the rounds of an experiment, the TSR will be 1; otherwise it will be less than 1.

However, TSR is only based on complete and balanced paired comparison data. When the paired data is incomplete with missing edges, HodgeRank will give us a general treatment of inconsistency where both triangular and global cycles are considered.

3. HODGERANK ON RANDOM GRAPHS

In this section, we propose a new random design principle to conduct paired comparison and analyze data for a reliable global ranking and inconsistency. Our sampling mechanism exploits the Erdös-Rényi random graphs, a simple but efficient approach in our study. Other random graphs can be applied into the framework, which are left for future studies. HodgeRank is a particularly suitable tool to analyze paired comparison data in such graphs by adapting to their topological structures. We first explain how to develop a statistical ranking model based on Hodge theory on general graphs, and then describe the principles that the random selection must adhere to.

3.1 HodgeRank on Graphs

Let $\wedge = \{1, ..., m\}$ be a set of participants and $V = \{1, ..., n\}$ be the set of videos to be ranked. Paired comparison data is collected as a function on $\wedge \times V \times V$, which is *skew-symmetric* for each α, i.e. $Y_{ij}^\alpha = -Y_{ji}^\alpha$ representing the degree that α
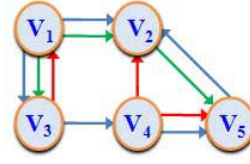

Figure 1: An example of paired comparison hypergraph for 5 videos.

prefers i to j. The simplest setting is the binary choice, where[1]

$$Y_{ij}^\alpha = \begin{cases} 1 & if\ \alpha\ prefers\ i\ to\ j, \\ -1 & otherwise. \end{cases} \quad (1)$$

General Y_{ij}^α can be used to represent paired comparison grades, *e.g.* $Y_{ij}^\alpha > 0$ refers to the degree that α prefers i to j and the vice versa $Y_{ji}^\alpha = -Y_{ij}^\alpha < 0$ measures the dispreference degree [18]. In this paper we shall focus on binary choice, to avoid the scale ambiguity issue discussed early. However the theory can be applied to the general case.

Such paired comparison data can be represented by a directed graph, or hypergraph, with n nodes, where each directed edge between i and j refers the preference indicated by Y_{ij}^α. Figure 1 shows an illustration of such hypergraph.

A nonnegative weight function $\omega : \wedge \times V \times V \longrightarrow [0, \infty)$ is defined as,

$$\omega_{ij}^\alpha = \begin{cases} 1 & if\ \alpha\ makes\ a\ comparison\ for\ \{i, j\}, \\ 0 & otherwise. \end{cases} \quad (2)$$

It may reflect the confidence level that a participant compares $\{i, j\}$ by taking different values, which is however not pursued in this paper.

Following [18], the statistical rank aggregation problem is looking for some global ranking score $s : V \to R$ such that

$$\min_{s \in \mathrm{R}^{|V|}} \sum_{i,j,\alpha} \omega_{ij}^\alpha (s_i - s_j - Y_{ij}^\alpha)^2, \quad (3)$$

which is equivalent to the following weighted least square problem

$$\min_{s \in \mathrm{R}^{|V|}} \sum_{i,j} \omega_{ij} (s_i - s_j - \hat{Y}_{ij})^2, \quad (4)$$

where $\hat{Y}_{ij} = (\sum_\alpha \omega_{ij}^\alpha Y_{ij}^\alpha)/(\sum_\alpha \omega_{ij}^\alpha)$ and $\omega_{ij} = \sum_\alpha \omega_{ij}^\alpha$.

A graph structure arises naturally from ranking data as follows. Let $G = (V, E)$ be a paired ranking graph whose vertex set is V, the set of videos to be ranked, and whose edge set is E, the set of video pairs which receive some comparisons, i.e.

$$E = \left\{ \{i, j\} \epsilon \binom{V}{2} \mid \sum_\alpha \omega_{i,j}^\alpha > 0 \right\}. \quad (5)$$

In classical statistical paired ranking theory, a paired ranking is called *complete* if each participant α in $\wedge$ gives a total judgment of all videos in V; otherwise it is called *incomplete*. It is *balanced* if the paired comparison graph is k-regular with equal weights $\omega_{ij} = \sum_\alpha \omega_{ij}^\alpha \equiv c$ for all $\{i, j\} \in E$; otherwise it is called *imbalanced*. A complete and balanced ranking induces a complete graph with equal weights on all

---

[1]If ties occur, we randomly split them into either preference.

edges. The existing paired comparison methods in VQA often assume complete and balanced data [8]. However, this is an unrealistic assumption for real world data. For example, in crowdsourcing, different video pairs might receive different number of comparisons, which leads to different edge weights ω_{ij}. Nevertheless, as to be shown below, it is efficient to utilize a simple random design based on Erdös-Rényi random graph theory where for each participant some video pairs are chosen randomly. The HodgeRank approach adopted in this paper enables us a unified scheme which can deal with incomplete and imbalanced data.

The minimization problem (4) can be generalized to a family of *linear models* in paired comparison methods [9]. To see this, we first rewrite (4) in another simpler form. Assume that for each edge as video pair $\{i, j\}$, the number of comparisons is n_{ij}, among which a_{ij} participants have a preference on i over j and a_{ji} vice versa. So $a_{ij} + a_{ji} = n_{ij}$ if no tie occurs. Therefore, for each edge $\{i, j\} \in E$, we have a preference probability estimated from data $\hat{\pi}_{ij} = a_{ij}/n_{ij}$. With this definition, the problem (4) can be rewritten as

$$\min_{s \in \mathrm{R}^{|V|}} \sum_{\{i,j\} \in E} n_{ij} (s_i - s_j - (2\hat{\pi}_{ij} - 1))^2, \qquad (6)$$

since $\hat{Y}_{ij} = (a_{ij} - a_{ji})/n_{ij} = 2\hat{\pi}_{ij} - 1$ due to equation (2).

General *linear models*, which are firstly formulated by G. Noether [25], assume that the true preference probability can be fully decided by a linear scaling function on V, i.e.

$$\pi_{ij} = \mathrm{Prob}\{i \text{ is preferred over } j\} = F(s_i^* - s_j^*), \quad (7)$$

for some $s^* \in \mathrm{R}^{|V|}$. F can be chosen as any symmetric cumulated distributed function. When only an empirical preference probability $\hat{\pi}_{ij}$ is observed, we can map it to a skew-symmetric function by inverse of F,

$$\hat{Y}_{ij} = F^{-1}(\hat{\pi}_{ij}), \qquad (8)$$

where $\hat{Y}_{ij} = -\hat{Y}_{ji}$. However, in this case, one can only expect that

$$\hat{Y}_{ij} = s_i^* - s_j^* + \varepsilon_{ij}, \qquad (9)$$

where ε_{ij} accounts for the noise. The case in (6) takes a linear F and is often called a *uniform model*. Below we summarize some well known models which have been studied extensively in literature [9].

1. *Uniform* model:

$$\hat{Y}_{ij} = 2\hat{\pi}_{ij} - 1. \qquad (10)$$

2. *Bradley-Terry* model:

$$\hat{Y}_{ij} = \log \frac{\hat{\pi}_{ij}}{1 - \hat{\pi}_{ij}}. \qquad (11)$$

3. *Thurstone-Mosteller* model:

$$\hat{Y}_{ij} = F^{-1}(\hat{\pi}_{ij}). \qquad (12)$$

where F is essentially the Gauss error function

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-x/[2\sigma^2(1-\rho)]^{1/2}}^{\infty} e^{-\frac{1}{2}t^2} dt. \qquad (13)$$

Note that unknown constants σ and ρ will only contribute to a rescaling of the solution of (4).

4. *Angular transform* model:

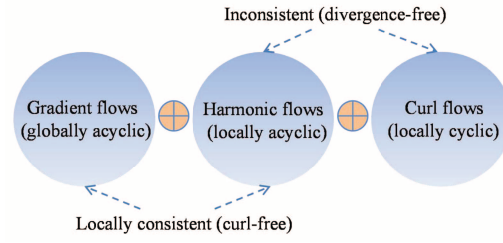$$\hat{Y}_{ij} = \arcsin(2\hat{\pi}_{ij} - 1). \qquad (14)$$

Figure 2: **Hodge decomposition of paired rankings.**

Different models will give different $\hat{Y}_{ij}$ from the same observation $\hat{\pi}_{ij}$, followed by the same weighted least square problem (4) for the solution. Therefore, a deeper analysis of problem (4) will disclose more properties about the ranking problem.

HodgeRank on graph $G = (V, E)$ provides us such a tool, which characterizes the solution and residue of (4), adaptive to topological structures of G. The following theorem adapted from [18] describes a decomposition of $\hat{Y}$, which can be visualized as edge flows on graph G with direction $i \to j$ if $\hat{Y}_{ij} > 0$ and vice versa. Before the statement of the theorem, we first define the triangle set of G as all the 3-cliques in G.

$$T = \left\{ \{i, j, k\} \epsilon \binom{V}{3} | \{i, j\}, \{j, k\}, \{k, i\} \epsilon E \right\}. \qquad (15)$$

Equipped with T, graph G becomes an abstract simplicial complex, the clique complex $\chi(G) = (V, E, T)$.

Theorem 1 [Hodge Decomposition of Paired Ranking] Let $\hat{Y}_{ij}$ be a paired comparison flow on graph $G = (V, E)$, i.e. $\hat{Y}_{ij} = -\hat{Y}_{ji}$ for $\{i, j\} \in E$ and 0 otherwise. There is a unique decomposition of $\hat{Y}$ satisfying

$$\hat{Y} = \hat{Y}^{(1)} + \hat{Y}^{(2)} + \hat{Y}^{(3)}, \qquad (16)$$

where

$$\hat{Y}_{ij}^{(1)} = \hat{s}_i - \hat{s}_j, \quad \text{for some } \hat{s} \in \mathrm{R}^V, \qquad (17)$$

$$\hat{Y}_{ij}^{(2)} + \hat{Y}_{jk}^{(2)} + \hat{Y}_{ki}^{(2)} = 0, \quad \text{for each } \{i, j, k\} \in T, \qquad (18)$$

$$\sum_{j \sim i} \omega_{ij} \hat{Y}_{ij}^{(2)} = 0, \quad \text{for each } i \in V. \qquad (19)$$

The decomposition above is *orthogonal* under the following inner product on $\mathrm{R}^{|E|}$, $\langle u, v \rangle_\omega = \sum_{\{i,j\} \in E} \omega_{ij} u_{ij} v_{ij}$.

The following provides some remarks on the decomposition.

1. $\hat{Y}_{ij}^{(1)}$ is a rank two skew-symmetric matrix and gives a linear score function $\hat{s} \in \mathrm{R}^V$ up to a translation constant. We thus call $\hat{Y}^{(1)}$ a *gradient flow* since it is given by the difference (discrete gradient) of the score function $\hat{s}$ on graph nodes,

$$\hat{Y}_{ij}^{(1)} = (\delta_0 \hat{s})(i, j) := \hat{s}_i - \hat{s}_j, \qquad (20)$$

where $\delta_0 : \mathrm{R}^V \to \mathrm{R}^E$ is a finite difference operator (matrix) on G. $\hat{s}$ can be chosen as any least square solution of (4), where we often choose the minimal norm solution,

$$\hat{s} = \Delta_0^\dagger \delta_0^* \hat{Y}, \qquad (21)$$

where $\delta_0^* = \delta_0^T W$ ($W = \text{diag}(\omega_{ij})$), $\Delta_0 = \delta_0^* \cdot \delta_0$ is the unnormalized graph Laplacian defined by $(\Delta_0)_{ii} = \sum_{j \sim i} \omega_{ij}$ and $(\Delta_0)_{ij} = -\omega_{ij}$, and $(\cdot)^\dagger$ is the Moore-Penrose (pseudo) inverse. On a complete and balanced graph, (21) is reduced to the $\hat{s}_i = \frac{1}{n-1} \sum_{j \neq i} \hat{Y}_{ij}$, often called *Borda Count* as the earliest preference aggregation rule in social choice [18].

2. $\hat{Y}^{(2)}$ satisfies two conditions (18) and (19), which are called *curl-free* and *divergence-free* conditions respectively. The former requires the triangular trace of $\hat{Y}$ to be zero, on every 3-clique in graph G; while the later requires the total sum (inflow minus outflow) to be zero on each node of G. These two conditions characterize a linear subspace which is called *harmonic flows*.

3. The residue $\hat{Y}^{(3)}$ actually satisfies (19) but not (18). In fact, it measures the amount of intrinsic (local) inconsistancy in $\hat{Y}$ characterized by the triangular trace. We often call this component *curl flow*. In particular, the following relative curl,

$$\text{curl}_{ijk}^r = \frac{|\hat{Y}_{ij} + \hat{Y}_{jk} + \hat{Y}_{ki}|}{|\hat{Y}_{ij}| + |\hat{Y}_{jk}| + |\hat{Y}_{ki}|} \in [0, 1], \qquad (22)$$

can be used to characterize triangular intransitivity; $\text{curl}_{ijk}^r = 1$ iff $\{i, j, k\}$ contains an intransitive triangle of $\hat{Y}$. Note that computing the percentage of $\text{curl}_{ijk}^r = 1$ is equivalent to calculating the Transitivity Satisfaction Rate (TSR).

Figure 2 illustrates the Hodge decomposition for paired comparison flows. The readers may refer to [18] for the detail of theoretical development. Below we just make a few comments on the application of HodgeRank in our setting.

1. To find a global rankng $\hat{s}$ in (21), recent developments of Spielman-Teng [31] and Koutis-Miller-Peng [23] give fast (almost linear in $|E|\text{Poly}(\log|V|)$) algorithms for this purpose.

2. Inconsistency of $\hat{Y}$ has two parts: global inconsistency measured by harmonic flow $\hat{Y}^{(2)}$ and local inconsistency measured by curls in $\hat{Y}^{(3)}$. Due to the orthogonal decomposition, $\|\hat{Y}^{(2)}\|_\omega^2 / \|\hat{Y}\|_\omega^2$ and $\|\hat{Y}^{(3)}\|_\omega^2 / \|\hat{Y}\|_\omega^2$ provide percentages of global and local inconsistencies, respectively.

3. A nontrivial harmonic component $\hat{Y}^{(2)} \neq 0$ implies the fixed tournament issue, i.e. for any candidate $i \in V$, there is a paired comparison design by removing some of the edges in $G = (V, E)$ such that i is the overall winner.

4. One can control the harmonic component by controlling the topology of clique complex $\chi(G)$. In a loop-free clique complex $\chi(G)$ where $\beta_1 = 0$, harmonic component vanishes. In this case, there are no cycles which traverse all the nodes, e.g. 1>2>3>4> ... >n>1. All the inconsistency will be summarized in those triangular cycles, e.g. i>j>k>i.

Theorem 2. The linear space of harmonic flows has dimension equal to β_1, i.e. the number of independent loops in clique complex $\chi(G)$, which is called the first order Betti number.

Fortunately, for Erdös-Rényi random graphs, it is not hard to obtain graphs whose β_1 are zero.

3.2 Erdös-Rényi Random Graphs

Erdös-Rényi random graphs $G(n, p)$ start from n vertices and draw their edges independently according to a fixed probability p. Such random graph model is chosen to meet the scenario that in crowdsourcing ranking raters and videos come in an unspecified way. Among various models, Erdös-Rényi random graph is the simplest one equivalent to I.I.D.
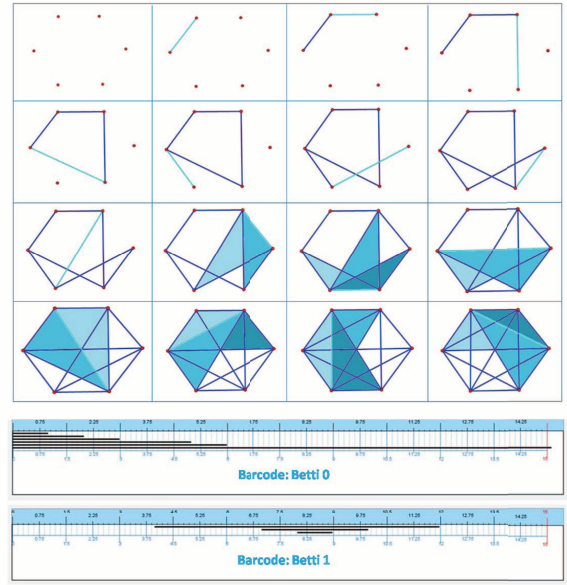

Figure 3: Persistence barcodes of Betti numbers.

sampling. Therefore, such a model is systematically studied in the paper. Other random graph models (e.g. k-regular, preference attachment) can also be developed with HodgeRank on general graphs, which is however left to our future pursuit.

However, to exploit Erdös-Rényi random graphs in crowdsourcing experimental designs, one has to meet some conditions depending on our purpose:

1) *The resulting graph should be connected, if we hope to derive global scores for all videos in comparison*;

2) *The resulting graph should be loop-free in its clique complex, if we hope to get rid of the global inconsistency in harmonic component*.

The two conditions can be easily satisfied for large Erdös-Rényi random graphs.

Theorem 3. Let $G(n, p)$ be the set of Erdös-Rényi random graphs with n nodes and edge appearance probability p. Then the following holds as $n \to \infty$,

1. [Erdös-Rényi 1959] [12] if $p \succ \log n/n$, then $G(n, p)$ is almost always connected; and if $p \prec \log n/n$ then $G(n, p)$ is almost always disconnected;

2. [Kahle 2009] [19] if $p = O(n^\alpha)$, with $\alpha < -1$ or $\alpha > -1/2$, then the expected β_1 of the clique complex $\chi(G(n, p))$ is almost always equal to zero, i.e. loop-free.

These theories imply that when p is large enough, Erdös-Rényi random graphs will meet the two conditions above with high probability. In particular, almost linear $O(n \log n)$ edges suffice to derive a global ranking, and with $O(n^{3/2})$ edges harmonic-free condition is met.

However, it remains a question how to ensure that a given graph instance satisfies the two conditions? This issue is important since in the future we might even develop further experimental designs beyond merely Erdös-Rényi random graphs, e.g., considering confidence levels on edges. Fortunately, recent development in computational topology provides us such a tool, persistent homology.

3.3 Persistence Homology Barcodes

Persistence homology is firstly introduced by [11] in computational topology, and later developed by [35] into an algebraic theory. Roughly speaking, it provides us an online algorithm to compute the Betti numbers when simplexes enter in a sequential way. For more details of persistent homology, readers may refer to the surveys in [7, 10]. Here we just discuss in brief the application of persistent homology to monitor the number of connected components (β_0) and loops (β_1) .

To use persistent homology, we will put the nodes, edges and triangles in $\chi(G) = (V, E, T)$ in a linear order, such that a node appears no later than its associated edge and an edge no later than its associated triangle. For example, in random graph designs for video comparisons, we can assume the videos (nodes) come in a certain order (e.g. production time, or all created in the same time), after that pairs of videos (edges) are presented to us independently one by one. A triangle $\{i, j, k\}$ is created whenever all the three associated edges appeared. Persistent homology may return the evolution of the number of connected components (β_0) and the number of independent loops (β_1) at each time when a new node/edge/triangle is born. Figure 3 illustrates a birth process of clique complex and its associated Betti numbers (β_0 and β_1) that are computed and plotted by JPlex [29].

With the aid of persistent homology, one can compute the mean Betti numbers for random graphs. For example, with 16 videos ($n = 16$), the expected β_0 and β_1 (with 100 random graphs) are plotted in Figure 4(a). Note that with $p > 0.7$ with high probability the expected β_1 for $G(16, p)$ equals to 0. This phase transition probability will drop as the number of nodes increases, and this can be seen from the cases of $n = 32$ ($p > 0.5$) and $n = 64$ ($p > 0.4$), as plotted in Figure 4(b) and (c). As [19] shows, this probability asymptotically drops at the rate $p \sim n^{-1/2}$.

4. EXPERIMENTS

In this section, we systematically evaluate the effectiveness of our proposed HRRG method for subjective VQA. First, the dataset used for the experiments is briefly explained, followed by the experimental design of obtaining paired comparison data. Next, a procedure is presented to obtain the ground-truth as a set of complete data. Finally, the results with incomplete data are demonstrated with three random sampling schemes.

4.1 Dataset

We adopt the publicly-accessible database for VQA, LIVE Database [3], which includes 10 different reference videos and 15 distorted versions of each reference, for a total of 160 videos. The distorted videos are obtained using four different distortion processes—MPEG-2 compression (4 distorted videos per reference), H.264 compression (4 distorted videos per reference), lossy transmission of H.264 compressed bitstreams through simulated IP networks (3 distorted videos per reference) and lossy transmission of H.264 compressed bitstreams through simulated wireless networks (4 distorted videos per reference). Nine out of ten videos are 10 seconds long, while the 10th video is 8.68 seconds long. Seven sequences have a frame rate of 25 frames per second, while the remaining three have a frame rate of 50 frames per second. The videos are diverse in content and include a wide range
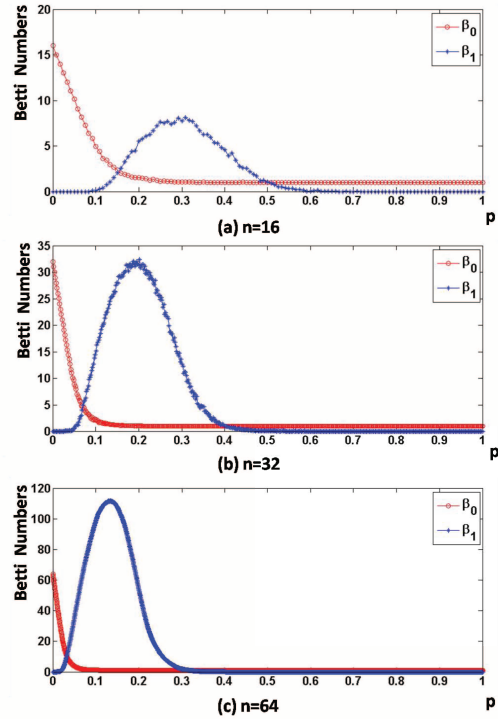

Figure 4: Average Betti numbers of random graphs.

of objects, textures, motions and camera movements. In subjective testing, the observers are asked to provide their opinion of video quality on a continuous scale. In other words, the MOS is adopted to analyze the perceived quality of each video. Note that we do not use the subjective scores in LIVE [3], we only borrow the video sources it provides. Different from LIVE [3], we propose to assess video quality with paired comparison.

4.2 Paired Data Collection

We now present our experiment design for collecting the set of paired data. The complete comparisons of this video database will require $10 \times \binom{16}{2} = 1200$ decisions. Considering that the order of presentation may bias final results, we need to balance them out at the design stage. A complete balancing-out would be achieved by repeating the whole experiment with the order that each pair reversed. However, this is too expensive and time-consuming. Therefore, our playlists will be based on a random permutation of 1200 test pairs with a random within-pair order. Moreover, we hope to avoid the situation with successive pairs of test videos from the same reference, to avoid contextual and memory effects in their judgments of quality. For this purpose, after the playlist 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.

A benefit of such a random presentation scheme is to make it impossible for participants to cheat our system by inputting "smart" answers. This is because the order of each pair and the order within each pair are totally random in
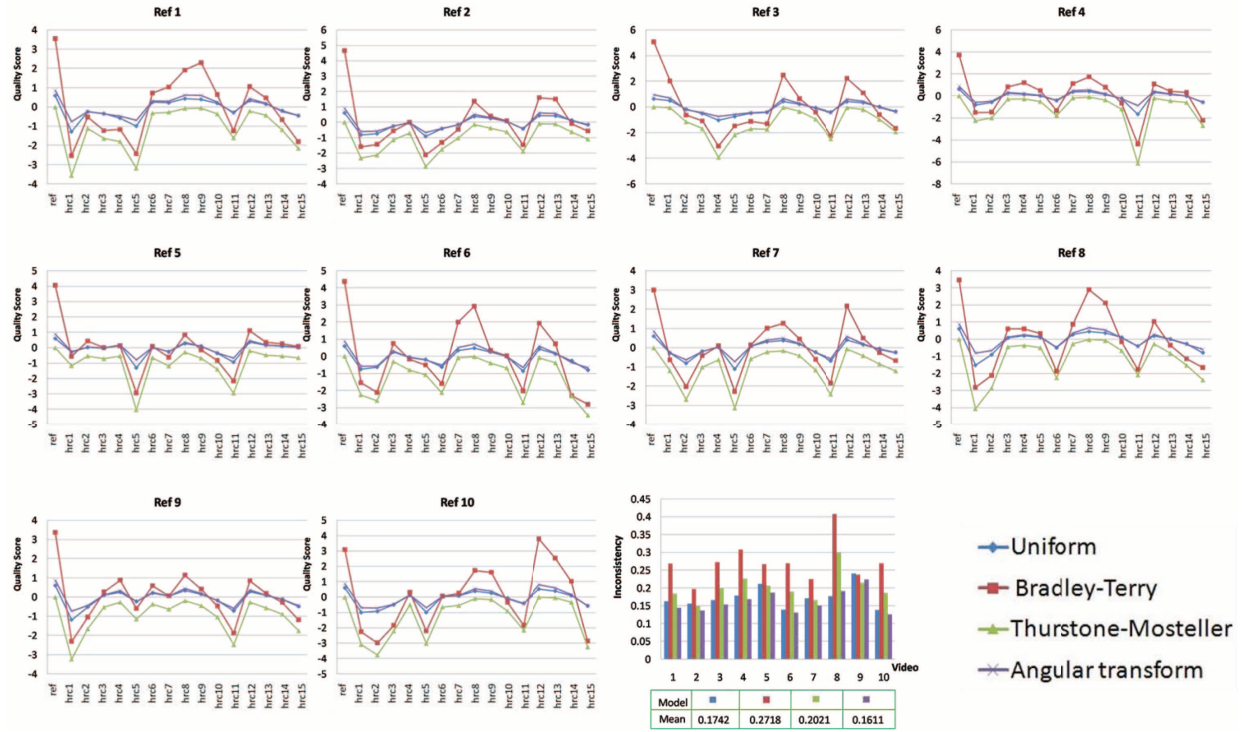
Figure 5: Global ranking scores and inconsistency distribution of 32 complete rounds on Live Database [3]. The *Angular transform* model has the smallest mean inconsistency at 0.1611 shown by the purple line.

each experiment, and the order is not disclosed to the participants before the test.

Before starting the experiment, each participant is briefed about the goal of the experiment and given a short training session to familiarize themselves with the testing procedure. In the testing process, videos are displayed at their native resolutions to prevent any distortions due to scaling operations performed by software or hardware. As each comparison takes approximately 30-40 seconds, the total time for each subjective experiment will vary from 10 up to 14 hours. According to [2], the execution time of one experiment by each observer should not exceed 30 minutes. Thus, we split the playlist into 30 sessions where each session consists of 40 video pairs and thus will not exceed a half hour. Finally 209 random observers, each of whom perform varied number of comparisons, provide 41800 paired comparisons in total. These paired comparisons can compose 32 rounds of complete comparisons. Because each round needs 1200 paired comparisons, the total number of comparisons for 32 rounds is $38400 = 32 \times 1200$.

The results of paired comparisons can be collectively summarized by $Y_{ij}^{\alpha,r}$, where for each reference video $r = 1, \ldots, 10$, $Y_{ij}^{\alpha,r}$ follows the same definition in Section 3.1 with $\alpha = 1, \ldots, 32$ for round (group) index. Note that the experiments below will run in the same way for each reference video, whence the index r will be omitted when we don't need to specify it. For each reference video, such paired comparison data can be represented by a directed graph (or hypergraph) with 16 nodes, and between every pair of nodes there are 32 directed edges indicating the preferences.

4.3 Experimental Results

4.3.1 Ground-truth

The purpose of this paper is to show that with some random samplings, incomplete data could provide good approximation of the results from the complete data. In other words, HodgeRank with incomplete data could well approximate the global ranking derived from complete data. Therefore, results obtained from 32 rounds of complete comparisons are treated as the ground-truth in our experiment. HodgeRank with such a complete and balanced data will be reduced to the Borda Count following (21). The *total inconsistency* is measured by

$$\text{Inc.Total}(\hat{Y}) = \frac{\|\hat{Y} - \hat{Y}^{(1)}\|_\omega^2}{\|\hat{Y}\|_\omega^2} = \frac{\sum_{ij} \omega_{ij}(\hat{s}_i - \hat{s}_j - \hat{Y}_{ij})^2}{\sum_{ij} \omega_{ij}\hat{Y}_{ij}^2},$$
(23)

which equals to the sum of $\|\hat{Y}^{(2)}\|_\omega^2/\|\hat{Y}\|_\omega^2$ (*global/harmonic inconsistency*) and $\|\hat{Y}^{(3)}\|_\omega^2/\|\hat{Y}\|_\omega^2\|$ (*local/curl inconsistency*). We also define the *harmonic percentage* as the ratio

$$\text{Percentage.Harm}(\hat{Y}) = \frac{\|\hat{Y}^{(2)}\|_\omega^2}{\text{Inc.Total}(\hat{Y})}.$$
(24)

With complete and balanced data, global/harmonic inconsistency vanishes according to Section 3.1.

Global ranking scores $\hat{s}$ for each reference and inconsistency distribution are given in Figure 5, where "ref" represents 10 different reference videos in Live Database [3] and "hrc1-15" are 15 distorted versions of each reference. It can

be seen that Hodge decomposition with *Angular transform* model has the smallest mean inconsistency. Therefore, we will adopt this model in the following experiments for incomplete data.

4.3.2 Results of Incomplete Data

In the following experiments, we adopt the Kendall rank correlation (τ) coefficient [20] to measure the rank correlation between global rankings from complete and incomplete data. Given two global 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

$$\tau(x,y) = \frac{\sum_{\{i,j\}\in E} X_{ij}Y_{ij}}{\sqrt{\sum X_{ij}^2 \cdot \sum Y_{ij}^2}}, \qquad (25)$$

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 coefficient is normalized such that two identical rankings produce a correlation of $+1$, while a ranking and its perfect inverse gives a -1, and the expected correlation of two random rankings is 0.

As illustrated in Section 3.3 Figure 4, for $n = 16$, if more than 25% random edges are added, the resulting graph is connected with high probability, and with more than 70% edges, the resulting clique complex is loop-free with high probability. Connectivity is necessary if we would like to derive a global score on all videos. The existence of harmonic ranking may jeopardize the global score by incurring the fixed tournament issue. To illustrate this point, Figure 6 shows the Kendall's τ, total inconsistency, harmonic inconsistency, and percentage of harmonic over total inconsistency, against the edge sampling percentage ranging from 20% to 80%. In this example, harmonic inconsistency accounts for more than 50% total inconsistency before 25% edges, and rapidly drops to zero after 70% edges (where Kendall's τ coefficient goes beyond 0.9 and total inconsistency stabilizes below 0.2).

Therefore in the following experiments, to avoid the possible issue of harmonic ranking, we choose an upper bound for the thresholding probability above, *i.e.* 75% ($120 \times 0.75 = 90$) for $n = 16$ node graphs. In this case, with high probability the total inconsistency will be fully characterized by the local inconsistency. For general large Erdös-Rényi graphs, we can choose any upper bound for $p = O(n^{-1/2})$ with $O(n^{3/2})$ edges. Note that such a choice is only a sufficient condition to avoid harmonic ranking. In the cases where harmonic inconsistency is small enough, one can choose much smaller thresholding probability, up to $p = O(n^{-1} \log n)$ with almost linear $O(n \log n)$ edges which is the lower bound to guarantee connectivity.

In the following, we study three random sampling schemes, guided by this principle from Erdös-Rényi random graphs. These sampling schemes exhibit an increasing order of freedom, which is thus increasingly adaptive to crowdsourcing applications.

The following sampling schemes are the same for each r of the reference video. The first sampling scheme, called here *group balanced random sampling*, draws 75% random pairs from each round $\alpha \in \{1, \dots, 32\}$ of complete comparisons. Note that each group thus has the same number of pairs. The second scheme, called here *group imbalanced ran-*
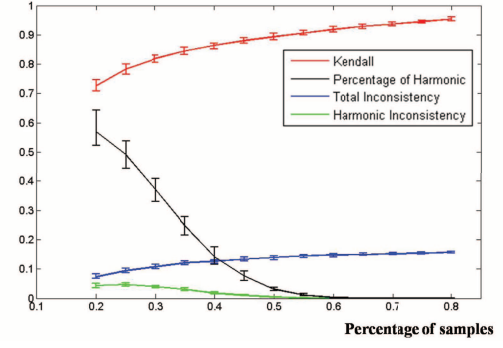

Figure 6: Inconsistency decomposition and Kendall's τ to ground truth, versus percentage of edge samples. Results are averaged over 10 reference videos with 100 bootstrapped samples.

Table 1: Kendall's τ and inconsistency of Exp-I.

	min	mean	max	std
Kendall's τ	0.9567	0.9716	0.9917	0.0058
Inconsistency	0.1643	0.1740	0.1817	0.0032

dom sampling, draws 75% random paired comparisons from the whole collection over $\alpha = 1, \dots, 32$, with some groups covering possibly more than 75% of edges while others less. The third one, called here *sufficient-coverage random sampling*, simply draws arbitrary number of paired comparisons from the whole collection, with the only requirement that at least 75% distinct pairs are covered. We will examine the average performance of each sampling. Other complicated sampling schemes are possible. However, we focus on the three schemes here due to their simplicity and thus can be regarded as a foundation for further developments.

All the three sampling schemes above could guarantee that 75% distinct pairs of 16 videos are compared, whence with high probability the harmonic ranking induced from such data vanishes.

Exp-I: group balanced random sampling

The first sampling scheme is studied here. For each of the 10 reference videos, and for each of the 32 groups of complete paired comparisons, 75% (i.e. 90) video pairs are drawn independently without replacement from 120 edges. The name, *group balanced random sampling*, refers to the fact that each group contains the same number of pairs. Then, HodgeRank (4) is applied to obtain quality scores of each video from this incomplete dataset. To ensure the statistical stability, we run the random sampling process 100 times.

Table 1 shows the mean Kendall's τ and inconsistency results of 100 times achieved by this scheme. From these experimental results, we make the following comments.

First, it is shown that this sampling approach could provide good approximate results of the complete data, with an average Kendall's τ of 0.9716.

Second, we observe that the mean inconsistency of this

Table 2: Kendall's τ and inconsistency of Exp-II.

	min	mean	max	std
Kendall's τ	0.9550	0.9699	0.9867	0.0066
Inconsistency	0.1661	0.1734	0.1812	0.0031

sampling scheme (0.1740) is only slightly higher than the complete data (0.1611), which is within our acceptable level.

In a summary, these results show that the random sampling mechanism based on random graph theory is good in providing approximations of complete data via HodgeRank.

Exp-II: group imbalanced random sampling

Different from Exp-I, for each of the 10 reference videos, in this experiment we randomly select 75% of all paired comparisons, i.e. $32 \times 120 \times 0.75 = 2880$ comparisons. The number of pairs selected from each of the 32 rounds in this case might be different, where the name, *group imbalanced random sampling*, comes from. However, the total number of the selected samples for each reference video is the same to Exp-I.

Table 2 shows the mean Kendall's τ-coefficients and inconsistencies which indicates that this method could produce comparable results of Exp-I. The Kendall's τ between the ground-truth and the 100 random experiments has an average of 0.9699 ± 0.0066 (in contrast to 0.9716 ± 0.0058 in Exp-I), while the average inconsistency is 0.1734 ± 0.0031 (in contrast to 0.1740 ± 0.0032 in Exp-I).

Exp-III: sufficient-coverage random sampling

Our last experiment is to examine the performance under *sufficient-coverage random sampling* settings. That is, we randomly sample arbitrary number of pairs from each of 32 rounds, with the only requirement that at least 75% (90) edges of the graph are covered, i.e. at least a paired comparison occurs on 75% (90) distinct video pairs. Due to the multiple comparisons between a pair of videos, a natural question is raised that how many percentage of samples are needed to reach this 75% coverage. Similar to Erdös-Rényi random graph theory on simple graphs, Figure 7 draws percentage of samples versus median number of edges (or distinct pairs) covered. As we can see, after 4% of samples on this hypergraph, with high probability 75% distinct pairs will be covered.

Moreover, Table 3 shows the mean Kendall's τ-coefficient and inconsistency of 100 random experiments. The high average performance (0.9337 ± 0.0415 for Kendall's τ and 0.2256 ± 0.0606 for inconsistency) shows the promising application of this simple sampling scheme.

More complicated sampling schemes can be developed based on various combinations of the prototype methods above. Considering a more practical situation, the available time that can be used for testing of each participant may be distinctly different. From this point of view, the number of pairs (k) shown to participants can be adjusted according to their time limit. That is, when a participant's time is comparatively adequate, k can be set up as a larger value and thus could provide opportunity for him/her to view more pairs. But if he/she is under the pressure of limited time, k should be smaller. In the extreme case for online crowd-
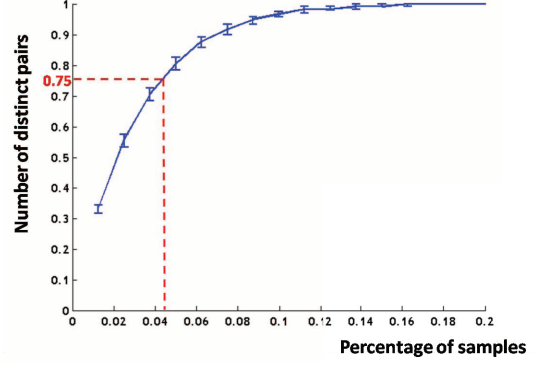

Figure 7: Percentage of samples versus number of distinct pairs. For each percentage level, the experiments are repeated 100 times and the median number of distinct pairs with $[0.25, 0.75]$ confidence interval are plotted in the figure.

Table 3: Kendall's τ and inconsistency of of Exp-III.

	min	mean	max	std
Kendall's τ	0.8067	0.9337	0.9857	0.0415
Inconsistency	0.1623	0.2256	0.3777	0.0606

sourcing experiments, a complete random sampling scheme can be used to select both participants and video pairs randomly. As shown above, HodgeRank and random graph theory can guide us to find the global ranking efficiently.

5. CONCLUSIONS

In this paper, we have proposed an efficient approach called **HodgeRank on Random Graphs** towards subjective VQA. Our approach is based on random graph theory and HodgeRank on graphs. In particular, we study three random sampling schemes inspired by Erdös-Rényi random graph theory, followed by HodgeRank to analyze the incomplete and imbalanced data collected in these ways. In these sampling schemes, participants only need to perform a random fraction of all possible paired comparisons. But with a sufficiency of coverage satisfied, HodgeRank may give reliable results without jeopardizing the accuracy of the result. In contrast to the traditional deterministic incomplete block designs, our random design is not only suitable for traditional laboratory and focus-group studies, but also fit for crowdsourcing experiments on Internet.

Additionally, we would like to point out here that inconsistencies are not necessarily due to untrustworthy inputs of the participants but may very well be an inherent characteristic of the data. In future, we plan to focus on this problem to investigate into more sophisticated techniques to detect the part of the inconsistencies only resulting from the careless inputs of participants. Moreover, with the rapid advent of technologies on rich user interface, the proposed framework can be extended to assess users' experience in interactive applications with an online learning setting where random graph models may take into account of sampling order (e.g. preference attachment graphs), which will also be part of our future work.

6. ACKNOWLEDGMENTS

7. REFERENCES

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

[2] *ITU-R Methodology for the Subjective Assessment of the Quality of Television Pictures.* 2002.

[3] LIVE video quality assessment database. http://live.ece.utexas.edu/research/quality/, 2008.

[4] O. Alonso, D. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.

[5] K. Arrow. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58(4):328–346, 1950.

[6] D. Brabham. Crowdsourcing as a model for problem solving: an introduction and cases. *Convergence*, 14(1):75, 2008.

[7] G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

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

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

[10] H. Edelsbrunner and J. Harer. Computational topology : an introduction. 2010.

[11] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete and Computational Geometry*, 28(4):511–533, 2002.

[12] P. Erdos and A. Renyi. On random graphs i. *Publ. Math. Debrecen*, 6:290–297, 1959.

[13] Y. Freund, R. Iyer, R. Shapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4(6):933–969, 2004.

[14] T. Hastie and R. Tibshirani. Classification by pairwise coupling. *Annals of Statistics*, 26(2):451–471, 1998.

[15] A. Hekstra, J. Beerends, D. Ledermann, F. de Caluwe, S. Kohler, R. Koenen, S. Rihs, M. Ehrsam, and D. Schlauss. PVQM-A perceptual video quality measure. *Signal Processing: Image Communication*, 17:781–798, 2002.

[16] R. Herbrich, T. Graepel, and K. Obermayer. *Large margin rank boundaries for ordinal regression.* MIT Press, 2000.

[17] J. Howe. The rise of crowdsourcing. *Wired Magazine*, 14(6):176–183, 2006.

[18] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 2010.

[19] M. Kahle. Topology of random clique complexes. *Discrete Mathematics*, 309:1658–1671, 2009.

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

[21] M. Kendall and B. Smith. On the method of paired comparisons. *Biometrika*, 31(3-4):324–345, 1940.

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

[23] I. Koutis, G. Miller, and R. Peng. Approaching optimality for solving sdd systems. FOCS '10 51st Annual IEEE Symposium on Foundations of Computer Science, 2010.

[24] J. Lubin. The use of psychophysical data and models in the analysis of display system performance. pages 163–178. Digital Images and Human Vision, 1993.

[25] G. Noether. Remarks about a paired comparison model. *Psychometrika*, 25:357–367, 1960.

[26] S. Nowak and S. Ruger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. International conference on Multimedia information retrieval, 2010.

[27] M. Pinson and S.Wolf. A new standardized method for objectively measuring video quality. *IEEE Transactions on Broadcasting*, 50(3):312–322, September 2004.

[28] T. Saaty. A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, 15(3):234–281, 1977.

[29] 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.

[30] A. Sorokin and D. Forsyth. Utility data annotation with Amazon Mechanical Turk. pages 1–8. Computer Vision and Pattern Recognition Workshops, June 2008.

[31] D. Spielman and S.-H. Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. STOC '04 Proceedings of the thirty-sixth annual ACM symposium on Theory of computing, 2004.

[32] L. Thurstone. The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology*, 27:384–400, 1927.

[33] C. van den Branden Lambrecht and O. Verscheure. Perceptual quality measure using a spatiotemporal model of the human visual system. *SPIE Conference on Human Vision and Electronic Imaging*, 2668(1):450–461, March 1996.

[34] Z. Wang, L. Lu, and A. Bovik. Video quality assessment based on structural distortion measurement. *Signal Processing: Image Communication*, 19(2):121–132, February 2004.

[35] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete and Computational Geometry*, 33(2):249–274, 2005.