HodgeRank on Random Graphs for Subjective Video Quality Assessment

Qianqian Xu, Qingming Huang, *Senior Member, IEEE,* Tingting Jiang, Bowei Yan, Weisi Lin, *Senior Member, IEEE,* and Yuan Yao*

***Abstract*—This paper introduces a novel framework, *HodgeRank on Random Graphs* (HRRG), based on paired comparison, for subjective video quality assessment. Two types of random graph models are studied, *i.e.*, Erdös-Rényi random graphs and random regular graphs. Hodge decomposition of paired comparison data may 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 video database. Both of the two random designs are promising sampling methods without jeopardizing the accuracy of the results. In particular, due to balanced sampling, random regular graphs may achieve better performances when sampling rates are small. However, when the number of videos is large or when sampling rates are large, their performances are so close that Erdös-Rényi random graphs, as the simplest I.I.D. (independent and identically distributed) sampling scheme, could provide good approximations to random regular graphs, as a dependent sampling scheme. In contrast to the traditional deterministic incomplete block designs, our random design is not only suitable for traditional laboratory studies, but also for crowdsourcing experiments on Internet where the raters are distributive and it is hard to control with fixed designs.**

***Index Terms*—Video Quality Assessment, Paired Comparison, HodgeRank, Random Graphs, Persistence Homology**

I. 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 and averaged to evaluate the quality of each video sequence. This process is labor-intensive and time-consuming. Therefore, there has been an increasing demand to build intelligent, objective quality measurement models (see [1] for a survey) 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 [2], individuals are asked to give a rating from Bad to Excellent (Bad-1, Poor-2, Fair-3, Good-4, and Excellent-5) to evaluate the quality of a video. However, such a test may suffer from the following problems [3]:

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 [3]. 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, the paired comparison approach leaves a heavier burden on participants with a larger number of comparisons. For example, if we are given 15 distorted versions of 1 reference video, 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 direct paired comparison method. When the number of videos to be judged is large, it may be practically impossible, or at least unacceptable from the viewpoint of the participants. In addition, if the test time for a single participant lasts too long [4], participants may lose patience and thus may input random decisions carelessly or intentionally. Therefore, how to make paired comparison method efficient, reliable and applicable in reality has become an urgent issue in the VQA research.

A natural strategy to address this issue is to expose every participant with only a fraction of all possible paired comparisons. Hence it raises a question: how to choose the pairs that will be viewed by participants? There has been a

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, in part by National Natural Science Foundation of China: 61025011, 61071157 and 60833006.

*Corresponding author.

Q. Xu is with the Graduate University of Chinese Academy of Sciences, Beijing 100190, China, (e-mail: qqxu@jdl.ac.cn).

Q. Huang is with the Graduate University of Chinese Academy of Sciences, and the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, (e-mail: qmhuang@jdl.ac.cn).

T. Jiang is with the National Engineering Lab. for Video Technology, Key Lab of Machine Perception (MoE), School of EECS, Peking University, Beijing 100871, China, (e-mail: ttjiang@pku.edu.cn).

B. Yan is with the School of Mathematical Sciences, Peking University, Beijing 100871, China, (e-mail: yanbowei@gmail.com).

W. Lin is with the School of Computer Engineering, Nanyang Technological University, Nanyang 639798, Singapore, (e-mail: wslin@ntu.edu.sg).

Y. Yao is with the School of Mathematical Sciences, LMAM and LMP, Peking University, Beijing 100871, China, (e-mail: yuany@math.pku.edu.cn).

large literature in statistics on deterministic incomplete block design [5]. However, these designs may not be suitable for crowdsourcing on Internet where raters are distributive over Internet with varied backgrounds and it is hard to control with traditional experimental designs. To meet this challenge, the work in [6] proposes a randomized paired comparison method which randomly selects small subsets of pairs for each assessor to view; the work shows that randomization is effective in reducing costs of complete design without losing the main effects. However it leaves some open problems arising from randomization: (1) how to systematically deal with the resulting imbalanced and incomplete data; (2) how many samples one needs to achieve certain performance.

In this paper, we propose a general framework to analyze the imbalanced and incomplete data in randomized paired comparison experiments, ***HodgeRank on Random Graphs*** (HRRG). In this framework, every item (*e.g.* video) in comparison is regarded as a graph node and an assessor collects random samples of node pairs or edges, independently and with an identical distribution (I.I.D.), or in a more complicated way. Two particular random graph models are investigated in this paper: (1) Erdös-Rényi random graphs which model the I.I.D. sampling scheme; (2) random k-regular graphs which can be sampled in a more complicated way but result in balanced paired comparison data, *i.e.* every video receives the same number of comparisons. Paired comparison data are then mapped to edge flows on such random graphs. Equipped with a recent new development of Hodge theoretical approach to statistical ranking [7], we can infer a reliable global ranking from such data with a much less sampling complexity than complete design. The proposed methodology is not only suitable for traditional laboratory studies, but also fits for crowdsourcing experiments. It provides an answer to the two open problems above.

HodgeRank [7] is a general framework to decompose paired comparison data on graphs, possibly imbalanced (where different video pairs may receive different number of comparisons) and incomplete (where every participant may only give partial comparisons), into three orthogonal components. In these components HodgeRank not only provides us a mean to determine a global ranking from paired comparison data under various statistical models (*e.g.* Thurstone-Mosteller and Bradley-Terry etc.), but also measures the inconsistency of the global ranking obtained. The inconsistency shows the validity of the ranking obtained and can be further studied in terms of its geometric scale, namely whether inconsistency in the ranking data arises locally or globally. Local inconsistency can be fully characterized by triangular cycles, while global inconsistency involves cycles consisting nodes more than three, which may arise due to data incompleteness and once presented with a large component indicates some serious conflicts in ranking data. However through random graphs, we can efficiently control global inconsistency.

Although HodgeRank can be applied to general graphs, two particular random graph models are studied in this paper due to their potential importance in crowdsourcing experiments. They are Erdös-Rényi random graphs and random regular graphs. In Erdös-Rényi random graphs with n nodes, every edge will be sampled with probability p in an I.I.D. way. For large Erdös-Rényi random graphs, asymptotic theoretical results tell us that it is necessary to have $p \succ O(n^{-1}\log n)$ such that the graph is connected and global ranking is thus possible; to avoid global inconsistency, it suffices to have larger sampling rates at $p \succ O(n^{-1/2})$. Random k-regular graphs can be generated in a more complicated way with dependent edge sampling process. They are incomplete Round-Robin tournaments with balanced paired comparisons. Such a balanced feature is important to HodgeRank as it leads to small condition numbers of graph Laplacians, whence with stabler or faster solutions. For sparse graphs where a small number of paired comparisons are made, random regular graphs will lead to a better performance than Erdös-Rényi random graphs; but for large graphs with large k, random regular graphs will converge to Erdös-Rényi random graphs asymptotically as observed in both theory and experiments.

We further demonstrate the effectiveness and generality of the proposed framework on LIVE video database [8], which includes 10 different reference videos and 15 distorted versions of each reference, for a total of 160 videos. Experimental results show that the proposed framework is promising and has potentially wide applications in subjective VQA.

The main contributions of our work include the following:

1. We propose a novel framework of HodgeRank with random graphs to quantify the quality of video. Hodge (Helmholtz) decomposition on graphs is introduced to derive, from incomplete and imbalanced data, quality scores of videos and the inconsistency of participants' judgments. The rating procedure is efficient, labor-saving, and more importantly, without jeopardizing the accuracy of the results.

2. To conduct paired comparisons, two random design schemes are proposed based on Erdös-Rényi random graphs and random regular graphs with sampling complexity studies. For large random graphs, $O(n\log n)$ distinct random edges are needed to guarantee graph connectivity and thus to achieve any global ranking, but $O(n^{3/2})$ distinct random edges are sufficient to avoid global inconsistency. For sparse random graphs, random regular graphs may lead to better performance in HodgeRank than Erdös-Rényi random graphs due to the balanced property.

This paper is an extension of our conference paper [9], which only studies HodgeRank with Erdös-Rényi random graphs. The following distinctions are made in this paper: a minor one is to show by an example how HodgeRank can be applied to select assessors according to their inconsistency, while the major one lies in a systematic treatment with two different but closely related types of random graphs, Erdös-Rényi random graphs and random regular graphs. The reason to choose random regular graphs lies in their balanced feature, *i.e.*, every video is compared with the same number of alternatives. Such graphs have small condition numbers in their graph Laplacians and thus affect stability of HodgeRank. The following outlines the main results about HodgeRank of the two random graph models.

- Similar qualitative topological phase transitions are observed for these two types of random graphs and it shows that random regular graphs are easier to satisfy the loop-

free condition than Erdös-Rényi random graphs when adding the same number of distinct pairs;
- When random graphs are sparse (the number of edges added is small), random k-regular graphs have better performance in HodgeRank than Erdös-Rényi random graphs in HodgeRank due to balanced nature;
- When k is large or in an overall performance measure, HodgeRank with Erdös-Rényi random graphs provides good approximations to that with random regular graphs, which meets the theoretical conjecture that as the number of videos grows and edges are dense enough, Erdös-Rényi random graphs asymptotically converge to random regular graphs [10].

Therefore, both models are good candidates for researchers depending on their specific applications.

The remainder of this paper is organized as follows. Section II contains a review of related works. Then Section III establishes the Hodge decomposition theory, as well as the principles for random sampling grounded in random graph theory. The detailed experiments are demonstrated in Section IV. Section V presents the conclusive remarks along with a discussion for future work.

II. RELATED WORK

A. Paired Comparison

Paired comparison 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, psychological, statistical, and computer science [5], [11]–[14]. It has also drawn increasing attention from the machine learning community as it may be adapted to classification problems [15]–[17].

There have been studies on the design of subjective tests to evaluate video quality in paired comparison method. One such example is [3], which proposed a crowdsourceable framework based on paired comparison. However, one major shortcoming of [3] lies in that it makes a strong assumption that all paired comparison data collected are complete which is impossible for a large number of videos. For example, the way to evaluate Transitivity Satisfaction Rate (TSR) depends on such complete design assumption. To address this issue, the work in [6] suggests a Randomised Pair Comparison method in which a random subset of all pairs are chosen for different participants to reduce the number of comparisons. However, this work does not address how to deal with the imbalanced and incomplete data arisen in random sampling, and also leaves open the issue of how much samples one needs.

In this paper, we present a new framework based on HodgeRank [7] on random graphs, which deals with incomplete and imbalanced data distributed on random graphs and further derives the constraints on sampling complexity in crowdsourcing experiment that the random selection must adhere to.

B. Crowdsourcing

Crowdsourcing is the act of outsourcing tasks, traditionally performed by an employee or contractor, to an undefined, large group of people or community (a "crowd"), through an open call [18]. The difference between crowdsourcing and ordinary outsourcing is that a task or problem is outsourced to an undefined Internet public rather than a specific group of people.

With the growth of crowdsourcing platforms, such as Amazon Mechanical Turk (MTurk) [19], it becomes more and more popular to ask an Internet crowd to conduct experiments on their personal computers. For example, researchers can seek help from the Internet crowd to conduct user studies on image annotation [20], [21], document relevance [22], and document evaluation [19], etc.

C. Inconsistency Checking

After collecting the paired comparison 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 [3] proposes Transitivity Satisfaction Rate (TSR) to measure the consistency of a participant's judgments, which checks all the intransitive (circular) triangles such that $A \succ B \succ C \succ A$ where $\succ$ indicates preference. The TSR is defined as the number of transitive judgment triplets 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 comparison data is incomplete with missing edges, it does not suffice to check triangular cycles. In this case, Hodge decomposition on graphs will give us a general treatment of inconsistency which considers not only triangular cycles but also global cycles like $A \succ B \succ C \succ D \ldots \succ A$.

D. Random Graphs

In this paper, we consider a random graph as a graph generated by some random process [23], [24]. 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 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 [25] which is a stochastic process that starts with n vertices and no edges, and at each step adds one new edge uniformly. Such models can be viewed as a sampling process of video pairs or edges independently and identically distributed (I.I.D.). Another popular model, which is called random regular graph, can be regarded by taking a graph uniformly at random from the set of all simple regular graphs on n vertices [26]. There are various processes to generate random regular graphs, among which the most popular approach is perhaps by random matching [27]. In paired comparison methods, regular graphs occur in the designs of incomplete Round-Robin-Tournaments where

every competitor receives the same number of comparisons, often called balanced designs [5]. Such balanced designs with incomplete blocks are believed to be important to create a relatively fair scenario for all the participants without calling the complete game. In HodgeRank they are also important because regular graphs have small condition numbers in graph Laplacians which leads to stable solutions.

There are some other kinds of random models, such as preferential attachment random graph [28], small world random graph [29], and geometric random graph [30], which may also play important roles in HodgeRank under certain circumstances. The general principle of HodgeRank can be applied to all these different models. However, in this paper we particularly focus on the first two types of random graphs, Erdös-Rényi and regular graphs, leaving other models for future studies.

III. HodgeRank on Random Graphs

In this section, we propose two new random design principles to conduct paired comparison and analyze data for a reliable global ranking and inconsistency. Our sampling mechanism exploits the Erdös-Rényi random graph and random regular graph, two different but closely related models. 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.

A. 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 participant α, *i.e.*, $Y_{ij}^{\alpha} = -Y_{ji}^{\alpha}$ representing the degree that α prefers i to j. The simplest setting is the binary choice, where

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

In 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 [7]. This includes the following additional scales often used in VQA [6]: (1) 3-point Likert scale which contains a neutral element in addition to a preference for i or j; (2) 4-point Likert scale which provides choices for weak and strong preference for either i or j, without a neutral element; (3) 5-point Likert scale which adds a neutral element into the 4-point likert scale.

In this paper we shall focus on the binary choice, which is the simplest setting and the data collected in this paper belongs to this case. However the theory can be applied to the more general case with multiple choices above.

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.

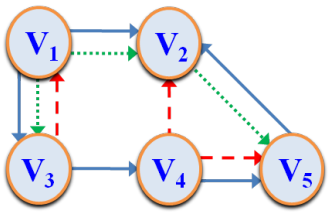


Fig. 1. An example of paired comparison hypergraph for 5 videos.

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

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

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

Our statistical rank aggregation problem is to look for some global ranking score $s : V \to R$ such that

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

which is equivalent to the following weighted least square problem

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

where $\hat{Y}_{ij} = (\sum_{\alpha} \omega_{ij}^{\alpha} Y_{ij}^{\alpha}) / (\sum_{\alpha} \omega_{ij}^{\alpha})$ and $\omega_{ij} = \sum_{\alpha} \omega_{ij}^{\alpha}$. For the principles behind such a choice, readers may refer [7].

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.*,

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

A pairwise 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 called *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 edges. The existing paired comparison methods in VQA often assume complete and balanced data [3]. However, this is an unrealistic assumption for real world data, *e.g.* randomized experiments [6]. Moreover in crowdsourcing, raters and videos come in an unspecified way and it is hard to control the test process with precise experimental designs. Nevertheless, as to be shown below, it is efficient to utilize some random sampling design based on random graph theory where for each participant a fraction of 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 emerged from random sampling in paired comparisons.

The minimization problem (4) can be generalized to a family of *linear models* in paired comparison methods [5]. 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 (a_{ji} carries the opposite meaning). 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

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

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 [31], assume that the true preference probability can be fully decided by a linear scaling function on V, *i.e.*,

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

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 the inverse of F,

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

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

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

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 [5].

1. *Uniform* model:

$$\hat{Y}_{ij} = 2\hat{\pi}_{ij} - 1. \tag{10}$$

2. *Bradley-Terry* model:

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

3. *Thurstone-Mosteller* model:

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

where F is essentially the Gauss error function

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

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

4. *Angular transform* model:

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

This model is created for the so called variance stabilization property: asymptotically $\hat{Y}_{ij}$ has variance only depending on number of ratings on edge $\{i, j\}$ or the weight ω_{ij}, but not on the true probability p_{ij}.

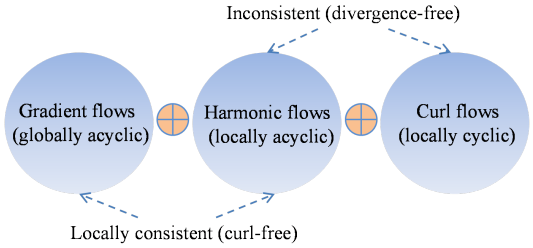


Fig. 2. Hodge decomposition (three orthogonal components) of paired rankings [7].

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 [7] 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.

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

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 $\hat{Y}_{ij} = 0$ otherwise. There is a unique decomposition of $\hat{Y}$ satisfying

$$\hat{Y} = \hat{Y}^g + \hat{Y}^h + \hat{Y}^c, \tag{16}$$

where

$$\hat{Y}^g_{ij} = \hat{s}_i - \hat{s}_j, \quad \text{for some } \hat{s} \in \mathrm{R}^V, \tag{17}$$

$$\hat{Y}^h_{ij} + \hat{Y}^h_{jk} + \hat{Y}^h_{ki} = 0, \text{ for each } \{i, j, k\} \in T, \tag{18}$$

$$\sum_{j \sim i} \omega_{ij} \hat{Y}^h_{ij} = 0, \text{ for each } i \in V. \tag{19}$$

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

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

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,

$$\hat{s} = \Delta_0^\dagger \delta_0^* \hat{Y}, \tag{21}$$

where $\delta_0^* = \delta_0^T W$ ($W = \mathrm{diag}(\omega_{ij})$), $\Delta_0 = \delta_0^* \cdot \delta_0$ is the unnormalized graph Laplacian defined by $(\Delta_0)_{ii} = \sum_{j\sim i} \omega_{ij}$ and $(\Delta_0)_{ij} = -\omega_{ij}$, and $(\cdot)^\dagger$ is the Moore-Penrose (pseudo) inverse. On a complete and balanced graph, (21) is reduced to $\hat{s}_i = \frac{1}{n-1}\sum_{j\neq i}\hat{Y}_{ij}$, often called *Borda Count* as the earliest preference aggregation rule in social choice [7]. For expander graphs like regular graphs, graph Laplacian Δ_0 has small condition numbers and thus the global ranking is stable against noise on data.

Algorithm 1: Procedure of Hodge Decomposition in Matlab Pseudocodes

Input: A paired comparison hypergraph G provide by assessors.
Output: Global score $\hat{s}$, gradient flow $\hat{Y}^g$, curl flow $\hat{Y}^c$, and harmonic flow $\hat{Y}^h$.

Initialization:
$\hat{Y}$ (a numEdge-vector consisting $\hat{Y}_{ij}$ defined),
W (a numEdge-vector consisting ω_{ij}).
Step 1:
Compute δ_0, δ_1; // δ_0 = gradient, δ_1 = curl
$\delta_0^* = \delta_0^T * diag(W)$; // the conjugate of δ_0
$\triangle_0 = \delta_0^* * \delta_0$; // Unnormalized Graph Laplacian
$div = \delta_0^* * \hat{Y}$; // divergence operator
$\hat{s} = lsqr(\triangle_0, div)$; // global score
Step 2:
Compute 1st projection on gradient flow: $\hat{Y}^g = \delta_0 * \hat{s}$;
Step 3:
$\delta_1^* = \delta_1^T * diag(1./W)$;
$\triangle_1 = \delta_1 * \delta_1^*$;
$curl = \delta_1 * \hat{Y}$;
$z = lsqr(\triangle_1, curl)$;
Compute 3rd projection on curl flow: $\hat{Y}^c = \delta_1^* * z$;
Step 4:
Compute 2nd projection on harmonic flow: $\hat{Y}^h = \hat{Y} - \hat{Y}^g - \hat{Y}^c$.

2. $\hat{Y}^h$ 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 $\hat{Y}^c$ 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,

$$\mathrm{curl}^r_{ijk} = \frac{|\hat{Y}_{ij} + \hat{Y}_{jk} + \hat{Y}_{ki}|}{|\hat{Y}_{ij}| + |\hat{Y}_{jk}| + |\hat{Y}_{ki}|} = \frac{|\hat{Y}^c_{ij} + \hat{Y}^c_{jk} + \hat{Y}^c_{ki}|}{|\hat{Y}_{ij}| + |\hat{Y}_{jk}| + |\hat{Y}_{ki}|} \in [0,1], \tag{22}$$

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

Figure 2 illustrates the Hodge decomposition for paired comparison flows and Algorithm 1 shows how to compute global ranking and other components. The readers may refer to [7] 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 ranking $\hat{s}$ in (21), the recent developments of Spielman-Teng [32] and Koutis-Miller-Peng [33] suggest fast (almost linear in $|E|\mathrm{Poly}(\log|V|)$) algorithms for this purpose.

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

3. A nontrivial harmonic component $\hat{Y}^h \neq 0$ 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.*, $1 \succ 2 \succ 3 \succ 4 \succ \ldots \succ n \succ 1$. All the inconsistency will be summarized in those triangular cycles, *e.g.*, $i \succ j \succ k \succ i$.

Theorem 2. The linear space of harmonic flows has the 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, with the aid of some random sampling principles, it is not hard to obtain graphs whose β_1 are zero.

B. Random Graphs

In this section, we first describe two classical random models: Erdös-Rényi random graph and random regular graph; then we investigate the relation between them.

1) Erdös-Rényi Random Graph: Erdös-Rényi random graph $G(n,p)$ starts from n vertices and draws its 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. sampling. Therefore, such a model is to be systematically studied in the paper.

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

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

2. *The resultant 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 graph.

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] [25] 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;

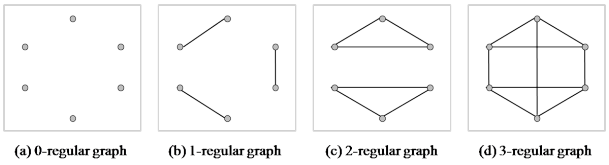


Fig. 3. Examples of k-regular graphs.

2. [Kahle 2009] [34] 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 graph 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.

Despite such an asymptotic theory for large random graphs, it remains a question how to ensure that a given graph instance satisfies the two conditions? Fortunately, the recent development in computational topology provides us such a tool, persistent homology, which will be illustrated in Section III-C.

2) Random Regular Graph: In data collection of paired comparisons, *balance* is often a desired property in the sense that every video has the same number of comparisons against others. This requires the graph to be regular. For example, Round-robin tournaments in sports are one of the most popular paired comparison method, which has balanced data in the sense that every participant has an equal chance against all other participants. This is the fairest way to determine a champion as the element of luck is seen to be reduced compared to a knockout system. A participant's final record can represent his/her true athletics level more accurately since it was generated by equal competition with all the participants.

However, complete Round-robin tournaments needs $\binom{n}{2}$ paired comparisons, *i.e.*, a complete paired comparison graph, which is a heavy burden when n is large. To provide a reliable result using fewer rounds than a complete Round-robin, k-regular graphs are adopted in this paper, as incomplete designs of Round-robin tournaments with less amount but still balanced data. In k-regular graphs, each node has the same number of neighbors; *i.e.*, every node has the same degree k, *e.g.* in Figure 3.

To meet the stochastic situation in crowdsourcing ranking, we exploit random regular graphs in this paper. A random matching algorithm in [27] is used to generate random regular graphs. Random regular graphs can thus be regarded as taking a graph uniformly at random from the set of all simple regular graphs on vertices. Here, we use $\zeta(n,k)$ to denote the uniform probability space of k-regular graphs on the n vertices $\{1,2,...,n\}$.When $k = n-1$, the graph generated is a complete and balanced graph.

3) Connections between Erdös-Rényi Random Graph and Random Regular Graph: There are close connections between Erdös-Rényi random graphs and random regular graphs. An easy observation is that when $p \succ \log n/n$, $G(n,p)$ has all vertex degrees tightly concentrated around their mean $p(n-1) \sim pn$ by Chernoff inequality for binomial random variables. So Erdös-Rényi random graphs are expected to behave like k-regular graphs asymptotically as $n \to \infty$ and the graph is dense enough. Although it is still an open question in theory to precisely characterize such asymptotic equivalence, some results can be found *e.g.* in [10]. However for small or sparse graphs, their different influences on HodgeRank are still unknown. In the next section we will introduce some tools from computational topology to compute the number of connected components and holes in clique complexes of finite random graphs. From simulations we will see again the similarity between two random graph models. However, due to the balanced nature of random regular graphs, some performances of HodgeRank will be improved in random regular graphs as will be shown in the experimental section.

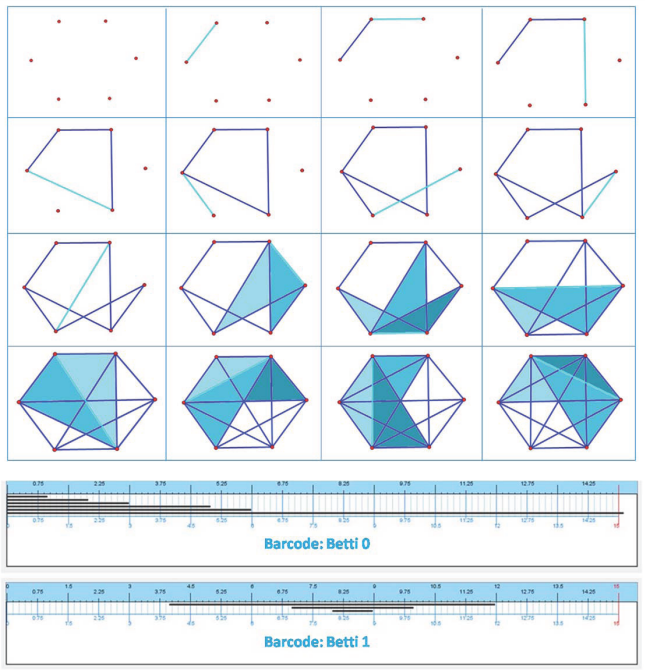


Fig. 4. Persistence barcodes of Betti numbers.

C. Persistence Homology Barcodes

Persistence homology is firstly introduced by [35] in computational topology, and later developed by [36] 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 [37], [38]. 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 that the videos (nodes) come in a certain order (*e.g.*, production time,

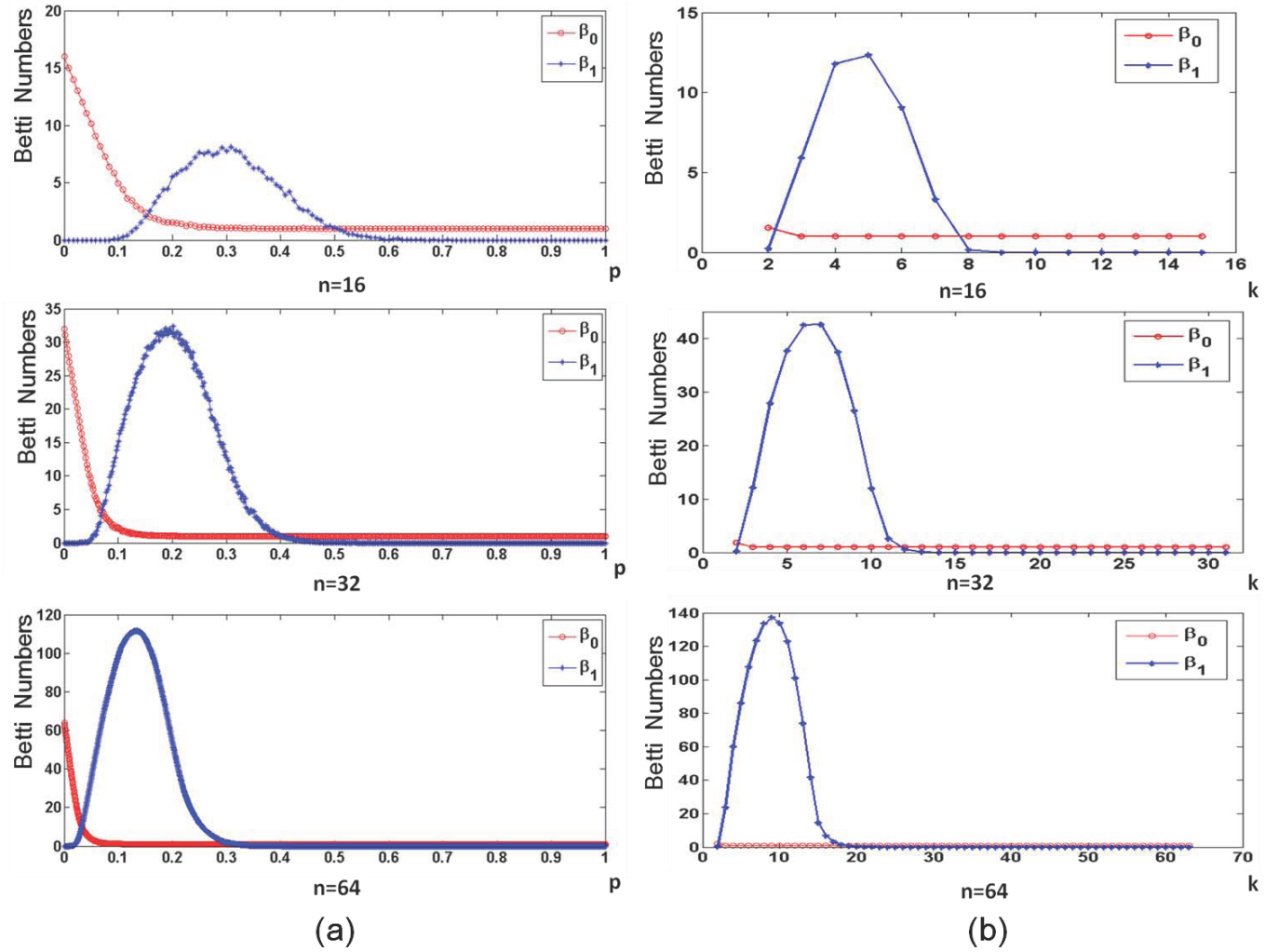


Fig. 5. Average Betti numbers of clique complexes for random graphs. (a) Erdös-Rényi random graphs; (b) random regular graphs.

or all created in the same time), after that pairs of videos (edges) are presented to us 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 4 illustrates a birth process of clique complex and its associated Betti numbers (β_0 and β_1) that are computed and plotted by JPlex [39]. At the first frame (say $t = 0$), 6 videos as 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. For more details on the algorithmic aspects of persistent homology, readers may refer to [35].

With the aid of persistent homology, one can compute the mean Betti numbers for random graphs. For example, for Erdös-Rényi random graphs with 16, 32, and 64 nodes ($n = 16, 32, 64$), the expected β_0 and β_1 (with 100 random graphs) are plotted in Figure 5 (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 5 (a). As [34] shows, this probability asymptotically drops at the rate $p \sim n^{-1/2}$.

Moreover, Figure 5 (b) shows the average β_0 and β_1 (with 100 random graphs) of random *k*-regular graphs with $n = 16, 32, 64$. The phase transitions are qualitatively similar to Erdös-Rényi random graphs with some subtle distinctions. It can be seen from this figure that when $k > 9$ with high probability the expected β_1 for $\zeta(16, k)$ equals to 0 which corresponds to $16 \times 9/2 = 72$ distinct pairs. However, as illustrated in Figure 5(a), for Erdös-Rényi random graph, it needs $p > 0.7$ corresponding to $120 \times 0.7 = 84$ distinct pairs. This reflects that balanced data is easier to satisfy the loop-free

condition than imbalanced data when adding the same number of distinct pairs.

IV. 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, we first show how to apply the inconsistency measures in Hodge decomposition to evaluate assessor reliability; then build up a base-line with complete paired comparison design with model selections. Finally, the results with incomplete data are demonstrated with two random sampling schemes.

A. Dataset

We adopt the publicly-accessible database for VQA, LIVE video database [8], which includes 10 different reference videos and 15 distorted versions of each reference, for a total of 160 videos. In the subjective test, 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 [8], we only borrow the video sources it provides. Different from LIVE [8], we propose to assess video quality with paired comparison.

B. Paired Comparison Data Collection

We now present our experiment design for collecting the set of paired comparison 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 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 [4], 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. After discarding some inconsistent/invalid data, 38400 paired comparisons (32 rounds complete data) are treated as our experimental data.

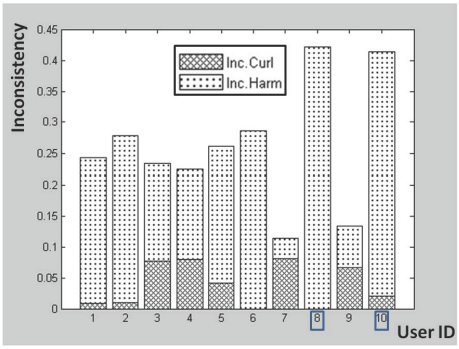


Fig. 6. Experimental results of 10 assessors' reliability.

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 III-A with $\alpha = 1, \ldots, 32$ for round (group) index. 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.

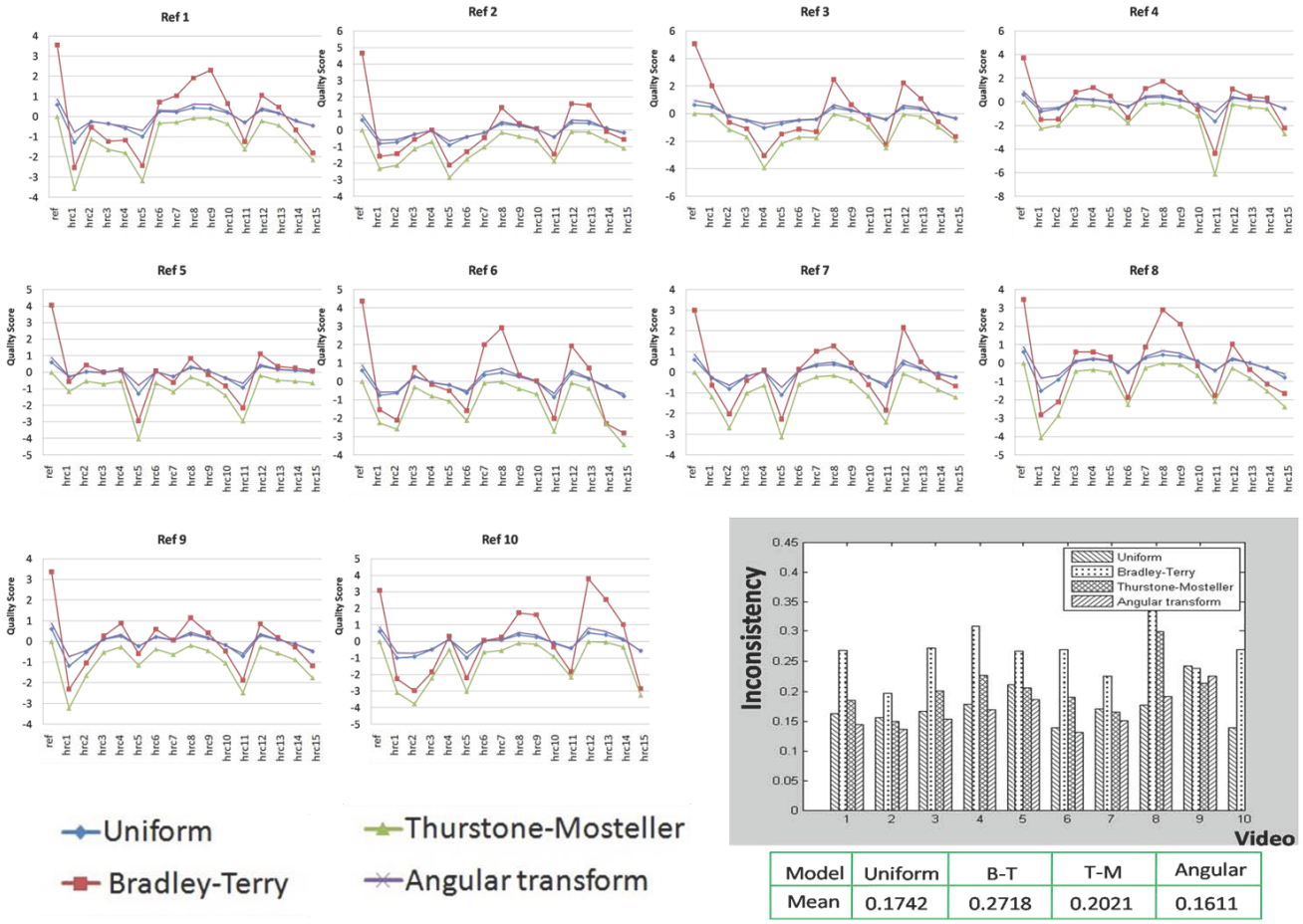
Finally, we note that the data collection above suffices to study approximations of global ranking from incomplete random samples in this paper. However, it can not be used to investigate the effects of single or multiple experimental design variables. For the latter purpose, one has to collect samples under various experimental controls which will be left for future studies.

C. Experimental Results

1) Assessment of Assessor's Reliability: We give a preliminary example of HodgeRank on evaluating assessors' reliability. The following shows how to pick out the volunteers that are giving inaccurate/dishonest results. After receiving the paired comparison results of each assessor, HodgeRank can be used to derive the total inconsistency (Inc.Total), curl inconsistency (Inc.Curl) and harmonic inconsistency (Inc.Harm). The *total inconsistency* is measured by

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

which equals to the sum of $\mathrm{Inc.Harm}(\hat{\mathrm{Y}}) = \|\hat{Y}^h\|_\omega^2 / \|\hat{Y}\|_\omega^2$ and $\mathrm{Inc.Curl}(\hat{\mathrm{Y}}) = \|\hat{Y}^c\|_\omega^2 / \|\hat{Y}\|_\omega^2$. We also define the *har-*



Model	Uniform	B-T	T-M	Angular
Mean	0.1742	0.2718	0.2021	0.1611

Fig. 7. Global ranking scores and inconsistency distribution of 32 complete rounds on Live video database [8]. The *Angular transform* model has the smallest mean inconsistency at 0.1611.

monic percentage as the ratio

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

Figure 6 shows the inconsistencies of 10 participants who provide their opinions for the 16 different quality videos in ref1, each one with incomplete data. It can be seen from this figure that assessors 8 and 10 have extraordinarily large total inconsistency. A closer inspection on the two assessor's record shows that the two assessors are careless or exposed to a larger noisy environment. In this way, we can identify and discard some inconsistent data provided by unreliable assessors. We note that the example above is preliminary for illustration, and a systematic treatment of this topic needs a distribution model of harmonic or curl flows under various experimental conditions, which is left for future pursuit.

2) Complete Design with Model Selection: 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. Therefore, results obtained from 32 rounds of complete comparisons are treated as the baseline in our experiment. HodgeRank with such a complete and balanced data will be reduced to the Borda Count following (21). With complete and balanced data, global/harmonic inconsistency vanishes according to Section III-A.

Global ranking scores $\hat{s}$ for each reference and inconsistency distribution are given in Figure 7, where "ref" represents 10 different reference videos in Live video database [8] 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 and the uniform model is the second best with a slightly worse inconsistency. Therefore, we will adopt the Angular transform model in the following experiments for incomplete data.

3) Results of Incomplete Data: In the following experiments, we shall focus on the performance of incomplete data under different sampling complexity in two random graph designs, i.e. Erdös-Rényi random graphs and random regular graphs. In particular we will sample the complete data at different rates measured by the number of distinct edges in the graph and study the following performance measurements: (1) Kendall's τ correlation [40] of HodgeRank on sampled data vs. the complete data; (2) Inconsistency measured by percentage of total inconsistency and harmonic inconsistency; (3) Statistical stability. The first two experiments study Erdös-Rényi random graphs and random regular graphs under different sampling ratios for each of 10 groups, respectively. The third experiment compares their overall performances on Live

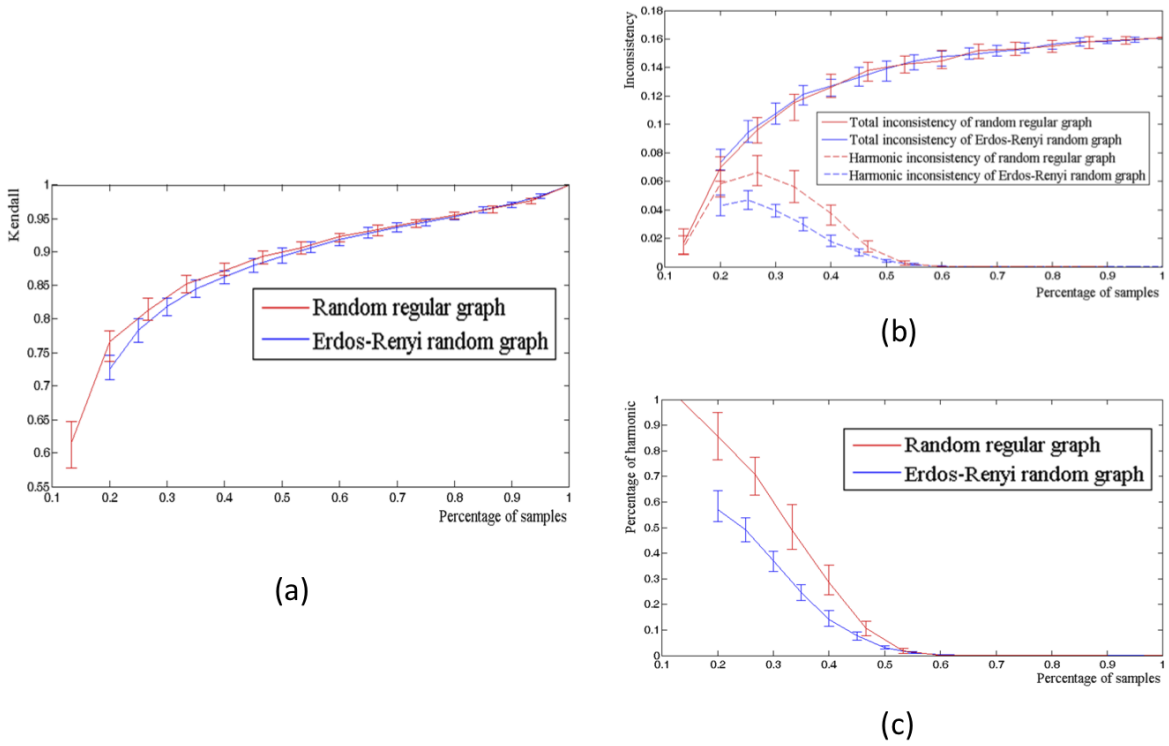


Fig. 8. Comparisons of Kendall's τ correlation to ground truth and Inconsistency decomposition, versus sampling rates, in Exp-I and Exp-II. (a) Kendall's τ correlation to the ground truth; (b) Total inconsistency and harmonic inconsistency; (c) Percentage of harmonic component over total inconsistency. Results are averaged over 10 reference videos with 100 bootstrapped samples.

data [8].

Exp-I: Erdös-Rényi random sampling

As illustrated in Section III-C Figure 5(a), for $n = 16$, if more than 25% random edges are added, the resultant graph is connected with high probability; and with more than 70% edges, the resultant 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. Here we use sampling complexity (number of distinct edges in the paired comparison graph) to control the harmonic ranking component. By the two phase transitions observed in the last section, the projection on harmonic ranking only exists when the number of distinct edges is in medium range. To illustrate this point, we design this experiment. In this experiment, the sampling schemes are the same for each r of the reference video. And we randomly draw $p\%$ pairs from each round $\alpha \in \{1, \ldots, 32\}$ of complete comparisons. Note that we extract the same pairs from each of the 32 groups. 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. Blue curves in Figure 8 shows the Kendall's τ, total inconsistency, harmonic inconsistency, and percentage of harmonic over total inconsistency, against the edge sampling rates ranging from 20% to 100%. 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, to avoid the possible issue of harmonic ranking, we can 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 a 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.

Exp-II: random regular sampling

As illustrated in Section III-C Figure 5(b), for $n = 16$, when $k > 9$, the resultant clique complex is loop-free with high probability and thus we can derive a reliable global ranking. To illustrate this point, we first establish a random k-regular graph. Then, for each of the 10 reference videos, and for each of the 32 groups of complete paired comparisons, the graph generated in the previous step is used to sample pairs. Just the same as Exp-I, each group contains the same pairs. Red curves in Figure 8 shows the Kendall's τ, total inconsistency, harmonic inconsistency, and percentage of harmonic over total inconsistency, against the sampling rates corresponding to the

TABLE I
KENDALL'S τ AND INCONSISTENCY OF ERDÖS-RÉNYI RANDOM SAMPLING.

	min	mean	max	std
Kendall's τ	0.9350	0.9536	0.9750	0.0085
Harmonic Inconsistency	0	0	0	0
Total Inconsistency	0.1770	0.1895	0.2009	0.0048

TABLE II
KENDALL'S τ AND INCONSISTENCY OF RANDOM k-REGULAR SAMPLING.

	min	mean	max	std
Kendall's τ	0.9333	0.9539	0.9667	0.0078
Harmonic Inconsistency	0	0	0	0
Total Inconsistency	0.1779	0.1891	0.1994	0.0047

number of k ranging from 2 to 15.

Just the same as Erdös-Rényi random graph, in random k-regular graphs, for general n, we can also choose an upper bound for k, *i.e.*, $k = 9$ for $n = 16$ node graphs. In this case, with a high probability the resultant paired comparison graph is connected and its associated clique complex is loop-free. Therefore, the inconsistency is attributed to merely local inconsistency where the global ranking will not suffer the global inconsistency issue. Note that such a threshold varies with the total number of videos.

The comparisons of two random sampling schemes in Figure 8 also disclose the following: when sampling rates are small (say $< 40\%$), random k-regular graphs will lead to better performances in terms of higher Kendall's correlation with the ground truth and lower total inconsistency. This shows the benefit of balanced sampling in regular graphs. However, due to structural properties of k-regular graphs with small k, harmonic components will contribute more for random regular graphs than Erdös-Rényi random graphs in this range. We note that the distinction between two sampling schemes rapidly decreases as sampling rates increase. In order to make a comparative study of the performance of these two random sampling schemes in a global manner, an additional experiment is conducted in the following.

Exp-III: Erdös-Rényi random sampling vs. random regular sampling

This experiment shows the average performance of these two sampling schemes under arbitrary number of samples. Specifically, for each reference, we randomly sample 32 k-regular graphs from 32 rounds independently. Meanwhile, for Erdös-Rényi random sampling, we randomly sample the same number of pairs from each reference in 32 rounds. To ensure the statistical stability, we run the random sampling process 100 times. Tables I and II show the mean Kendall's τ and inconsistency results of 100 times achieved by these two schemes. Results are averaged over 10 reference videos. From these experimental results, we make the following comments.

First, it is shown that both of these two sampling approaches could provide good approximate results of the complete data, with an average Kendall's τ of 0.9539 ± 0.0078 (random k-regular) and 0.9536 ± 0.0085 (Erdös-Rényi), respectively. Although the gaps between their performances are small, there is a slightly better overall performance in random regular graphs than Erdös-Rényi random sampling with higher Kendall's τ-coefficients, lower total inconsistency (0.1891 ± 0.0047 vs. 0.1895 ± 0.0048), and better stability (std of Kendall's τ and total inconsistency in random regular sampling are both smaller than in Erdös-Rényi random sampling). We note that ignoring the stochastic fluctuations, one may qualitatively conclude that the two random graphs bear similar overall performances in such an experiment.

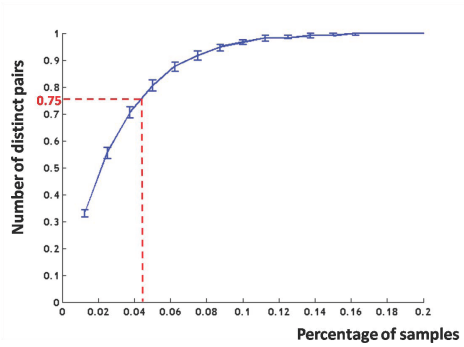


Fig. 9. Percentage of samples versus number of distinct pairs. For each percentage level, the experiments are repeated 100 times and the median number of distinct pairs with $[0.25, 0.75]$ confidence interval are plotted in the figure.

Second, we observe that the harmonic inconsistency in both of these two sampling is always 0. As mentioned above, as long as the resultant clique complex is loop-free, the harmonic inconsistency will vanish. Due to the multiple comparisons between a pair of videos, a natural question is raised that how many percentage of samples are needed to satisfy the loop-free condition? Similar to Erdös-Rényi random graph theory on simple graphs, Figure 9 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 and thus can induce a loop-free complex. That is to say, it is easy to meet this requirement and thus can avoid the possible issue of harmonic inconsistency.

Third, in such an overall performance measure, Erdös-Rényi random graphs are I.I.D. sampling on paired comparison experiments, which provides reasonably good approximations of random k-regular, a kind of dependent sampling scheme, despite their distinctions when sampling rates are low in Exp-II. Therefore, both of these have their advantages and disadvantages and are both promising sampling methods. So which one to choose depends on specific application requirements. For example, when raters need more flexibility to finish this task, maybe Erdös-Rényi random sampling is a good choice. After all, it does not jeopardize much the accuracy of the results. However, when organizers require that each video should be compared in a fair and equitable manner, *i.e.*, every video has the same number of paired comparisons with other videos, random regular graphs can take up this mission.

V. 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 Hodge decomposition of paired comparison data on graphs. In particular, we study two random sampling schemes inspired by Erdös-Rényi and random regular graph theory, followed by HodgeRank to analyze the incomplete and imbalanced data collected in these experiments. 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 where the raters are distributive over Internet and it is hard to control with traditional experimental designs.

There are both distinctions and similarities in the two random graph models. Erdös-Rényi random graphs are the simplest random sampling design bearing the I.I.D. property, while the random regular graphs belong to a sort of dependent sampling with balanced paired comparisons. The balanced nature of random k-regular graphs makes it with better performance when sampling rates are low in our experiments. However, such a distinction rapidly vanishes when sampling rates are high or measured in an overall way. In theory, for large n, Erdös-Rényi random graphs may provide good asymptotic approximations for random regular graphs. Therefore, one may choose suitable models depending on the specific circumstances.

With the rapid advent of technologies on rich user interface, in future, we plan 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). Besides, a dataset with more reference videos, distorted videos, and statistically significant subjective scores will be of great value to the VQA research community, which will also be part of our future work.

REFERENCES

[1] 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.

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

[3] 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.

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

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

[6] A. Eichhorn, P. Ni, and R. Eg, "Randomised pair comparison: an economic and robust method for audiovisual quality assessment." NOSSDAV, 2010, pp. 63–68.

[7] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye., "Statistical ranking and combinatorial Hodge theory," *Mathematical Programming*, vol. 127, no. 1, pp. 6470–6481, 2011.

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

[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] J. H. Kim and V. Vu, "Sandwiching random graphs," *Advances in Mathematics*, no. 188, pp. 444–469, 2004.

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

[12] M. Kendall and B. Smith, "On the method of paired comparisons," *Biometrika*, vol. 31, no. 3-4, pp. 324–345, 1940.

[13] T. Saaty, "A scaling method for priorities in hierarchical structures," *Journal of Mathematical Psychology*, vol. 15, no. 3, pp. 234–281, 1977.

[14] K. Arrow, "A difficulty in the concept of social welfare," *Journal of Political Economy*, vol. 58, no. 4, pp. 328–346, 1950.

[15] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," *Annals of Statistics*, vol. 26, no. 2, pp. 451–471, 1998.

[16] Y. Freund, R. Iyer, R. Shapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of Machine Learning Research*, vol. 4, no. 6, pp. 933–969, 2004.

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

[18] J. Howe, "The rise of crowdsourcing," *Wired Magazine*, vol. 14, no. 6, pp. 176–183, 2006.

[19] 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.

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

[21] 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.

[22] O. Alonso, D. Rose, and B. Stewart, "Crowdsourcing for relevance evaluation," *SIGIR Forum*, vol. 42, no. 2, pp. 9–15, 2008.

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

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

[25] P. Erdos and A. Renyi, "On random graphs i," *Publ. Math. Debrecen*, vol. 6, pp. 290–297, 1959.

[26] N. Wormald, "Models of random regular graphs." In Surveys in Combinatorics, 1999, pp. 239–298.

[27] J. H. Kim and V. H. Vu, "Generating random regular graphs." The Thirty-fifth Annual ACM Symposium on Theory of Computing (STOC), 2003, pp. 213–222.

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

[29] D. Watts and S. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, no. 393, pp. 440–442, 1998.

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

[31] G. Noether, "Remarks about a paired comparison model," *Psychometrika*, vol. 25, pp. 357–367, 1960.

[32] D. Spielman and S.-H. Teng, "Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems." The Thirty-sixth Annual ACM Symposium on Theory of Computing (STOC), 2004.

[33] I. Koutis, G. Miller, and R. Peng, "Approaching optimality for solving sdd systems." The Fifty-first Annual IEEE Symposium on Foundations of Computer Science (FOCS), 2010.

[34] M. Kahle, "Topology of random clique complexes," *Discrete Mathematics*, vol. 309, pp. 1658–1671, 2009.

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

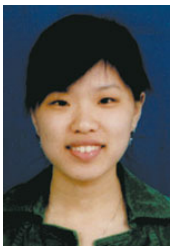
[36] A. Zomorodian and G. Carlsson, "Computing persistent homology," *Discrete and Computational Geometry*, vol. 33, no. 2, pp. 249–274, 2005.

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

[38] H. Edelsbrunner and J. Harer, "Computational topology : an introduction," 2010.

[39] 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.

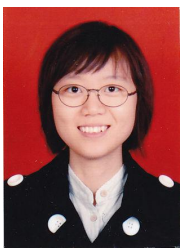
[40] Kendall, Maurice, and J. Gibbons, *Rank Correlation Methods*. Oxford University Press, 1990.



Qianqian Xu is currently pursuing the Ph.D. degree in Graduate University of Chinese Academy of Sciences. Her research interests include video processing and video quality assessment.

Qingming Huang (SM'08) received the B.S. degree in computer science and Ph.D. degree in computer engineering from Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively. He is currently a Professor with the Graduate University of the Chinese Academy of Sciences (CAS), Beijing, China, and an Adjunct Research Professor with the Institute of Computing Technology, CAS. He has authored or coauthored nearly 200 academic papers in prestigious international journals and conferences. His research areas include multimedia video analysis, video adaptation, image processing, computer vision, and pattern recognition. Dr. Huang is a reviewer for IEEE Trans. on Multimedia, IEEE Trans. on Circuits and Systems for Video Technology, and IEEE Trans. on Communications. He has served as program chair, track chair and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, PSIVT, etc.

Tingting Jiang received the BS degree in computer science from University of Science and Technology of China in Hefei, China, in 2001 and the PhD degree in computer science from Duke University, Durham, North Carolina, USA, in 2007. She is now an assistant professor of computer science at Peking University, Beijing, China. Her research interests include computer vision, image and video quality assessment.



Bowei Yan is currently working toward the M.S. degree majoring in statistics at the School of Mathematical Science, Peking University, Beijing, China. Her research interests include statistical learning, data science and topological and geometric methods for high-dimension data analysis.

Weisi Lin (M'92-SM'98) received the B.Sc. degree in electronics and the M.Sc. degree in digital signal processing from Zhongshan University, Guangzhou, China, and the Ph.D. degree in computer vision from King's College, London University, London, U.K. He taught and conducted research at Zhongshan University, Shantou University (China), Bath University (U.K.), the National University of Singapore, the Institute of Microelectronics (Singapore), and the Institute for Infocomm Research (Singapore). He has been the Project Leader of over ten major successfully-delivered projects in digital multimedia technology development. He also served as the Lab Head, Visual Processing, and the Acting Department Manager, Media Processing, for the Institute for Infocomm Research. Currently, he is an Associate Professor in the School of Computer Engineering, Nanyang Technological University, Singapore. His areas of expertise include image processing, perceptual modeling, video compression, multimedia communication and computer vision. He has published over 190 refereed papers in international journals and conferences.

Dr. Lin is a Chartered Engineer (U.K.), a fellow of Institution of Engineering Technology, and an Honorary Fellow, Singapore Institute of Engineering Technologists. He organized special sessions in IEEE International Conference on Multimedia and Expo (ICME 2006, 2012), IEEE International Workshop on Multimedia Analysis and Processing (2007), IEEE International Symposium on Circuits and Systems (ISCAS 2010), Pacific-Rim Conference on Multimedia (PCM 2009), SPIE Visual Communications and Image Processing (VCIP 2010), Asia Pacific Signal and Information Processing Association (APSIPA 2011), and MobiMedia 2011. He gave invited/keynote/panelist talks in International Workshop on Video Processing and Quality Metrics (2006), IEEE International Conference on Computer Communications and Networks (2007), SPIE VCIP 2010, and IEEE Multimedia Communication Technical Committee (MMTC) Interest Group of Quality of Experience for Multimedia Communications (2011), and tutorials in PCM 2007, PCM 2009, IEEE ISCAS 2008, IEEE ICME 2009, APSIPA 2010, and IEEE International Conference on Image Processing (2010). He is currently on the editorial boards of IEEE Trans. on Multimedia, IEEE SIGNAL PROCESSING LETTERS and Journal of Visual Communication and Image Representation, and four IEEE Technical Committees. He cochairs the IEEE MMTC Special Interest Group on Quality of Experience. He has been on Technical Program Committees and/or Organizing Committees of a number of international conferences.

Yuan Yao received the B.S.E and M.S.E in control engineering both from Harbin Institute of Technology, China, in 1996 and 1998, respectively, M.Phil in mathematics from City University of Hong Kong in 2002, and Ph.D. in mathematics from the University of California, Berkeley, in 2006. Since then he has been with Stanford University and in 2009, he joined the School of Mathematical Sciences, Peking University, Beijing, China, as a professor of statistics in the Hundred Talents Program. His current research interests include topological and geometric methods for high dimensional data analysis and statistical machine learning, with applications in computational biology, computer vision, and information retrieval. Dr. Yao is a member of American Mathematical Society (AMS), Association for Computing Machinery (ACM), Institute of Mathematical Statistics (IMS), and Society for Industrial and Applied Mathematics (SIAM). He served as area or session chair in NIPS and ICIAM, as well as a reviewer of Foundation of Computational Mathematics, IEEE Trans. Infomation Theory, J. Machine Learning Research, and Neural Computation, etc.