REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggesstions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any oenalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY) 24-04-2020	2. REPORT TYPE Final Report	3. DATES COVERED (From - To) 25-Apr-2019 - 24-Jan-2020

4. TITLE AND SUBTITLE
Final Report: Assessing and Expanding the Limits of Collective Intelligence

5a. CONTRACT NUMBER
W911NF-19-1-0260

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Arizona State University
ORSPA
P.O. Box 876011
Tempe, AZ 85287 -6011

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
74113-CS-II.6

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Adolfo Escobedo
UU	UU	UU	UU		19b. TELEPHONE NUMBER 480-965-5248

RPPR Final Report
as of 06-May-2020

Agency Code:

Proposal Number: 74113CSII **Agreement Number: W911NF-19-1-0260**

INVESTIGATOR(S):

Name: PHD Adolfo Raphael Escobedo
Email: adRes@asu.edu
Phone Number: 4809655248
Principal: Y

Organization: **Arizona State University**
Address: ORSPA, Tempe, AZ 852876011
Country: USA

DUNS Number: 943360412 EIN: 860196696

Report Date: 24-Apr-2020 Date Received: 24-Apr-2020

Final Report for Period Beginning 25-Apr-2019 and Ending 24-Jan-2020

Title: Assessing and Expanding the Limits of Collective Intelligence

Begin Performance Period: 25-Apr-2019 **End Performance Period:** 24-Jan-2020

Report Term: 0-Other

Submitted By: PHD Adolfo Escobedo Email: adRes@asu.edu
Phone: (480) 965-5248

Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: **STEM Participants:** 4

Major Goals: The major goals for the project, expanded from the Statement of Work, are as follows:
1) Construct specialized optimization tools that are capable of aggregating and gleaning insights from continuous, ordinal, and multimodal judgement-elicitation data using different choices of distance functions. These models and algorithms will be equipped to solve problems with over 100 evaluation entities and agents efficiently (i.e., in the order of minutes).
2) Develop customized statistical models for generating nontrivial problem instances with objectively defined degrees of collective similarity/dissimilarity from an underlying ground-truth set. The goal of the related tasks are to incorporate a variety of characteristics relevant to complex estimation tasks: incompleteness, sparsity, noise/error, imperfect subtask overlap, and high dimensionality.
3) Test the efficacy of the proposed aggregation methodologies, specifically to derive empirical parametric descriptions of the relationships of varying elicitation choices and statistical distribution configurations with the similarity between the aggregate estimate and the underlying ground truth; tests may also be performed on existing benchmarks for other applications; for example for wireless sensor network estimation in contested environments.
4) Validate the proposed tools in practice through the implementation of crowd based estimation tasks. These will be guided by standard crowd wisdom benchmarks of varying difficulties such as ranking the difficulties of puzzles, counting randomly distributed dots in images, and/or box-office prediction.
5) Leverage multidisciplinary collaborations to demonstrate the practical value of the developed tools.
6) Prepare a final technical report summarizing the outcomes of the project. Prepare and submit manuscripts to peer-reviewed conferences and journals.

Accomplishments: 1) Developed a multimodal aggregation framework that represents a principled unification of data from different contrasting sources such as cardinal (ratings, assigns a scalar value to each of the objects) and ordinal (rankings, objects are ordered from most preferred to least preferred in the form of an ordered list) evaluations. The incorporation of these multimodal evaluations achieved better results than using single modality data separately. Additionally, proposed a convex relaxation of the Cardinal and Ordinal aggregation model capable of solving instances with large number of evaluation entities efficiently (i.e., small amount of time). Generalized distances used in the Cardinal and Ordinal Aggregation models to deal with evaluations that are highly incomplete, high-dimensionality, and contain ties.
2) Performed a literature review of statistical models for generating synthetic data sets appropriate for both cardinal and ordinal aggregation models. Generated data from adaptations to the Mallows models on rankings data, in particular to generate rankings with ties and with
3) Formulated exact mathematical models for Ordinal aggregation based on three different metric functions,

namely, Spearman’s Footrule, Chevyshev’s distance, and Hamming distance and compared their performance (i. e., ability to recover ground truth) against Kemeny Snell Distance for ordinal aggregation. Tested different aggregation functions: L1 norm, L2 norm and normalized projected Cook Kress distance for Cardinal Aggregation in the Joint Aggregation model to compare their respective abilities to tackle challenges associated with the extraction of collective intelligence from complete evaluations. A report document outlining sets of tests for comparing the efficacy of the featured models on synthetic complete ranking and rating data is herein attached. Initial testing on synthetic complete ranking and rating data has been performed.
4) Created a flexible web application using JavaScript, HTML and CSS that can be utilized to crowdsource tasks involving ranking and numerical estimation. Implemented REST API functionality using Node.js. Developed database schema using mongoDB and wrote scripts to manage and pull data using python. Utilized this web application to develop and deploy a crowdsourced activity involving ranking and numerically estimating the number of dots in images. Adjusted the activity to explore the effects of varying problem difficulty, varying problem sizes, and the distribution of incomplete problems. The dots activity was deployed on Amazon MTurk, a crowdsourcing website, to collect data from a total of 521 people across three different studies. The first study focused on the effects of different problem difficulty. The second study focused on the effects of distributing incomplete problems. The third activity explored the effects of distributing incomplete problems, along with the effects of varying problem sizes. The most notable results were found in the third study, which employed 300 participants.
5) Established a collaboration with Dr. Lauren Huie at Air Force Research Labs, who heads a program on Wireless Sensor Networks in Contested Environments. Using models developed through this project, my PhD student Kyle Skolfield (at an internship in AFRL in NY), and I and my PhD student Romena Yasmin at ASU worked over the summer on related WSN problems. A related conference paper was recently accepted for the upcoming IEEE 6th World Forum on Internet of Things (WF-IoT 202), to be held virtually in June.
Established a collaboration with Dr. Ross Maciejewski, associated professor in the Computer Science Engineering program at ASU. This collaboration was instrumental for deploying the proposed methodology to crowd based estimation tasks.
6) Support from this award has resulted in an accepted journal publication to appear in the European Journal of Operational Research and in a submitted journal publication under review in Information Sciences. It also led to one conference paper for the upcoming IEEE 6th World Forum on Internet of Things (WF-IoT 202), to be held virtually in June 2020. Additionally, it resulted in an undergraduate honor’s thesis, which was defended on 04/24/2020. This thesis will be expanded into a conference paper submission for the Thirty-Fifth AAAI Conference on Artificial Intelligence to be held in February 2021. Lastly, outputs from this project are part of a working PhD dissertation and other working manuscripts. The uploaded final report addendum contains this unpublished material.

Training Opportunities: A PhD student (Romena Yasmin) was recruited and hired to perform research during the reporting period. The student was trained in the following:

- Principled methods for aggregating ordinal and/or cardinal data.
- Statistical models for generating artificial data sets parameterized by ground truth and dispersion (i.e., noise) parameters
- Experimental designs for assessing the capability of aggregation methods to recover the ground truth from artificially generated data sets
- Working with the Linux operating system and writing bash scripts to run computational jobs related to these experiments on ASU's HPC cluster.
- Applications of these concepts in Wireless Sensor Networks.

A second PhD student (Kyle Skolfield) was trained in principled aggregation to apply the tools developed through this project to Wireless Sensor Networks state estimation problems. The student was involved in the collaboration with Dr. Lauren Huie at Air Force Research Labs during summer of 2019, which was made possible through this award.

A third PhD student (Yeawon Yoo) was hired to perform research during the second half of the reporting period. The student was trained in the following:

- Principled methods for aggregating ordinal and/or cardinal data.
- Advanced methodologies for solving the consensus ranking problem with ties.
- Converting data from the standard PrefLib format.
- Applications of these concepts for crowd-based human computation.

An undergraduate student (Ryan Kemmer) was recruited and hired to perform research for the second half of the project. The student was trained in the following:

- Design crowdsourced estimation tasks from which to gather data that can be used to test the multimodal aggregation models.
- RESTful API development using Node.js and express.
- Web design and development using javascript, HTML and CSS, with an emphasis on using the javascript d3 library for user facing activites.
- Database development using MongoDB NoSQL databases.
- Development and testing of object-oriented python scripts for data analysis, data pulling, and data cleaning utilizing pymongo, pandas, numpy, and json libraries.
- Statistical methods for evaluating the performance of aggregation methods.
- Techniques for obtaining quality data through crowdsourcing.

Results Dissemination: Support from this award has resulted in an accepted journal publication to appear in the European Journal of Operational Research and in a submitted journal publication under review in Information Sciences. It also led to one conference paper for the upcoming IEEE 6th World Forum on Internet of Things (WF-IoT 202), to be held virtually in June 2020 (we will give a presentation of the paper therein). Additionally, it resulted in an undergraduate honor's thesis, which was defended on 04/24/2020. This thesis will be expanded into a conference paper submission for the Thirty-Fifth AAAI Conference on Artificial Intelligence to be held in February 2021. Results from this work will also be presented at the INFORMS Annual Meeting in late 2020.

Honors and Awards: The application to crowd-based human computation became the subject of an undergraduate honor's thesis, which was defended on 04/24/2020.

Protocol Activity Status:

Technology Transfer: In pursuit of a novel application of the methodologies developed in this project, we established a collaboration with Dr. Lauren Huie at Air Force Research Labs, who heads a program on Wireless Sensor Networks in Contested Environments. This results in a conference paper to be published and presented as part of the upcoming IEEE 6th World Forum on Internet of Things (WF-IoT 202), to be held virtually in June 2020.

PARTICIPANTS:

Participant Type: Graduate Student (research assistant)
Participant: Romena Yasmin

Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Kyle Skolfield
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Undergraduate Student
Participant: Ryan Kemmer
Person Months Worked: 5.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Yeawon Yoo
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

ARTICLES:

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 4-Under Review
Journal: Information Sciences
Publication Identifier Type: DOI Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 4/23/20 12:00AM Date Published:
Publication Location:
Article Title: A new binary programming formulation and social choice propertyfor expediting the solution process to Kemeny rank aggregation
Authors: Yeawon Yoo, Adolfo R. Escobedo
Keywords: Group decision-making, Rank aggregation, Computational social choice
Abstract: We introduce a binary programming formulation for generalized Kemeny rank aggregation—whose ranking inputs may be complete and incomplete, with and without ties. The new formulation provides comparative advantages over a related formulation including reduced memory requirements and faster computing times when solving large problems. Moreover, we develop a new social choice property, the Non-strict Extended Condorcet Criterion (NECC), which can be regarded as a natural extension of the Condorcet criterion and the Extended Condorcet criterion; unlike its parent properties, the NECC is adequate for handling complete rankings with ties. The property is leveraged to develop a structural decomposition algorithm through which certain instances with hundreds of alternatives can be solved exactly in seconds.To test the practical implications of the new formulation and social choice property, we work with instances drawn from a probabilistic distribution and benchmark instances from PrefLib.
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: European Journal of Operational Research
Publication Identifier Type: DOI Publication Identifier: 10.1016/j.ejor.2020.02.027
Volume: Issue: First Page #:
Date Submitted: 4/23/20 12:00AM Date Published: 2/1/20 2:00PM
Publication Location:
Article Title: A new correlation coefficient for comparing and aggregating non-strict and incomplete rankings
Authors: Yeawon Yoo, Adolfo R. Escobedo, J. Kyle Skolfield
Keywords: Group decisions and negotiations; Robust ranking aggregation; Correlation and distance functions; Non-strict incomplete rankings
Abstract: We introduce a correlation coefficient for non-strict (tied) and incomplete (unknown) rankings. The new measure enforces a neutral treatment of incompleteness and is shown to be connected with a recently developed distance function for Kemeny aggregation. The work proves the equivalence of an additional distance and correlation pairing in the space of non-strict incomplete rankings, which reinforces the suitability of the featured measure to solve the general consensus ranking problem. These connections induce new exact optimization methodologies: a specialized branch and bound algorithm and an exact integer programming formulation. Associated experiments with the branch and bound algorithm demonstrate that, as data becomes noisier, the featured correlation coefficient yields relatively fewer alternative optimal solutions and that the aggregate rankings tend to be closer to an underlying ground truth shared by a majority. Specialized instance sampling techniques are also introduced.
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 2-Awaiting Publicat
Conference Name: IEEE 6th World Forum on Internet of Things (WF-IoT 2020)
Date Received: 23-Apr-2020 Conference Date: 02-Jun-2020 Date Published: 01-Jul-2020
Conference Location: Virtual (Online)
Paper Title: A Comparison of Axiomatic Distance-Based Collective Intelligence Methods for Wireless Sensor Network State Estimation in the Presence of Information Injection
Authors: J. Kyle Skolfield, Romena Yasmin, Adolfo R. Escobedo, Lauren M. Huie
Acknowledged Federal Support: **Y**

DISSERTATIONS:

Publication Type: Thesis or Dissertation
Institution: Arizona State University
Date Received: 24-Apr-2020 Completion Date: 4/24/20 2:00PM
Title: Input-Elicitation Methods for Crowdsourced Human Computation
Authors: Ryan Kemmer
Acknowledged Federal Support: **Y**

Final Report Addendum for ARO Award 74113NSII

Overall, this award led to many productive outcomes by the research team, who was able to achieve the large majority of the stated goals. The PI feels encouraged about the possibilities opened up by this research award. Results from this project will be used to apply to future funding opportunities including the Newton Award for Transformative Ideas during the COVID-19 Pandemic and possibly the ARMY Young Investigator Award and the AFOSR Young Investigator Program. These opportunities would have not been possible without this ARMY STIR award, for which the PI is extremely grateful.

Support from this award has resulted in an accepted journal publication to appear in the *European Journal of Operational Research* and to a submitted journal publication under review in *Information Sciences*. It also led to one conference paper for the upcoming *IEEE 6th World Forum on Internet of Things (WF-IoT 202)*, to be held virtually in June 2020. Additionally, it resulted in an undergraduate honor's thesis, which was defended on 04/24/2020. This thesis will be expanded into a conference paper submission for the *Thirty-Fifth AAAI Conference on Artificial Intelligence* to be held in February 2021. All of these manuscripts have been uploaded to the ARO portal (see Products).

Lastly, outputs from this project are part of a working PhD dissertation and other working manuscripts. This addendum contains this unpublished material. The first section deals with the methodology developed during the project. The second section deals with the implementation of this theory to better harness crowd wisdom in a human computation study associated with this project; the study obtained IRB approval from ASU (STUDY00010770: Estimating the number of dots in an image).

1 Aggregation Models

The first part of this addendum consists of the outcomes from four computational experiments: The first experiment evaluates the minimum number of judges required to achieve the ground truth for the three models described in Escobedo et al. (2020). The second and third set of experiments evaluates different ranking and rating metrics to determine which one provides a better consensus. The last experiment tests how well each aggregation framework enhance the collective wisdom through a human computation experiment.

Experiment Set 1:

This experiment set seeks to determine at what point the group consensus is closest to the ground truth within a specified range. In this set three models are compared: Convex relaxation of Cardinal and Ordinal Aggregation(COAr) model, Exact MILP formulation of COA and the Correlation based version of Ordinal Aggregation(OA) model. Each of these models are described in detail in Escobedo, Hochbaum and Moreno-Centeno. The Normalized projected Cook-Kress distance(NPCK) was used as the distance function for the complete ratings and the normalized projected Kemeny-Snell distance (NPKS) was used for the rating vectors. Based on these distance measures the formulation of each of the models is given below:

Model 1.1: COAr: Convex relaxation of Cardinal and Ordinal Aggregation:

$$
\begin{array}{lll}
\text{minimize} & \displaystyle\sum_{k \in K} 2C^k \sum_{(i,j) \in A^k} t_{ij}^k + \sum_{k \in K} \frac{1}{2} D^k \sum_{(i,j) \in B^k} h_{ij}^k & \\
\text{subject to,} & t_{ij}^k \geq \mu(x_i - x_j) - p_{ij}^k & (i,j) \in A^k, k = 1, 2....m \\
& t_{ij}^k \geq -[\mu(x_i - x_j) - p_{ij}^k] & (i,j) \in A^k, k = 1, 2....m \\
& h_{ij}^k \geq x_i - x_j + 1 & (i,j) \in B^k, k = 1, 2....m; \text{s.t. } \operatorname{sign}(b_j^k - b_i^k) = -1 \\
& h_{ij}^k \geq x_i - x_j & (i,j) \in B^k, k = 1, 2....m; \text{s.t. } \operatorname{sign}(b_j^k - b_i^k) = 0 \\
& h_{ij}^k \geq -x_i + x_j & (i,j) \in B^k, k = 1, 2....m; \text{s.t. } \operatorname{sign}(b_j^k - b_i^k) = 0 \\
& h_{ij}^k \geq -x_i + x_j + 1 & (i,j) \in B^k, k = 1, 2....m; \text{s.t. } \operatorname{sign}(b_j^k - b_i^k) = 1 \\
& h_{ij}^k \geq 0 & (i,j) \in B^k, k = 1, 2....m \\
& h_{ij}^k \text{ unrestricted} & \\
& 0 \leq x_i \leq (U - L)/\mu & i = 1, 2.......n
\end{array}
$$

where,

$$
\begin{aligned}
n &= \text{number of objects} \\
m &= \text{number of judges} \\
A^k &= \text{set of objects rated by judge k} \\
B^k &= \text{set of objects ranked by judge k} \\
x_i &= \text{aggregate rating of object i} \\
p_{ij}^k &= a_i^k - a_j^k \\
a_i^k &= \text{the rating value given by judge } k \text{ to object } i \\
b_i^k &= \text{the ranking value given by judge } k \text{ to object } i
\end{aligned}
$$

Model 1.2: COA: Exact MILP formulation of of Cardinal and Ordinal Aggregation:

$$
\begin{array}{lll}
\text{maximize} & \displaystyle -\sum_{k \in K} 4C^k \sum_{(i,j) \in A^k} t_{ij}^k + \sum_{i=1}^{n} \sum_{j=1}^{n} 2\hat{B}_{ij} y_{ij} & \\
\text{subject to,} & t_{ij}^k \geq \mu(x_i - x_j) - p_{ij}^k & (i,j) \in A^k, k = 1, 2....m \\
& t_{ij}^k \geq -[\mu(x_i - x_j) - p_{ij}^k] & (i,j) \in A^k, k = 1, 2....m \\
& x_i - x_j \leq (M + \mu) y_{ij} - \mu & i, j = 1, 2...n \\
& x_i - x_j \geq M(y_{ij} - 1) & i, j = 1, 2...n \\
& y_{ij} + y_{ji} \geq 1 & i, j = 1, 2...n \\
& y_{ij} - y_{kj} - y_{ik} \geq -1 & i, j, k = 1, 2...n \\
& 0 \leq x_i \leq (U - L)/\mu & i = 1, 2.......n \\
& y_{ij} \in \{0, 1\} & i, j = 1, 2...n
\end{array}
$$

where,

$$y_{ij} = \begin{cases} 1, & \text{if object i is assigned an equal and higher rating than object j} \\ 0, & \text{if object i is assigned a lower rating than object j} \end{cases}$$

$$b_{ij} = \begin{cases} 1, & \text{if } b_i \leq b_j \\ -1, & \quad \text{if } b_i > b_j \\ 0, & \text{if } i = j, \text{ or, } b_i \text{ or, } b_j \text{ not defined} \end{cases}$$

$$\hat{B}_{ij} = \sum_{k=1}^{m} \frac{b_{ij}^k}{|V_b^k|(|V_b^k| - 1)}$$

Model 1.3: OA: Correlation based version of Ordinal Aggregation:

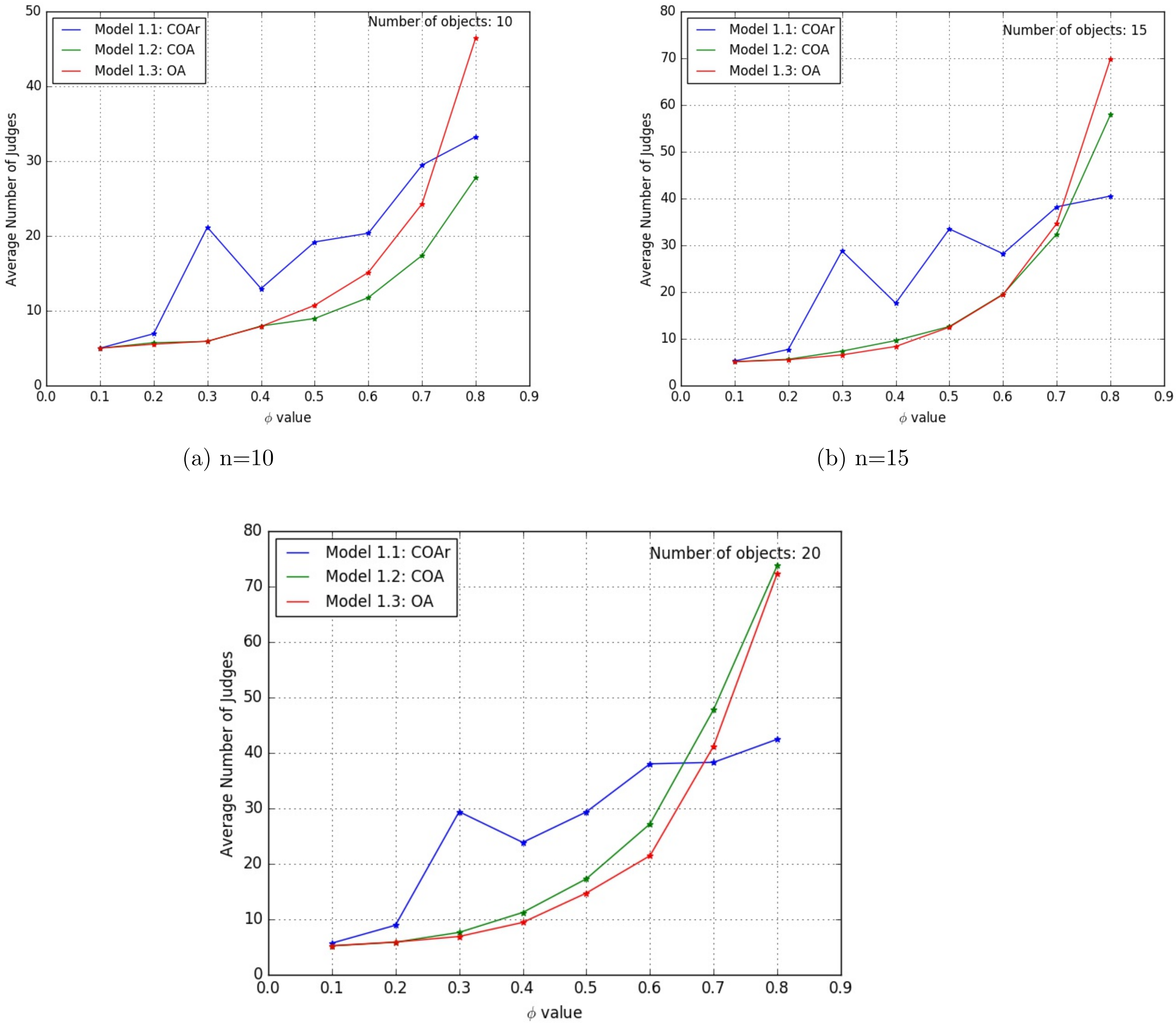
$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} 2\hat{B}_{ij}y_{ij} \\ \text{subject to,} \quad & y_{ij} + y_{ji} \geq 1 && i,j = 1,2...n; && i \neq j \\ & y_{ij} - y_{kj} - y_{ik} \geq -1 && i,j,k = 1,2...n; && i \neq j \neq k \\ & y_{ij} \in \{0,1\} && i,j = 1,2...n \end{aligned}$$

For each of the above three models the Kemeny Snell distance between the ground truth and the aggregate ranking is calculated for n = 10, 15 and 20, where n is the number of objects. To determine the average number of judges required for each object set the KS distance is evaluated for number of judges(k) between 5 and 100 and the value for which the distance obtains the minimum distance is recorded. In each of these experiments the rating and ranking values are homogeneous and complete. The reference ratings are assigned using the Mallow's ϕ-distribution and eight possible values for $\phi \in (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8)$ is used for all k.

Table 1.1 shows the average solution statistics obtained over 30 repetitions for each ϕ value. The solution statistics includes the average number of judges and the minimum KS distance obtained.

Table 1.1: Average solution statistics for complete rating and ranking evaluations with homogeneous errors

	n = 10						n = 15					
ϕ	k$_{min}$			d$_{KS}$			k$_{min}$			d$_{KS}$		
	COAr	COA	OA	COAr	COA	OA	COAr	COA	OA	COAr	COA	OA
0.1	5	5	5	0.000	0.000	0.000	5	5	5	0.000	0.000	0.000
0.2	7	6	6	0.000	0.000	0.000	8	6	6	0.000	0.000	0.000
0.3	21	6	6	0.000	0.000	0.000	29	7	7	0.000	0.000	0.000
0.4	13	8	8	0.024	0.000	0.000	18	10	8	0.018	0.000	0.000
0.5	19	9	11	0.042	0.000	0.000	34	13	13	0.032	0.000	0.000
0.6	20	12	15	0.066	0.000	0.000	28	20	19	0.053	0.000	0.000
0.7	29	17	24	0.094	0.000	0.000	38	32	35	0.082	0.000	0.000
0.8	33	28	47	0.132	0.000	0.000	41	58	70	0.121	0.000	0.001
	n = 20						n = 25					
ϕ	k$_{min}$			d$_{KS}$			k$_{min}$			d$_{KS}$		
	COAr	COA	OA	COAr	COA	OA	COAr	COA	OA	COAr	COA	OA
0.1	6	5	5	0.000	0.000	0.000						
0.2	9	6	6	0.000	0.000	0.000						
0.3	29	8	7	0.000	0.000	0.000						
0.4	24	11	9	0.013	0.000	0.000						
0.5	29	17	15	0.025	0.000	0.000						
0.6	38	27	21	0.045	0.000	0.000						
0.7	38	48	41	0.071	0.000	0.000						
0.8	42	74	72	0.110	0.002	0.002						



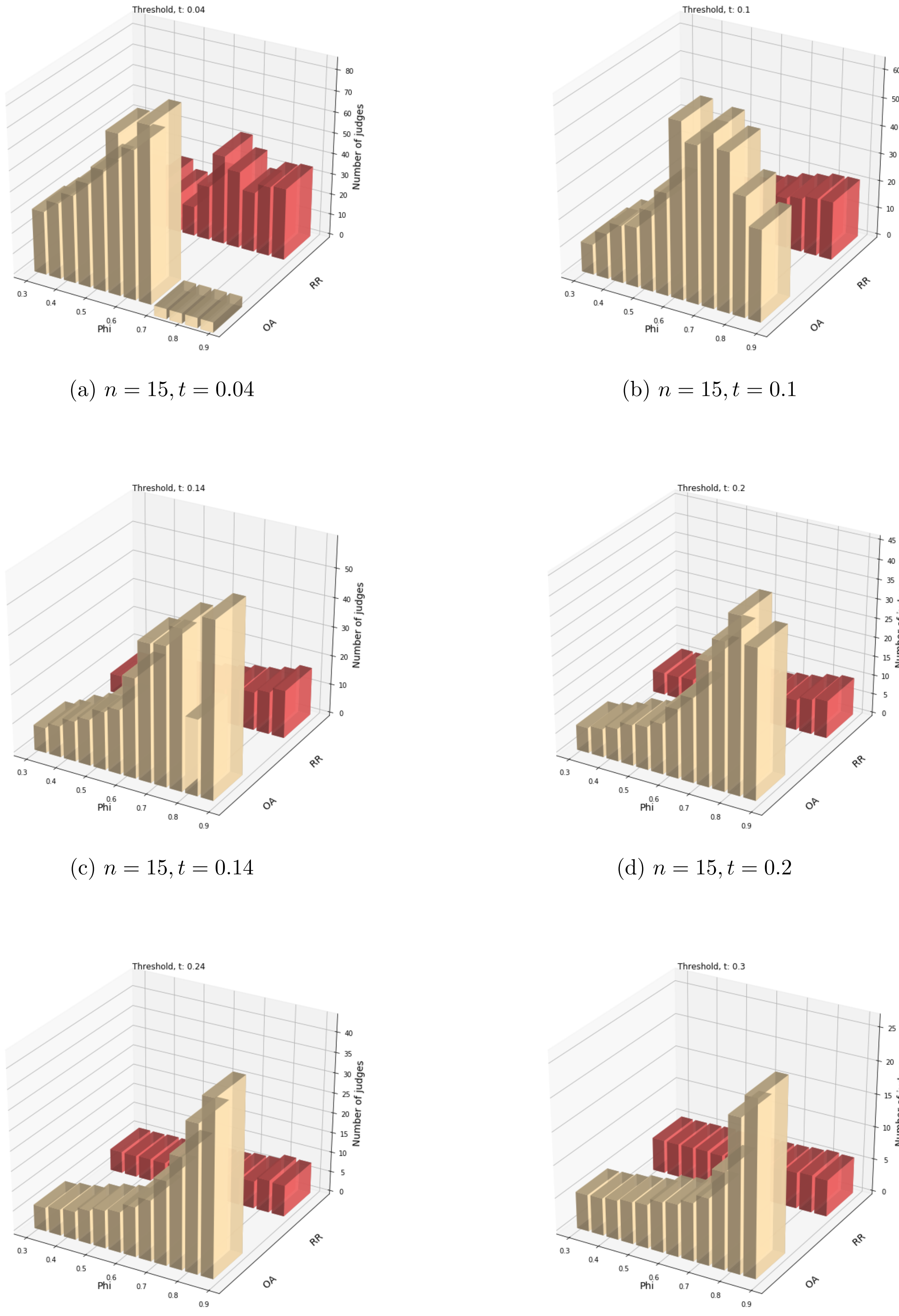
(a) n=10

(b) n=15

(c) n=20

Figure 1: Average number of judges required for min value of d_{KS}

For incomplete ranking and rating only the OA and COA model was used. In this set, the size of the objects evaluated was drawn from the Uniform distribution $U(4, 8)$ for each judge, k and the specific objects in the evaluation set was selected at random. Here, similar to the previous set the reference ratings were assigned using the Mallow's ϕ-distribution but instead of eight twelve possible values for ϕ between 0.3 and 0.85 was used with an increment of 0.05 for all k and the maximum number of judges were limited to 50. The following graphs shows the number of judges required for each model to reach a certain distance value, t.



(a) $n = 15, t = 0.04$

(b) $n = 15, t = 0.1$

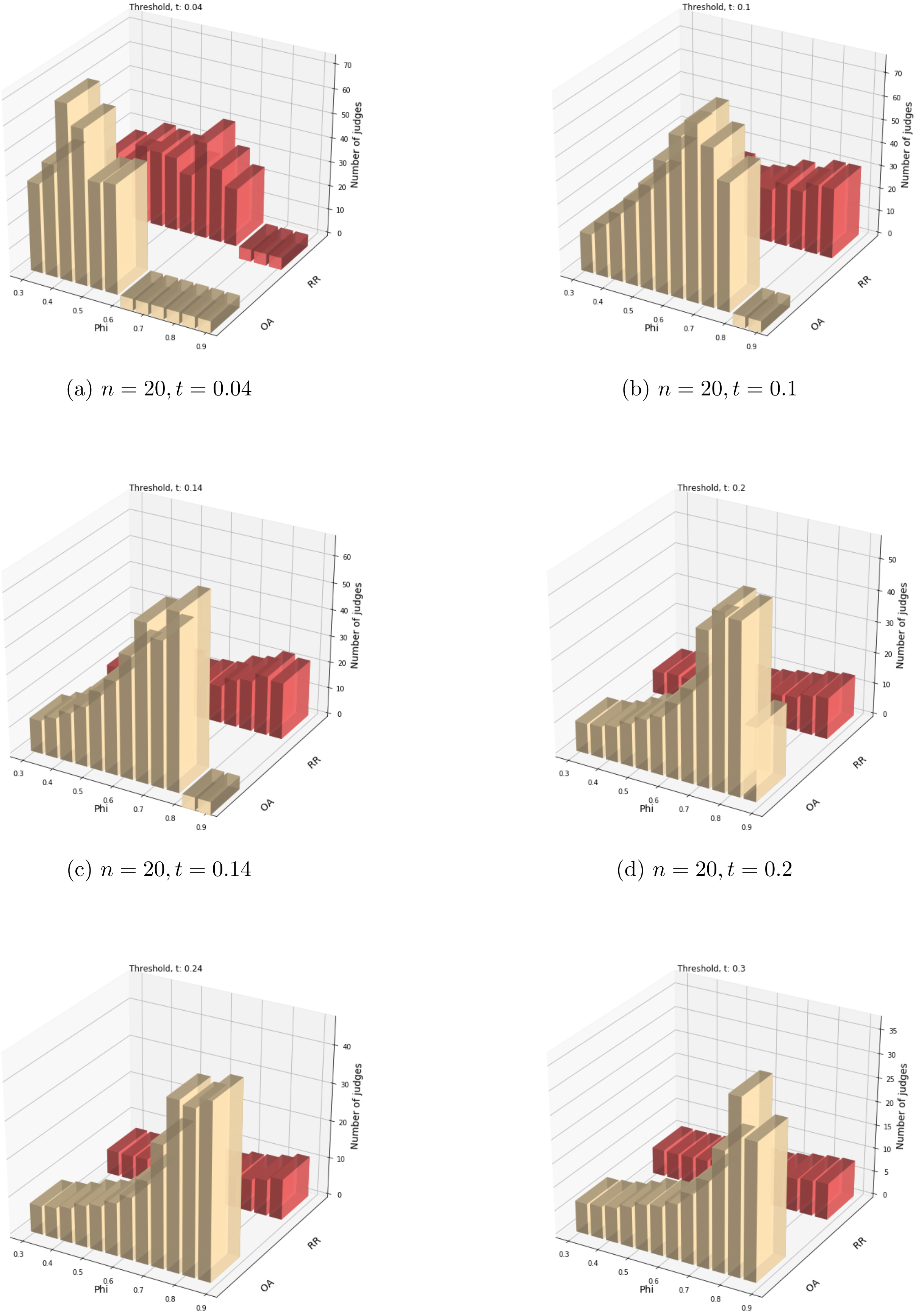
(c) $n = 15, t = 0.14$

(d) $n = 15, t = 0.2$

(e) $n = 15, t = 0.24$

(f) $n = 15, t = 0.3$

Figure 2: Graphs for incomplete ranking, $n = 15$



(a) $n = 20, t = 0.04$

(b) $n = 20, t = 0.1$

(c) $n = 20, t = 0.14$

(d) $n = 20, t = 0.2$

(e) $n = 20, t = 0.24$

(f) $n = 20, t = 0.3$

Figure 3: Graphs for incomplete ranking, $n = 20$

2. **Experiment Set 2:**

In this experiment set three other metrics for rank distances are used besides the Kendall-Tau correlation coefficient to evaluate it's effectiveness. The distances used are: Spearman footrule distance, Chevyshev's distance and the Hamming distance.

Given two complete ranking vectors σ and τ, The Spearman footrule distance between these two full lists can be written as:

$$SF(\sigma, \tau) = \sum_{i=1}^{n} |\sigma(i) - \tau(i)|$$

Proposition 2.1: Given a set V of n objects and a valid ranking-matrix $Y \in \mathbf{B}^{n\times n}$ for a non strict complete ranking, the rank of the ith object, $\sigma(i)$ can be obtained by using the following equation:

$$\sigma(i) = n - \sum_{j \in V, j \neq i} y_{ij} \quad \forall i \in V$$

Proof: From the definition of the ranking-matrix, $y_{ij} = 1$, when object j is ranked higher or is tied with object i. Therefore we have,

$$\begin{aligned}\sum_{j \in V, j \neq i} y_{ij} &= \text{number of objects ranked higher and is tied with object } i \\ &= n - (\text{number of objects ranked lower than object } i + 1) \\ &= n - \text{rank of object } i \\ &= n - \sigma(i)\end{aligned}$$

Based on Proposition 2.1 using Spearman footrule as the metric function we can formulate the Ordinal Aggregation problem as follows:

Model 2.1: SF-OA: OA using Spearman Footrule distance:

$$\begin{aligned}
\text{minimize} \quad & \sum_{k \in K} \sum_{i \in R^k} h_i^k \\
\text{subject to,} \quad & y_{ij} + y_{ji} \geq 1 && i, j = 1, 2...n \\
& y_{ij} - y_{kj} - y_{ik} \geq -1 && i, j, k = 1, 2...n \\
& h_i^k \geq n - \sum_{j \in V, i \neq j} y_{ij} - b_i^k && k = 1, 2...m, i \in B^k \\
& h_i^k \geq -n + \sum_{j \in V, i \neq j} y_{ij} + b_i^k && k = 1, 2...m, i \in B^k \\
& h_i^k \text{ unrestricted} \\
& 0 \leq y_{ij} \leq 1 && i, j = 1, 2...n \\
& y_{ij} \in \mathbb{Z}
\end{aligned}$$

The Spearman footrule distance basically calculates the L_1 norm between two complete rankings whereas the Chevyshev's distance determines the L_∞ norm between two full lists. The Chevyshev's distance between two complete ranking vectors can be given as:

$$C(\sigma, \tau) = \max\{|\sigma_1(i) - \tau_1(i)|, |\sigma_2(i) - \tau_2(i)|.........|\sigma_n(i) - \tau_n(i)|\}$$

Based on the definition of Chevyshev's distance and using Proposition 2.1 to determine the rank position of each object the Ordinal Aggregation model can be reformulated as:

Model 2.2: C-OA: OA using Chevyshev's distance:

$$\begin{aligned}
\text{minimize} \quad & \sum_{k \in K} h^k & \\
\text{subject to,} \quad & y_{ij} + y_{ji} \geq 1 & i, j = 1, 2 ... n \\
& y_{ij} - y_{kj} - y_{ik} \geq -1 & i, j, k = 1, 2 ... n \\
& h^k \geq n - \sum_{j \in V, i \neq j} y_{ij} - b_i^k & k = 1, 2 ... m, i \in R^k \\
& h^k \geq -n + \sum_{j \in V, i \neq j} y_{ij} + b_i^k & k = 1, 2 ... m, i \in R^k \\
& h^k \text{ unrestricted} & \\
& 0 \leq y_{ij} \leq 1 & i, j = 1, 2 ... n \\
& y_{ij} \in \mathbb{Z} &
\end{aligned}$$

The last distance function evaluated is the Hamming Distance, which is defined for two complete ranking vectors as:

$$H(\sigma, \tau) = n - \sum_{i=1}^{n} \sum_{j=1}^{n} I(\sigma(i) = j) I(\tau(i) = j)$$

From the above definition the hamming distance calculates the number of dissimilarities between two rankings. So if we maximize the number of similarities between the aggregate rank and the reference rankings we will be able to formulate a model that uses hamming distance as the distance function for Ordinal Aggregation. Therefore the model can be reformulated as:

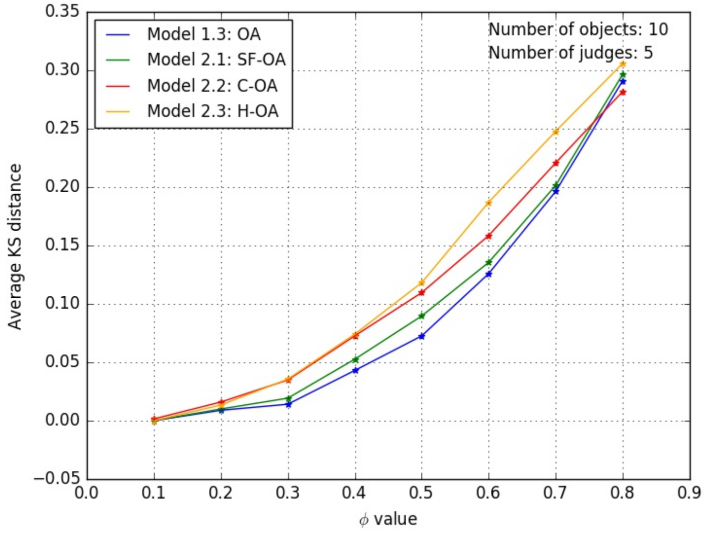
Model 2.3: H-OA: OA using Hamming distance:

$$\begin{aligned}
\text{maximize} \quad & \sum_{k \in K} \sum_{i \in V} z_i^k & \\
\text{subject to,} \quad & y_{ij} + y_{ji} \geq 1 & i, j = 1, 2 ... n \\
& y_{ij} - y_{kj} - y_{ik} \geq -1 & i, j, k = 1, 2 ... n \\
& h_i^k \leq (n - 1)(1 - z_i^k) & k = 1, 2 ... m, i \in R^k \\
& h_i^k \geq -(n - 1) z_i^k + 1 & k = 1, 2 ... m, i \in R^k \\
& h_i^k \geq n - \sum_{j \in V, i \neq j} y_{ij} - b_i^k & k = 1, 2 ... m, i \in R^k \\
& h_i^k \geq -n + \sum_{j \in V, i \neq j} y_{ij} + b_i^k & k = 1, 2 ... m, i \in R^k \\
& 0 \leq z_i^k \leq 1 & i, j = 1, 2 ... n \\
& h_i^k \text{ unrestricted} & \\
& 0 \leq y_{ij} \leq 1 & i, j = 1, 2 ... n \\
& y_{ij}, z_i^k \in \mathbb{Z} &
\end{aligned}$$

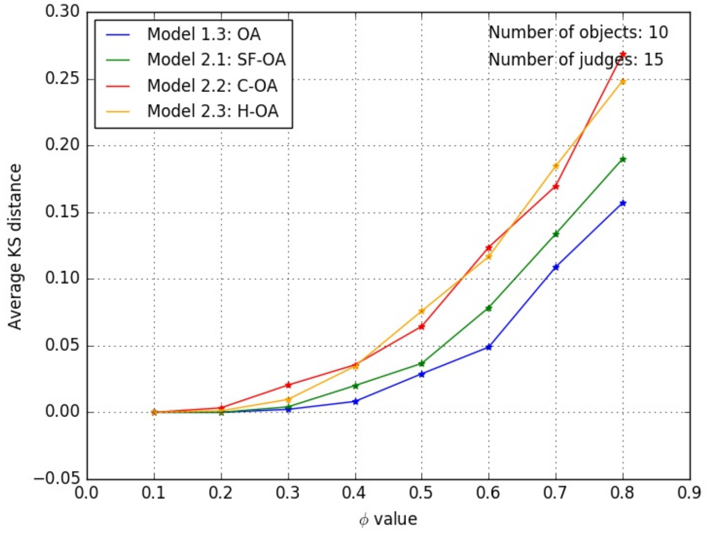
The objective of this experiment set is to determine which model provides a lower distance value for Ordinal Aggregation. The values of n are used for each model (10, 15, 20) and the value of m are determined using the following equation:

$$m = \lfloor p * n \rfloor$$

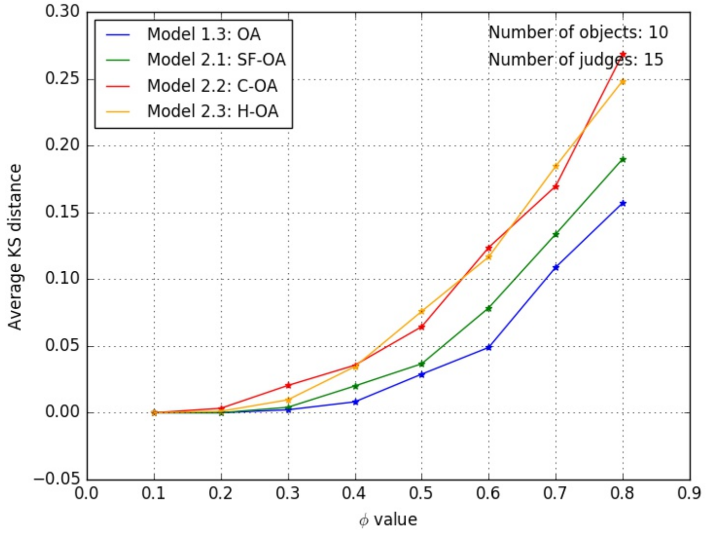
where, p = 0.5, 1.0, 1.5. The p values helps to evaluate the cases when $m < n$, $m = n$ and $m > n$. Similar to the first set each experiment is repeated 30 times for different values of ϕ and complete reference rankings.



(a) n=10, p=0.5



(b) n=10, p=1.0



(c) n=10, p=1.5

Figure 4: Average KS distance for different rating distances when n=10

3. Experiment Set 3:

For the experiments in this set the Kendall Tau correlation coefficient is used as the distance metric for the ordinal aggregation, but for cardinal aggregation the L_1 and L_2 norms are used. Both models are given below:

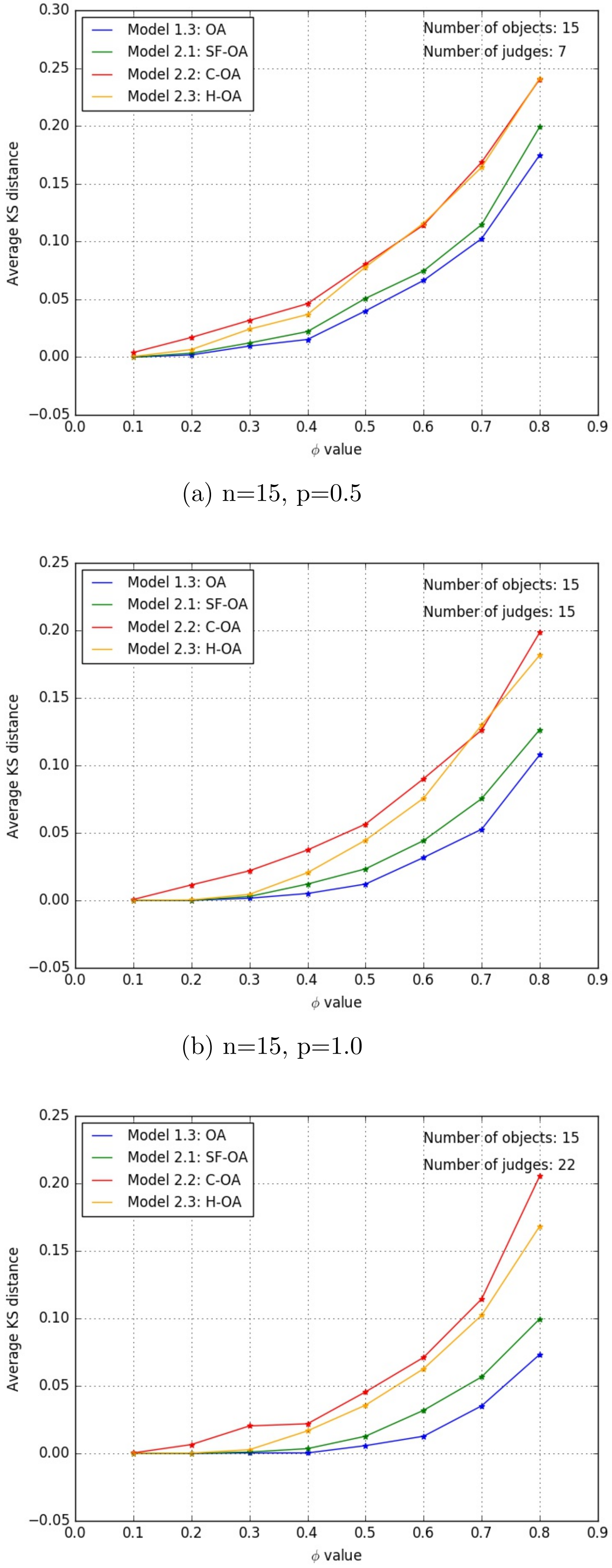
Model 3.1: L_1 norm:

$$\begin{array}{lll}
\text{maximize} & \displaystyle\sum_{k\in K}\sum_{i\in V^k} -\frac{1}{R|V^k|}t_i^k + \sum_{i=1}^{n}\sum_{j=1}^{n} 2\hat{B}_{ij}y_{ij} & \\
\text{subject to,} & t_i^k \geq \mu x_i - a_i^k & i \in V^k, k = 1, 2....m \\
 & t_i^k \geq -\mu x_i + a_i^k & i \in V^k, k = 1, 2....m \\
 & x_i - x_j \leq (M+\mu)y_{ij} - \mu & i, j = 1, 2...n \\
 & x_i - x_j \geq M(y_{ij} - 1) & i, j = 1, 2...n \\
 & y_{ij} + y_{ji} \geq 1 & i, j = 1, 2...n \\
 & y_{ij} - y_{kj} - y_{ik} \geq -1 & i, j, k = 1, 2...n \\
 & 0 \leq x_i \leq \dfrac{(U-L)}{\mu} & i = 1, 2.......n \\
 & 0 \leq y_{ij} \leq 1 & i, j = 1, 2...n
\end{array}$$

Model 3.2: L_2 norm:

$$\begin{array}{lll}
\text{maximize} & \displaystyle\sum_{k\in K}\sum_{i\in V^k} -\frac{\mu^2}{R^2|V^k|^2}x_i^2 + \sum_{k\in K}\sum_{i\in V^k} 2\frac{\mu a_i^k}{R^2|V^k|^2}x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} 2\hat{B}_{ij}y_{ij} & \\
\text{subject to,} & x_i - x_j \leq (M+\mu)y_{ij} - \mu & i, j = 1, 2...n \\
 & x_i - x_j \geq M(y_{ij} - 1) & i, j = 1, 2...n \\
 & y_{ij} + y_{ji} \geq 1 & i, j = 1, 2...n \\
 & y_{ij} - y_{kj} - y_{ik} \geq -1 & i, j, k = 1, 2...n \\
 & 0 \leq x_i \leq \dfrac{(U-L)}{\mu} & i = 1, 2.......n \\
 & 0 \leq y_{ij} \leq 1 & i, j = 1, 2...n
\end{array}$$

The instance and solution statistics for this set is same as the ones in experiment set 2.

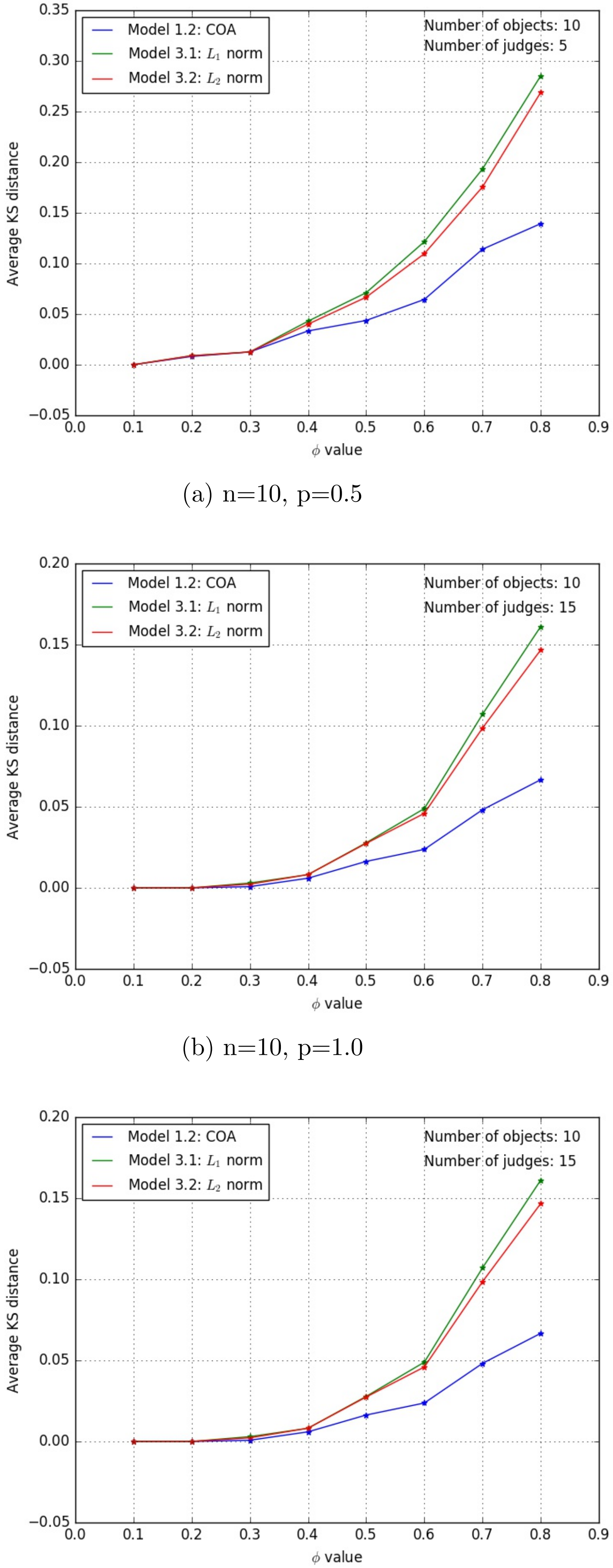


(a) n=15, p=0.5

(b) n=15, p=1.0

(c) n=15, p=1.5

Figure 5: Average KS distance for different ranking distances when n=15

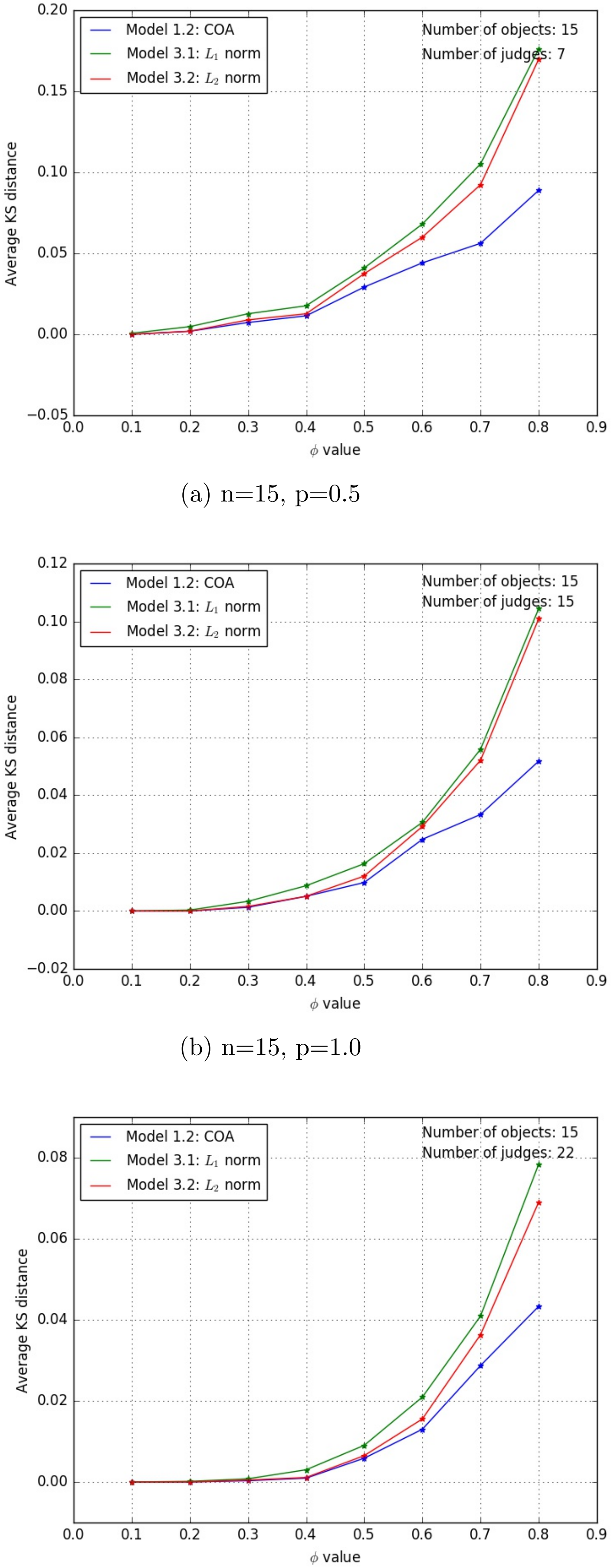


(a) n=10, p=0.5

(b) n=10, p=1.0

(c) n=10, p=1.5

Figure 6: Average KS distance for different rating distances when n=10

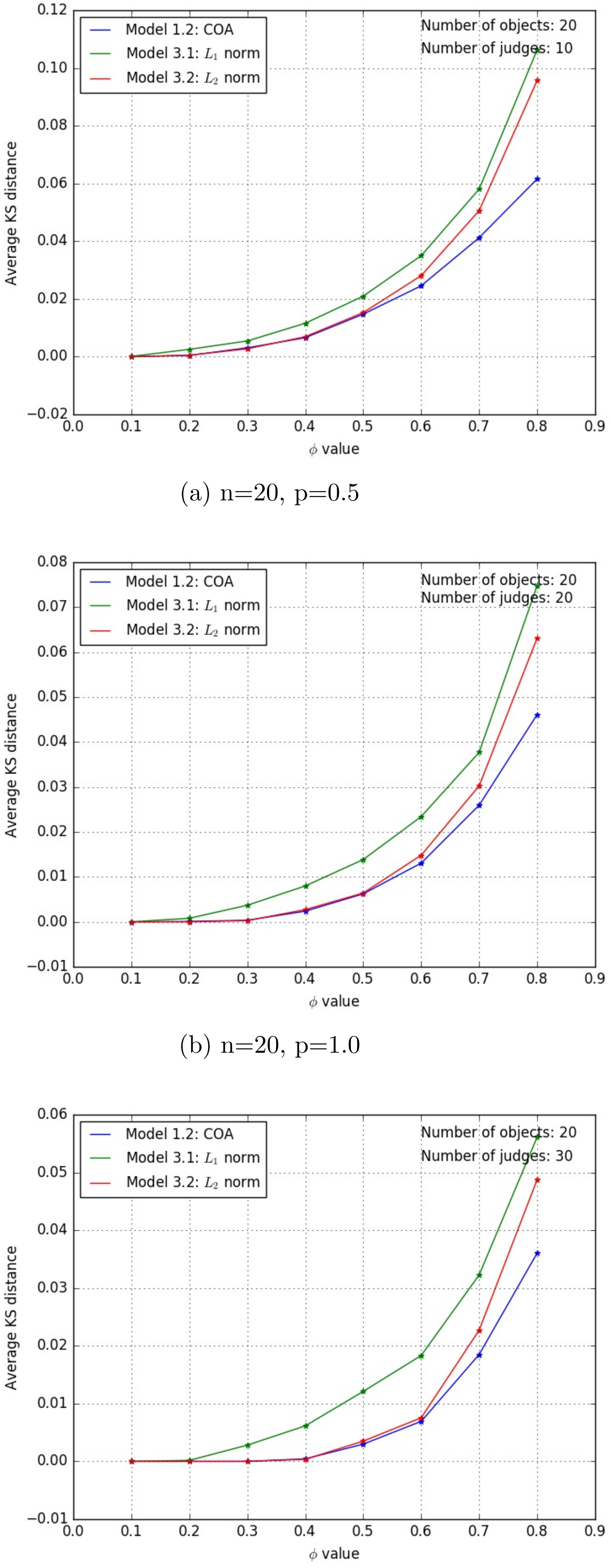


(a) n=15, p=0.5

(b) n=15, p=1.0

(c) n=15, p=1.5

Figure 7: Average KS distance for different rating distances when n=15



(a) n=20, p=0.5

(b) n=20, p=1.0

(c) n=20, p=1.5

Figure 8: Average KS distance for different rating distances when n=20

From the graphs we can see that for each case the COA model provides a lower distance value. The worst performance is given by the L_1 norm.

2 Crowd Wisdom

The second part of this addendum employs a set of aggregation methods to harness crowd wisdom. Crowdsourcing has made a significant deal of impact on many fields. It has a proved potential for making discoveries, raising public awareness, developing research with more accurate results, and developing innovations by harnessing the collective efforts of multiple people. For example, the UK Newspaper "the Guardian" used crowdsourcing to classify hundreds and thousands of expense claims of parliament, and identify a lot of fraudulent activity (Wazny, 2017). It has also recently been utilized to determine collective human ethics and drive ethical decision making behind self-driving cars (Noothigattu et al., 2018); model gene networks by aggregating all predicted interactions between proteins and its target genes (Marbach et al., 2012); improve the accuracy of forecasting epidemic diseases (Li et al., 2016). Because this technology is still in its infancy, new applications are being discovered every day, and it is likely that more advanced ways of applying crowdsourcing to new scenarios are yet to be discovered. Among some of the most popular crowdsourcing activities are human computation tasks such as digitization of books through reCAPTCHA, which verifies human by asking users to read a scanned text or images that is easy for humans but for robots (Von Ahn et al., 2008), to collect labels for machine learning (Dekel and Shamir, 2009; Raykar et al., 2010). Human computation tasks usually require single tasks to be done by many people, where the combined results are aggregated together (Mao et al., 2013). Because of varying subjective scales among humans, it is usually unreasonable to trust a single person to provide a result. Hence, more reliable outcome can be obtained by aggregating input from many individuals and utilizing group decisions. The opinions of groups tend to outperform the opinions of individuals. This is a principle commonly referred to as the "wisdom of crowds", which theorizes that aggregated information from large groups of people generally results in better outcomes than that from any individual.

However, crowds are not always wise; according to Surowiecki (2005), the following four conditions are required to make a crowd wise: *independence, diversity, decentralization,* and *aggregation.* In detail, each person in the crowd should be independent, so that they pay attention mostly to their own information. Also, crowds needs to be diverse, so that people are bringing different pieces of information and not worrying about what everyone around them thinks. Moreover, crowds needs to be decentralized, so that no one is dictating the crowd's answer. Lastly, it needs a way of aggregating people's opinions into one collective verdict.

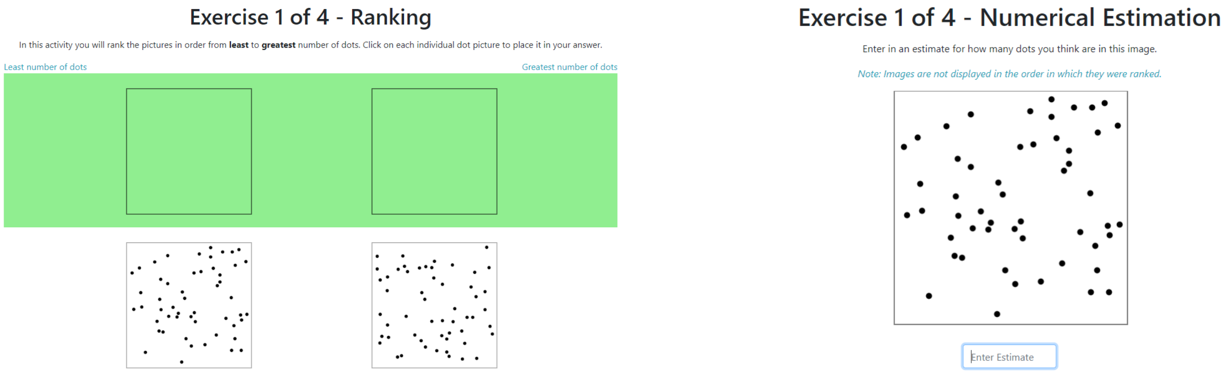
Rating and rank aggregation is one of the pertinent areas that could strengthen the effect of the wisdom of crowds. The essence of rank aggregation is combining individual preferences into collective preferences, which has the same vein as the idea of the wisdom of crowds. Hence, this work applies rank aggregation frameworks to enhance the wisdom of crowds in human computation tasks. The existing works in computational social choice have shown that depending on the input-elicitation and aggregation methodologies used, group consensus is significantly affected (Mao et al., 2013). For instance, it is shown that using a ranking scale to compare objects may potentially yield different results from asking users to provide a rating, which tends to have higher variability (Rankin and Grube, 1980). Recent studies also show that collecting and aggregating multiple sources of information together, such as individual ratings and rankings, could better represent opinions of individual judges (Escobedo et al., 2020). Furthermore, the size of the task at hand given to each individual user has an impact on the overall accuracy and efficiency. Crowdsourced comparison tasks have found that using larger problem sizes saves crowdsourcers a significant amount of money, and that there is a big trade-off between problem size and effort (Wilber et al., 2014).

Human computation is a new and evolving research area that studies how to harness human intelligence to solve computational problems that are difficult for computers to process (Law and Ahn, 2011). For example, image recognition is trivial for humans, but it is still challenging for computer programs. Hence, a user labels images through the online game to help image labeling for a more accurate image search (Von Ahn and Dabbish, 2004; Von Ahn, 2008).

Our motivation is to show how rank information can increase quality of numerical estimation and how rating information can increase quality of ranking estimation. Moreover, we would like to leverage these multiple modalities to make it possible to obtain wisdom from smaller crowds, which will make crowd wisdom more practical for companies to implement. Lastly, we want to see how consensus-based optimization models could help achieve these goals.

A *dot-guessing game* is one of the human computation tasks, which has been suggested by Horton (2010). The basic idea of the dot-guessing game is that a subject estimates the number of dots in the image. Due to its pseudosubjective property which allows humans to provide subjective assessments using objective metrics (Horton, 2010; Janowski et al., 2011), a dot-guessing game is an appropriate human computation experiment on which to explore the application of the tools developed in this research.

Our human computation experiment is extended from the basic concept of the dot-guessing game. Rather than asking the estimates of the number of dots only, we first ask users to provide the rank position of each image among the set of images (i.e., ordinal evaluation) and then the number of dots in each image (i.e., cardinal evaluation).The goal of our human computation experiment was to utilize crowd wisdom to rank and provide numerical estimates to 30 images of dots. Note that each image is shown one at a time in the numerical estimation, while multiple images are shown at a time in ranking estimation. The following figures show the interface of each task.


Exercise 1 of 4 - Ranking

In this activity you will rank the pictures in order from **least** to **greatest** number of dots. Click on each individual dot picture to place it in your answer.

(a) Dot guessing game: rankings

Exercise 1 of 4 - Numerical Estimation

Enter in an estimate for how many dots you think are in this image.

Note: Images are not displayed in the order in which they were ranked.

(b) Dot guessing game: ratings

Figure 9: Interface of the dot-guessing experiment

This activity was published on Amazon MTurk to distribute tasks to real online workers. Amazon MTurk is a popular crowdsourcing marketplace that allows individuals to post human intelligent tasks and these tasks are fulfilled by human on this website (Ipeirotis, 2010). Each MTurk worker who accepted our activity was shown an intro page with a briefing of the activity, and then was prompted to put in their MTurk ID to enter the activity. Once a new user entered the activity, the user was assigned 4 questions. These questions contained images from each of the four problem sets, and involved a question that involved evaluating 2, 3, 5, and 6 images at a time. The order in which the questions appeared was randomized, and the images that appeared in each question were assigned randomly from their corresponding problem set. Images were assigned to users in a way where each image was evaluated the exact same number of times.

Each image contains dots ranging from 50 to 79. For each exercise, a user is looking at 2, 3, 5, and 6 images at a time. The mages shown to users are called a *frame* throughout this section. For example, Figure (9a) displays the interface when two frames are shown to the user.

To ensure that all the images are seen by exactly the same number of users, a set of images are preassigned in the problem set, called a *batch*. For each question of varying size, it has a different number of batches: frames of 2 have 4 batches, frames of 3 have 6 batches, frames of 5 have 10 batches, and frames of 6 have 12 batches. Therefore, each exercise for users is drawn from the batches. At the end of the study, users are asked to fill out a brief survey about their experience with different kinds of questions and basic user demographics. A total of 300 participants completed the study. Data from participants that started the study, but did not finish all of the questions were removed. Out of the 300 participants that completed the activity, 288 participants completed the demographic survey.

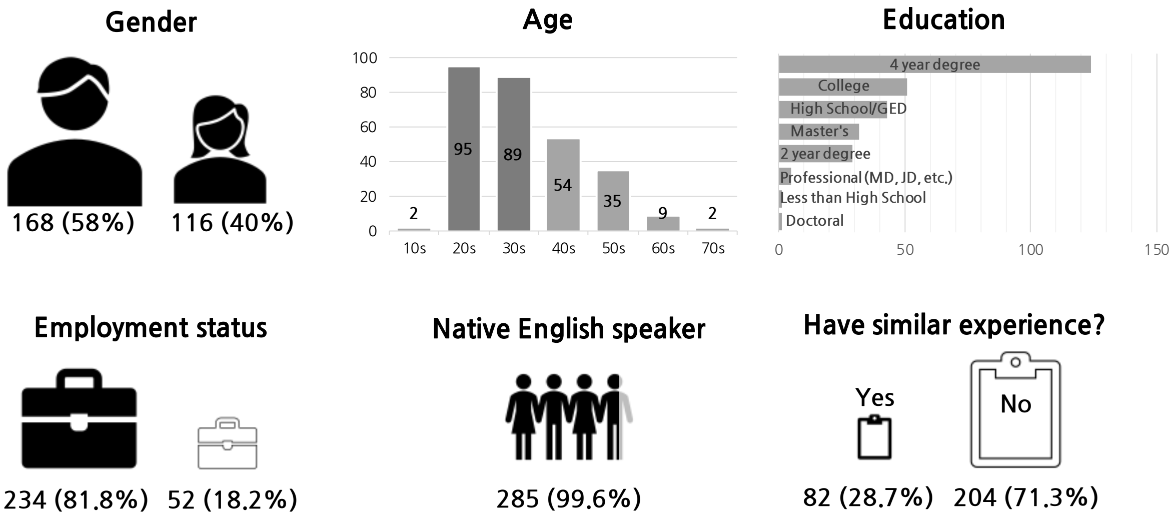


Figure 10: Summary of participants demographics

Note that four requirements mentioned earlier are met in the experiment setting: Firstly, each person provides ratings and rankings independently, and each person provides diverse opinions with any reasonable positive integer. Moreover, there is no dictator who imposes the crowd's opinions. Furthermore, rating and ranking estimates from each user are aggregated via consensus-based aggregation frameworks, which is the main focus of this experiment.

Through the experiment, the following research questions are addressed: Does information from more participants yield a better aggregated outcome? Does having more information improve the quality of the crowd wisdom? Does the number of images shown to the users increase the accuracy of estimation? Do the different types of multimodal information (i.e., both ratings and ranking) help achieve better crowd wisdom?

In addition to aforementioned aggregation frameworks, other aggregation models are used for this experiment. Similar to CA, the separation-deviation model (SD) can be used where the input is given as pairwise comparison preferences of alternative and pointwise score evaluation (Hochbaum, 2010). The two major components of this model are: *separation* and *deviation*. The separation term takes into account the difference between the pairwise comparison of two alternatives v_i and v_j in the aggregated outcome and each judge's evaluations, as d_{CK} and d_{NPCK} did. The deviation term considers the difference between the value of alternative v_i in the aggregated outcome and in each judge's evaluation. In the optimization model, the separation is penalized by s_{ij}^k and the deviation is penalized by d_i^k, with respect to a judge k's evaluation on the alternatives v_i and v_j. The model can be mathematically formulated as follows:

$$\underset{\boldsymbol{r}}{\operatorname{argmin}} \sum_{\ell=1}^{|L|} \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij}^k((r_i - r_j) - (a_i^k - a_j^k)) + \sum_{i=1}^{n} \sum_{j=1}^{n} d_i^k(r_i - a_i^k), \tag{1}$$

where r_i represents the score of alternative v_i in the aggregated outcome and a_i^k represents the score of alternative in the k-th judge's evaluation. Here, r_i is constrained to be integer and the upper and lower bounds of r_i are $\max(a_i^k)$ and $\min(a_i^k)$, respectively. Note that the model is solvable in polynomial time because the constraint coefficient matrix is totally unimodular.

The Borda count (de Borda, 1781) is a well-known scoring method that assigns a score to each candidate, calculates a final score for each candidate by summing the scores earned over all evaluations and chooses the winner with the highest score. Assume that when there are n alternatives, the most preferred alternative receives a score of n, the second most preferred alternative receives a score of $n-1$, and so on. This can be mathematically formulated as follows:

$$\arg\max_{v_i \in V} Borda(v_i) = \arg\max_{v_i \in V} \sum_{\ell=1}^{|L|} (n - a_i^\ell + 1). \tag{2}$$

This method can yield inconsistent outcomes due to vulnerability to error and manipulation (Dummett, 1998; Favardin et al., 2002), especially when the rankings are incomplete (Moreno-Centeno and Escobedo, 2016).

The plurality rule selects an alternative with the most first-place votes (i.e., which has the largest plurality score). Let a function f of determining if an alternative v_i is chosen as the winner be:

$$f(a_i^\ell) = \begin{cases} 1 & \text{if } a_i^\ell = 1, \\ 0 & \text{otherwise,} \end{cases}$$

then, the plurality rule can be written as follows:

$$\arg\max_{v_i \in V} Plurality(v_i) = \arg\max_{v_i \in V} \sum_{\ell=1}^{|L|} f(a_i^\ell). \tag{3}$$

The Copeland rule chooses an alternative with the highest Copeland score, which can be mathematically formulated as (Brandt et al., 2016):

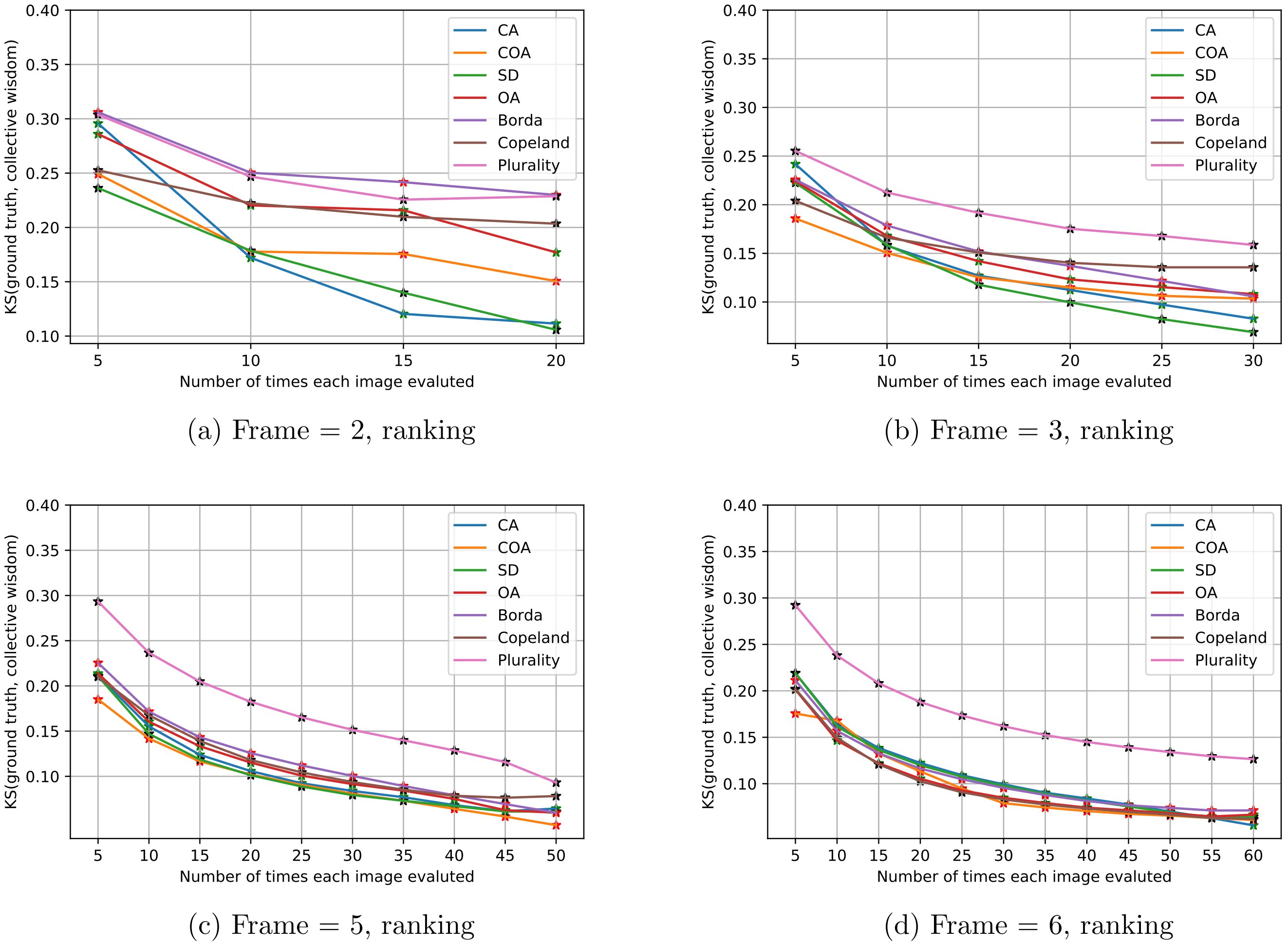
$$\arg\max_{v_i \in V} Copeland(v_i) = \arg\max_{v_i \in V} (|\{v_j \in V \backslash \{v_i\} : a_i < a_j\}| - |\{v_j \in V \backslash \{v_i\} : a_i > a_j\}|). \tag{4}$$

The aforementioned aggregation frameworks are used to enhance the wisdom of crowds. First, we used optimization-based aggregation methods: cardinal aggregation (CA), ordinal aggregation (OA), joint aggregation (COA) and the separation-deviation model. In addition, we used non-optimization-based aggregation methods, which are the traditional voting methods: the Borda rule, Copeland rule, plurality rule. These methods have been previous used for dot-guessing tasks (Mao et al., 2013). Because these voting methods are social choice functions, they only return the winner, not the alternative-ordering (i.e., ranking). To obtain a ranking, alternatives are sorted by their scores (e.g., Borda score, Copeland score, plurality score) in non-increasing order.

To examine how close the collective rankings are to the ground truth, the distance between the ground truth and the collective ranking is quantified using the Kemeny-Snell distance. Similarly, the closeness between the ground truth and the collective rating is quantified using the Euclidean distance. Note that the Euclidean distance is normalized by the difference of the maximum number and the minimum number of dots (here, the distance is divided by 29).

Experiment results This section summarizes how close the collective ranking and rating are to the ground truth for each aggregation framework. The left column in Figure 2 and 2 shows the results of 60 participants and the right column shows the results of 300 participants.

Figure 2 summarizes how close the collective ranking is to the ground truth for each rank aggregation framework.



(a) Frame = 2, ranking

(b) Frame = 3, ranking

(c) Frame = 5, ranking

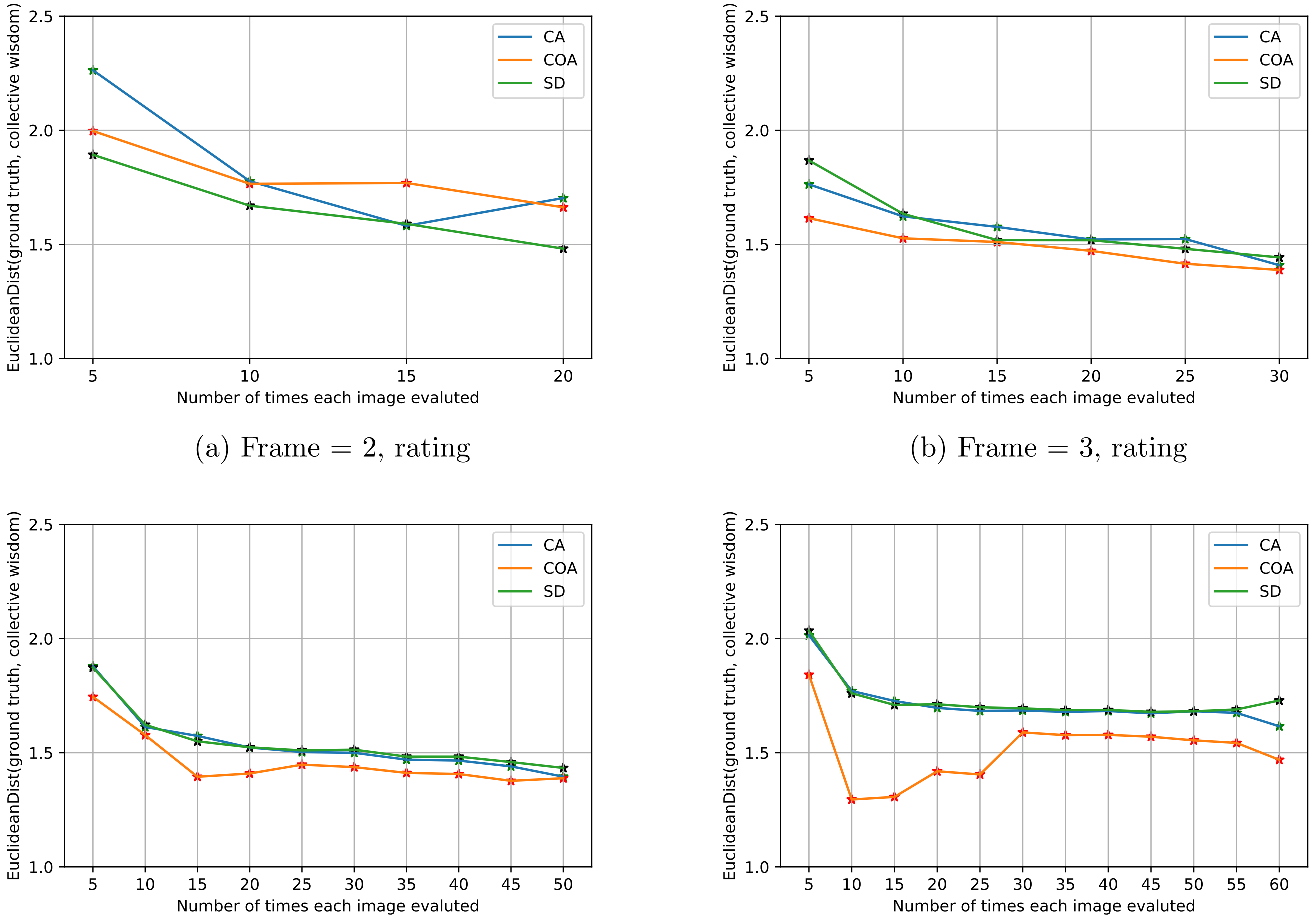
(d) Frame = 6, ranking

Figure 11: As the number of evaluations on each image increases, the aggregated outcome is closer to the ground truth.

As shown in the above figures, the higher number of times each image is evaluated, the shorter distance from the collective rankings to the ground truth is; that is, having more information also helps recover the ground truth ranking. These results align with the idea of the wisdom of crowds.

Moreover, showing more images at a time (i.e., higher size of frames) helps humans estimate a ranking close to the ground truth. It can be presumed that having more images can help users to give a correct order of at least some of the images. Specifically, when two images are seen, there are only two possibilities of ordering them: giving a correct order or not. In contrast, when six images are seen, there are 720 possibilities of ordering them. Although it is very difficult to order images perfectly, users may order part of them correctly due to the implicit pairwise comparisons in a ranking. Hence, despite the higher cognitive load to order more images, the collective ranking from the higher number of frames returns the shorter distance to the ground truth ranking.

Furthermore, the joint aggregation model (i.e., aggregation using both ratings and rankings) outperforms other models when users are displayed more images at a time. This means that having different types of multimodal information helps recover the ground truth ranking for a larger individual problem sizes.



(a) Frame = 2, rating

(b) Frame = 3, rating

(c) Frame = 5, rating

(d) Frame = 6, rating

Figure 12: The joint aggregation model outperforms other models.

Figure 2 summarizes how close collective numerical estimations (ratings) are to the ground truth for rating aggregation frameworks. The rating aggregation, the separation-deviation, and the joint aggregation models are only tested for the collective ratings because rank aggregation and other traditional voting methods are not able to find collective ratings.

Similar to rank aggregation, the higher number of times each image is evaluated, the closer the collective ratings are to the ground truth; this result supports that gathering more information yields better aggregated outcome. However, contrary to collective rankings, there is no specific pattern as the changes in the number of images shown. Because numerical estimation is done one image at a time, the number of times each image is shown does not seem to significantly affect the quality of solutions.

Furthermore, with respect to the model performance, although there is no big difference between three models, the joint aggregation model performs slightly better than the separation-deviation model and ratings-only model, in general. This indicates different types of multimodal information (both rankings and ratings) can be leveraged to achieve better crowd wisdom.

References

Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. *Handbook of computational social choice.* Cambridge University Press, 2016.

Jean C de Borda. Mémoire sur les élections au scrutin. 1781.

Ofer Dekel and Ohad Shamir. Vox populi: Collecting high-quality labels from a crowd. In *COLT*, 2009.

Michael Dummett. The borda count and agenda manipulation. *Social Choice and Welfare*, 15(2):289–296, 1998.

Adolfo R Escobedo, Dorit S Hochbaum, and Erick Moreno-Centeno. An axiomatic distance methodology for aggregating multimodal evaluations. *under review*, 2020.

Pierre Favardin, Dominique Lepelley, and Jérôme Serais. Borda rule, copeland method and strategic manipulation. *Review of Economic Design*, 7(2):213–228, 2002.

Dorit S Hochbaum. The separation, and separation-deviation methodology for group decision making and aggregate ranking. In *Risk and Optimization in an Uncertain World*, pages 116–141. INFORMS, 2010.

John J Horton. The dot-guessing game: A 'fruit fly'for human computation research. *Available at SSRN 1600372*, 2010.

Panagiotis G Ipeirotis. Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2):16–21, 2010.

Lucjan Janowski, Mikołaj Leszczuk, Zdzisław Papir, and Piotr Romaniak. The design of an objective metric and construction of a prototype system for monitoring perceived quality (qoe) of video sequences. *Journal of Telecommunications and Information Technology*, pages 87–94, 2011.

Edith Law and Luis von Ahn. Human computation. *Synthesis lectures on artificial intelligence and machine learning*, 5(3):1–121, 2011.

Eldon Y Li, Chen-Yuan Tung, and Shu-Hsun Chang. The wisdom of crowds in action: Forecasting epidemic diseases with a web-based prediction market system. *International journal of medical informatics*, 92: 35–43, 2016.

Andrew Mao, Ariel D Procaccia, and Yiling Chen. Better human computation through principled voting. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Andrej Aderhold, Richard Bonneau, Yukun Chen, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796, 2012.

Erick Moreno-Centeno and Adolfo R Escobedo. Axiomatic aggregation of incomplete rankings. *IIE Transactions*, 48(6):475–488, 2016.

Ritesh Noothigattu, Snehalkumar S Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D Procaccia. A voting-based system for ethical decision making. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

William L Rankin and Joel W Grube. A comparison of ranking and rating procedures for value system measurement. *European Journal of Social Psychology*, 10(3):233–246, 1980.

Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322, 2010.

James Surowiecki. *The wisdom of crowds.* Anchor, 2005.

Luis Von Ahn. Human computation. In *2008 IEEE 24th international conference on data engineering*, pages 1–2. IEEE, 2008.

Luis Von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, 2004.

Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008.

Kerri Wazny. "crowdsourcing" ten years in: A review. *Journal of global health*, 7(2), 2017.

Michael J Wilber, Iljung S Kwak, and Serge J Belongie. Cost-effective hits for relative similarity comparisons. In *Second AAAI conference on human computation and crowdsourcing*, 2014.