



# BargCrEx: A System for Bargaining Based Aggregation of Crowd and Expert Opinions in Crowdsourcing

Ana Vukicevic<sup>1,2</sup> · Milan Vukicevic<sup>1</sup> · Sandro Radovanovic<sup>1</sup> · Boris Delibasic<sup>1</sup>

Accepted: 21 April 2022 / Published online: 21 May 2022  
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

## Abstract

Crowdsourcing and crowd voting systems are being increasingly used in societal, industry, and academic problems (labeling, recommendations, social choice, etc.) due to their possibility to exploit “wisdom of crowd” and obtain good quality solutions, and/or voter satisfaction, with high cost-efficiency. However, the decisions based on crowd vote aggregation do not guarantee high-quality results due to crowd voter data quality. Additionally, such decisions often do not satisfy the majority of voters due to data heterogeneity (multimodal or uniform vote distributions) and/or outliers, which cause traditional aggregation procedures (e.g., central tendency measures) to propose decisions with low voter satisfaction. In this research, we propose a system for the integration of crowd and expert knowledge in a crowdsourcing setting with limited resources. The system addresses the problem of sparse voting data by using machine learning models (matrix factorization and regression) for the estimation of crowd and expert votes/grades. The problem of vote aggregation under multimodal or uniform vote distributions is addressed by the inclusion of expert votes and aggregation of crowd and expert votes based on optimization and bargaining models (Kalai–Smorodinsky and Nash) usually used in game theory. Experimental evaluation on real world and artificial problems showed that the bargaining-based aggregation outperforms the traditional methods in terms of cumulative satisfaction of experts and crowd. Additionally, the machine learning models showed satisfactory predictive performance and enabled cost reduction in the process of vote collection.

**Keywords** Crowd-voting · Expert knowledge · Matrix-factorisation · Machine learning · Bargaining models

---

✉ Ana Vukicevic  
ak20195017@student.fon.bg.ac.rs

Extended author information available on the last page of the article

## 1 Introduction

Crowdsourcing has gained increased popularity in recent years. The benefits of inclusion of crowd opinions (ranks, grades, labels, polls, etc.) have led to many successful applications in a wide range of industry applications: retail (Devari et al. 2017), internet sales (Kleemann et al. 2008), social networks (Jiang et al. 2019), web search (Alonso et al. 2008), contest winner selection (Ong et al. 2017), ranking (Liu and Moitra 2020), etc. Additionally, crowds are frequently engaged in social processes: participatory budgeting (Haltofová 2018), crowd voting and sensing (Jiang et al. 2020), etc. Potential benefits of crowdsourcing include:

1. The collection of crowd votes (opinions) is less costly and more time-effective than the collection of expert opinions.
2. “Wisdom of the crowd” (Keuschnigg and Ganser 2017) may lead to correct and timely solutions, even if experts are unable to induce them (Garcia and Klein 2017).
3. The decisions based on the opinion of a crowd majority may lead to higher social welfare and satisfaction (Aitamurto et al. 2017).
4. The information from a crowd may have a high value in tasks with no “ground truth” (Luther et al. 2015).

The value of collection and aggregation of a large number of opinions has been known for years. Hogarth (1978) showed that, in nominal groups, the occurrence of crowd wisdom is a mathematical fact: with an increasing size, a collection of autonomous judges free of social interaction, apart from some aggregation rule, will almost always be more accurate than the expected value of a random draw from individual opinions. Many kinds of research showed that adding diversity to individual judgments in many cases leads to the cancellation of individual biases through the aggregation of opinions (Keuschnigg and Ganser 2017; Larrick and Soll 2006; Lorenz et al. 2011). Thus, even if error-prone and/or uncertain judges are included in the crowd, the wisdom of crowds exploits the law of large numbers and cancellation of contradictory errors, and can outperform homogeneous expert judgments (Grofman et al. 1983; Hong and Page 2004). In addition to that, many platforms for crowdsourcing data are available and crowd opinions may be collected at a low cost. All this has led to many successful applications for many industry problems like choosing innovative ideas that should be adopted (Ghezzi et al. 2018); giving feedback on creative works (Chen et al. 2020); making recommendations based on users’ critical rating (Isinkaye et al. 2015); stock market predictions (Hong et al. 2016); selecting winners in competitions, etc. Additionally, crowd voting often leads to greater satisfaction and welfare, and thus it is successfully applied in many societal problems like democratic participating in political elections and policymaking (e.g., law regulation Aitamurto et al. 2017); budget allocation (Goel et al. 2019), etc. Crowdsourcing has gained an increased interest and value in the critical fields such as healthcare (Meyer et al. 2016),

especially in the recent COVID-19 pandemics (Desai et al. 2020), where fast answers are needed and the crowd may provide valuable information promptly.

However, the potentials of crowdsourcing in many real world applications may not be fulfilled, especially the ones with a large number of alternatives (i.e., grading or labeling) where the crowdsourcing budget is limited and the number of alternatives is large. Even though votes are crowdsourced (small price per vote), the number of votes per alternative is often small (e.g., 1–3) (Keuschnigg and Ganser 2017). This situation is also found in a real life experiment from this research. A small number of votes per alternative is contradictory to the idea of exploiting “crowd wisdom” having in mind an uncertain quality of crowd votes. Therefore, vote aggregation often leads to wrong conclusions. This problem is well-known and addressed by many research efforts (e.g., Liu and Moitra 2020; Haltofová 2018; Jiang et al. 2020; Keuschnigg and Ganser 2017) by two major approaches: collection of additional votes, and task routing (both with an additional cost).

The collection of additional votes may lead to the reduction of variance and higher quality solutions due to the “crowd wisdom” phenomenon (Singh et al. 2020). However, adding new votes also increases the cost of data collection and may introduce additional biases like bias in ordinal voting systems (Lees and Welty 2019). Moreover, adding additional crowd opinions cannot guarantee an increase in the quality of aggregated decisions, especially in cases when there is no ground truth (Srinivasan and Chander 2019), or in the situations where experts cannot achieve a consensus (Desai et al. 2020). These situations are frequently reflected as uniform or multimodal vote distributions, where aggregations with central tendency measures (even the weighted ones) perform poorly. Additionally, Keuschnigg and Ganser (2017) showed that diversity is the key only in continuous estimation tasks (averaging) and much less important in discrete choice tasks (voting), in which agents’ abilities (expertise) remain crucial and collective decision-making must adapt to the predictive situation at hand.

Task routing demands collection of external data (e.g., labeled data, description of problem, description of voters, questionnaires) and additional (in most cases complex) procedures for correlating external data with voters and tasks at hand (Keuschnigg and Ganser 2017). Many task routing solutions are criticized (Keuschnigg and Ganser 2017) due to strict assumptions: workers are willing to wait patiently for a task to be assigned, or the quality of a worker’s output can be evaluated instantaneously. In many practical applications, these assumptions are not fulfilled and (Keuschnigg and Ganser 2017) states that an ideal task router should be unsupervised since labeling ‘gold’ data is expensive.

Motivated by the aforementioned problems and the idea to develop a method for vote estimation that may be adopted in real world scenarios (simple, cost efficient, and requires minimum or no involvement of the user), we propose the estimation of grades of all voters towards all alternatives by exploitation of ML models. The proposed method is based on matrix factorization and regression. It is able to estimate voter affinities and build predictive models based on sparse voting data. The proposed method allows exploitation of both crowd votes (low cost and high quality uncertainty) and expert votes (high cost and low quality uncertainty). Additionally, the proposed

method is unsupervised in the sense that no additional ('gold') data or complex routing procedures are necessary.

In this research, we also address the problem of bias in vote aggregation process (even if enough votes are available) (Lees and Welty 2019). More specifically, in the cases of uniform (high dispersion of votes) or multimodal vote distribution, central tendency measures that are traditionally used for aggregation lead to the solutions that in many cases are not correct, or do not satisfy the majority (or dense part) (Singh et al. 2020) of voters. Furthermore, in the problems with ordinal votes (which is also the case in our real world experiment), the traditional aggregation methods often lead to indistinguishable ratings (Lees and Welty 2019). This problem is even more emphasized having in mind the uncertainty quality of crowdsourced (non-expert) votes due to task complexity, incompetence, lack of interest, favoritism, manipulation of the crowd (malicious workers) for the problem at hand (Singh et al. 2020; Dodevska et al. 2020).

In order to reduce or avoid the bias of the crowd, often present in traditional aggregation methods such as weighted average (Lees and Welty 2019), we propose a method for aggregation of crowd and expert votes as a bargaining solution. The crowd and expert voters are acting as agents that try to maximize their satisfaction with a final (aggregated) solution (grade, judgment). The basic intuition here is that in the cases of multimodal or uniform crowd vote distributions, aggregation should converge to an expert opinion. In order to achieve this, we contrast crowd and expert voters and exploit Kalai–Smorodinsky (Kalai and Smorodinsky 1975) and Nash (Rachmilevitch 2019) bargaining solutions that are mostly used in game theory. The proposed model is trying to maximize both expert and crowd satisfactions, with respect to the level of agreement (LOA) or homogeneity of opinion within groups. This way, the aggregation of final grades is posed as an optimization problem and implemented within a framework that allows a completely unsupervised modeling of problems with no ground truth, or historical or external data about crowd expertise.

The main contributions of this research are threefold:

- We propose a framework for aggregation of crowd and expert opinions based on bargaining theory and optimization.
- We propose several measures for quantification of voter satisfaction.
- We propose exploiting machine learning (matrix factorization and regression) methods for estimation of the crowd and expert opinions based on a limited number of crowd and expert (sparse) voting data.

It is important to note that the proposed vote estimation and aggregation procedures may be used independently or synergetically.

## 2 Related Work

The popularity of crowdsourcing problems has caused many kinds of research in recent years and recently several review papers have been published on this topic including (Suran et al. 2020; Dodevska et al. 2020). Thus, in this review, we focus only on the papers closest to our research.

In BargCrEx framework, we propose extraction of voter preferences (affinities) based on a limited number of collected votes (sparse data). The extraction of preferences based on sparse data is a well-studied topic with a high impact on many applications and on many algorithms—e.g., Alternating Least Squares (ALS) (Takács and Tikk 2012), autoencoders, Word2Vec (Mikolov et al. 2013), Glove (Pennington et al. 2014), and similar algorithms that showed a cutting edge performance in NLP (Natural Language Processing) problems. However, in this research, we use matrix factorization approach, since it allows a straightforward extraction of both voter preferences and alternative characteristics. In general, different techniques for latent factors may be used, but this comparison is out of the scope of this research.

Luther et al. (2015) analyzed the problem of accessing and exploitation of design critique outside a firm. They provided a piece of evidence that aggregated crowd critique approached the quality of expert critique. Additionally, they showed that the designers who got crowd critique improved their design process and were enthusiastic about the integration of the critique in their designs. The authors Luther et al. (2015) reported that Crowd-Crit used visual support for aggregation of rich critiques data that includes text comments, graphical annotations, valence (positive or negative), and expertise. However, aggregation is based on semi-manual analyses of experts (supported by visualization and drill-down capabilities of Crowd-Crit). Our proposed model is similar to Luther et al. (2015), but allows automation of expert and crowd vote aggregation and allows identification of the situations where aggregated crowd votes may lead to the correct solution even if they are not aligned with expert opinions.

Keuschnigg and Ganser (2017) analyzed the influence of judges' number, ability (expertise), and diversity on the accuracy of aggregate predictions in crowd voting setting. They highlight that the samples of heterogeneous agents outperform the same-sized homogeneous teams of high ability in the case of continuous estimation task (averaging). Additionally, they show that in case of discrete choice tasks (voting), individual abilities remain crucial for the groups with less than 16 members. Still, Keuschnigg and Ganser (2017) emphasize that modeling of the tradeoff between expertise and diversity remains an open problem and highly dependent on a specific application. This was one of the motivations for our research to propose modeling of this tradeoff by employing bargaining solutions from game theory.

The problem of crowdsourcing cost is addressed in Singh et al. (2020) by using the expectation maximization algorithm for estimating the crowd expertise (quality) and the complexity of tasks. Singh et al. (2020) propose a method for allocation of a limited number of crowd voters to the tasks based on crowd expertise and task complexity. The authors observe that the final answer in most crowdsourcing systems is derived as a consensus, and that even simple questions demand a larger sample size if the variance between the answers is high. Compared to Singh et al. (2020), in this research, we try to exploit machine learning models to estimate the opinion of each voter for each alternative based on a limited number of allocated resources (crowd) to specific alternatives. Thus, it is possible to reduce variance in opinions for a fixed cost (number of voters).

Matrix factorization has already been considered in crowdsourcing settings since it naturally models the sparse nature of the voter-task data. In Jung (2014), the

author exploits matrix factorization to estimate the quality of a worker for a specific task. This approach is technically very similar to the proposed estimation procedure. However, the method proposed from Jung (2014) is used for task routing estimation of quality (not votes) and demands external (“gold”) data. Our proposed approach does not demand any additional data and tries to estimate votes while using simple and unsupervised routing methods (i.e., round robin).

Another important problem that needs to be addressed in crowdsourcing platforms is noise in data (Procaccia and Shah 2016; Srinivasan and Chander 2019). They consider the situations where voters are uncertain about their preferences and model uncertain votes as distributions over rankings. The results show that ignoring uncertainty can lead to suboptimal outcomes. Compared to Procaccia and Shah (2016), Srinivasan and Chander (2019), the model proposed in this paper does not consider uncertainty on the level of voter-alternative pairs, and instead, it models uncertainty based on voting deviations within and between expert and crowd groups. Thus, there is no need for the collection of additional data (pairwise comparisons). Additionally, each voter needs to evaluate a small percentage of all the alternatives available.

The integration of crowd and expert votes in crowd sourcing settings is not frequent in literature. The research presented in Snow et al. (2008) analyzes cost effectiveness of non-expert voters for annotation problems and concludes that at least four such voters may be sufficient for high quality labeling results. However, in contrast to our research, they employ experts for the validation of non-expert votes and not for building a model that will estimate and aggregate both expert and non-expert votes.

The idea of integration of crowd and expert votes with a limited number of voters was addressed by Kovacevic et al. (2020a, 2020b) who proposed a framework and several methods for estimation of voter preferences and aggregation by weighting and machine learning methods such as clustering and outlier detection. These methods show a good performance for contest ranking, but they are highly dependent on the number of hyper-parameters (i.e., the number of clusters, distance metric, outlier threshold, etc.) and do not guarantee the generation of solutions on the Pareto front of expert and crowd satisfaction. The aggregation methods based on the optimization procedure proposed in this research contain a single hyper-parameter that may be modeled by the user. Additionally, the modeling of that hyper-parameter may be automatized by applying the bargaining theory that guarantees the Pareto optimality.

The SmartCrowd framework proposed by Bhatt et al. (2019) is also worth mentioning. It allows (1) characterization of the participants by using their social media posts with summary word vectors, (2) clustering of the participants based on these vectors, and (3) sampling of the participants from these clusters, maximizing multiple diversity measures to form final diverse crowds. They show that SmartCrowd generates diverse crowds and that they outperform random crowds. They estimate the diversity based on external data (tweets). In a sense, this research also tries to estimate the diversity of crowds, but concerning both crowd and expert members and without external information.

The problem of bias in aggregation methods in crowd-systems settings with a focus on ordinal restaurant voting was addressed by Lees and Welty (2019). They

show that ordinal rankings (e.g., 1–5) often converge to an indistinguishable rating, since there is a trend in certain cities for the majority of restaurants to have a four-star rating. Based on their research and assumptions, they suggest explicit models for better personalization and more informative ratings. In this research, we try to avoid aggregation bias for ordinal ratings by proposing a model that is guided by the level of agreement between and within crowd and expert groups, and an optimization procedure that is robust to outliers. Thus, averaging bias in aggregation may be avoided.

### 3 Proposed System

In this research, we propose a system that addresses the problem of the lack of data under limited budget assumption, and bias of vote aggregation under uniform or multimodal vote distributions. An important characteristic of the proposed system is that it does not need any additional (external) data. Additionally, it requires minimal or no involvement of users for setting up hyper-parameters.

We assume the existence of  $n_c$  crowd voters,  $n_{cv}$  crowd votes,  $n_e$  of expert voters,  $n_{ev}$  of expert votes,  $n_a$  alternatives (tasks), and  $n_{av}$  of votes per alternative. Further, we assume that the voting data is sparse ( $n_{av} \ll n_c + n_v$ ). In other words, all voters are not evaluating each of the alternatives and this is a typical case in many crowdsourcing scenarios (Jung 2014). Further, we assume that expert votes are more expensive (but with a higher quality) and are under a limited budget assumption  $n_c \gg n_e$  and thus  $n_{cv} \gg n_{ev}$ . In the proposed system, votes (grades) may be ordinal or continuous.

A typical scenario is that the number of crowd votes is uniformly distributed over the alternatives (e.g., each alternative is judged by three different participants). For example, if 1000 alternatives should be evaluated with three votes from different judges, we need 3000 judges with one vote, or 1000 judges where each judge would give three votes. Even if the budget allows the collection of this number of votes, each alternative would still be evaluated by a small number of judges and thus aggregation would be highly susceptible to bias. Additionally, the collection of votes from expert voters is, in most cases, more expensive and it is not possible to collect expert votes for each alternative. This situation is also present in our real world experiment.

General data and process flow (Fig. 1) can be described in the following steps:

1. Collection and aggregation of crowd and expert votes.
2. Estimation of voter preferences and missing votes.
3. Aggregation of votes based on a bargaining based optimization procedure.

In the first step, crowd and expert votes are aggregated (bagged) in a single (sparse) data set. This allows the estimation of preferences (and/or grades) for all alternatives from all voters (both expert and crowd). The main idea of the second step is to estimate votes from all voters towards each alternative. However, the assumed sparse



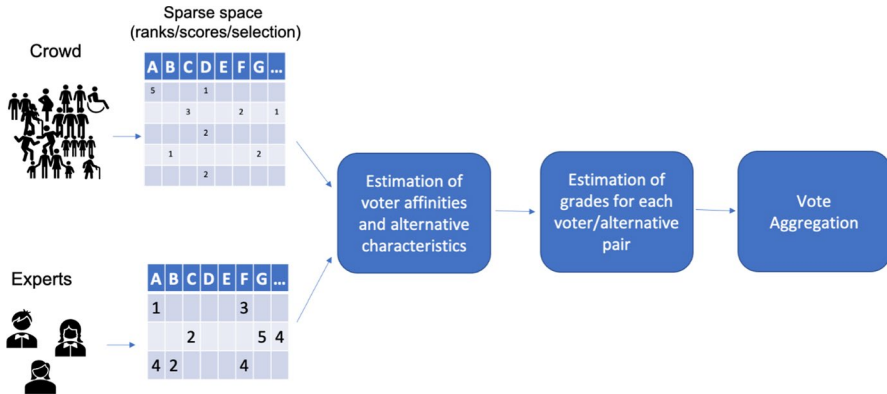


Fig. 1 General data flow of BargCrEx

nature of data does not allow direct application of traditional regression models. Thus, we propose the estimation of voter preferences (as an intermediate step) by exploitation of matrix factorization. Matrix factorization (in this research the ALS algorithm (Takács and Tikk 2012) is utilised) allows efficient estimation of latent factors (voter preferences and alternative characteristics) on sparse data. The collaborative nature of ALS models may exploit crowd votes to estimate votes from experts. For example, if an expert voter has similar (known) votes to a part of the crowd, ALS will assign similar factors (preferences) to them. This way, it is possible to estimate expensive expert votes by exploiting the “wisdom of crowd”. The same works for the crowd. The only difference is that the unknown preferences of the crowd will be mostly estimated based on the similarity to other crowd members (since the assumption is that more crowd votes are available). Even though matrix factorization has showed cutting-edge performance in estimating affinities and ranks in many collaborative (sparse) problems, it does not have good performance in predicting actual values. Thus, we propose a regression layer where votes (grades) are predicted. Factorization and regression layers are used to reduce costs of crowdsourcing and/or infer votes for all voter/alternative pairs under limited budget and, to the best of our knowledge, such procedure has not been used in crowdsourcing settings. This task is very important, since it enables the estimation of expert grades for each alternative as well as an increase in the number and diversity of crowd voters. It is important to note that the described procedure may be used independently from the proposed aggregation method and even without splitting the crowd and expert voters (i.e., estimating only crowd votes and aggregating with traditional procedures). However, since our proposed aggregation procedure assumes the inclusion of (expensive) expert votes, the estimation of such votes for all alternatives may lead to significant cost reductions in the real world applications. It is clear from the description of the estimation procedure that the aggregation task may be done based only on expert voters, since their grades are estimated for each alternative. However, such approach may potentially lead to biased solutions if ignoring the “crowd wisdom”. Additionally, expert voters may not have a consensus and thus no aggregation



of expert votes should be considered as ground truth. The proposed approach also allows aggregation based on high weight of crowd votes if experts are not in consensus. We think that this is especially useful for the problems with no ground truth, e.g., in the best song selection, experts may be biased towards some specific detail (i.e., rhythm, harmony, visual appearance, etc.). Latent features identified by matrix factorization may also be used to derive crowd weights or identify expert lookalikes in the crowd, without the need for collecting historical data. In the experimental section, we use this approach as a benchmark for bargaining-based aggregation.

Assuming that experts' votes are available for all alternatives and that a crowd diversity is achieved, we address the problem of aggregation of voting data. In order to reduce or avoid biases that may be introduced from both expert and crowd members (e.g., experts are biased to some solutions, crowd does not have enough expertise or shows intentional malicious behavior, etc.) and the influence of outliers in the aggregated solution, we propose an aggregation based on bargaining solution between the experts and the crowd. This is achieved by contrasting experts and crowd and assigning them bargaining power (weights) proportional to their vote homogeneity (level of agreement). The bargaining power (weights) are further used in the optimization process, where the total (expert and crowd) satisfactions is maximized. The described procedure avoids some problems of traditional aggregation schemes based on weighting votes (or aggregation based solely on central tendency measures). For example, if expert votes have much higher weights compared to crowd votes (which is reasonable to expect), but they have completely different opinions (votes), it should be reasonable to expect that one of them has the correct answer. However, traditional aggregation procedures will tend to average their votes (leading to the incorrect answer), and high weight of expert voters will prevent them from exploiting a part of crowd votes that may be grouped around the correct answer. The optimization of total satisfaction (instead of weighting within group aggregates) allows modeling of such situations even though the assigned weights are the same within groups. Practically, this allows the procedure to find solutions with respect to different densities within groups (multimodal distributions) that would not be possible with within-group aggregation. Different cases that elaborate this basic intuition and define expected aggregation outcome of a method are described in Table 1 and experimental evidence is described in Fig. 3.

### 3.1 Estimation of Preferences and Grades

For the estimation of grades that are not collected through the crowdsourcing procedure we propose a two-step process:

1. Estimation of latent variables (embeddings) based on sparse voting data, and
2. Estimation of grades for each voter-alternative pair.

In the first step, we propose a matrix factorization approach for identification of latent features of both voters (preferences or affinities) and alternatives (latent characteristics of features) without the need for inclusion of external or historical data.

**Table 1** Defined level of agreement between groups

Case	Expert (within) LOA	Crowd (within) LOA	Joint (total) LOA	Description
1	High	High	High	All participants have similar opinions ES: High, CS: High, OS: High
2	High	High	Low	Opinions within groups are similar but not between groups. The optimal grade cannot satisfy both groups. ES: High or Low CS: High or Low, OS: Medium
3	High	Low	High	Not possible (one group has a high within disagreement)
4	High	Low	Low	Expert opinion should be supported ES: High, CS: Low, OS: Medium
5	Low	High	High	Not possible (one group has a high within disagreement)
6	Low	High	Low	Crowd opinion should be supported: ES: Low, CS: High, OS: Medium
7	Low	Low	High	Not possible (both groups have high within disagreement cannot lead to high satisfaction)
8	Low	Low	Low	All opinions are different. No solution is trustworthy? ES: Low CS: Low, OS: Low

Matrix factorization assumes that each user can be described with  $k$  attributes (factors), and each alternative can be described with an analogous set of  $k$  attributes (factors). Thus, it is possible to use identified latent features as predictors for estimation of unknown grades as well as for the definition of similarity notion within and between voters and alternatives. Enabling similarity calculation between voters allows the assignment of weights to experts and this approach was used as one of the benchmarks in the experimental section. In this research, we used Alternating Least Squares (ALS) (Takács and Tikk 2012) for matrix factorization. Similarly, to other factorization models, ALS model may be represented as (1):

$$\hat{r}_{va} = x_v^T \cdot y_a = \sum_k x_{vk} y_{ak} \quad (1)$$

where  $\hat{r}_{va}$  represents an estimation of the grade for voter-alternative pair  $r_{va}$ ,  $x_v^T$  represents voter factors (voter affinities), and  $y_a$  represents alternative factors (characteristics). Voter and alternative factors are represented as low dimensional embeddings with a dimension of  $k$ , where  $k$  is optimized as a hyper-parameter of ALS. Optimization of the ALS model was done by minimization of the loss function presented in (2).

$$L = \sum_{v,a \in D} (r_{va} - x_v^T \cdot y_a)^2 + \lambda_x \sum_v \|x_v\|^2 + \lambda_y \sum_a \|y_a\|^2 \quad (2)$$

This loss function is based on the minimization of the square of differences between real and estimated grades. To prevent overfitting, ridge regularization terms and hyper-parameters ( $\lambda_x$  and  $\lambda_y$ ) are used. ALS algorithm was selected because of its cutting edge performance in terms of ranking quality, but also because of its scalability that enables big data processing.

However, direct application of matrix factorization model (1) for the estimation of grades for each voter-alternative pair is not convenient. This is because matrix multiplication results in a similarity score (cosine similarity between voter and alternative factors) that is not bounded by a minimum and maximum possible grades. Thus, after the identification of factors, we add a predictive layer (second step of the estimation approach). In the predictive layer, we use factors (embeddings) as predictors (features) for the estimation of grades for each voter-alternative pair. Since voter grade judgments are assumed to be ordered and, in general, on different scales (i.e., 1-5, 1-10, etc.) we use regression models for the estimation of the unknown grades. It would be possible to pose this problem as a classification task, but that would result in loss of information due to discrete outcomes, which would further propagate errors in the aggregation phase. Additionally, due to a multilevel scale, the classification problem would have to be posed as a multiclass one. That would add unnecessary complexity in learning and evaluation of the estimation models. Thus, in our experimental section, we employ and compare the performance of popular regression models, such as Random Forests (Breiman 2001) and Gradient Boosting trees (Friedman 2002) that showed outstanding performance in many application areas including crowdsourcing (Semanjski and Gautama 2015; Sutton et al. 2019) and represent the industry standard. The estimation of voter affinities with ALS, as well as building regression models need hyper-parameter tuning. However, both ALS and regression problems are guided by error minimization (difference model estimation and existing votes), and so  $\lambda_x$  and  $\lambda_y$  and number of factors in ALS (as well as hyper-parameters if regression algorithms) may be automated with hyper-parameter tuning methods such as: grid search, random search (Bergstra and Bengio 2012), or metaheuristic-based search (Chan et al. 2013). We used grid search for the automation of hyper-parameter search. Even though matrix factorization is very efficient in modeling sparse data, it is important to emphasize some limitations of this approach. First, ALS models have the “cold-start” problem, meaning that each alternative has to be evaluated by at least one voter and each voter has to evaluate at least one alternative. This has to be considered in the calculation of the necessary budget. Further, even without the “cold start” problem, the level of sparsity as well as the number of alternatives and voters influence the model quality. Regarding maximum sparsity estimation, it complies with standard recommendations for ALS. In particular, sparsity should be less than 95% for most problems (Idrissi and Zellou 2020). Additionally, an extensive study by Jung (2014) evaluated matrix factorization models against matrix size, sparsity, and similarity of tasks and showed that a satisfactory solution is achieved when matrix density is only 10%. Additionally,

it showed that significant gains in matrix reconstruction are achieved when matrix density is around 30% for the tasks with medium similarity and high correlation.

### 3.2 Vote Aggregation by Bargaining Optimization

After estimating both expert and crowd preferences, we assume that diversity of crowd is achieved and expert grades for each alternative are estimated. In this phase, only expert votes (true and/or estimated) may be adopted for aggregation, since they are considered as the informed decision-makers. However, in many cases, experts do not mutually agree on a decision and thus aggregations should not be considered as the grounded truth. Additionally, if only expert votes are considered for making the final decision, the crowd opinion may be completely disregarded and thus satisfaction of the crowd may be low (this is very important in the social application of crowdsourcing). Moreover, disregarding crowd votes may lead to ignorance of crowd wisdom and lead towards incorrect decisions. The same applies if only crowd votes are used. More specifically, crowd votes may have, and most often do have, high variance and cannot be used as the ground truth. Also, if experts' votes are omitted, experts' satisfaction may be low.

We introduced the notions of satisfaction and level of agreement (LOA) in order to develop an aggregation method that avoids the aforementioned problems for the situations where weights of voters are not known. In this research, we model satisfaction as a cumulative distance of votes from an aggregated solution. Satisfaction is modeled for the experts (ES), the crowd (CS), and the overall satisfaction (OS) that represents an aggregation of ES and CS (i.e., sum or product).

Further, we introduce notions within group and the total level of agreement (LOA). Within group, LOA provides an upper bound on the maximum satisfaction of each group and may be represented with some notion of variance within a single group (e.g. standard deviation, cumulative distance from the median). In this research, we use the average absolute distance from the median of the group. Total LOA may be defined as the distance from all grades to the global median. However, total LOA is not explicitly calculated and serves just for the definition of possible outcomes and explanation of the proposed method's intuition.

The main intuition for the aggregation of the final grade is the following: if experts have a high LOA, then experts' opinion may be considered as the "ground truth" and the final grade (aggregation) should lead to a high experts' satisfaction (ES) regardless of the crowd opinion. In contrast, if experts have a low LOA, but the crowd has a high LOA, then it is possible that the crowd "uncovered the ground truth" and the aggregated grade should lead to a high crowd satisfaction, regardless of the expert satisfaction. In both cases (high crowd satisfaction or high expert satisfaction), the overall (total) satisfaction (OS) may be high if there is a high LOA between the crowd and the expert groups. To explain the intuition of the proposed approach, we simplified the LOA to have only two values. More specifically, high LOA signifies a low variance of group votes, and a low LOA signifies a high variance of group votes. In Table 1, possible scenarios with respect to different combinations of within group and total LOA are enumerated

and explained. This enumeration is used to estimate the expected outcomes and desirable solutions of the proposed aggregation method (or any other) as well as the “sanity check” of the results of the described experiments.

It can be seen from Table 1 that Case 1 and Case 8 represent the extreme cases where, in Case 1, all participants are in agreement (within groups and total) or in disagreement, and that leads to extreme values of ES, CS, and OS. These situations are trivial, and any aggregation method would yield a satisfactory solution.

Case 2 describes a situation where agreement exists within groups, but without a high level of total agreement. In these situations, either crowd or expert group can be satisfied. Thus, the final decision should support one of the groups while disregarding the other (i.e., final decision should lead to a high expert satisfaction). One can adopt that experts are more informed about the problem at hand and select the solution obtained solely from the experts. This will result in the low crowds’ satisfaction, but it will yield a solution that experts agree upon.

Cases 4 and 6 represent the situations where one of the groups has a high LOA, but there is no high level of the total agreement. In this case, the opinion of the group with a high LOA should be supported to avoid degradation of OS through averaging. More specifically, if experts have an agreement on the problem at hand, while the crowd does not, then the aggregation procedure should result in the solution that experts agree on. This solution will satisfy experts’ opinions and partially the opinions of the crowd. Therefore, the overall satisfaction will be high. If another solution is selected, it would lower the satisfaction of the experts, while not increasing the satisfaction of the crowd. Mathematically, the relative satisfaction gain of expert group and crowd group is different. While the satisfaction of the crowd is expected to be close to constant (due to the low level of agreement – a high variance in votes), the satisfaction of the experts will decrease rapidly if another solution is selected.

Case 8 represents a situation where LOA within groups is low as well as the total LOA. This scenario represents a situation where votes are uniformly or multimodally distributed in both groups. In this case, it is not possible to achieve high overall satisfaction. In the case of a uniform distribution of both groups, high satisfaction cannot be expected for any subgroup of voters. However, in the case of multimodality, the final decision should satisfy the part of the subgroup that is dense (i.e., the supporting part of the crowd and one expert), while disregarding the others.

To implement the intuition described in Table 1, we propose the following general model for vote aggregation based on ES and CS defined in (3).

$$\max[\lambda * ES(x) + (1 - \lambda) * CS(x)] \quad (3)$$

where *ES* and *CS* represent expert and crowd satisfaction (respectively) that depend on the grade *x*. The value  $\lambda$  represents the parameter that models the importance of crowds’ and experts’ votes concerning their agreement. The main idea is that  $\lambda$  reduces the impact of a single group (expert or crowd) if there is no within-group agreement. This way, it is possible that the final solution satisfies the group with a high LOA and avoids distortion in the aggregation phase caused by the group

with a low LOA. Therefore, it is possible to recognize the situations where the crowd “unveils the true solution even if experts do not agree”. Moreover, the final solution may support one or several experts that are supported by crowd opinion.

In general, expert and crowd satisfaction may be modeled with any similarity or distance metric. Since we assume a different number of crowd and expert voters, we propose a model that measures corresponding disagreements as an average value of individual disagreements from specific solutions to average out the difference in the number between expert and crowd groups as shown in (4).

$$\max \left[ \lambda * \frac{1}{n_e} \sum_{i=1}^{n_e} s(e_i, x) + (1 - \lambda) \frac{1}{n_c} \sum_{j=1}^{n_c} s(c_j, x) \right] \tag{4}$$

where  $s$  represents an arbitrary similarity metric,  $n_e$  the number of expert members,  $n_c$  the number of crowd members, and  $x$  the proposed grade. To employ a distance metric instead of similarity and keep a notion of similarity (and maximization objective), the objective function may be defined as (5):

$$\max \lambda \left( x_{max} - \frac{1}{n_e} \sum_{i=1}^{n_e} d(e_i, x) \right) + (1 - \lambda) \left( x_{max} - \frac{1}{n_c} \sum_{j=1}^{n_c} d(c_j, x) \right) \tag{5}$$

where  $x_{max}$  represents the best possible grade. Even though any distance (or similarity) metric may be used, outlier resilience is a desirable property for modeling the satisfaction of the majority. To satisfy this property, we propose measuring satisfaction as the distance between the maximum possible grade (maximum possible satisfaction) and the average absolute distance from the proposed grade. Absolute distance ( $L_1$ , Manhattan distance) as presented in (6) is less susceptible to the influence of outliers compared to exponentiated distances (i.e., Euclidean distance).

$$\max \lambda \left( x_{max} - \frac{1}{n_e} \sum_{i=1}^{n_e} |x - e_i| \right) + (1 - \lambda) \left( x_{max} - \frac{1}{n_c} \sum_{j=1}^{n_c} |x - c_j| \right) \tag{6}$$

However, due to the presence of absolute values in (6), we transform it into a constrained problem that can be solved with standard linear optimization methods represented in (7).

$$\max \lambda * \frac{1}{n_e} \sum_{i=1}^{n_e} E_i + (1 - \lambda) \frac{1}{n_c} \sum_{j=1}^{n_c} C_j \tag{7}$$

$$\begin{aligned} & s.t \\ & x - E_i \geq e_i, \quad i = 1, \dots, n_e \\ & -x - E_i \geq -e_i, \quad i = 1, \dots, n_e \\ & x - C_j \geq c_j, \quad j = 1, \dots, n_c \\ & -x - C_j \geq -c_j, \quad j = 1, \dots, n_c \end{aligned}$$

Based on the goal function from (7) it is clear that by changing  $\lambda$  parameter it is possible to generate the Pareto front of the solutions that model the tradeoff between ES and CS. However, the analysis of the Pareto front and manual selection of the final solution is a cumbersome and time-consuming task. To automate the selection of the final solution, based on intuition presented in Table 1, we propose the modeling of ES and CS tradeoff based on bargaining the problems that are used in game theory. In this research, we adapted Kalai-Smordinski and Nash bargaining solutions to the problem of aggregation of the crowd and expert votes (Thomson 1994; Samuelson 2016).

Kalai-Smordinski bargaining solution (Kalai and Smorodinsky 1975) selects such a solution that equalizes the ratios of maximal gains between two players (in this case expert and crowd groups). In other words, the Kalai-Smordinski solution for  $\lambda$  is optimal in terms that the unit change of  $\lambda$  will lead to a smaller increase of relative satisfaction of one group, compared to the reduction of relative satisfaction of another group. If we recall the scenarios 4 and 6 where one group has a high LOA, while the other does not, then the Kalai-Smorodinsky bargaining solution would yield a solution for  $\lambda$  on the Pareto front that will be closer to the group having high LOA. In other words, for the Kalai-Smorodinsky solution, the value  $\lambda$  is interpreted as the point of agreement between the two groups that equalize relative gains in the groups' satisfaction. Thus, this approach will satisfy the intuition presented in the previous section. Kalai-Smorodinsky solution has additional properties that are of interest to the problem at hand. First, this approach works if experts and crowds are switched due to symmetry property. This means that the group ordering and the order of votes inside a group are irrelevant, and regardless of how the groups and the votes are ordered, the solution would be the same. Then, the Kalai-Smorodinsky solution is scale-invariant. Since gain presents a relative measure of performance (in this case satisfaction), the process of obtaining the solution is independent if the ranges of satisfaction for one group are different from the other. Satisfaction is normalized to the best possible satisfaction a group can obtain, thus making the best possible satisfaction equal to one. However, if an additional group (e.g., experts in another field) is added, then the resulting solution will differ greatly. More specifically, due to the process of equalization of gains, the new group will interfere with the gains between any other two groups.

Mathematically, for the case finding  $\lambda$  in the context of the tradeoff between CS and ES Kalai-Smorodinsky bargaining solution may be defined as (8).

$$\frac{ES_1 \lambda + ES_2(1 - \lambda)}{\max(ES_1, ES_2)} = \frac{CS_1 \lambda + CS_2(1 - \lambda)}{\max(CS_1, CS_2)} \quad (8)$$

where  $ES_1$  represents Experts satisfaction in case of  $\lambda = 1$ ,  $ES_2$  represents Experts satisfaction in case of  $\lambda = 0$ ,  $CS_1$  represents Crowd satisfaction in case of  $\lambda = 1$ ,  $CS_2$  represents Crowd satisfaction in case of  $\lambda = 0$

This problem may be solved analytically and the solution for  $\lambda$  may be represented in a convenient form (9):



$$\lambda = \frac{\max(ES_1, ES_2)CS_2 - \max(CS_1, CS_2)ES_2}{\max(CS_1, CS_2)(ES_1 - ES_2) - \max(ES_1, ES_2)(CS_1, CS_2)} \tag{9}$$

It is interesting to note that the Kalai-Smordinski bargaining problem is not dependent on the goal function, but only on the maximum and minimum CS and ES values. This leads to an efficient linear time calculation of both  $\lambda$  and the optimization problem (7). Our experimental evaluation showed that Kalai-Smordinsky based optimization of  $\lambda$  leads to satisfactory results that fulfill intuition presented in Table 1. Another bargaining solution that we analyzed was the Nash bargaining solution (Van Damme 1986; Rachmilevitch 2019). From all Pareto solutions of (7), the Nash bargaining solution is the one that maximizes the area (product) between experts' and crowd's satisfaction. In general, the Nash bargaining solution may be represented by the formula (10):

$$\max \prod_{j=1}^m \lambda_j f_s(x, g_s_j) \tag{10}$$

where  $f_s$ —aggregated satisfaction (expert and crowd),  $m$ —number of groups,  $\lambda_j$ —group weight (in this case, it is  $\lambda$  and  $1 - \lambda$  for the experts and the crowd, respectively),  $g_s$ —group satisfaction (expert or crowd).

To be able to solve this problem with linear optimization procedures, this objective may be more conveniently written as (11):

$$\max \sum_{j=1}^m \ln(\lambda_j + \ln(f_s(x, g_s_j))) \tag{11}$$

In terms of ES and CS, tradeoff may be written as (12):

$$\begin{aligned} \max \ln(\lambda) + \ln \left( ES_{max} - \frac{1}{n_e} \sum_{i=1}^{n_e} |x - e_i| \right) \\ + \ln(1 - \lambda) + \ln \left( CS_{max} - \frac{1}{n_c} \sum_{j=1}^{n_c} |x - c_j| \right) \end{aligned} \tag{12}$$

where  $ES_{max}$  represents the maximum possible ES (that depends on the scale of possible grades) and  $CS_{max}$  maximum crowd satisfaction.

The problem stated in (7) is convex and may be efficiently solved by simultaneous optimization of  $\lambda$  and  $x$ . Since the goal function seeks the greatest area of both experts' and crowd's satisfaction, it can be deduced that  $\lambda$  should be equal to 0.5. Any other solution would yield a lower satisfaction area. Similar to the Kalai-Smorodinsky bargaining solution, the Nash bargaining solution yields a Pareto optimal solution. Also, the ordering of the groups and the votes inside a group is irrelevant (as the solution will be the same). Additionally, the Nash bargaining solution has one benefit compared to the Kalai-Smorodinsky solution. Due to the multiplication of the group satisfactions, the addition of another group would not interfere with the satisfaction of the other groups.

A Nash bargaining solution presents an egalitarian justice point of view on the resource allocation problem. In this case, the allocated resource is the satisfaction of a group with the aim to provide a fair allocation of satisfaction. In other words, the vote that will be equally just to both the experts and the crowd would be chosen. However, providing an equally just solution is the result of averaging (Nash bargaining solution is similar to the geometric mean of the votes), thus it will lower the satisfaction of one group even if that group has a high LOA leading to not fulfilling the intuitions presented in cases 4 and 6.

#### 4 Measures for Model Selection and Evaluation

In order to evaluate bargaining-based and benchmark models, we propose several evaluation (model selection) measures based on the notions of Expert satisfaction (ES) and Crowd satisfaction (CS). We define ES and CS with the following formulas described in (13).

$$ES = x_{max} - \frac{1}{n_e} \sum_{i=1}^{n_e} |x - e_i|; \quad CS = x_{max} - \frac{1}{n_c} \sum_{j=1}^{n_c} |x - c_j| \quad (13)$$

where  $x$ —proposed grade,  $x_{max}$ —maximum possible grade on voting scale,  $e_i$ —expert vote,  $c_j$ —crowd vote.

Based on the formulas, it is evident that both ES and CS have an upper boundary equal to the maximal possible grade.

Further, we propose measuring a cumulative satisfaction of crowd and expert voters as the sum and the product (area) between the groups (14):

$$\begin{aligned} S_{sum} &= ES + CS \\ S_{area} &= ES * CS \end{aligned} \quad (14)$$

In order to make satisfaction measures comparable between different problems, we define relative satisfaction as (15):

$$\begin{aligned} S_{rel\_sum} &= \frac{ES}{ES^*} + \frac{CS}{CS^*} \\ S_{rel\_area} &= \frac{ES}{ES^*} * \frac{CS}{CS^*} \end{aligned} \quad (15)$$

where  $ES^*$  and  $CS^*$  stand for the maximum possible satisfactions of expert and crowd groups, respectively. The maximum possible satisfaction of a group is achieved if the aggregated (final) grade minimizes the cumulative distance from the final grade within a group. Specifically, in the proposed model the maximum group satisfaction is achieved in the median of votes of the group. So maximum group satisfactions may be represented as (16):

$$\begin{aligned}
 ES^* &= x_{max} - \frac{1}{n_e} \sum_{i=1}^{n_e} |x_e^* - e_i| \\
 CS^* &= x_{max} - \frac{1}{n_c} \sum_{j=1}^{n_c} |x_c^* - c_j|
 \end{aligned}
 \tag{16}$$

where  $x_e^*$ —optimal grade for expert group,  $x_c^*$ —optimal grade for crowd group.

It is trivial to see from (15) that the relative sum satisfaction is bounded between 0 and 2, while the relative area satisfaction is bounded between 0 and 1. For both measures, the maximum values are achieved if the medians of both groups are the same. Even though the proposed measures give a composite notion of satisfaction, and are bounded and allow the comparison between different problems, they do not show an obvious relationship between the achieved satisfaction of groups and their heterogeneity/homogeneity. This is important for measuring the achievement of intuition described in Table 1; the group with a higher homogeneity should achieve higher satisfaction. In terms of bargaining theory, the group with a better negotiation position should achieve a higher gain. In order to measure this intuition, we propose the following measure of Negotiation Position Gain (NPG) (17):

$$NPG = \frac{S_k}{S_k^*} - \frac{S_{-k}}{S_{-k}^*}, \text{ where } k = \operatorname{argmax}(ES_{max}, CS_{max})
 \tag{17}$$

NPG shows the difference between the satisfaction of the group with a better negotiation position and the group with a worse negotiation position. If the NPG is positive, it means that the group with a better negotiation position achieved a higher satisfaction. It is important to note that NPG cannot be used as a standalone measure for model quality because it ignores the overall satisfaction and leads to extreme solutions where the group with a better negotiation position achieves the maximum satisfaction, while the other group's satisfaction is neglected. Thus, this measure should be used in addition to composite measures (sum and area) and this will be shown in the experimental section.

## 5 Experimental Evaluation

### 5.1 Data

In this research, we used 2 real-world datasets for the evaluation of the proposed approach. The datasets were collected from Credibility Coalition (<https://credibilitycoalition.org/>) and contain crowd and expert grades about article relevance. In the first dataset, the experts are scientists, while in the other, the experts are journalists. Both datasets contain 2472 environment/climate-related articles created in 2019. A smaller sample of articles is evaluated by all experts and the majority of voters. The rest of the articles are mostly annotated by 3 crowd raters each. These datasets are publicly available and can be found in <https://data.world/credibilitycoal>

[tion/credibility-factors2020/](#). As a part of crowd evaluation, each rater filled out a demographic survey followed by committing to an Annotator Code of Conduct of performing their duties in as accurate and diligent a manner as possible provided in their informed consent. Once qualified, the selected crowd received reading and rating tasks which included their credibility perception per article on a 5-point Likert scale, ranging from very low (1) to very high (5). The instructions to fill out the seven-question survey across a 7–10 day period (estimated at 10 h) were provided in a handbook with a recommended limit of 10–15 min per article. It is important to note that all rated articles were written in English and represented a range of liberal to conservative positions or attitudes towards climate problems. To collect articles, the team used the Buzzsumo social media research tool in late 2018 in order to find the most popular articles over the previous year with the keywords of “climate change,” “global warming,” “environment,” and “pollution.” Additionally, we created several artificial datasets in order to inspect whether the proposed aggregation methods followed the intuitions described and enumerated in Table 1. Each of the datasets represented three hypothetical experts and 100 crowd members (since we assumed a larger number of crowd voters in crowd voting setting). Each participant was represented with a grade in 1–10 interval in order to emphasize the agreement and disagreement between these two groups. Detailed visualizations of artificial datasets and solutions provided by aggregation methods are presented in the results section.

## 5.2 Experimental Setup

The experiments were divided into two parts. The first - for the evaluation of the grade estimation methodology based on latent factors and predictive models, and the second—for the evaluation of the aggregation methods. The proposed framework was implemented in Python programming language with the use of python's machine learning and optimization stack: *numpy* for linear algebra calculations, *SciPy* for optimization, *sklearn* for building machine learning models, and *matplotlib* and *seaborn* for visualization. The experiments were conducted on a quad core processor and 16Gb of RAM.

### 5.2.1 Estimation Method

In order to find the latent factors (embeddings) of voters and alternative spaces, and consequently define the similarities between voters, we utilized Alternating Least Squares (ALS) that we evaluated based on mean squared error (MSE), since it reduces the influence of outliers. This is important because the models for vote aggregation tend to adapt the final (aggregated) solution to the majority of participants, and thus it is more important to get accurate predictions for the homogenous participants that represent the majority. Expert and crowd votes were used together and a part of their votes was masked and used for measuring

error on the test set. Several hyper-parameters were optimized in order to minimize errors on the test data. Grid search of parameters is shown in Table 2. After obtaining latent factors for participants (voters), we created a dataset for learning the regression model that predicts unknown grades (votes). The predictors (input features) are constructed on the level of participant-item. So, each instance of the dataset consists of participant embeddings, item embeddings, and outcome (grade) variable. The dimension of such dataset depends on the number of expert voters with the dimension size showed in (18).

$$N_d = ((n_c + n_e) * n_a) \times (2 * n_f) \quad (18)$$

where  $n_c$  represents the number of crowd voters in a dataset,  $n_e$  represents the number of experts in a dataset,  $n_a$  represents the number of all alternatives given in a dataset,  $n_f$  represents represent the number of latent (hidden) features obtained by ALS algorithm.

For learning regression models, the data was divided into train and test sets with 70:30 ratio. ALS hyperparameters were tuned by applying 10-fold cross-validation on the training data. All models were evaluated by MAE because of the reasons described above. In Table 3 the utilized regression algorithms and the hyperparameter ranges are shown. After selecting the best regression model, the predictions are made for each unknown voter-alternative pair and a new dataset is formed with alternatives as rows, and voter grades as columns. This dataset was used for the evaluation of the proposed aggregation models and benchmarks.

### 5.2.2 Aggregation Methods

Based on the datasets provided from the described estimation process, we evaluated bargaining-based aggregation methods, against traditional aggregation methods: majority vote, mean, and median. We measured the traditional aggregation method performance on global (on the whole dataset) and local (within group) levels. Additionally, as a benchmark, we used weighted average methods proposed by Kovacevic et al. (2020b) based on the similarity of crowd members with experts, resulting in 10 benchmark aggregation methods described in Table 4. The aggregations based on mean, median, and majority aggregation from Table 4 are self-explanatory. The weighted aggregation procedures are defined as follows: based on embedded features, for each crowd voter, a weight is assigned with respect to the closest expert (higher weights are assigned to the crowd members that are close to experts). After weight assignment, the weighted average is used for the calculation of final grades. In case of Weighted (crowd), weighted crowd grades are used, and in case of Weighted (crowd and expert), both crowd and expert weighted opinions are used.

**Table 2** Hyperparameter grid search optimization of ALS

Parameter	Hyperparameter range
Number of latent factors	[20, 30, 40, 50, 70, 100]
Regularizations	[0., 0.1, 0.3, 0.5, 0.7, 1., 10., 100.]

**Table 3** Hyperparameter grid search optimization of regression algorithms

Algorithm	Parameters	Values
Linear regression	–	–
Random forest	n_estimators	[1, 3, 5, 7, 10, 50, 100, 200, 500]
	max_depth	[5, 10, 15]
Gradient boosted regressor	n_estimators	[1, 3, 5, 7, 10, 50, 100, 200, 500]
	max_depth	[1, 3, 5, 7, 10, 15, 30]
	llearning_rate	[0.01, 0.1, 0.05, 0.25, 0.5, 1]

It is important to note that, in most cases, global benchmarks represent crowd opinion because of the assumption that the number of crowd participants is much larger compared to the number of expert participants. On the other hand, in some cases, it is possible to have a similar number of expert and crowd votes, and global benchmarks may show additional information about model performance.

## 6 Results and Discussion

Following the described experimental setup, we report the results for estimation and aggregation methods separately, since in general, these methods may be used independently from each other.

### 6.1 Estimation Results

The results of ALS algorithm for finding embeddings for both Journal and Scientist datasets are reported in Table 5.

**Table 4** Description of benchmark aggregation models

Majority (global)	Majority vote of all voters (experts and crowd)
Median (global)	Median vote of all voters (experts and crowd)
Mean (global)	Mean vote of all voters (experts and crowd)
Majority (expert)	Majority vote of expert voters (experts and crowd)
Median (expert)	Median vote of expert voters (experts and crowd)
Mean (expert)	Mean vote of expert voters (experts and crowd)
Majority (crowd)	Majority vote of expert voters (experts and crowd)
Median (crowd)	Median vote of expert voters (experts and crowd)
Mean (crowd)	Mean vote of expert voters (experts and crowd)
Weighted (crowd)	Weighted average sum of crowd votes
Weighted (crowd and experts)	Weighted average of crowd and expert votes

**Table 5** Results of ALS algorithm for data embedding

Data set	N_factors	Regularization	Train MSE	Test MSE
Journal	50	0	0.65	0.89
Science	50	0	0.64	0.88

**Table 6** Estimation results for Science dataset

DataSet	Model	best_params	Train MSE	Test MSE
Science	LinearRegression	{}	0.764	0.952
Science	RandomForest	{'max_depth': 15, 'n_estimators': 500}	0.549	<b>0.574</b>
Science	GradientBoostingRegressor	{'learning_rate': 0.05, 'max_depth': 7, 'n_estimators': 200}	0.546	0.591

Minimal MSE on test set is showed in bold font

The results from Table 5 show that the optimal number of embedding features was 50 for both datasets, while the regularization parameter was 0. Since regularization was not used, we assume that the selection of the reduced number of embedding features (50 out of maximum 100) resulted in a model with a decreased complexity and thus additional regularization was not needed. As described in the methodology section, MSE resulted from the matrix factorization model is effective for model selection and alternative ranking, but it cannot be easily interpreted as prediction error. In this case, interpretation is not important, since embeddings are used as an input for regression models. Additionally, the differences between the train and test sets are reasonably small.

The performance of optimized predictive (regression) models based on feature embeddings and their corresponding hyper-parameters are reported in Table 6 for Science data set and in Table 7. It can be seen that MSE on the test dataset was ~ 0.575. Since the range of possible grades was 1-5, we argue that this error is on a satisfactory level, since it is on the level of 10–15% of the range.

## 6.2 Aggregation Results

In order to inspect the compliance of proposed (bargaining) aggregation methods with the intuitions described in Table 1, we generated two groups of artificial datasets. The first group are the datasets with extreme cases that are described in Fig. 2, and the second one (shown in Fig. 3) are less extreme cases. In the upper part of Fig. 2 the distributions of grades for each case are represented by boxplots of crowd and expert groups. For the sake of clarity, in the lower part, distributions are represented by swarm plots (all instances). In case 1, both experts and crowd agree that the grade should be 1. In case 2, there is no variance within groups, but votes between the groups are completely opposite (crowd votes for grade 10 and experts for grade 1). In case 4, experts agree on the grade 1 (without variance), but crowd has a high variance (grades between 1 and 10). Case 1 is trivial (no disagreement

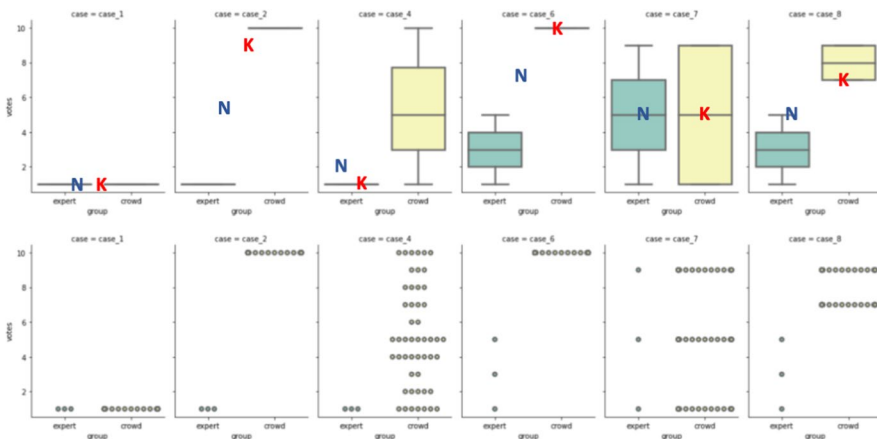


**Table 7** Estimation results for Journal dataset

DataSet	Model	best_params	Train MSE	Test MSE
Journal	LinearRegression	{}	0.760	0.958
Journal	RandomForest	{'max_depth': 15, 'n_estimators': 200}	0.542	<b>0.576</b>
Journal	GradientBoostingRegressor	{'learning_rate': 0.05, 'max_depth': 5, 'n_estimators': 500}	0.559	0.598

Minimal MSE on test set is showed in bold font

within, or between groups) and both methods provided the same solution 1. In Case 2, there is no within group variance, but between group variance is the maximum possible. In this case, the Nash solution (5.5) is between the two groups, while Kalai solution (9.1) favored one group (in this case - experts). However, this is an extreme case and in real-life usage, in the cases of no within variance and maximum between variance, a user should define the default group (i.e., experts) or to alternate the lambda parameter of the model. In Case 4, there is a high LOA within experts and a low LOA within the crowd. Additionally, the crowd grades are spread over the complete space of possible grades. In this case, Kalai-Smordisky matched exactly the expert solution (as desired by the intuition described in Table 1, while Nash provided a solution that is moved towards the crowd median, which is not a desired behavior when one group has such a low LOA. This situation is even more pronounced in Case 6 where one group (crowd) has a maximal LOA, and the other group has a low LOA and an opposite opinion. Here Nash aggregation method found grade 6.5 that is in between the opposite ends, but does not satisfy either group. In contrast, Kalai provided the solution 10 that satisfies the group with high LOA (in this case - crowd). Case 7 is trivial, since both groups have a low LOA and grades are spread over the complete space. Case 8 represents a situation where groups have opposite opinions, but do not have maximal LOA. In this case, the crowd LOA is a bit smaller compared to the expert LOA and Kalai-Smordinsky provided a solution



**Fig. 2** Artificial “extreme” cases

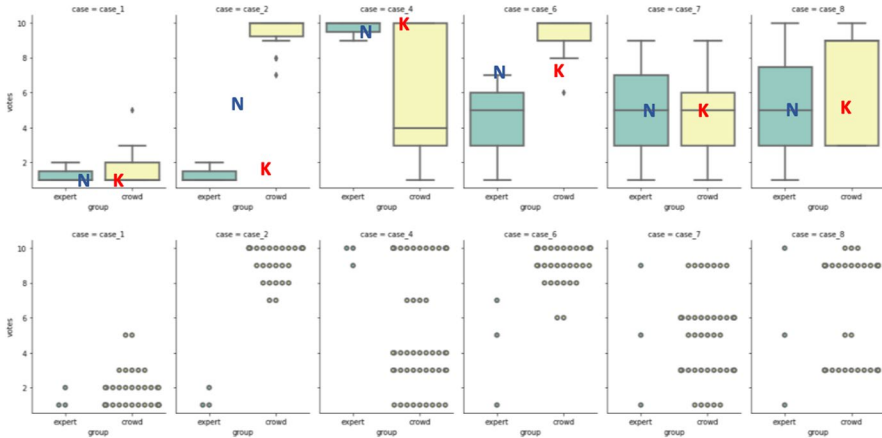


Fig. 3 Artificial “less extreme” cases

that represents the lowest grade of the crowd and this solution is in compliance with the intuition represented in Table 1. In the dataset with less extreme cases (Fig. 3), the results are similar compared to the extreme cases. In Case 1, groups have a high within and between LOA and both solutions agree on the solution. In Case 2, Kalai-Smordinsky solution favored the expert opinion, since it has a bit less variance compared to the crowd. On the other hand, the Nash solution is between two distributions. In the last 3 cases, both methods agree on the optimal solution. It is interesting to note that in the cases 6 and 8, both solutions satisfy most of the crowd and majority of the experts. Additionally, the outliers present in the crowd did not cause deviation from the optimal solutions. Based on this, we can conclude that Kalai-Smordinsky solution complies with the intuition presented in Table 1.

In the final part of the experimental evaluation, we applied benchmark and bargaining-based aggregation methods on the real-world data (Journalist and Scientist datasets described at the beginning of this section). Due to a large number of alternatives that are evaluated, we show only the aggregated performance for each method in terms of cumulative satisfaction (satisfaction sum and area) and Negotiation Position Gain (NPG). Since, in many cases, the medians between expert and crowd grades are either the same or very close (trivial cases), we analyzed the results for the subsets of problems with different median differences:

- All problems;
- Problems where expert-crowd median difference  $\geq 0.5$ ;
- Problems where expert-crowd median difference  $\geq 1$ .

In Fig. 4, cumulative results for the 1st dataset (expert scientists) is shown. On X-axis, the average gain (relative satisfaction) difference is shown. On Y-axis, the average satisfaction sum and area (cumulative satisfaction) are presented on the upper and lower part of Fig. 4, respectively. It can be seen that the Kalai and

Nash methods dominate all benchmark methods by both cumulative criteria (sum and area). Additionally, both Kalai and Nash methods perform similarly by aforementioned criteria. On the other hand, Kalai has a larger NPG, meaning that Kalai achieved more satisfaction for groups with more homogeneity (a higher maximum satisfaction). This complies with the intuitions described and evaluated on artificial data. It is interesting to note that most of the benchmark aggregation models have a negative negotiation position gain, meaning that these models achieved more satisfaction of the less homogenous group (group with smaller maximum satisfaction). On the other hand, expert mean, expert median, and expert majority benchmarks achieved the highest negotiation position gain (but lower product and sum compared to Kalai and Nash solutions). Detailed inspection of the results showed that experts were more homogenous than crowd voters in the large majority of the cases and thus, the aforementioned benchmarks achieved the maximum satisfaction of expert voters, while disregarding crowd voter satisfaction. It is interesting to note that even though solutions of these methods were extreme (maximum expert satisfaction and disregard of crowd satisfaction) these methods achieved a higher overall satisfaction by means of both sum and area, compared to other benchmark methods that include weighting. This insight conforms with the intuition that aggregation of all votes (even if they are weighted) often leads to solutions that do not satisfy either of the groups. Figure 5 shows the results for the 2<sup>nd</sup> dataset (journalist experts) in the same form as for the scientist experts. It can be seen from the figure that the results show very similar behavior as for the 1st dataset, and that the same conclusions may be drawn. From both real-world examples, it can be seen that bargaining-based optimization methods achieve higher composite satisfaction compared to benchmark models. Additionally, they give higher NPG than benchmark methods, except in extreme cases where the satisfaction of one group is neglected (expert\_mean and expert\_median). Kalai-Smordynsky and Nash aggregations achieve similar composite scores. Nash achieves better result in terms of satisfaction area, while Kalai-Smordinski achieves better result in terms of NPG. Therefore, Kalai-Smordinski complies better with the intuitions of this research. However, Nash solutions are better if the goal is to equalize satisfactions of groups. It is interesting to note that

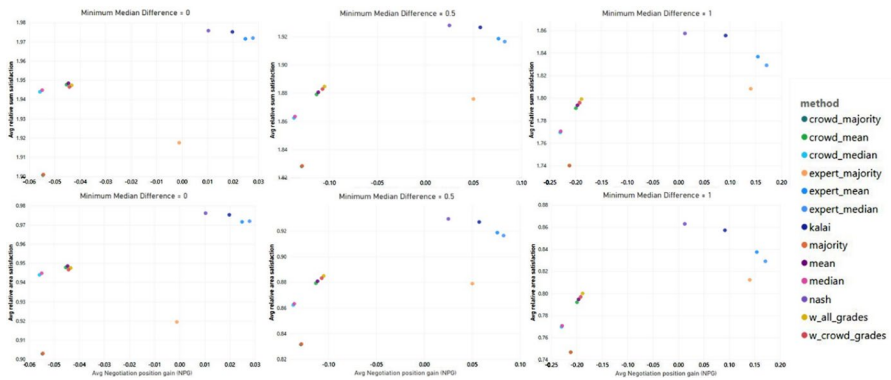


Fig. 4 Results on Science dataset

the weighted average approach based on similarity with experts did not give satisfactory results (it is comparable with the basic aggregation methods such as mean or median). We hypothesize that the bargaining-based optimization methods achieved better results, since they model each task (alternative) individually, while weighting methods assign weights generally - each voter has the same weight for each task (alternative) at hand. These results show a promising performance for exploiting BargCrEx in social choice and industry applications.

### 7 Conclusion and Future Research

In this research, we proposed a system for the estimation of crowd and expert votes and their aggregation in a crowdsourcing setting. The proposed system has several advantages. The inclusion of crowd and experts as separate groups allows modeling of vote aggregation procedures as a bargaining problem. Thus, the intuitions about the “optimal” solution may be implemented based on the bargaining power (in this case, the within group level of agreement). The bargaining based aggregation overcomes problems of the traditional (central tendency based) aggregation, especially in the situations of multimodal or uniform vote distributions. The process of aggregation is unsupervised and does not make any assumptions about the vote/voter quality (and thus needs no external data, or complex routing procedures). Additionally, aggregation is based on optimization and does not involve tedious tuning of hyperparameters, post-processing analyses, or manual selection of final solutions from the Pareto front. The estimation process based on matrix factorization and regression allows modeling of sparse voting data and does not require any processing and/or labeling of historical or external data. It enables the reduction of costs in vote collection as well as the estimation of both expert and crowd votes without the collection of external data. Moreover, hyper-parameter tuning of machine learning algorithms demands minimal or no user interference.

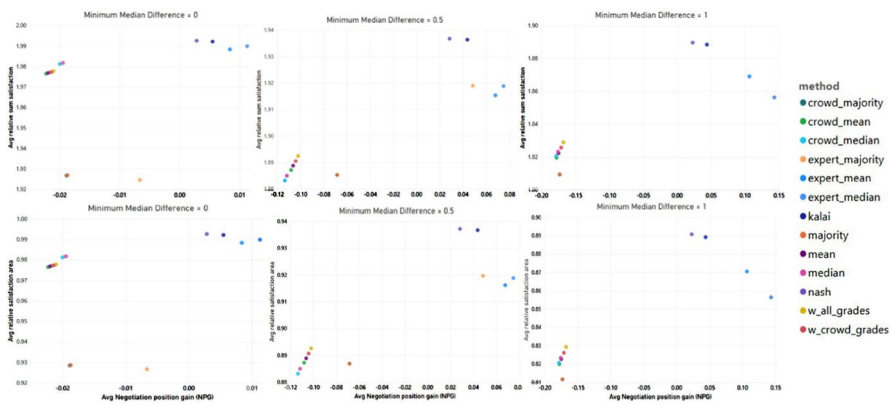


Fig. 5 Results on Journal dataset

It is important to note that proposed method can be seamlessly applied to other real life crowdsourcing applications (besides estimation of article relevance showed in experimental section). Since estimation method is using only crowd and expert grades as inputs (without the need for collection of domain specific data) it can be applied on any crowdsourcing problem, in which expert votes may be collected for relatively small portion of alternatives (compared to the number of crowd votes). Talent and song contests are typical example where both crowd and expert juries are involved and traditional aggregation methods are frequently lead to ambiguous solutions causing high dissatisfaction on both crowd and expert sides. Further, the system based on modeling satisfaction of different groups may be efficiently used in social choice applications (e.g. budget allocation, policy making etc.) where in many cases crowd may be biased or insufficiently informed about the potential impact of decisions that should be made based on voting. The proposed system may also significantly reduce costs and increase quality of decisions in expensive and labour intensive large scale industry applications such as labeling, ranking, or recommendations.

Further, estimation and aggregation procedures are independent and they may be integrated into the existing crowdsourcing real-world solutions seamlessly (for estimation of voter affinities and/or bargaining based aggregation). Finally, different bargaining intuitions may be included into the existing framework. Additionally, we proposed several metrics for quantification of voters satisfaction with final grades.

Our experiments on both real world and synthetic data showed that Kalai–Smorodinsky bargaining solution achieves the best results with respect to the intuition that the group with a higher level of agreement (LOA) should achieve more satisfaction. Another insight is that both bargaining aggregation procedures (Kalai–Smorodinsky and Nash) achieve a higher total satisfaction (for both crowd and expert groups) compared to benchmark models including the models based on weighted average. Finally, the experiments showed that the proposed estimation procedure led to satisfactory error.

One of the limitations of the proposed system is that it does not automatically determine the number of voters (experts or crowd) in order to achieve an adequate level of accuracy. However, this is a domain-specific and very challenging task and will be addressed as a part of our future work. Additionally, matrix factorization algorithms possess the “cold start” problem and demand that each of the alternatives needs to have at least one vote and each of the voters needs to give at least one vote. Furthermore, the factorization performance is also dependent on the level of sparsity, the number of alternatives, and voters.

In the future work, we will try to expand the aggregation model on multiple groups and allow a bargaining process between different types of crowd (e.g., crowd with different interests) and/or expert voters (e.g., journalists and scientists). Another line of the future work will be the integration of bargaining solutions with the existing argumentation and crowd weighting approaches. Another limitation of this framework is that it assumes only the existence of two groups (crowd and expert).

**Acknowledgements** This paper is a result of the project ONR - N62909-19-1-2008 supported by the Office for Naval Research, the United States: Aggregating computational algorithms and human decision-making preferences in multi-agent settings.

## References

- Aitamurto T, Landemore H, Saldivar Galli J (2017) Unmasking the crowd: participants' motivation factors, expectations, and profile in a crowdsourced law reform. *Inf Commun Soc* 20(8):1239–1260
- Alonso O, Rose DE, Stewart B (2008) Crowdsourcing for relevance evaluation. *ACM SigIR Forum* 42(2):9–15
- Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. *J Mach Learn Res* 13(2):281–305
- Bhatt S, Chen K, Shalin VL, Sheth AP, Minnery B (2019) Who should be the captain this week? leveraging inferred diversity-enhanced crowd wisdom for a fantasy premier league captain prediction. In: *Proceedings of the international AAAI conference on Web and Social Media*, vol 13, pp 103–113
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Chan S, Treleaven P, Capra L (2013) Continuous hyperparameter optimization for large-scale recommender systems. In: *2013 IEEE international conference on Big Data*. IEEE, pp 350–358
- Chen L, Xu P, Liu D (2020) Effect of crowd voting on participation in crowdsourcing contests. *J Manag Inf Syst* 37(2):510–535
- Desai A, Warner J, Kuderer N, Thompson M, Painter C, Lyman G, Lopes G (2020) Crowdsourcing a crisis response for covid-19 in oncology. *Nat Cancer* 1(5):473–476
- Devari A, Nikolaev AG, He Q (2017) Crowdsourcing the last mile delivery of online orders by exploiting the social networks of retail store customers. *Transp Res Part E Logist Transp Rev* 105:105–122
- Dodevska ZA, Kovacevic A, Vukicevic M, Delibašić B (2020) Two sides of collective decision making—votes from crowd and knowledge from experts. In: *International conference on decision support system technology*. Springer, Cham, pp 3–14
- Friedman JH (2002) Stochastic gradient boosting. *Comput Stat Data Anal* 38(4):367–378
- Garcia AC, Klein M (2017) pbot: an idea filtering method based on negative multi-voting and pareto aggregation. Available at SSRN 3175329
- Ghezzi A, Gabelloni D, Martini A, Natalicchio A (2018) Crowdsourcing: a review and suggestions for future research. *Int J Manag Rev* 20(2):343–363
- Goel A, Krishnaswamy AK, Sakshuwong S, Aitamurto T (2019) Knapsack voting for participatory budgeting. *ACM Trans Econ Comput: TEAC* 7(2):1–27
- Grofman B, Owen G, Feld SL (1983) Thirteen theorems in search of the truth. *Theor Decis* 15(3):261–278
- Haltofová B (2018) Fostering community engagement through crowdsourcing: case study on participatory budgeting. *Theor Empir Res Urban Manag* 13(1):5–12
- Hogarth RM (1978) A note on aggregating opinions. *Organ Behav Hum Perform* 21(1):40–46
- Hong L, Page SE (2004) Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proc Natl Acad Sci* 101(46):16385–16389
- Hong H, Du Q, Wang G, Fan W, Xu D (2016) Crowd wisdom: the impact of opinion diversity and participant independence on crowd performance. *AMCIS 2016 Proceedings*. 11 <https://credibilitycoalition.org/>. Accessed 10/03/2021
- <https://data.world/credibilitycoalition/credibilityfactors2020/>. Accessed 10/03/2021
- Idrissi N, Zellou A (2020) A systematic literature review of sparsity issues in recommender systems. *Soc Netw Anal Min* 10(1):1–23
- Isinkaye FO, Folajimi YO, Ojokoh BA (2015) Recommendation systems: principles, methods and evaluation. *Egypt Inform J* 16(3):261–273
- Jiang J, An B, Jiang Y, Zhang C, Bu Z, Cao J (2019) Group-oriented task allocation for crowdsourcing in social networks. *IEEE Trans Syst. Man. Cybernetics Syst.* 51(7):4417–4432. <https://doi.org/10.1109/TSMC.2019.2933327>
- Jiang N, Xu D, Zhou J, Yan H, Wan T, Zheng J (2020) Toward optimal participant decisions with voting-based incentive model for crowd sensing. *Inf Sci* 512:1–17


- Jung HJ (2014) Quality assurance in crowdsourcing via matrix factorization based task routing. In: Proceedings of the 23rd international conference on World Wide Web, pp 3–8
- Kalai E, Smorodinsky M (1975) Other solutions to nash's bargaining problem. *Econom J Econom Soc*, pp 513–518
- Keuschnigg M, Ganser C (2017) Crowd wisdom relies on agents' ability in small groups with a voting aggregation rule. *Manag Sci* 63(3):818–828
- Kleemann F, Voß GG, Rieder K (2008) Un (der) paid innovators: the commercial utilization of consumer work through crowdsourcing. *Sci Technol Innov Stud* 4(1):5–26
- Kovacevic A, Vukicevic M, Radovanovic S, Delibasic B (2020a) Crex-wisdom framework for fusion of crowd and experts in crowd voting environment—machine learning approach. In: ADBIS, TPDFL and EDA 2020 common workshops and doctoral consortium. Springer, Cham, pp 131–144
- Kovacevic A, Vukicevic M, Jovanovic M (2020b) Fusion of crowd and expert knowledge based on feature embeddings and clustering in crowd voting setting. In: Proceedings of the XVII international symposium SymOrg. Zlatibor, Serbia, September 7–10, pp 270–277
- Larrick RP, Soll JB (2006) Intuitions about combining opinions: misappreciation of the averaging principle. *Manag Sci* 52(1):111–127
- Lees A, Welty C (2019) Discovering user bias in ordinal voting systems. In: Companion proceedings of the 2019 World Wide Web Conference, pp 1106–1110
- Liu A, Moitra A (2020) Better algorithms for estimating non-parametric models in crowd-sourcing and rank aggregation. In: Conference on learning theory. PMLR, pp 2780–2829
- Lorenz J, Rauhut H, Schweitzer F, Helbing D (2011) How social influence can undermine the wisdom of crowd effect. *Proc Natl Acad Sci* 108(22):9020–9025
- Luther K, Tolentino JL, Wu W, Pavel A, Bailey BP, Agrawala M, Dow SP (2015) Structuring, aggregating, and evaluating crowdsourced design critique. In: Proceedings of the 18th ACM conference on computer supported cooperative work and social computing, pp 473–485
- Meyer AN, Longhurst CA, Singh H (2016) Crowdsourcing diagnosis for patients with undiagnosed illnesses: an evaluation of crowdmed. *J Med Internet Res* 18(1):12
- Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781
- Ong JJ, Bilardi JE, Tucker JD (2017) Wisdom of the crowds: crowd-based development of a logo for a conference using a crowdsourcing contest. *Sex Transm Dis* 44(10):630
- Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
- Procaccia A, Shah N (2016) optimal aggregation of uncertain preferences. In: Proceedings of the AAAI conference on artificial intelligence, vol 30, no 1
- Rachmilevitch S (2019) Egalitarianism, utilitarianism, and the nash bargaining solution. *Soc Choice Welfare* 52(4):741–751
- Samuelson L (2016) Game theory in economics and beyond. *J Econ Perspect* 30(4):107–30
- Semanjski I, Gautama S (2015) Smart city mobility application—gradient boosting trees for mobility prediction and analysis based on crowdsourced data. *Sensors* 15(7):15974–15987
- Singh R, Héliouët L, Miklos Z (2020) Reducing the cost of aggregation in crowdsourcing. In: International conference on Web Services, Springer, Cham, pp 77–95
- Snow R, O'Connor B, Jurafsky D, Ng A (2008) Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In: Proceedings of the 2008 conference on empirical methods in natural language processing. Hawaii, Association for Computational Linguistics, Honolulu, pp 254–263
- Srinivasan R, Chander A (2019) Crowdsourcing in the absence of ground truth—a case study. arXiv:1906.07254
- Suran S, Pattanaik V, Draheim D (2020) Frameworks for collective intelligence: a systematic literature review. *ACM Comput Surv: CSUR* 53(1):1–36
- Sutton C, Ghiringhelli LM, Yamamoto T, Lysogorskiy Y, Blumenthal L, Hammerschmidt T, Scheffler M (2019) Crowd-sourcing materials-science challenges with the nomad 2018 kaggle competition. *NPJ Comput Mater* 5(1):1–11
- Takács G, Tikk D (2012) Alternating least squares for personalized ranking. In: Proceedings of the sixth ACM conference on recommender systems, pp 83–90
- Thomson W (1994) Cooperative models of bargaining. In: Handbook of game theory with economic applications, vol 2, pp 1237–1284. Elsevier



Van Damme E (1986) The nash bargaining solution is optimal. *J Econ Theory* 38(1):78–100

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Authors and Affiliations

**Ana Vukicevic<sup>1,2</sup>**  · **Milan Vukicevic<sup>1</sup>** · **Sandro Radovanovic<sup>1</sup>** · **Boris Delibasic<sup>1</sup>**

Milan Vukicevic  
vukicevic.milan@fon.bg.ac.rs

Sandro Radovanovic  
sandro.radovanovic@fon.bg.ac.rs

Boris Delibasic  
boris.delibasic@fon.bg.ac.rs

<sup>1</sup> Faculty of Organizational Sciences, University of Belgrade, Jove Ilica 154, 11000 Beograd, Serbia

<sup>2</sup> Saga New Frontier Group Ltd, Belgrade, Serbia