The Wisdom of Crowds in the Recollection of Order Information

Authored by:

Mark Steyvers
Michael D. Lee
Brent Miller
Pernille Hemmer

Abstract

When individuals independently recollect events or retrieve facts from memory, how can we aggregate these retrieved memories to reconstruct the actual set of events or facts? In this research, we report the performance of individuals in a series of general knowledge tasks, where the goal is to reconstruct from memory the order of historic events, or the order of items along some physical dimension. We introduce two Bayesian models for aggregating order information based on a Thurstonian approach and Mallows model. Both models assume that each individuals reconstruction is based on either a random permutation of the unobserved ground truth, or by a pure guessing strategy. We apply MCMC to make inferences about the underlying truth and the strategies employed by individuals. The models demonstrate a wisdom of crowds" effect, where the aggregated orderings are closer to the true ordering than the orderings of the best individual."

1 Paper Body

When individuals independently recollect events or retrieve facts from memory, how can we aggregate these retrieved memories to reconstruct the actual set of events or facts? In this research, we report the performance of individuals in a series of general knowledge tasks, where the goal is to reconstruct from memory the order of historic events , or the order of items along some physical dimension. We introduce two Bayesian models for aggregating order information based on a Thurstonian approach and Mallows model. Both models assume that each individual's reconstruction is based on either a random permutation of the unobserved ground truth, or by a pure guessing strategy. We apply MCMC to make inferences about the underlying truth and the strategies employed by individuals. The models demonstrate a "wisdom of crowds " effect, where the

aggregated or derings are closer to the true ordering than the orderings of the best individual.

1

I nt ro duc t io n

Many demonstrations have shown that aggregating the judgments of a number of individuals results in an estimate that is close to the true answer, a phenomenon that has come to be known as the ?wisdom of crowds? [1]. This was demonstrated by Galton, who sho wed that the estimated weight of an ox, when averaged across individuals, closely approximated the true weight [2]. Similarly, on the game show Who Wants to be a Millionaire, contestants are given the opportunity to ask all members of the audience to answer multiple choice questions. Over several seasons of the show, the modal response of the audience corresponded to the correct answer 91% of the time. More sophisticated aggregation approaches have been developed for multiple choice tasks, such as Cult ural Consensus Theory, that additionally take differences across indi viduals and items into account [3]. The wisdom of crowds idea is currently used in several real-world applications, such as prediction markets [4], spam filtering, and the prediction of consumer preferences through collaborative filtering. Recently, it was shown that a form of the wisdom of crowds phenomenon also occurs within a single person [5]. Averaging multiple guesses from one person provides better estimates than the individual guesses. We are interested in applying this wisdom of crowds phenomenon to human memory involving situations where individuals have to retrieve information more complex than single numerical estimates or answers to multiple choice questions. We will focus here on memory for order information. For example, we test individuals on their ability to reconstruct from memory the order of historic events (e.g., the order of US presidents), or the magnitude along some physical dimension (e.g., the order of largest US cities). We then develop computational models that infer distributions over orderings to explain the observed orderings across individuals. The goal is to demonstrate a wisdom of crowds effects where

the inferred orderings are closer to the actual ordering than the orderings produced by the majority of individuals. Aggregating rank order data is not a new problem. In social choice theory, a number of systems have been developed for aggregating rank order preferences for groups (Marden, 1995). Preferential voting systems, where voters explicitly rank order their candidate preferences, are designed to pick one or several candidates out of a field of man y. These systems, such as the Borda count, perform well in aggregating the individuals' rank order data, but with an inherent bias towards determining the top members of the list. However, as voting is a means for expressing individual preferences, there is no ground truth. The goal for these systems is to determine an aggregate of preferences that is in some sense ?fair? to all members of the group. The rank aggregation problem has also been studied in machine learning and information retrieval [6,7]. For example, if one is presented with a ranked list of webpages from several search engines, how can these be combined to create a single ranking that is more accurate and less sensitive to spam? Relatively little research has been done on the rank order aggregation problem with the goal

of approximating a known ground truth. In follow-ups to Galton's work, some experiments were performed testing the ability of individuals to rank-order magnitudes in psychophysical experiments [8]. Also, an informal aggregation model for rank order data was developed for the Cultural Consensus Theory, using factor analysis of the covariance structure of rank order judgments [3]. This was used to (partially) recover the order of causes of death in the US on the basis of the individual orderings. We present empirical and theoretical research on the wisdom of crowds phenomenon for rank order aggregation. No communication between people is allowed for these tasks, and therefore the aggregation method operates on the data produced b y independent decisionmakers. Importantly, for all of the problems there is a known ground truth. We compare several heuristic computational approaches?based on voting theory and existing models of social choice?that analyze the individual judgments and provide a single answer as output, which can be compared to the ground truth. We refer to these synthesized answers as ?group? answers because they capture the collective wisdom of the group, even though no communication between group members occurred. We also apply probabilistic models based on a Thurstonian approach and Mallows model. The Thurstonian model represents the group knowledge about items as distributions on an interval dimension [9]. Mallows model is a distance-based model that represents the group answer as a modal ordering of items, and assumes each individual to have orderings that are more or less close to the modal ordering [10]. Although Thurstonian and Mallows type of models have often been used to analyze preference rankings [11], they have not been applied, as far as we are aware, to ordering problems where there is a ground truth. We also present extensions of these models that allow for the possibility of different response strategies?some individuals might be purely guessing because they have no knowledge of the problem and others might have partial knowledge of the ground truth. We develop efficient MCMC algorithms to infer the latent group orderings and assignments of individuals to response strategies. The advantage of MCMC estimation procedure is that it gives a probability distribution over group orderings, and we can therefore assess the likelihood of any particular group ordering.

2

Expe r i me nt

2 .1

M e t ho d

Participants were 78 undergraduate students at t he University of California, Irvine. The experiment was composed of 17 questions involving general knowledge regarding: population statistics (4 questions), geography (3 questions), dates, such as release dates for movies and books (7 questions), U.S. Presidents, material hardness, the 10 Commandments, and the first 10 Amendments of the U.S. Co nstitution. An interactive interface was presented on a computer screen. Participants were instructed to order the presented items (e.g., ?Order these books by their first release date, earliest to most recent?), and responded by dragging the individual items on the screen to the desired location in the ordering. The initial ordering of the 10 items within a question was randomized

across all questions and all participants.

Table 1: Unique orderings for each individual for the states and presidents ordering problems A B C D E F G H I J 0 2

A B C D E F G I H J 1 1
A AA B BB C CC DED F D E E F F G GH HH I I I G J J J 1 1 2 5 1 1
A B C D F E G I H J 2 1
A AA B BB C CD E EC DF E F D F G GH I H G H I I J J J 2 2 2 1 1 1
A B D C F E G H I J 2 3
A AA B BB C CC DD D E E E F GH I H F H I G G F J J J I 3 3 3 2 1 1
A B C D F E I G H J 3 1
A AA B BB C CC E F F F DD D E E G GH I I G HH I J J J 3 3 3 1 1 1
A B D C F E H G I J 3 1
A B D F C E G H I J 3 1
AA B B CD E C F F GE DH I I HG J J 4 4 1 1
A B D C H E F G I J 4 1
A C B D E F H I J G 4 1
AA E B B C CG D F F D G E I H H I J J 4 5 1 1
A B D F C H E G I J 5 1
A B D F E C G I H J 5 1
AA B B D E F F ED C C HG GH I I J J 5 5 1 1
A B C D F H I G E J 6 1
A B C E D G J F I H 6 1
AA B B DD CH H C F E E F I I GG J J 6 6 1 1
A B F C D E I H G J 6 1
A B F D C E H I G J 6 1
AA B B C C DH H F I E F D EG G I J J 7 7 1 1
A B F C E D I H G J 7 1
A C B F D I E H G J 7 1
ABB F AC B CD DEA CH F EDH H I E I F G GG J J J I 7 7 7 1 1 1
A B C H F D I E G J 8 1
A AB B GA C CC I B F F D J E F D DE E G I G HH H J J I 8 8 8 1 1 1
A B C H I D F E G J 9 1
A BA B AB D F D C CH HH I F D C I I F J E E E GG G J J 9 9 10 1 1 1
A B E I D F C H G J
A B C D E F G H I J 0 5
A B C D E F G H J I 1 1
A AA B BC C CB DED ED E GF F F GG HH H I I I J J J 1 1 1 2 1 1
A B C D E F H G J I 2 1
A AA B BC C CB E ED DD E F G F HF G GH H I I J J J I 2 2 2 1 3 1
A B C D E F G J I H 3 1
A AA B BC C CB E E E DD D F G F G F G I I H J H J H J I 3 3 3 1 1 1
A B C D E F J G I H 4 1
A AA B BB C CC D E E EDD H F G GH F F I J J J H I G I 4 4 4 1 1 1
A B D C E F I G J H 4 1
A B D E C F I G H J 4 1
AA B B C C DE I D E F G J F H J I HG 6 6 1 1

A B C E D F J I G H 6 1
A B C E D H G J F I 6 1
AA B B C C E E D F I D G J HG F I J H 6 6 1 1
A B E C D I G F H J 6 1
A C B D E G I J F H 6 1
AA CD B C D B E E H F F I J H I G G J 6 6 1 1
A B C D E J I F G H 7 1
A B C E D J G F I H 7 1
AA B B CD E C G E D J I F J I F G HH 7 7 1 1
A C B E D F I J H G 7 1
A C B E G D I F J H 7 1
AA C C B E E B I G DD G I F F HH J J 7 7 1 1
A B C E G H F I J D 8 1
A B D C J E G F I H 8 1
A AA C E E B CD E BB GF C HD G D I F J H J F GH I J I 8 8 8 1 1 1
A F C D B G H E I J 8 1
A AA B BB C CC E E J DD D H J F J H G I I E GF I F GH 9 9 9 1 1 1
A C B F D G J I E H 9 1
A AA C CB D EC B B E ED I I J H J G F G F G F I D HH J 9 9 10 1 1 1
A C B F E J H D G I
2 .2
A AA F B F C F E B J C I D D D CB E E I HH G G GH J I J
D F B C E A G I H J
ABA B AC DEG I I F F F B EH J C CD J D H HG I G J E
B A E I H C D J F G
E AD F H E C C I G EH AD A D F B B J F H I C I BG J G J
C I G A D F H E B J
B J H C I A D F E G
E C HB G I BG J H A J F F DD C E I A
I E A H G D F C B J
J G B H I F E A C D
J H G J I I DG H E E F CD B C F B AA
A = Oregon B = Utah C = Nebras ka D = Iowa E = Alabama F = Ohio G = Virginia H = Delaware I = Connecticut J = Maine
10 10 11 11 11 12 13 14 14 14 16 18 20 22 24 26 26 33 37 42
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
A AA C C E B F B F EC GD F H BH E I I J H G D G J I J D
A B I G C E D F H J
A AA CH B B BC E EG J C J GD D H J I DF H I I E F G F
A B D J E G C I H F
A AA B CB G F C C EH F B J I J I H GD J I E DD G EH F
A C E B I F J H G D
A C F H E D G J B I
AA E E DC I B G J B I CD HG F H J F
A C F B J H I E G D
A E G D C J F I H B

CA G J J G A I ED B C I H D B F F H E

H C D I J E G F A B

A = George Wash ington B = John Adams C = Thomas Jefferson D = James Monroe E = Andrew Jackson F = Theodore Roosevelt G = Wood row Wilson H = Franklin D. Roosevelt I = Harry S. Truman J = Dwight D. Eisenhower

10 10 10 10 11 12 12 13 13 13 13 14 14 14 14 15 17 18 19 26 28

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

R e s ul t s

To evaluate the performance of participants as well as models, we measured the distance between the reconstructed and the correct ordering. A commonly used distance metric for orderings is Kendall?s ?. This distance metric counts the number of adjacent pairwise disagreements between orderings. Values of ? range from: 0 ? ? ? ?(? ? 1)/2, where N is the number of items in the order (10 for all of our questions). A value of zero means the ordering is exactly right, and a value of one means that the ordering is correct except for two neighboring items being transposed, and so on up to the maximum po ssible value of 45. Table 1 shows all unique orderings, by column, that were produced for two problems: arranging U.S. States by east-west location, and sorting U.S. Presidents by the time they served in office. The correct ordering is shown on the right. The columns are sorted by Kendall's ? distance. The first and second number below each ordering correspond to Kendall's ? distance and the number of participants who produced the ordering respectively. These two examples show that only a small number of participants reproduce d the correct ordering (in fact, for 11 out of 17 problems, no participant gave the correct answer). It also shows that very few orderings are produced by multiple participants. For 8 out of 17 problems, each participant produced a unique ordering. To summarize the results across participants, the column labeled PC in Table 2 shows the proportion of individuals who got the ordering exactly right for each of the ordering task questions. On average, about one percent of participants recreated the correct rank ordering perfectly. The column ?, shows the mean ? values over the population of participants for each of the 17 sorting task questions. As this is a prior knowledge task, it is interesting to note the best performance overall was achieved on the Presidents, States from west to east, Oscar movies, and Movie release dates tas ks. These four questions relate to educational and cultural knowledge that seems most likely to be shared by our undergraduate subjects. Finally, an important summary statistic is the performance of the best individual. Instead of picking the best individual separately for each problem, we find the individual who scores best across all problems. Table 2, bottom row, shows that this individual has on average a ? distance of 7.8. To demonstrate the wisdom of crowds effect, we have to show that the synthesized group ordering outperforms the ordering, on average, of this best individual.

3

Mo de li ng

We evaluated a number of aggregation models on their ability to reconstruct the ground truth based on the group ordering inferred from individual order-

ings. First, we evaluate two heuristic methods from social choice theory based on the mode and Borda counts. One drawback of such heuristic aggregation models is that they create no explicit representation of each individual's working knowledge. Therefore, even though such methods can aggregate

Table 2: Performance of the four models and human participants Problem books city population europe city population us city population world country landmass country population hardness holidays movies releasedate oscar best-movies oscar movies presidents rivers states westeast superbowl ten amendments ten commandments AVERAGE BEST INDIVIDUAL

Humans ? PC .000 12.3 .000 16.9 .000 15.9 .000 19.3 .000 10.9 .000 14.6 .000 15.3 .051 8.9 .013 7.3 .013 11.2 .000 11.9 .064 7.5 .000 16.1 .026 8.2 .000 18.6 .013 14.0 .000 16.8 .011 13.3 0 7.8

Thurstonian Model C ? Rank 0 5 91 0 11 81 0 7 96 0 16 73 0 5 95 0 12 74 0 14 64 0 4 78 0 2 95 0 4 90 0 1 100 0 2 87 0 13 77 0 2 88 0 16 65 0 2 97 0 8 90 .00 7.29 84.8

Mallows Model C ? Rank 0 5 91 0 12 77 0 7 96 0 16 73 0 5 95 0 11 82 0 14 64 0 5 77 0 2 95 0 4 90 0 1 100 0 1 94 0 14 67 0 2 88 0 15 71 0 3 96 0 7 91 .00 7.29 85.1

Borda Counts C ? Rank 0 7 82 0 11 81 0 12 67 0 15 77 0 5 95 0 11 82 0 11 91 0 4 78 0 2 95 0 3 97 0 2 96 0 3 79 0 11 91 0 3 78 0 10 96 0 5 90 0 12 74 .00 7.47 85.3

Mode C ? Rank 0 12 40 0 17 42 0 16 45 0 19 44 0 7 76 0 15 53 0 15 46 1 0 100 0 2 95 0 3 97 0 2 96 1 0 100 0 16 42 0 1 97 0 19 40 0 4 95 0 17 51 .12 9.67 68.2

the individual pieces of knowledge across individuals, they cannot explain why individuals rank the items in a particular way. To address this potential weakness, we develop two simple probabilistic models based on the Thurstonian approach [9] and Mallows model [10]. 3 .1

H e uri s ti c Mo del s

We tested two heuristic aggregation models. In the simplest heuristic, based on the mode, the group answer is based on the most frequently occurring sequence of all observed sequences. In cases where several different sequences correspond to the mode, a ra ndomly chosen modal sequence was picked. The second method uses the Borda count method, a widely used technique from voting theory. In the Borda count method, weighted counts are assigned such that the first choice ?candidate? receives a count of N (where N is the number of candidates), the second choice candidate receives a count of N-1, and so on. These counts are summed across candidates and the candidate with the highest count is considered the ?most preferred?. Here, we use the Borda count to create an ordering over all items by ordering the Borda counts. Table 2 reports the performance of all of the aggregation models. For each, we checked whether the inferred group order is correct (C) and measured Kendall's ?. We also report in the rank column the percentage of participants who perform worse or the same as the group answer, as measured by ?. With the rank statistic, we can verify the wisdom of crowds effect. In an ideal model, the aggregate answer should be as good as or better than all of the individuals in the group. Table 1 shows the

results separately for each problem, and averaged across all the problems. These results show that the mode heuristic leads to the worst performance overall in rank. On average, the mode is as good or better of an estimate than 68% of participants. This means that 32% of participants came up with better solutions individually. This is not surprising, since, with an ordering of 10 items, it is possible that only a few participants will agree on the ordering of items. The difficulty in inferring the mode makes it an unreliable method for constructing a group answer. This problem will be exacerbated for orderings involving more than 10 items, as the number of possible orderings grows combinatorially. The Borda count method performs relatively well in terms of Kendall's ? and overall rank performance. On average, these methods perform with ranks of 85%, indicating that the group answers from these methods score amongst the best individuals . On average, the Borda count has an average distance of 7.47, which outperforms the best individual over all problems.

Thurstonian model (z = 1)
Guessing model (z = 0)
A B
x1 x2
A
C
A
x3
C
B C
x4
B
C C
A
B A
B
y1 : A ¡ B ¡ C
y3 : C ¡ B ¡ A
y2 : A ¡ C ¡ B
y4 : C ¡ A ¡ B

Figure 1. Illustration of the extended Thurstonian Model with a guessing component

3 .2

A Th ur s to ni an M o del

In the Thurstonian approach, the overall item knowledge for the group is represented explicitly as a set of coordinates on an interval dimension. The interval representation is justifiable, at least for some of the problems in our study that involve one -dimensional concepts, such as the relative timing of events, or the lengths of items. We will introduce an extension of the Thurstonian approach where the orderings of some of the individuals are drawn from a Thurstonian model and others are drawn are based on a guessing process with no relation to the underlying interval representation. To introduce the basic Thurstonian

approach, let N be the number of items in the ordering task and M the number of individuals ordering the items. Each item is represented as a value ? ? along this dimension, where ? ? 1,? , ? . Each individual is assumed to have access to this group-level information. Ho wever, individuals might not have precise knowledge about the exact location of each item. We model each individual's location of the item by a single sample from a Normal distribution, centered on the item?s group location. Specifically, in our model, when determining the order of items, a person ? ? 1, ? , ? samples a value from each item ?, ??? N ?? , ?? . The standard deviation ?? captures the uncertainty that individuals have about item ? and the samples ??? represent the mental representation for the individual. The ordering for each individual is then based on the ordering of their samples. Let ?? be the observed ordering of the items for individual j so that ?? = (? 1, ? 2, ? , ? ?) if and only if ??1 ? ¡ ??2? ¡ ? ¡ ??? ? . Figure 1 (left panel) shows an example of this basic Thurstonian model with group-level information for three items, A, B, and C and two individuals. In the illustration, there is a larger degree of overlap between the representations for B and C making it likely that items B and C are transposed (as illustrated for the second individual). We extend this basic Thurstonian model by incorporating a guessing component. We found this to be a necessary extension because some individuals in the ordering tasks actually were (a)

(b)

?

?

?0 ?0

xj

?

?

zj

yj

zj j=1,?,M

yj j=1,?,M

Figure 2. Graphical model of the extended Thurstonian Model (a) and Mallows model (b)

Presidents

Country Landmass

George Washington (1)

Russia (1)

John Adams (2)

Canada (4)

Thomas Jefferson (3)

China (2)

James Monroe (5)

United States (3)

Andrew Jackson (4)

Brazil (7)

Theodore Roosevelt (6)

Australia (5)
Woodrow Wilson (7)
India (6)
Franklin D. Roosevelt (9)
Argentina (8)
Harry S. Truman (8)
Kazakhstan (10)
Dwight D. Eisenhower (10)

Freedom of speech & religion (1) Right to bear arms (2) No quartering of soldiers (4) No unreasonable searches (3) Due process (5) Trial by Jury (6) Civil Trial by Jury (7) No cruel punishment (8) Right to non-specified rights (10)

Sudan (9) First
Last
Ten Amendments
Power for the States & People (9) Largest
Smallest

Figure 3. Sample Thurstonian inferred distributions. The vertical order is the ground truth ordering, while the numbers in parentheses show the inferred group ordering not familiar with any of the items in the ordering tasks (such as the Ten Commandments or ten amendents). In the extended Thurstonian model, the ordering of such cases are assumed to originate from a single distribution, ??? N ?0 , ?0 , where no distinction is made between the different items?all samples come from the same distribution with parameters ? 0, ?0. Therefore, the orderings produced by the individuals under this model are completely random. For example, Figure 1, right panel shows two orderings produced from this guessing model. We associate a latent state ?? with each individual that determines whether the ordering from each individual is produced by the guessing model or the Thurstonian mo del: ??? — ?? , ?? , ? 0, ?0

N ?? , ?? N ?0, ?0
?? = 1 ?? = 0.
(1)

To complete the model, we placed a standard prior on all normal distributions, ? ?, ? ? 1 ? 2 . Figure 2a shows the graphical model for the Thurstonian model. Although the model looks straightforward as a hierarchical model, inference in this model has proven to be difficult because the observed variable ?? is a deterministic ranking function (indicated by the double bordered circle) of the underlying latent variable ?? . The basic Thurstonian model was introduced by Thurstone in 1927, but only recently have MCMC methods been developed for estimation [12]. We developed a simplified MCMC procedure as described in the supplementary materials that allows for efficient estimation of the underlying true ordering, as well as the assignment of individuals to response strategies. The results of the extended Thurstonian model are shown in Table 2. The model performs approximately as well as the Borda count method. The model does not recover the exact answer for any of the 17 problems, based on the knowledge provided by the 78 participants. It is possible that a larger sample size is needed in order to achieve perfect reconstructions of the ground truth. However, the

model, on average, has an distance of 7.29 from the actual truth, which is better than the best individual over all problems. One advantage of the probabilistic approach is that it gives insight into the difficulty of the task and the response strategies of individuals. For some problems, such as the Ten Commandments, 32% of individuals were assigned to the guessing strategy (?? = 0). For other problems, such as the US Presidents, only 16% of individuals were assigned to the guessing strategy, indicating that knowledge about this domain was more widely distributed in our group of individuals. Therefore, the extension of the Thurstonia n model can eliminate individuals who are purely guessing the answers. An advantage of the representation underlying the Thurstonian model is that it allows a visualization of group knowledge not only in terms of the order of items, but also in terms of the uncertainty associated with each item on the interval scale. Figure 3 shows the inferred distributions for four problems where the model performed relatively well. The crosses correspond to the mean of ? ? across all samples, and the error bars represent the standard

deviations ?? based on a geometric average across all samples. These visualizations are intuitive, and show how some items are confused with others in the group population. For instance, nearly all participants were able to identify Maine as the easternmost state in our list, but many confused the central states. Li kewise, there was a large agreement on the proper placement of "the right to bear arms " in the amendments question ? this amendment is often popularly referred to as ?The Second Amendment?. 3 .3

M al l ow s Mo del

One drawback of the Thurstonian model is that it gives an analog representation for each item, which might be inappropriate for some problems. For example, it seems psychologically implausible that the ten amendments or Ten Commandments are mentally represented as coordinates on an interval scale. Therefore, we also applied probabilistic models where the group answer is based on a pure rank ordering. One such a model is Mallows model [7, 9, 10], a distance-based model that assumes that observed orderings that are close to the group ordering are more likely than those far away. One instantiation of Mallows model is based on Kendall's distance to measure the number of pairwise permutations between the group order and the individual order. Specifically, the probability of any observed order ?, given the group order ? is: ? ?—?, ? =

1 ? ?

? ???

?,?

(2)

where ? is the Kendall's distance. The scaling parameter ? determines how close the observed orders are to the group ordering. As described by [7], the normalization function ? ? does not depend on ? and can be calculated efficiently by: ???

? 1?? ?=1 1????

?? =

.

(3)

The model as stated in the Eqs. (2) and (3) describe that standard Mallows model that has often been used to model preference ranking data. We now introduce a si mple variant of this model that allows for contaminants. The idea is that some of the individuals orderings do not originate at all from some common group knowledge, and instead are based on a guessing process. The extended model introduces a latent state ?? where ?? = 1 if the individual j produced the ordering based on Mallows model and ?? = 0 if the individual is guessing. We model guessing by choosing an ordering uniformly from all possible orderings of N items. Therefore, in the extended model, we have 1

? ??—?,?, ?? =

? ?

? ??? 1/?!

?,?

?? = 1 ?? = 0.

(4)

To complete the model, we place a Bernoulli(1/2) prior over ?? . The MCMC inference algorithm to estimate the distribution over ?, ? and ? given the observed data is based on earlier work [6]. We extended the algorithm to estimate ? and also allow for the efficient estimation of ?. The details of the inference procedure are described in the supplementary materials. The result of the inference algorithm is a probability distribution over group answers ?, of which we take the mode as the single answer for a particular problem. Note that the inferred group ordering does not have to correspond with an ordering of any particular individual. The model just finds the ordering that is close to all of the observed orderings, except those that can be better explained by a guessing process. Figure 4 illustrates the model solution based on a single MCMC sample for the Ten Commandments and ten amendment sorting tasks. The figure shows the distribution of distances from the inferred group ordering. Each circle corresponds to an individual. Individuals assigned to Mallows model and the guessing model are illustrated by filled and unfilled circles respectively. The solid and dashed red lines show the expected distributions based on the model parameters. Note that although Mallows model describes an exponential falloff in probability based on the distance from the group ordering, the expected distributions also take into account the number of orderings that exist at each distance (see [11], page 79, for a recursive algorithm to compute this).

Ten Commandments Number of Individuals

6 5 4 3

zj ?1

2

zj ? 0

1 0

0

5

10

15

20

25
30
35
40
45
?
Ten Amendments
Number of Individuals
8 6
zj ?1 zj ? 0
4 2 0
0
5
10
15
20
25
30
35
40
45
?
d (y j , ?)

Figure 4. Distribution of distances from group answer for two example problems.

Figure 4 shows the distribution over individuals that are captured by the two routes in the model. The individuals with a Kendall's ? above below 15 tend to be assigned to Mallows route and all other individuals are assigned to the the guessing route. Interestingly, the distribution over distances appears to be bimodal, especially for the Ten Commandments. The middle peak of the distribution occurs at 22, which is close to the expected value of 22.5 based on guessing. This result seems intuitively plausible – not everybody has studied the Ten Commandments, let alone the order in which they occur. Table 2 shows the results for the extended Mallows model across all 17 problems. The overall performance, in terms of Kendall 's ? and rank is comparable to the Thurstonian model and the Borda count method. Therefore, there does not appear to be any overall advantage of this particular approach. For the Ten Commandments and ten amendment sorting tasks, Mallows model performs the same or better than the Thurstonian model. This suggests that for particular ordering tasks, where there is arguably no underl ying analog representation, a pure rank-ordering representation such as Mallows model might have an advantage.

4

Co nc l us io ns

We have presented two heuristic aggregation approaches, as well as two probabilistic approaches, for the problem of aggregating rank orders to uncover a

ground truth. For each problem, we found that there were individuals who performed better than the aggregation models (although we cannot identify these individuals until after the fact). However, across all problems, no person consistently outperformed the model. Therefore, for all aggregation methods, except for the mode, we demonstrated a wisdom of crowds effect, where the average performance of the model was better than the best individual over all problems. We also presented two probabilistic approaches based on the classic Thurstonian and Mallows approach. While neither of these models outperformed the simple Borda count heuristic models, they do have some advantages over them. The Thurstonian model not only extracts a group ordering, but also a representation of the uncertainty associated with the ordering. This can be visualized to gain insight into mental representations and processes. In addition, the Thurstonian and Mallows models were both extended with a guessing component to allow for the possibility that some individuals simply do not know any of the answers for a particular problem. Finally, although not explored here, the Bayesian approach potentially offers advantages over heuristic approaches because the probabilistic model can be easily expanded with additional sources of knowledge, such as confidence judgments from participants and background knowledge about t he items.

R e fer e nces [1] Surowiecki, J. (2004). The Wisdom of Crowds. New York, NY: W. W. Norton & Company, Inc. [2] Galton, F. (1907). Vox P opuli. Nature, 75, 450-451. [3] Romney, K. A., Batchelder, W. H., Weller, S. C. (1987). Recent Applications of Cultural Consensus Theory. American Behavioral Scientist, 31, 163-177. [4] Dani, V., Madani, O., Pennock, D.M., Sanghai, S.K., & Galebach, B. (2006). An Empirical Comparison of Algorithms for Aggregating Expert Predictions. In Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI). [5] Vul, E & Pashler, H (2008). Measuring the Crowd Within: Probabilistic representations Within individuals. Psychological Science, 19(7) 645-647. [6] Lebanon, G. & Lafferty, J. (2002). Cranking: Combining Rankings using Conditional Models on Permutations. P roc. of the 19th International Conf erence on Machine Learning. [7] Lebanon, G., & Mao, Y. (2008). Non-Parametric Modeling of Partially Ranked Data. Journal of Machine Learning Research, 9, 2401-2429. [8] Gordon, K. (1924). Group Judgments in the Field of Lifted Weights. Journal of Experimental Psychology, 7, 398-400. [9] Thurstone, L. L. (1927). A law of comparative judgement. Psychological Review, 34, 273? 286. [10] Mallows, C.L. (1957). Non-null ranking models, Biometrika, 44:114? 130. [11] Marden, J. I. (1995). Analyzing and Modeling Rank Data. New York, NY: Chapman & Hall USA. [12] Yao, G., & B?ckenholt, U. (1999). Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler. British Journal of Mathematical and Statistical Psychology, 52, 79? 92.

2 References

NA