

# The Wisdom of Crowds with Informative Priors

**Pernille Hemmer (phemmer@uci.edu)**

**Mark Steyvers (msteyver@uci.edu)**

**Brent Miller (brentm@uci.edu)**

Department of Cognitive Sciences,  
University of California, Irvine  
Irvine, CA, 92697-5100

## Abstract

In some eyewitness situations, a group of individuals might have witnessed the same sequence of events. We consider the problem of aggregating the event sequences as recalled by a small number of individuals with the goal to approximate the true sequence of events as best as possible. We present a Bayesian analysis of this aggregation problem that incorporates individual differences on memory ability as well as informative prior knowledge about event sequences as measured in a separate norming experiment. We show how the added prior knowledge leads to improved model reconstructions especially in small groups of error-prone individuals. The Bayesian aggregation model also leads to a wisdom of crowds effect where the model's reconstruction is as good as some of the best individuals in the group.

**Keywords:** Eyewitness Testimony; Wisdom of Crowds; Rank Ordering; Bayesian Modeling; Serial Recall.

## Introduction

Studies of eyewitness testimony have shown that human memory can be incomplete and unreliable (e.g., Loftus, 1975). In real world situations, there might be multiple eyewitnesses who all witnessed the same set of events. This raises the possibility of recovering the true set of events by analyzing the commonalities in the recalled memories across a number of individuals. We focus in this research on the problem of reconstructing event sequences. The goal is to reconstruct the true ordering of a set of events by aggregating the recalled orderings from a small number of individuals who all witnessed the same event sequence.

Research on "The Wisdom of Crowds" (WoC) has shown that the aggregation of judgments often leads to an estimate that is closer to the ground truth than the vast majority of estimates (Surowiecki, 2004). In a typical WoC experiment, a number of individuals provide numerical estimates of some quantity and the average of the estimates, based on the mean or median, is often closer to the true quantity than the majority of the individual's estimates (Galton, 1907; Surowiecki, 2004). It has been shown that a form of the WoC phenomenon also occurs within a single person (Vul & Pashler, 2008). Averaging multiple guesses from one person provides better estimates than the individual guesses.

The WoC effect can also be demonstrated on tasks in which the ground truth is not a single number. Recently, the WoC effect has been demonstrated with solutions to problem-solving situations such as the traveling salesman

problem (Yi, Steyvers, Lee & Dry, submitted). Steyvers, Lee, Miller, and Hemmer (2009) showed that order information from semantic memory can also be combined across individuals to give high accuracy in reconstructing the true order of items along some physical or temporal dimension. For example, when individuals recall the order of US presidents, or the order of rivers according to length, many of the individual orderings are error-prone, but the aggregate ordering tends to be more accurate. In the research by Steyvers et. al. (2009), a number of aggregation models for order information were tested. It was found that Bayesian approaches outperformed heuristic aggregation approaches.

The best results in WoC experiments have been demonstrated with large numbers of individuals. When errors are uncorrelated (as they tend to be when individuals independently give their judgments) the errors will cancel out in the average. In eyewitness situations however, there might rarely be a "crowd" available to witness the same set of events. In these cases, we have to rely on a small number of individuals (in many cases, just one) and significant errors might not cancel out in the average. Therefore, it might not be sufficient to just analyze the commonalities across the witness reports. We propose that it is better to combine the witness reports with prior knowledge about the particular event sequence. Combining prior knowledge with noisy information has been shown in other domains to improve the recovered estimate (Hemmer & Steyvers, 2008; Konkle & Oliva, 2007; Kan, Alexander, Verfaelle, 2009). The novelty of this research is that we incorporate informative prior knowledge in an aggregation model for order information in order to improve the aggregate estimate. This is especially helpful when aggregating across a small number of error-prone individuals.

The plan for this paper is as follows. We will first illustrate the problem of combining prior knowledge with numerical judgment data such as Galton's original experiment. We then report on behavioral experiments where we test people's ability to reconstruct from episodic memory the order of stereotyped events (e.g., getting up in the morning), or random events (e.g., clay animation without a clear story line). We also conduct experiments where we measure prior knowledge about the same set of events. We then develop a Bayesian approach that aggregates the orderings across individuals while taking prior knowledge into account.

Table 1. Mean absolute error between true and estimated weight (lbs)

N	$\sigma_0$			
	75	150	300	$\infty$
1	35.4	39.2	42.2	43.7
2	27.4	29.3	30.5	30.8
3	23.3	24.4	24.9	25.2
5	18.7	19.1	19.3	19.5
10	13.4	13.6	13.8	13.9
20	9.6	9.8	9.7	9.8

### Galton's ox with prior knowledge

Galton (1907) asked a large number of British fair-goers (approx. 800) to estimate the weight of an ox. He showed that the aggregate estimate, calculated by the median, closely approximated the true weight of 1198 lbs; in fact it was off by only 9 lbs. Suppose that Sir Galton only had the benefit of observing a small number of fair-goers' estimates, but had some *a priori* knowledge about the weight distribution of oxen. How could the prior information have been used in the aggregation?

We will apply a Bayesian analysis to this (hypothetical) situation. To keep the analysis straightforward, we will make some simplifying assumptions about Galton's data. For example, let's assume that the weight of oxen is Normally distributed,  $\mu_{true} \sim N(\mu_0, \sigma_0^2)$ , where  $\mu_0$  and  $\sigma_0^2$  are the prior mean and variance of ox weight, and  $\mu_{true}$  is the weight of the particular ox that was sampled for the estimation experiment. Suppose that the fair-goers' guesses are drawn from a Normal distribution centered on the true ox weight  $y_i \sim N(\mu_{true}, \sigma_m^2)$ , where  $\sigma_m^2$  is the variance of the estimates based on perceptual and estimation noise. Having observed all the weight estimates, and having knowledge about the prior mean and variance of the estimates, what is the best guess about the true weight of the ox? Using standard Bayesian inference techniques, the best estimate is a weighted average of the mean of the individual estimates and the prior mean of the oxen:  $\hat{\mu}_{true} = w\mu_0 + (1-w)\bar{y}$ , where  $w = (1/\hat{\sigma}_0^2) / [(1/\hat{\sigma}_0^2) + (N/\sigma_m^2)]$ , and  $N$  is the number of guesses provided by fair-goers. Therefore, the prior mean,  $\mu_0$  is weighted more heavily when the prior has a higher precision ( $1/\hat{\sigma}_0^2$ ). Therefore, if the prior is strong, it will have a strong influence on the estimated weight. Similarly, if there are a few individuals (small  $N$ ), the prior will also exert a strong influence on the estimated weight. Note that we make a distinction between the true prior variance,  $\sigma_0^2$ , and the assumed prior variance  $\hat{\sigma}_0^2$  by the researcher who is aggregating the judgments. Ideally,  $\hat{\sigma}_0^2 = \sigma_0^2$ , but we will investigate cases where  $\hat{\sigma}_0^2 > \sigma_0^2$ . This corresponds to a situation where the researcher is using a prior that is weaker than might be warranted by real-world knowledge.

In our simulations, we assumed that a prior mean  $\mu_0 = 1150$  lbs and a prior standard deviation  $\sigma_0 = 75$  lbs (we are actually not sure if this corresponds even roughly to the distribution of ox weights as reported but it will serve as

a useful example). We varied the assumed prior standard deviation  $\hat{\sigma}_0$  from 75 lbs (corresponding to an informative prior) to infinity (uninformative prior). We also varied the number of individuals,  $N$  from 1 to 20. We set the memory variance  $\sigma_m = 55$ , based on estimates of Galton's original data. Each hypothetical Galton experiment was repeated 50,000 times. For each experiment, we first draw a true weight  $\mu_{true}$  and then draw observations  $y_i$ . For each simulated experiment, we measured  $|\hat{\mu}_{true} - \mu_{true}|$ , the absolute error between the estimated and true weight.

Table 1 shows the simulation results. The worst performance is observed for a single individual where the researcher uses an uninformative prior. As the strength of the prior increases, so does accuracy. A strong prior (equivalent to the true prior) gives an 8 lb improvement in the estimate of the single individual. The table shows diminishing returns for using an informative prior as  $N$  increases. Therefore, informative priors are only useful when combining estimates across a small number of individuals.

This example only serves to illustrate the effect of incorporating prior knowledge into an aggregation approach involving numerical estimates. We acknowledge that there are many differences between Galton's analysis and our example (e.g., Galton used the median, the guesses from his fair-goers were not exactly Gaussian distributed, and we made assumptions about the prior weight of Oxen and the variability of peoples' prior knowledge). However, it does demonstrate that one can guard against noisy estimates by combining the noisy information with prior knowledge. We have previously proposed that this is a strategy adopted by human memory in order to improve accuracy in recall (Hemmer & Steyvers, 2009).

### Empirical Study on Serial Recall

Much research on serial recall has been done on random word and letter sequences that do not have any obvious organization. In such experiments, individuals are shown a sequence of words or letters, and the task is to recall the original temporal order as best as possible during a later test. Typical errors in the recalled orderings are transposition errors where the orderings are locally perturbed (Estes, 1997; Nairne, 1992) -- two events nearby in time tend to be reconstructed as occurring nearby but the amount of perturbation noise depends on many factors such as time elapsed between study and test, stimulus characteristics and individual differences. Similar patterns have been observed in more naturalistic experiments, such as naming the day of the week an event occurred (Huttenlocher, Hedges, & Prohaska, 1990), as well as for autobiographical memory, such as ordering the events of September 11<sup>th</sup> (Altmann, 2003). With more naturalistic event sequences, prior knowledge about the event sequences can influence episodic memory. People have clear expectations for routine activities and are sensitive to the ordering of actions within an activity (Bower, Black & Turner, 1979).



$\tau=1$  indicates that one adjacent pair of items was swapped. When participants are using a random guessing strategy, their expected mean expected distance is  $\tau = (N-1)/4 = 22.5$ .

Figure 1 shows the raw data collected for "bus" video sequence. In the prior knowledge experiment, participants produced orderings that were much better than chance, suggesting that a priori, it is possible to guess the true ordering of events in these types of event sequences. In the memory experiment, 2 participants produced the correct ordering, and 15 more were within one swap of the true order. Note that very few identical orderings are produced between participants. We found that for all 3 random events, in both the prior knowledge experiment and the memory experiment, each participant produced a unique ordering. For the 3 stereotyped event sequences however, only one sequence led to unique orderings across all participants.

Figure 2 shows the distributions of the Kendall  $\tau$  distances for the serial recall and prior knowledge experiment. The top panel shows the distances for stereotyped event sequences and the bottom panel shows the distances for random event sequences. The dashed line shows the distribution of distances that can be expected from a random guessing strategy (this distribution can be calculated exactly, see Marden, 1995). For both the stereotyped and random event sequences, the distances are lower for the memory task than for the prior knowledge task. The distances are also lower for the stereotyped event sequences than for the random event sequences. Even when participants did not study the videos (the prior knowledge condition), they performed better than chance in the stereotyped condition, as compared to the random condition where prior knowledge performance led to a distribution of distances very similar to distances expected from chance performance. These results demonstrate that general knowledge about events can greatly contribute to the accuracy of recalling these events.

## Modeling

The conclusion that we can draw from our empirical study is that prior knowledge can lead to improved average performance in recall. When ordering scenes from an event with strong prior expectations, the resulting orderings are relatively close to the true ordering. Of course, performance improves on average after observing the true event sequence and later recalling the sequence from memory. This raises the question of how one might incorporate an informative prior in a model for aggregating rank-ordered recall. Such priors might guard against errors from a small number of poorly performing individuals. In this paper, to model the serial recall data from the memory experiment, we use a version of Mallows model similar to one proposed by Steyvers et al. (2009), but generalize the model to allow for individual differences. We then present a simple extension that allows for informative priors where the prior is estimated from the orderings produced in the prior norming experiment.

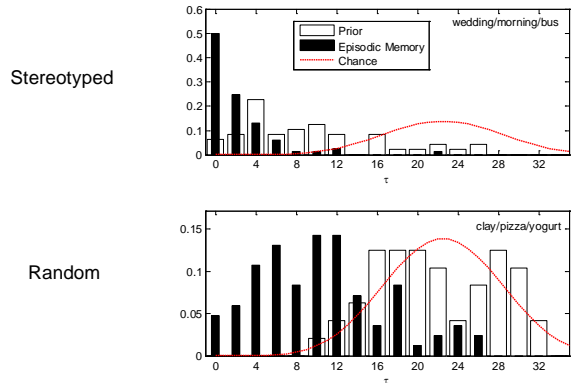


Figure 2. Distributions of Kendall  $\tau$  distances.

## Mallows Model with an Uninformative Prior

In a basic Mallows model (Marden, 1995), all individuals are assumed to derive their orderings from a single underlying ordering, that we will refer to as the *group knowledge*. The group knowledge is a latent variable in the model that can be estimated from the data. Importantly, Mallows model assumes that each individual produces orderings centered around the group ordering with distant orderings less likely than orderings close to the group ordering. Although Mallows-type models have often been used to analyze preference rankings (Marden, 1995), they have not been applied, as far as we are aware, to ordering data from serial recall experiments. We evaluated this aggregation model by comparing the estimated group ordering to the ground truth. If the model is able to tap into the collective wisdom of a group of individuals, the estimated group ordering should be close to the true ordering.

Specifically, let  $\mathbf{y}_j$  represent the ordering from individual  $j$ , and  $\omega$  the latent group ordering. In a Mallows model, the probability of each individual ordering given the group ordering is given by

$$p(\mathbf{y}_j | \omega, \theta_j) \propto e^{-d(\mathbf{y}_j, \omega)\theta_j} \quad (1)$$

where for simplicity we have omitted the normalization constant. The function  $d$  returns the Kendall  $\tau$  distance between two orderings. The scaling parameter  $\theta_j$  determines how close the observed order for individual  $j$  is to the group ordering. It can be interpreted as an individual (inverse) noise parameter -- good individuals tend to closer to the group consensus (high  $\theta$ ) whereas poor performing individuals return more idiosyncratic orderings further away from the group knowledge (low  $\theta$ ). We will assume a Gamma prior on the individual noise levels:  $\theta_j \sim \text{Gamma}(\theta_0\lambda, 1/\lambda)$ , where  $\lambda$  is a hyperparameter that sets of overall level of cohesion expected from the group. Importantly, in this first model, we have assumed a uniform prior over group orderings,  $\omega \sim \text{Uniform}(\Omega)$ , where  $\Omega$  is the set of all orderings. Therefore, a priori, the model assumes no preference for a particular group ordering.

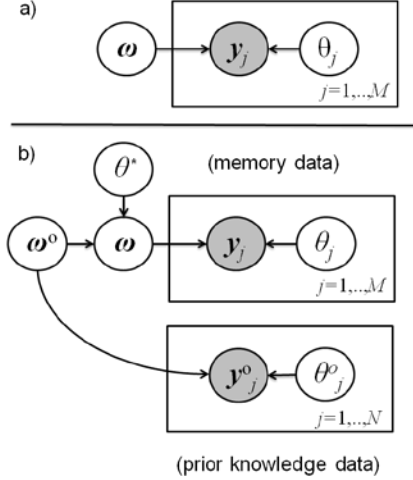


Figure 3. The graphical model representations for the Mallows model with an uninformative prior (a) and an informative prior about the group knowledge (b).

Figure 3, panel a, shows a graphical representation of the model. Shaded nodes represent observed variables while nodes without shading represent latent variables. The arrows indicate the conditional dependencies between the variables and the plate represents the repeated sampling steps across  $M$  subjects in the memory experiment.

### Mallows Model with an Informative Prior

We now introduce a simple variant of this model that allows for an informative prior. The idea is that the group knowledge is itself sampled from a Mallows model:

$$p(\omega | \omega^0, \theta^*) \propto e^{-d(\omega, \omega^0)\theta^*} \quad (2)$$

where  $\omega^0$  is the prior ordering from which the group ordering is derived, and  $\theta^*$  is a scaling parameter. This prior stage in Mallows model at first might not seem to gain any additional information because it is not clear how the prior ordering can be constrained. However, we have data in the prior norming experiment in which  $N$  participants tell us what orderings they expect from certain scenes. Let  $y_j^0$  represent the prior ordering given by individual  $j$  in the norming experiment. We assume that these are produced by a Mallows model with  $\omega^0$  as the "center":

$$p(y_j^0 | \omega^0, \theta_j^0) \propto e^{-d(y_j^0, \omega^0)\theta_j^0} \quad (3)$$

Figure 3, panel b, shows the corresponding graphical model. With this model, we are setting a prior on the group ordering -- when there is little data available from the memory experiment, the group ordering will be influenced by the data from the norming experiment leading to group orderings that are a priori deemed likely. When more data becomes available in the memory experiment, the norming data will have a diminishing influence on the group ordering -- this will be most determined by the memory data.

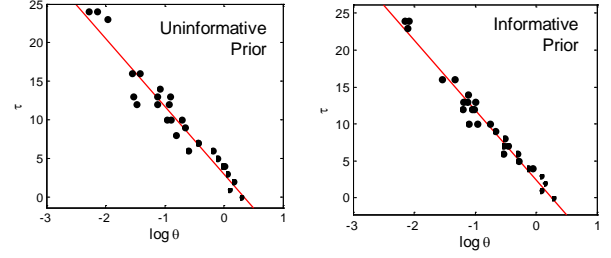


Figure 4. Calibration results for the two models for one event sequence.

### Modeling Results

All latent variables in the model were estimated using a MCMC procedure, separately for each event sequence. The result of the inference procedure is a probability distribution over group orderings, 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 to an ordering of any particular individual. The model just finds the ordering that is close to all of the observed memory orderings.

Figure 4 shows the calibration for the two models on a single event sequence (the clay animation video). Each panel shows the relationship between the inferred  $\theta$  (related to the distance of each individual to the group ordering) and the Kendall's  $\tau$  distance of the individual's answer to the ground truth. The plots show that individuals who are close to the group ordering tend to be closer to the ground truth. This means that the models can calibrate the performance levels of individuals, even in the absence of any explicit feedback or access to the ground truth.

Figure 5 shows the Kendall's  $\tau$  distance for each individual in the memory experiment averaged over the six event sequences. Note that there are substantial individual differences with some individuals coming relatively close to the ground truth. The figure also shows the average model performance and comparison between individual and model performance reveals a WoC effect. The model performs as well as some of the best individuals with only one individual outperforming the model. Therefore, we can conclude there is a weak WoC effect -- a strong WoC effect would correspond to a situation where the model outperforms all individuals in the group.

We now focus on applying the model to subsets of participants to mimic eyewitness situations. Specifically we are interested in analyzing model performance when the worst performing individuals are selected in the sample. In our sampling procedure, we sample the  $K$  worst individuals where we vary  $K$  from 1 (the single worst performing individual) to 28 (all individuals combined). Figure 6 shows model results for both models separated for stereotyped and random event sequences. For random event sequences, where the prior is weak, there is no improvement in the aggregation between the two models. In this case, the prior cannot be used to guard against any egregious errors committed by the worst individuals in the memory task as

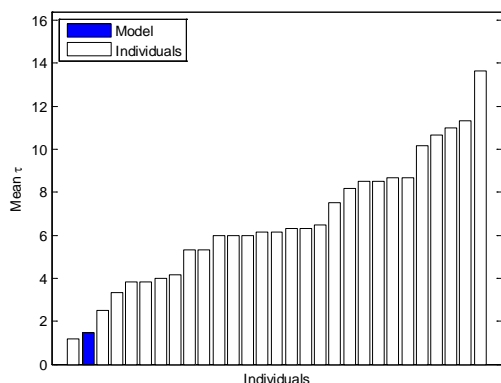


Figure 5. Performance of individuals and model (with informative prior) averaged over six event sequences.

orderings are all a priori equally likely. For stereotyped event sequences however, people have strong prior expectations about the true ordering of events and there is a marked improvement in the aggregate response in the model with the informative prior. This improvement is most pronounced with low sample sizes ( $K=1$  and  $K=2$ ) when the prior can still exert an influence on the inferred group orderings. In these cases, the worst individuals recall event sequences that are a priori unlikely and the prior "corrects for" the noise in the available data.

## Conclusions

We have presented two approaches for aggregating recalled event sequences in order to reconstruct the true event sequence as best as possible. Individuals are likely to differ in their ability to recall event sequences and pay attention to different parts on an event sequences. Therefore, by analyzing the consistencies in orderings across individuals, we can extract the collective wisdom in the group. We presented two aggregation approaches based on Mallows model that allow for individual differences. In the first approach, the model uses only the data from the individuals who all witnessed an event sequence. In the second approach, the model uses an additional source of data based on the prior knowledge about the events extracted from another group of individuals.

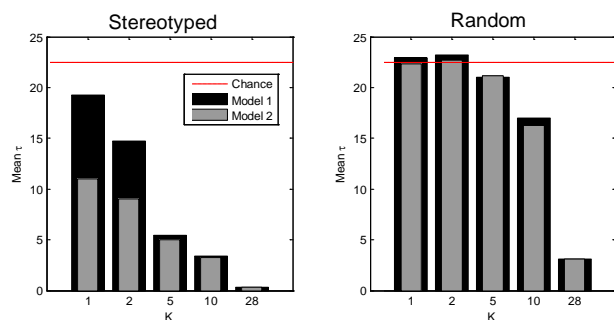


Figure 6. Results from the models with an uninformative prior (model 1) and informative prior (model 2) for subsets of the worst  $K$  individuals from the memory task.

We demonstrated a (weak) WoC effect, where the average performance of the model was better than every individual but one. Importantly, we have shown that the Mallows model with informative priors shows a marked improvement in reconstructing the ground truth in cases where the event sequences are highly stereotyped and a small sample of poorly performing individuals is used for aggregation. This is particularly relevant in eyewitness situations where we typically have only a small number of individuals available.

## References

- Altmann, E. M. (2003) Reconstructing the serial order of events: A case study of September 11, 2001. *Applied Cognitive Psychology*, **17**, 1067-1080.
- Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory. *Cognitive Psychology*, **11**, 177-220.
- Estes, W.K. (1997). Processes of Memory Loss, Recovery, and Distortion. *Psychological Review*, **104**, 148-169.
- Galton, F. (1907). Vox Populi. *Nature*, **75**, 450-451.
- Hemmer, P., Steyvers, M. (2009). Integrating Episodic Memories and Prior Knowledge at Multiple Levels of Abstraction. *Psychonomic Bulletin & Review*, **16**, 80-87.
- Hemmer, P. & Steyvers, M. (2008). A Bayesian Account of Reconstructive Memory. In V. Sloutsky, B. Love, and K. McRae (Eds.) *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.
- Huttenlocher, J., Hedges, L. V., & Prohaska, V. (1992). Memory for day of the week: A 5+2 day cycle. *Journal of Experimental Psychology: General*, **121**, 313-325.
- Kan, I.P., Alexander, M.P. & Verfaellie, M. (2009). Contribution of prior semantic knowledge to new episodic learning in amnesia. *Journal of Cognitive Neuroscience*, **21**, 938-944.
- Konkle, T., & Oliva, A. (2007). Normative representation of objects: Evidence for an ecological bias in perception and memory. In D. S. McNamara & J. G. Trafton (Eds.), *Proc.s of the 29th Annual Cognitive Science Society*, (pp. 407-413), Austin, TX: Cognitive Science Society.
- Loftus, E.F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology*, **7**, 560-572.
- Marden, J. I. (1995). *Analyzing and Modeling Rank Data*. New York, NY: Chapman & Hall USA.
- Nairne, J. S. (1992). The loss of positional certainty in long-term memory. *Psychological Science*, **3**, 199-202.
- Steyvers, M., Lee, M.D., Miller, B., & Hemmer, P. (2009). The Wisdom of Crowds in the Recollection of Order Information. In J. Lafferty, C. Williams (Eds.) *Advances in Neural Information Processing Systems*, **23**. MIT Press.
- Surowiecki, J. (2004). *The Wisdom of Crowds*. New York, NY: W. W. Norton & Company, Inc.
- Vul, E & Pashler, H (2008). Measuring the Crowd Within: Probabilistic representations Within individuals. *Psychological Science*, **19**(7) 645-647.
- Yi, S. K. M., Steyvers, M., Lee, M. D., Dry, M. J. (submitted) Wisdom of the Crowds in Traveling Salesman Problems.